Electronic Assessment of Free Text Document Using Non-Negative Matrix Factorization

Rufai M. M[1], Adigun J. O.[2] and Okikiola F. M.[3]
[1,2,3] *Department of Computer Technology, Yaba College of Technology, Lagos, Nigeria*

***Abstract*:-This study examines the performance of Non-Negative Matrix Factorization (NMF) technique in the electronic assessment of free text document. NMF is a low rank approximation technique. It has application in automated grading of free-text document by reducing the initial matrix generated from the set document into a low rank without compromising the semantic content. Our approach collects student and lecturers 'response in a particular test, converts them to document-term matrix and reduce them using NMF to a low rank approximation matrix. The technique was evaluated using Pearson correlation coefficient and mean divergence error. The results show that a correlation of 0.921728 was observed between the manual scores and the NMF graded scores while 0.88729 was observed between the manual scores and LSA which indicates that NMF is a better assessor when compared to LSA. NMF generates a closer result to the human grade when compared to the LSA. It also proves the ability of NMF as a suitable technique for representing a document in a semantic space without compromising the semantic content of the document.**

Keywords: **Non-Negative Matrix Factorization, Information Retrieval, LSA, Term Frequency, Inverse Document Frequency**

I. INTRODUCTION

In recent time, research efforts have been directed towards Electronic assessment of free text document because of the need to further automate the e-learning process with a view to further strengthen education quality and facilitate mobility among students (Martin, Martin, & Berry,2016). There has been a reasonable application of effort on the assessment of structured questioning such as Multiple-Choice Questions, Multiple Response Question and Hot Spot Question. The use of structured question in assessment has been observed to be in-adequate for assessing learning outcomes because it does not test a student ability to recall, organize and integrate ideas, the ability to express oneself in writing and the ability to supply merely than identify interpretation and application of data(Farrús, M., & Costa-jussà, M. R., 2013).

Free-text question is appreciated for its suitability in measuring high level skills and cognitive level of the learner (Maria De Marsico, Andrea & Marco, 2016).The question of subjectivity, reliability and consistency also feature in essay grading which are functions of the algorithm been used. A reliable assessment is obtained if it has close correlation to human grading. Two basic approaches to Electronic Assessment are Information Retrieval (IR) based and Linguistics-based. A popular example of the IR approach is Non-Negative Matrix Factorisation (NMF), where keywords and their co-occurrence statistics are used to reveal hidden semantic links between a gold standard and the essay to be assessed. Linguistic approaches emphasize the use of structures to decode the semantics. To address the challenges faced in the electronic assessment of free text documents, this paper develops an algorithm which applies NMF in its dimension reduction with a view to improve assessment performance.

II. RELATED WORKS

Non-Negative Matrix Factorisation (NMF)

The NMF is a computational technique for linear dimensionality reduction of a given data matrix X, which is able to explain data in terms of additive combination of nonnegative factors that represent realistic building blocks for the original data. Non-Negative Matrix factorization (NMF) is a low rank approximation technique with reduced storage and run-time requirements and reduced redundancy and noise (Casalino, Del Buono & Mencar 2016). It allows for additive parts-based, interpretable representation of the data. NMF approximates a matrix X by

$$V_{n\times m} \approx W_{n\times r} H_{r\times m} \qquad (3.2)$$

Where W and H are NMF factors and all entries in V, W and H are to be non-negative. r, m, n represent the rank of the matrices and r is chosen to satisfy (n+m)r<nm

The goal of NMF is to minimize the original matrix V. This resulted in the setting of two objective functions

$$min\|V - WH\|_F^2 = min \sum\sum \left(V_{ij} - (WH)_{ij}\right)^2 \qquad (3.3)$$

