Journal of Emerging Technologies (JET)
Volume 5, Issue 1 (2025)
ISSN: 2710-0189 (Print) | 2710-0197 (Online)
Website: https://journals.jozacpublishers.com/jet/
https://doi.org/10.5281/zenodo.14750865

Automatic marking of descriptive questions of online examinations using NLP

Rogers Bhalalusesa[1*]
[1]*Maths and ICT Department, The Open University of Tanzania, Tanzania, balalusesa@gmail.com*

*Corresponding author: balalusesa@gmail.com

Abstract: Online examinations are increasingly being integrated into universities using low-level questions on Bloom's taxonomy such as True-False and Multiple-Choice questions. Unfortunately, the summative assessments in Tanzania rely on descriptive questions that are higher in Bloom Taxonomy which are not easily marked by computers. To achieve a true online examination in Tanzania a system that can mark descriptive questions is necessary. This paper presents a Natural Language Processing (NLP) Technique employing Cosine Similarity that can automatically mark descriptive questions without human intervention. The Cosine Similarity compares between candidate answer and the marking scheme after having them transformed into mathematical vectors by the NLP technique called Term frequency–inverse document frequency (TF-IDF). Evaluation done by comparing the marking of short answers by NLP-based online exam system and Moodle as well as feedback from the Computer Science students from the Open University of Tanzania who used the NLP based online exam system indicates positive performance. 86% of questions were marked correctly by the NLP system and got a similar score to the one in Moodle. 66% of students to trusted the score returned by the system while only 4% did not trust the score which confirmed the acceptability of Cosine Similarity for marking mark descriptive questions automatically.

Keywords: Cosine similarity, E-assessment, NLP, Online examination, TF-IDF

1. Introduction

Assessments are done by the university to confirm if the students have achieved the goals set in the curriculum of the degree programme the students have registered for. Assessments are classified as formative or summative. Formative assessments are done during the teaching period as coursework and Summative assessments are usually done at the end of the teaching of the course to test the understanding of the students (Ghouali et al., 2020). Formative assessment can include closed questions (objective types of questions) such as Multiple-Choice Questions and True / False Questions only but a true summative assessment ought to have all types of questions including open and closed questions higher on Bloom Taxonomy to test the overall understanding of the course (Qasrawi & BeniAbdelrahman, 2020; Semlambo et al., 2022). Open-ended questions consisting of descriptive and text questions are highly favoured in Summative assessment because they test learners in all three domains of learning namely: Cognitive, Affective, and Psychomotor Learning (Aninditya & MA Hasibuan, 2019). With an increasing number of students in e-learning environments, Universities are adopting e-assessments to address the challenges of conducting assessments in a large number of students (Muzaffar et al., 2021).

Online examination is a form of E-assessment that is done using the Internet. E-assessment is a way of conducting examinations using information technology facilities. It can be called in different terminologies including Computer Based Test (CBT), Computer Based Assessment (CBA), E-examination, and E-

Assessment. Students read questions and give answers using computer input devices (Ghouali et al., 2020; Semlambo et al., 2022). Online examinations can be carried out anywhere and anytime provided that there is a secure examination environment. Students can be allowed to attempt exams at their time convenience unlike paper-based which is scheduled at one point in time (Stowell & Bennett 2010). Online Examination Reduces the time it takes to distribute exams and the time taken to mark and release the results to students. They are cost-saving because they reduce resources and human capital that is used to administer examinations. Everything in the online examination is kept securely in the information systems and can be retrieved easily and results can be released quickly (Sometimes instantly) so that the students can get feedback quickly which increases quality assurance and trustworthiness of assessments (Ozden, 2004; Semlambo et al., 2022).

Although online examinations are steadily being integrated into e-learning in other developed countries, their adoption is mainly for objective-type questions consisting of multiple choice and true answer questions (Ghouali et al., 2020) (Nandini & Uma Maheswari, 2020). In Tanzania however objective-type questions are usually done during the formative assessment period (coursework). Summative assessments usually contain questions higher on the Bloom taxonomy scale which mainly consist of descriptive questions. Since descriptive questions are difficult to mark by computers (Genelza, 2024) they are left to the lecturers to mark at a later time which introduces a delay in giving feedback to students. This results in online examination being only partially used in Summative assessments which in turn reduces its efficiency.

Marking the descriptive questions is difficult because computers being machines are not very effective in making textual comparisons between the student answer and the marking scheme. Automatic marking of descriptive questions is made difficult since not all possibilities of the way the candidate will answer the question are known. Therefore, usually, the marking scheme fails to capture the answers from the students and in turn marks many of the questions wrongly. To implement a fully online examination in summative assessment marking of descriptive textual questions ought to be done automatically by the computers.