The second objective function minimizes the Kullback-Leibler divergence

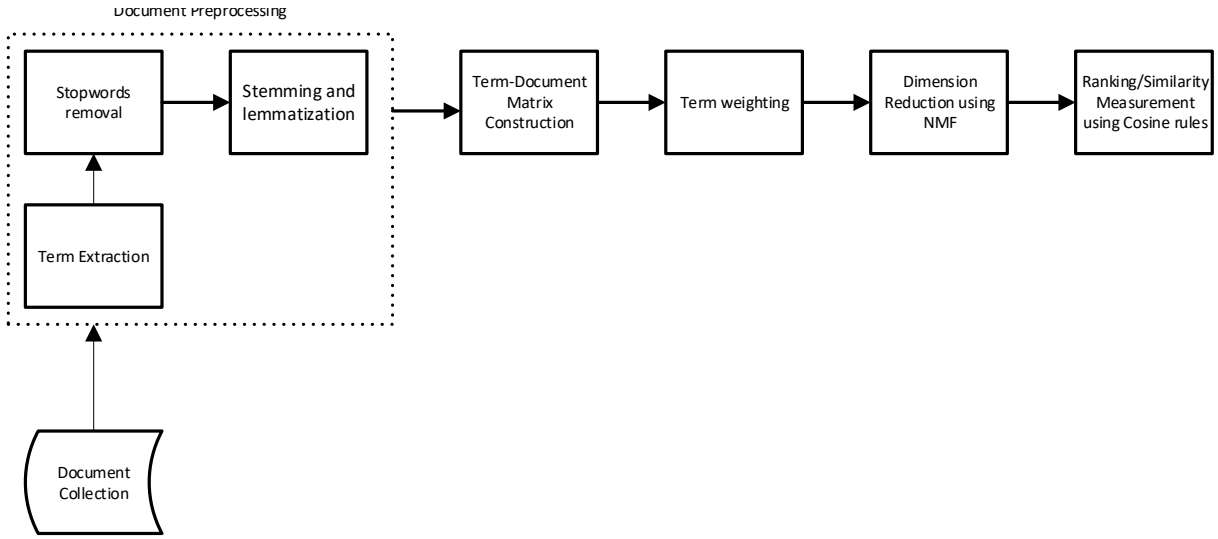
$$\min D_{KL}(V\|WH) = min \sum_i \sum_j \left(V_{ij} log \frac{V_{ij}}{(WH)_{ij}} - V_{ij} + (WH)ij\right) \qquad (3.4)$$

NMF perfectly fits in as a better alternative to SVD in dimension reduction because of its sparsity and non-negativity; reduction in storage and its interpretability.

However, its major challenge is its convergence issue because different NMF algorithm can converge to different local minima. This challenge is addressed by choosing the right initialization and update strategy.

III. METHODOLOGY

The steps involved in the application of NMF in electronic assessment of free text document is depicted in the block diagram below



The method adopted in this research is to first identify and collect documents that will be used for training and extraction of index terms. The extracted terms were subjected to document pre-processing where stop words are removed and the terms are stem to get their root words. The selected documents comprise of a reference text on the course, the lecturer's marking scheme and the answer scripts submitted by the students. The index term and the documents, from which they are generated, are used to create a document-term matrix, where documents tagged as document1, document2, document3..documentN (N= based on the number of documents) are used as the matrix column headings and the index terms are the rows headings. The entries to the matrix are the frequencies of occurrence of a term in a particular document and entries are weighted using Term Frequency - Inverse Document Frequency weighting scheme in order to give emphasis to terms with higher semantic value. The generated weighted matrix is dimensionally reduced in order to filter out noise and words with less semantic contribution. The similarity between two documents is obtained by evaluating similarity of their vectors using cosine similarity rule. The process is implemented using java programming language. The result is evaluated for validity using Pearson correlation statistics. The developed model consists of stopwords removal, stemming and lemmatization, document matrix construction, Term Weighting Generation, Dimension Reduction using NMF and Ranking/Similarity Measurement

i) Stop Words Removal

Stop words are eliminated with the objective of removing words with very low discrimination values for similarity purposes. The approach is to retrieve a list of stop words from an existing online dictionary and have these stop words removed during document preprocessing for both students answers and lecturers marking scheme.

ii) Stemming and Lemmatization

The goal of both stemming and lemmatization is to reduce inflectional forms and sometimes derivationally related forms of a word to a common base form. Lemmatization was done by having a mapping of a term base word and its various synonyms and inflectional forms. These terms and their synonyms are given the same weighting during term weighting generation.

iii) Document Matrix Construction

After document collection, the next stage is to construct the document-term matrix. In this matrix, each index word is a row and each document is a column. Each cell contains the number of times that index word occurs in the document.

iv) Term Weighting Generation

The essence of term weighting is to ensure that rare words are weighted more heavily than common words. For example, a word that occurs in only 5% of the documents should probably be weighted more heavily than a word that occurs in 90% of the documents. The reason for this is because rare words reveals better similarity features among documents. The term weighting approach adopted in this research is the

TFIDF (Term Frequency - Inverse Document Frequency). Under this method, the count in each cell is replaced by the following formula.

$$W_{t,d} = \left(1 + logtf_{t,d}\right).log\frac{N}{df_t} \quad (3.1)$$

Where:

$w_{t,d}$=weight of terms and documents

$tf_{t,d}$= the frequency of term t in document d

N=Number of documents

df_t= number of documents with term t

In other words, $w_{t,d}$ assigns to term t a weight in document d that is:

1. highest when t occurs many times within a small number of documents thus lending high discriminating power to those documents when similarity between documents is observed;
2. lower when the term occurs fewer times in a document, or occurs in many documents (thus offering a less pronounced relevance signal);
3. Lowest when the term occurs in virtually all documents.

v) Dimension Reduction using NMF

Let V be a non-negative matrix of dimension: n x m which is the term-document matrix constructed after term extraction and weighting.