The best way to mark descriptive questions automatically by computers is by comparing the student's answer with all the possible versions of the marking scheme answer mathematically. Such similarity comparison can be done using cosine similarity (Xia et al., 2015) after the words have been converted to mathematical vectors using Natural Language Process (NLP) Techniques. This paper presents one of the NLP techniques to mark the descriptive (textual) questions using a TF-IDF and Cosine Similarity. The paper will discuss various related work that has been done in online examinations, essay-type marking, and textual classification using NLP. The paper will discuss How TF-IDF is used to convert text into vectors of binary numbers and later on show how cosine similarity can compare the vectors to show the degree of similarity of the word vectors which in turn is the score of the student's answer. The research methodology of the paper will be described and a discussion and analysis of the results of experiments set to evaluate the NLP technique will be given at the end.

2. Literature review

2.1. Using NLP in Automatic Marking of Essay Questions

Automatic marking of essays involves working with textual information from users and can be solved by a branch of artificial intelligence called NLP. NLP in association with text mining can handle and analyse vast amounts of data from natural language and be used in comparisons between two distinct text paragraphs. (Ferreira-Mello et al., 2019; Jiang & Lu, 2020). NLP comprises the techniques that have been used to contextually evaluate and entail the educational data by providing an interpretive interface between humans and machines, (Barb & Kilicay-ergin 2020; Shaik et al. 2022). Computers cannot understand words and need numerical values of the texts to make it easier for them to compare the texts. Once the texts are converted into numerical values comparisons can then be made using Mathematical formulae.

There are different ways to convert text into numeric including simple data structures such as numbers and complex data structures such as vectors, sets of graphs(networks), and trees. Two documents may not be similar in length conversion of words that considers both magnitude and directions is ideal in representing text in numeric value. This is achieved through vectors. The type of numerical data that can represent text is Vectors. Vectors have both magnitude and direction. The text to be compared can easily be transferred into vectors with both magnitude and directions so that the mathematical formulae can be applied to the vectors.

Text similarity uses both semantic and lexical similarity. Lexical similarity is higher for words that have a similar character sequence. While semantic similarity is higher if words have the same meaning. NLP uses various mathematical formulae to calculate the lexical similarity measures between converted text such as between Manhattan Distance, Euclidean Distance, Dot Product, and Cosine Similarity (Wang & Dong, 2020).

The method with higher accuracy for textual similarity is Cosine Similarity because it is not affected by the magnitude of the texts to be compared. Cosine similarity compares both the direction and magnitude of vectors. Cosine Similarity can be thought of as Normalized Dot products because it normalizes the vectors so that their magnitude is between 0 and 1.

The high accuracy of Cosine Similarity has made marking of descriptive questions by comparing their vectors effective, however, there is the challenge of the way the text is converted into mathematical vectors. The process of converting the text into vectors is called word embedding. This could be done by converting the characters or words. Converting the characters would result in large (long) vectors (1 X N) that are not suitable for processing (since they require higher processing power). The efficient conversion is by using words and hence for Text similarity comparisons Word Embeddings are used. There are different variations of creating word embeddings to choose from but the best one must be the one that creates numeric vectors that bring about high accuracy during cosine similarity comparisons.

One method of creating word embeddings is One Hot encoding. This encoding is a representation of categorical variables as binary vectors. It is a way of transforming data into vectors where all components are 0, except for one component with a value of 1. All words making up a text are used to create a word vector representation for every position. Other words are given the value of 0 except the word being represented. This results in a dense vector since if we have a sentence with 10 words, we will end up having a vector of 10 X 10 as the representation of the sentence.

Another method of creating word embedding is the concept of a bag of words which simply means tokenizing all the words in the text to be learned(corpus) (Yan et al., 2020). Bag of words (BoW) simply means just taking each word and counting how many times it occurs in each document. Therefore, the string is transformed into a vector with the length as the number of what you have. This is widely known as Frequency-based embedding. This method uses a tokenization approach (tokenizing every word from the sentence with its frequency). Different improvements have been made to improve its performance like TF-IDF and Co-Occurrence Matrix.