NMF algorithm decomposes the matrix into a low rank approximation matrix

$$V_{n\times m} \approx W_{n\times r}H_{r\times m},$$

Where W and H are NMF factors and all entries in V, W and H are to be non-negative. r, m, n represent the rank of the matrices and r is chosen to satisfy (n+m)r<nm

vi) Ranking/Similarity Measurement

A linear combination of two distance functions was used to measure the degree of similarity between the marking scheme and student's answer. The distance functions used are the cosine similarity rules and Euclidean distance as shown in equation 3.17, 3.18 and 3.19 respectively.

$$cosSIM\left(\vec{A}.\vec{B}\right) = cos(\theta) = \frac{\vec{A}.\vec{B}}{\|\vec{A}\|\|\vec{B}\|} \quad (3.17)$$

$$ED\left(\vec{A}.\vec{B}\right) = 1 - \sqrt{\sum_{i=1}^{k}(A_i - B_i)^2} \quad (3.18)$$

$$SIM\left(\vec{A}.\vec{B}\right) = \alpha.ED\left(\vec{A}.\vec{B}\right) + (1-\alpha)cosSIM\left(\vec{A}.\vec{B}\right) \quad (3.19)$$

Where $\alpha \in (0,1)$

IV. RESULTS AND DISCUSSION

In this research work, 30 answer booklets were collected in an introductory course to Artificial Intelligence. Students' responses in selected questions classification such as define, explain, highlight, categories, list were selected and fed as input into the designed Electronic Assessment application. The lecturer's response as seen on the marking scheme for selected questions is also supplied as input. The scripts were photocopied and copies given to three different lecturers to mark using one of the lecturer's marking scheme and the average score on each question computed. This was done to minimize the effect of human emotional and cognitive scoring attribute which sometimes feature in human assessment.

Table 4.1 shows the assessment done on each student's script using LSA, NMF and manual assessment by the course lecturer. The manual assessment was done by two course lecturers and the average of the two was computed as the manual score. The obtainable marked is between 0 and 1. Figure 4.1 is a graph showing the correlation between the human grade and the machine grade while Figure 4.2 shows the comparative performance between the NMF, LSA and the Manual Assessment

Table 4.1: Assessment Result using LSA and NMF and Lecturer's Manual Assessment

	LSA	NMF	MANUAL	DIFF LSA	DIFF NMF
Lecturer	1	1	1	0	0
student001	1	1	1	0	0
student002	0.772565	0.591608	0.5	0.272565	-0.09161
student003	0.897386	0.707817	0.875	0.022386	0.167183
student004	0.793917	0.613572	0.625	0.168917	0.011428
student005	0.719928	0.454264	0.375	0.344928	-0.07926
student006	0.817719	0.65938	0.625	0.192719	-0.03438
student007	0.864511	0.717137	0.75	0.114511	0.032863
student008	0.900387	0.83666	0.875	0.025387	0.03834
student009	0.732727	0.613572	0.625	0.107727	0.011428

student010	0.844252	0.701646	0.75	0.094252	0.048354
student011	0.917098	0.671584	0.875	0.042098	0.203416
student012	0.728421	0.553399	0.5	0.228421	-0.0534
student013	0.811911	0.676123	0.875	-0.06309	0.198877
student014	0.946399	0.858116	0.875	0.071399	0.016884
student015	0.891975	0.590569	0.75	0.141975	0.159431
student016	0.823175	0.652929	0.625	0.198175	-0.02793
student017	0.918407	0.782624	0.75	0.168407	-0.03262
student018	0.955241	0.821584	0.875	0.080241	0.053416
student019	0.671288	0.357833	0.5	0.171288	0.142167
student020	0.832479	0.74162	0.875	-0.04252	0.13338
student021	0.900387	0.63666	0.875	0.025387	0.23834
student022	0.603723	0.389898	0.5	0.103723	0.110102
student023	0.535886	0.358569	0.375	0.160886	0.016431
student024	0.725549	0.632456	0.625	0.100549	-0.00746
student025	0.08473	0.105409	0.125	-0.04027	0.019591
student026	0.569862	0.438529	0.5	0.069862	0.061471
student027	0.725549	0.632456	0.625	0.100549	-0.00746
student028	0.863044	0.737865	0.75	0.113044	0.012135
student029	0.765307	0.667424	0.625	0.140307	-0.04242
student030	0.678177	0.447723	0.5	0.178177	0.052277
student031	0.597768	0.387298	0.375	0.222768	-0.0123
student032	0.150827	0.150693	0.25	-0.09917	0.099307
student033	0.630196	0.596285	0.625	0.005196	0.028715
MEAN DIFFERENCE				0.100612	0.043138
MEASUREMENT OF ACCURACY				89.93885	95.68618