TF-IDF has been used as the core background process in Word2Vec which is the latest NLP technique for creating word embedding. Word2Vec is a family of model architectures and optimizations that can be used to learn word embeddings from large datasets. Word2Vec uses Neural Networks to learn the relationship between words and create a binary vector of the text. It works by “vectorizing” words. Its input is a sentence and its output is a set of vectors: feature vectors that represent words in that corpus. Word2vec numeric vectors are shorted than one hot encoding and in turn reduce the complexity of numeric value. Additionally, the word2vec numeric value takes into consideration the position of the occurrence of the word in the sentence. Word2Vec creates word embedding by predicting words that are related to each other from the corpus (Set of all words available in the text). However, word2vec is more suitable for only text and phrases and does not perform well when it comes to transforming the whole sentence into a vector. A variation of word2vec which transforms whole sentences into vectors with considerable precision is called Doc2Vec (Huang et al., 2019). Doc2Vec uses paragraphs as input but just like its mother algorithm word2vec, the output vectors created are susceptible to negative values when compared to cosine similarity. This results in negative Cosine Similarity which makes it difficult to transform the actual score to the examination score between 0 and 100. Other methods of Word Embedding are used in machine learning such as Glove, Fast2Text, and Node2Text however they are also catered for word similarity rather than sentence and document similarity because of negative vectors and extra computation to derive the sentence vectors in the first place. Therefore TF-IDF still has the upper hand when transforming the sentence/paragraph into vectors compared to word2vec and other machine learning techniques and will the embedding that will be used in this study to transform the text into numeric vectors.

2.2. Related Work

The paper by Kumamoto University proposes the use of cosine similarity measurement by incorporating semantic checking between dimensions of two-term vectors to increase the similarity value (Kitasuka et al., 2012). It uses lexical semantics from WordNet Corpus to equalize Vector Dimensions for terms that are different in length. Cosine similarity is then applied to vectors with equal dimensions to get vectors that ought to boost their similarity measure. The Automatic marking discussed in our paper has employed a similar technique of increasing semantic meaning but has used its own local Database instead of the WordNet Corpus

In the automatic grading system in India, Cosine Similarity is used with another NLP technique called World Movers Distance (WMD) to mark handwritten short answers captured by OCR (Jain et al., 2022). The converted text is converted into vectors using the BERT algorithm and Word2Vec for WMD. The study used the average of the WMD and Cosine Similarity as the final answer but its accuracy may be questionable since it is not certain how correct the text is derived from OCR. This study however uses input keywords and does not use Word2Vec which does not perform well when it comes to converting full-sentence.

` The paper by Xia et al.(2015) explores a cosine similarity ensemble (CSE), that enhances the correctness of similarity measures by enlarging the angles between patterns by changing the initial point of the word vectors while keeping the terminal point constant. This way the similarity measure becomes the weighted average from the measure of similarities with different vectors produced from changing the initial position. The only drawback of this work is that the way the vectors are produced still relies on one hot encoding which may result in long vectors that will only work for paragraphs with few words.

In the research by Udayana University Badung, Word2Vec and cosine similarity are used to mark essay questions (Abimanyu & Sanjaya, 2020). The answers from students together with the correct answers are converted into vectors by Word2Vec and similarity between them is calculated using cosine similarity to produce the final grade. The final grade is compared to the instructor's score to show the accuracy of the research. The research however used Bahasa Indonesia Language which is different from Word2Vec since it was trained in English and therefore removal of common words which would increase similarity could not be easily done. In this research the language used was English and also the technique of removing the common words was used during preprocessing to create truly unique vectors between student answers and the marking scheme.

The research work by Odzen tried to explore the requirements for an online examination from learners' perspective and found out that most learners preferred an online examination system which has the following features (Ozden, 2004). These were used as the baseline for developing the online system. The researcher was able to develop an online examination system from scratch but it did not include automatic feedback for short answer and essay type questions. The research done in Tanzania indicated that most of the lecturers are unhappy with the activity of marking and require significant attention (Namabira et al., 2022). The study indicates that both students and teachers prefer online assessment because of flexibility and timely feedback among other reasons. The research calls for the adoption of e-assessment in Tanzania to lessen the burden on the teacher when it comes to marking. This shows that more effort is needed to improve the automatic marking so that the lecturers can focus on other areas.

A study by Semlambo et al from the Institute of Accountancy in Arusha, Tanzania highlighted the adoption of e-assessment in Tanzania for low-level questions of bloom taxonomy which results in just measuring factual knowledge rather than comprehension (Semlambo et al., 2022). The perception of lecturers as reported by the author considered the online exam to be less inferior to traditional exam because this kind of exam can result in students passing exams from just guesswork since most of the results are presented to them for selection. To the authors, this was one of the academic barriers to acceptance of online examination and other types of questions should be presented for the assessment to be complete.

2.3. Converting Document into Numeric Vectors using TF-IDF

Term frequency–inverse document frequency TF-IDF is an NLP method arising from Bag of Words (BOW) that takes a text document consisting of either a sentence, paragraph or several paragraphs and outputs a vector in numerical numbers representing all the text in the document. All words that appear in the sentence are put together in a corpus (Bag of Words). TF-IDF vectorization involves calculating the TF-IDF score for every word in your corpus relative to that document. In TF-IDF the importance of a term is inversely related to its frequency across documents. Term Frequency (TF) gives us information on how often a term appears in a document and Inverse Document Frequency (IDF) gives us information about the relative rarity of a term in the collection of documents. By multiplying these values together, we can get our final TF-IDF value. The lower the TF-IDF score the less important and less relevant the term is and vice versa. TF-IDF uses nonnegative vectors and therefore the Cosine similarity derived from its vectors is always between 0 and 1(Godfrey et al. 2014). This allows for a simple translation of exam question scores between 0 and 100%. Based on Wang & Dong (2020), The equation for TF-IDF calculation is given by

$$tf\text{-}idf\ (w, d, D) = tf\ (w, d) \times idf\ (w, D)$$

$$tf\ (w, d) = Freq\ (w, d)$$
$$idf\ (w,\ D\ )\ =\ log\frac{|D|}{N(w)} \quad \text{eq (1)}$$

whereby Freq (w, d) indicates how often a word is used in the document, |D| indicates the number of the document, and N (w) is the number of words that appear in the document (Wang & Dong, 2020).

2.4. Technical Challenges of Vector Representation of Textual Documents

TF-IDF can perform well in converting- sentences into vectors. However, there exist three ways in which the user input can result in writing answers whose conversion into vectors and similarity score may be impacted. The first common words that are used in English words can be found in both student answer and the marking scheme which do not carry any weight in comparison. Words like pronouns ('he', 'she' 'it'), connecting words ('and', 'in'), and many more. The full list of words is kept in Appendix 1. If these words are used abundantly in the answers, then the score returned will be based more on these words instead of the words that represent the true semantic meaning of the answer

Second, the interface used may contain some mark-up language that may add more unwanted information to the text entered as the correct answer or the student answer and marking scheme. Examples of such markers are ('<b>', '
', etc). The mark-up can result from the lecturers creating a marking scheme by copying answers from the internet or other sources with mark-up language. In regards to the student's answer. Pasting of information is disallowed in the text area to make the student's answer not contain mark-ups and reduce plagiarism

Fourth the words that are used may be different from the marking scheme and student answer but have the same meaning. When this happens the TF-IDF will generate different vectors and the cosine similarity score returned might be inaccurate. Take for example the marking scheme lecture used the word 'build' and the student used the word 'make'. The two words have the same meaning however computers would see them as different. This part takes care of the semantic similarity component which is lacking when only Cosine Similarity is used on its own.

Additionally, the descriptive questions require the students to give explanations and not just write unmeaningful phrases. Some students would try to write random words to cheat the comparison when they learn that the system marks by using only word comparisons.

2.5. Addressing the Technical Challenges of Vector Representation of Textual Documents

All the challenges above can impact the score returned by the online exam system. The online exam system needs to address the above issues for scores to be more accurate. The first approach is on the Interface to enter the answers used by the lecturer and the candidates. Limiting which texts can be entered in the Interface will reduce the number of mark-up letters that can be found in the answers. The length of the answer needs to be considered so that the students can attempt the question instead of writing random words. A minimum number of words needs to be considered so that the student can be considered to have attempted the question and below that number the answer will be discarded. Furthermore, the student's answer should contain some pronouns or connecting words. The answer is also checked to see if it contains common English words to make sure that the learner is actually entering a sentence and not keywords only. This is to prevent learners from entering only keywords. Limitation can be applied using jQuery and JavaScript.

The second approach is by pre-processing the input data before conversion and comparisons. Pre-processing of input data is the key activity in NLP since it is known that most of the input data in so many problems contain a lot of noise. The noise data need to be removed or transformed to increase the accuracy of the score returned. The first part of pre-processing is by removing common words. Examples of common words are 'a', 'the' and others. The availability of so many common words can make the texts appear to be similar while they are not. Common Words are well-defined in the Python Package from Scikit-learn. The second part is from the mark-up text such <br`> that could have been added automatically in the text entered by the students during typing or during the preparations of the marking scheme by the lecturers. The mark-up texts can be removed by adding them to the set of common words so that they cannot be used in the conversion of words to vectors.

The third approach handles the semantic relationship of words. When the lecturer enters the answer, all its synonyms are generated so that we can different varieties of the answers. That means for one answer with N words of a paragraph with each word having S synonyms we can have different N X S versions of answers for comparison. This will assist in semantic similarity measurements since cosine similarity only looks at lexical similarity between texts. This is similar to other researchers who have used synonyms from the WordNet corpus to enhance the similarity of text classification (Kitasuka et al., 2012).