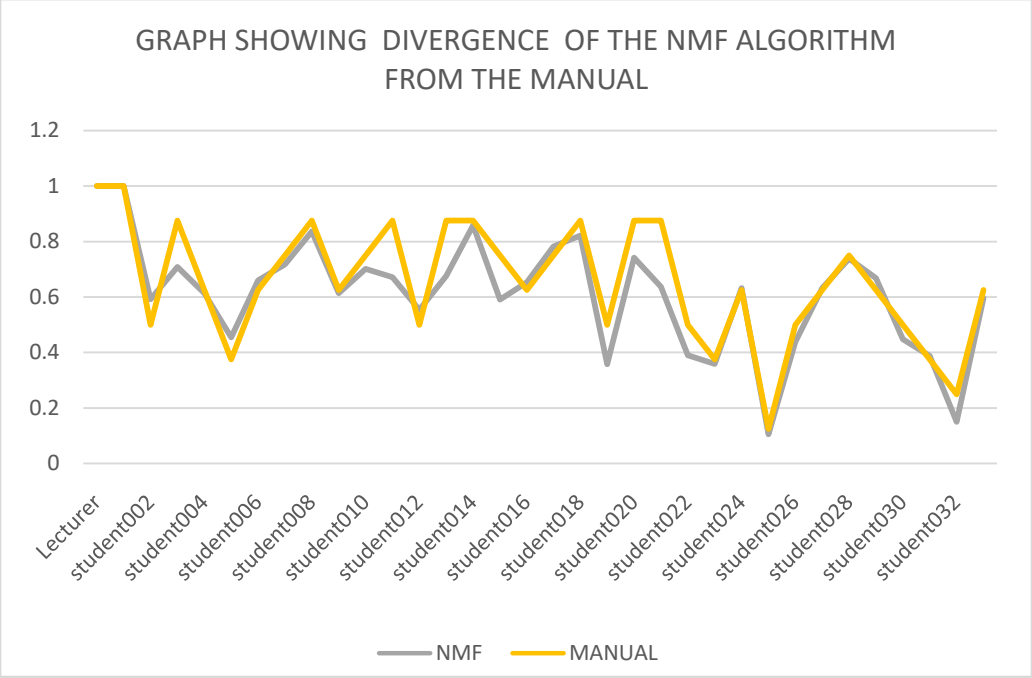


Figure 4.1: Graph showing divergence of the NMF Algorithm from The Manual Marking