2.6. Cosine Similarity for Text Classification

Cosine similarity is a popular NLP method in information retrieval and text mining that calculates the similarity between two documents represented by vectors using the cosine angle between them. Cosine similarity gets the value from the Cosine of the angle between the text when you represent them as vectors in a vector space. The cosine value is 1 for very similar text documents and 0 up to -1 for dissimilar text documents. Negative values indicate similarity in opposite directions which means the text documents are dissimilar. The formulae to calculate Cosine Similarity is given by.

$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}},$$ eq (2)

Source: SoftMaxAI (2023)

The example below shows how Cosine Similarity is used to calculate the similarity score between two sentences represented by doc_1 and doc_2.

doc_1 = "Data is the oil of the digital economy"

doc_2 = "Data is a new oil"

The Vector representation of the document will be as follows

doc_1_vector = [1, 1, 1, 1, 0, 1, 1, 2]

doc_2_vector = [1, 0, 0, 1, 1, 0, 1, 0]

	data	digital	economy	is	new	of	oil	the
doc_1	1	1	1	1	0	1	1	2
doc_2	1	0	0	1	1	0	1	0

$$A \cdot B = \sum_{i=1}^{n} A_i B_i$$
$$= (1 * 1) + (1 * 0) + (1 * 0) + (1 * 1) + (0 * 1) + (1 * 0) + (1 * 1) + (2 * 0)$$
$$= 3$$

$$\sqrt{\sum_{i=1}^{n} A_i^2} = \sqrt{1 + 1 + 1 + 1 + 0 + 1 + 1 + 4} = \sqrt{10}$$

$$\sqrt{\sum_{i=1}^{n} B_i^2} = \sqrt{1 + 0 + 0 + 1 + 1 + 0 + 1 + 0} = \sqrt{4}$$

$$cosine\ similarity = cos\theta = \frac{A \cdot B}{|A||B|} = \frac{3}{\sqrt{10} * \sqrt{4}} = 0.4743$$ eq (3)

Source: SoftmaxAI Pvt Ltd. (2023)

2.7. Effective Comparison of Text Using Cosine Similarity

Different settings are applied to short answers and Essay-type questions to make sure that the comparison between the text is done correctly. Short-answer questions usually have only one answer/point and essay-type questions contain many points/answers. For short answer questions, we have 1 paragraph from students and M sentences from the marking scheme each of which carries the weight of 100%. Lecturers are allowed to enter many variations of the marking scheme answer separated by sentence. Additionally, more versions of the marking scheme answers are generated using S synonyms. The total number of marking scheme

sentences will be M x S. Each sentence will be compared with the student's answer (student paragraph) and the best score is the one which will be taken as the score for that question. That means there will be M x S comparisons

For the essay-type question, all text entered by the students is split into paragraphs (P). There will be more than one marking scheme sentence (M) each marking scheme will generate a version of answers from synonyms (S) leading (M x S) marking scheme answers. Each paragraph is compared with all the marking scheme sentences (M x S) and the one with the highest score is taken as the score for the paragraph. All the scores are summed up and divided by the number of marking scheme scores to get the average mark of the essay question.

3. Methodology

Design science whose output is an artefact was used in this study (Gregor & Hevner, 2013). The output was the System developed in Python Django that could mark essay questions automatically. It was experimental research and a control experiment was used to test if the essay questions were marked with similar accuracy compared to Moodle LMS marking. The study population was ICT students at the Open University of Tanzania (OUT). OUT was chosen as it is the only full-time Distance-based university that uses E-Learning in all its major activities in Tanzania and intends to be an online university by 2025 based on its Strategic Plans. The use of online examination at OUT would be one of the steps taken to make it go fully online. The study used both primary and secondary data from OUT. Secondary data were collected from the Moodle database which contains questions and answers for courses available.

These were remarked again in the system developed (using cosine similarity in marking short answers and essay questions) and their scores were compared with Moodle scores to check the effectiveness of the system. Masters and Undergraduate ICT students from OUT attempted the exam and later on answered the questionnaire in regards to the correctness of the automatic marking. Primary data was the feedback from the students who attempted the Timed Test exam from the examination system developed for OUT that employs cosine similarity. A total of 57 Students (40 from B Sc ICT and Data Management and 17 from M Sc Computer Science and M Sc ITM attempted the exams which included essay-type questions and later evaluated the system based on the questionnaire). The questionnaire was quantitative and used a five-point Likert scale. A similar approach of using a 5-point Likert scale has been used widely such as the study done by researchers (Asaju & Ogar, 2022). Data was analysed in Python using Jupyter Notebook. Pandas and Matplotlib were used to provide visual feedback from the data collected (Lavanya et al., 2023).

4. Result and discussion

The questions and answers from Moodle for the Open University were downloaded and graded using the online examinations system. The scores from the online examination system were compared with the scores from Moodle LMS. Precision was calculated based on the number of questions that the scores matched between the Online examination system and Moodle against all scores. The full results are shown in Table 1. Precision stood at 75% and recall at 71% which brought the F-Score (Accuracy) to 73%. A similar technique was also used by Abimanyu & Sanjaya (2020) when they were validating their examination system in Indonesia.

Table 1: Summary of Performance of Automatic Marking of Short Answer Questions

No	Item	Count
1.	Total Number Questions/ Answer Pairs (samples)	105301
2.	Number of Questions/Answers graded correctly (True Positive)	39197
3.	Number of Questions/Answers graded correctly (True Negative)	37922
4.	Number of Questions/Answers graded wrong (False Negative)	15439
5.	Number of Questions/Answers graded wrong (False Positive)	12743
6.	Total Number of Questions/Answers graded correctly correct (True Positive + True Negative))	77119
7.	Number of Questions/Answers graded correctly wrong (False Negative + False Positive)	28182

Table 2: Accuracy of Online Examination System for Marking Short Answer Questions

1.	Item	Score (Ratio)	Score (Percentage)
2.	Precision	0.75466	75.47
3.	Recall	0.71742	71.74
4.	f_score	0.73557	73.56

After studying the questions and answers it was noted that most of the questions and answers where the online system graded differently from Moodle was when there was a possibility of a variation of the answers. Moodle failed to get the correct answer since Moodle does not have a Thesaurus installed but the online examination system uses Thesaurus and can capture the different words with similar meanings entered by the user. This shows an improvement in the online examination system but since there was no other metric to compare with the evaluation of its correctness was left to be done manually by students using the marking in a real examination environment.

The system was applied to live exams done by masters and undergraduate students of the ICT department of the Open University of Tanzania. Before looking at the various discussions regarding the feedback from the students who attempted the descriptive questions it was important to see the demography of the students who participated in the study. 90.5 % of the participants were male and 7.1 % were female. A total of 2% did not specify their gender. On the other hand, a large number of participants were aged between 31-40 which represents the group of people that grew up using computers in their tertiary education.

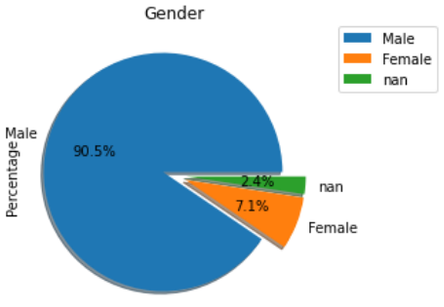

Figure 1: Distribution of Respondents Gender

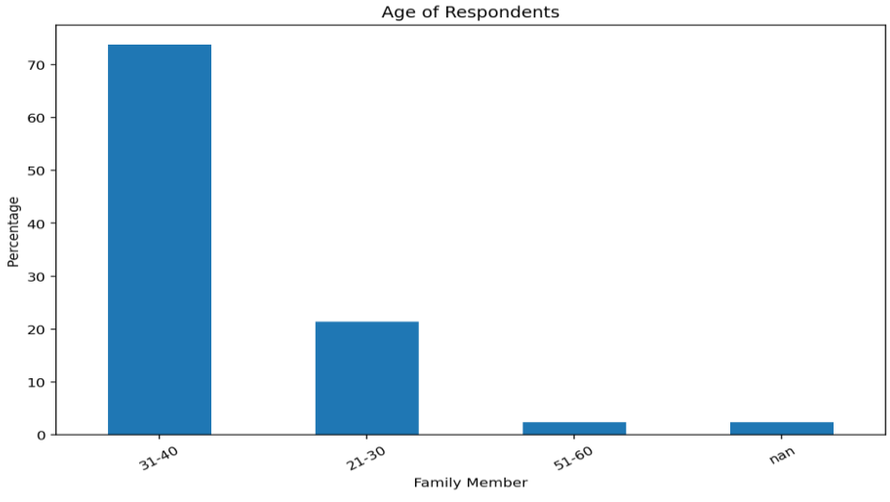

Figure 2: Distribution of Age of Respondents

It is equally important to know how they perceive the setting of online descriptive questions to relate their perception to the process of marking. The idea is that if the student finds it difficult to do the online descriptive question, then the student may not provide his best answer and not be happy at the end with the low score returned by the cosine similarity.