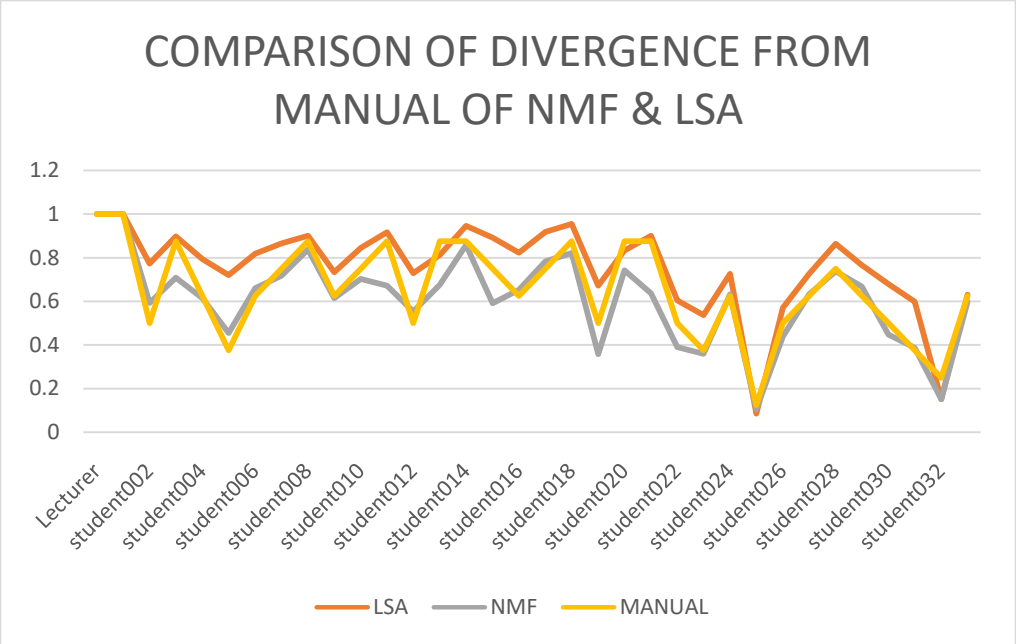


Figure 4.2: Graph showing between LSA and NMF in relation to Manual scoring based on divergence

V. PERFORMANCE EVALUATION

A performance evaluation of the system was conducted using the following performance metrics

a. Mean divergence
b. Measurement of Accuracy
c. Pearson Correlation Analysis

ii) Mean divergence and Measurement of Accuracy

Mean Divergence shows the ratio by which the automated system score differs from the manual score at ± value. The difference is as a result of human emotional and cognitive scoring attribute associated to the manual system. Therefore, the divergence variance V of result of a question number q for n computed as:

$$DF_{i,q} = |S_I - M_I|_q \quad (4.1)$$

$$V_q = \frac{\sum_i^n DF_{q,i}}{n} \quad (4.2)$$

Where DF is set of score differences, M is scores obtained from manual process, S is scores obtained from automated system respectively and i represents distinct student in set n.

The sum of the differences between the manual score and the system score $\sum_i^n DF_{q,i}$ = 0.486698687

The Mean Divergence $V_q = \frac{1.466699}{33}$ =0.044445

Accuracy of the system is calculated as:

Accuracy = 100 - (0.044445× 100) = 95.5554585255477

iii) Pearson Correlation Analysis

The performance of the new system compared to the old manual system was measured by carrying out Pearson correlation analysis. Pearson's correlation determines the degree to which two linearly dependent variables are related. This is done by computing the Pearson correlation coefficient r between the human grade scores and the NMF and LSA graded scores, using the obtained values from Table 4.1. A correlation of 0.921728 was observed between the manual scores and the NMF graded scores while 0.88729 was observed between the manual scores and LSA which indicates that NMF is a better assessor when compared to LSA and the scores obtained using NMF is closely related to the manual score.

Table 4.2: Pearson Correlations between Machine Grade and Human Grade

	LSA	NMF	MANUAL
LSA	1	0.911355	0.88729
NMF	0.911355	1	0.921728
MANUAL	0.88729	0.921728	1

VI. CONCLUSION

Electronic Assessment System that is based on comparisons between a lecturer's marking scheme and students' response to questions was developed. The paper also reported an experiment that evaluate the performance of NMF as a tool

for Electronic Assessment. A preliminary evaluation of the algorithm was carried out by comparing results of the human grader with that of the machine grader using the Pearson correlation statistics, mean divergence and measurement of accuracy. The results of the experiment showed that the mean difference between the human score and NMF generated score, the accuracy level and the Pearson Correlation Coefficient(r) are indications of a strong correlation between the manual scores and the NMF scores. Future evaluation could be geared towards investigating the performance of other techniques such as the hybridization of evolutionary algorithm with the information retrieval techniques on assessment of free text documents.

REFERENCES

[1]. Casalino G., Del Buono N., Mencar C. (2016) Nonnegative Matrix Factorizations for Intelligent Data Analysis. In: Naik G. (eds) Non-negative Matrix Factorization Techniques. Signals and Communication Technology. Springer, Berlin, Heidelberg

[2]. Farrús, M., & Costa-jussà, M. R. (2013). Automatic evaluation for e-learning using latent semantic analysis: A use case. The International Review of Research in Open and Distributed Learning, 14(1), 239-254.

[3]. Maria De Marsico, Andrea Sterbini and Marco Temperini (2016), "Grading Open-Ended Questions in an Educational Setting, via Non-exclusive Peer Evaluation", State-of-the-Art and Future Directions of Smart Learning p.357

[4]. Martin, D. I., Martin, J. C., & Berry, M. W. (2016). The Application of LSA to the Evaluation of Questionnaire Responses. In Unsupervised Learning Algorithms (pp. 449-484). Springer, Cham.

[5]. Senthil Kumaran, V., & Sankar, A. (2015). Towards an automated system for short-answer assessment using ontology mapping. International Arab Journal of e-Technology, 4, 17-25.