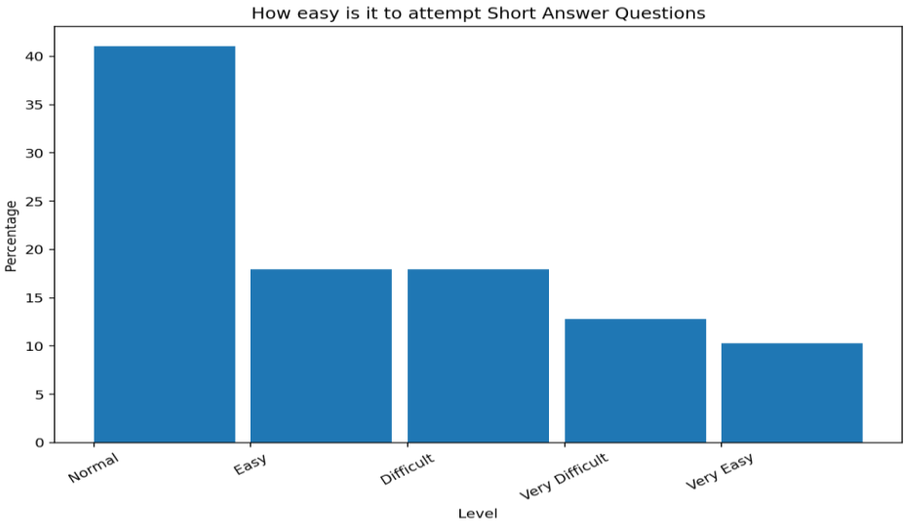

Figure 3: Feedback on the level of difficulty of attempting short answer questions

About 40% found it to be neither difficult nor easy to answer online short answer questions while 28% found it to be easy while 32% found it to be difficult. That means a large group chose to take a neutral ground and were undecided whether to commit whether it was simple or difficult to write answers in an online exam. Based on the fact that more students found it difficult than easier to write answers for short answer questions it was a little bit hard to conclude that the exam was easy. For essay-type questions almost 40% found it to be difficult while 22% found it to be easy and 38% found it to be neutral. This implies that it was a little bit difficult to write answers for essay-type questions and more assistance should be included to make sure that students can easily attempt the essay questions in an online examination setting. Both the results for short answer and essay type suggest that it is difficult to attempt descriptive answer questions in online examination settings.

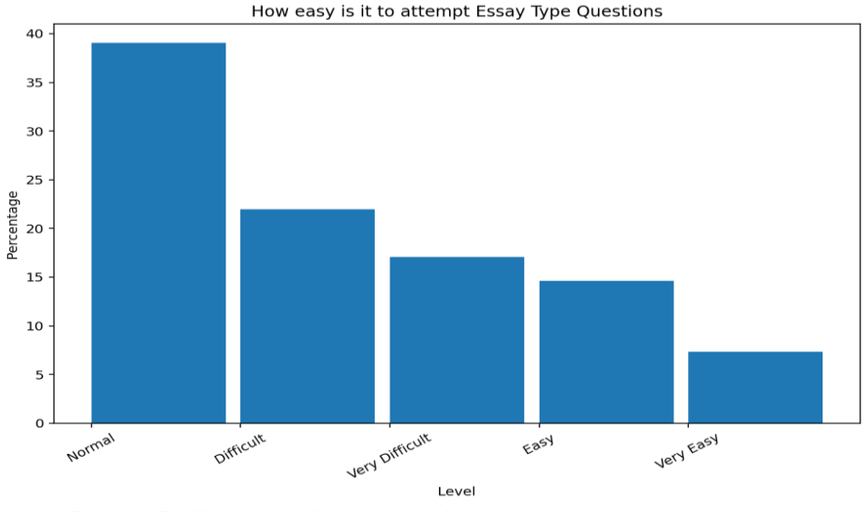

Figure 4: Feedback on the level of difficulty of attempting essay-type questions

Lastly, the learners were asked to rate the score returned by the online examination system. For short answer questions, the learners reported that the score graded by the system was largely correct. 66% of respondents reported the score to be correct while only 4 % said the score was not correct. 9.5% reported the score to be less than half correct. This shows a large part of the learners trusted the score returned by the online examination system.

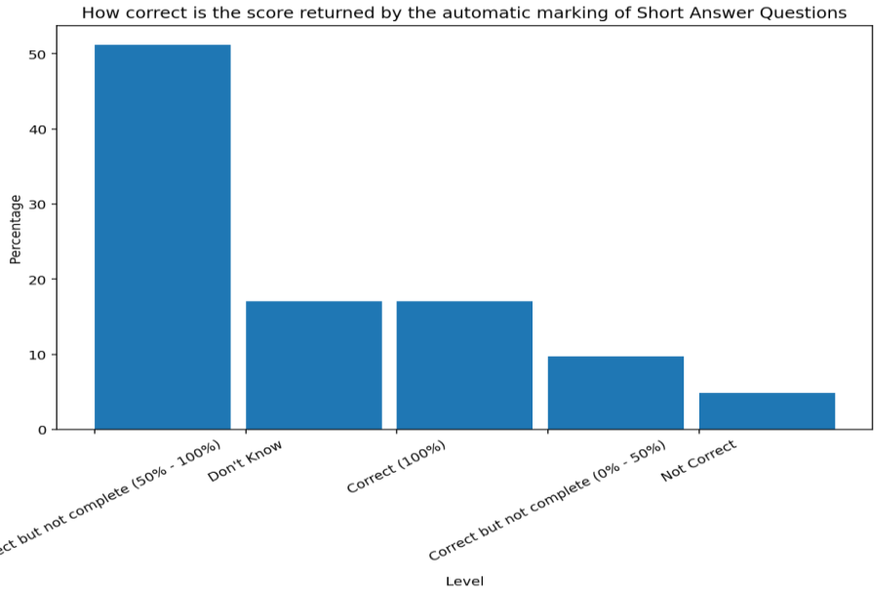

Figure 5: The correctness of the Online examination system in marking short answer question

Similarly, for essay-type questions the responses from the learners on the score returned by the online examination system were positive. More than 50% reported the score to be correct while only 4.8% reported the score to be incorrect. Of the people who reported the score to be correct only about 18% said it was completely correct while about 50 % percent said the score was correct but not 100% correct. This means that there is some level of acceptance of the score but not complete trust.

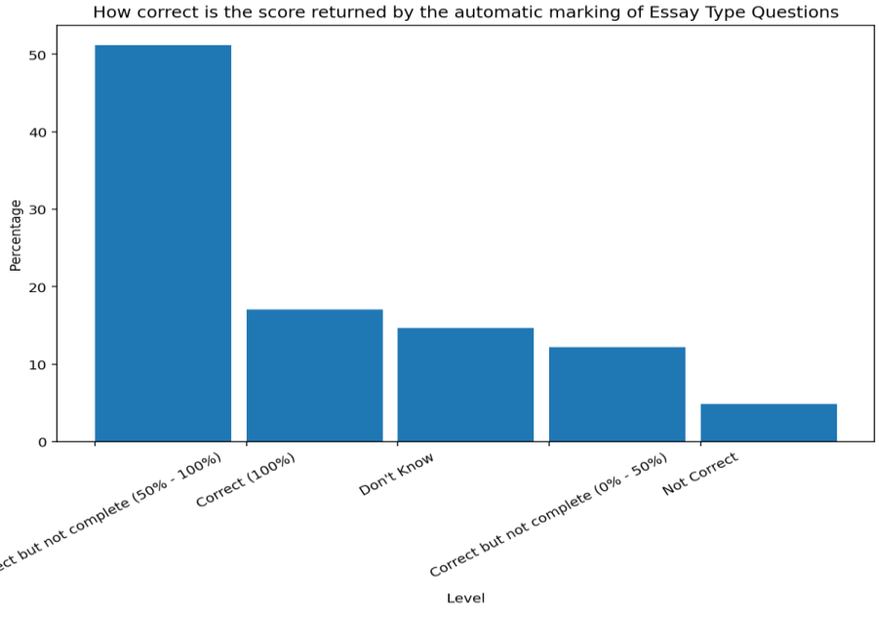

Figure 6: The correctness of the Online examination system in marking essay type question

Since short answer and essay type were set separately it was vital to see if the feedback from the students was consistent in both cases because a true online examination would contain both types of questions. Kendall correlation was used to check this since the data was on the Likert scale which is ordinal (De Raadt et al. 2021). Both the level of difficulty and feedback of the score were consistent for short answer and essay-type questions as they both returned a positive Kendall correlation of 0.768 and 0.891 respectively as shown in the table below.

Kendall's Tau Coefficient

$$\tau = \frac{n_c - n_d}{n_c + n_d} = \frac{n_c - n_d}{n(n-1)/2}$$

$$Where,\ n_c = number\ of\ concordant\ pairs$$
$$n_d = number\ of\ discordant\ pairs$$
$$n = number\ of\ pairs$$

Source: Kim et al (2014) eq (4)

Table 3: Correlation Coefficient between short answer questions and essay type question

Relation	Kendall Coefficient
The score returned for short answer questions against the score returned for essay question	0.768
The level of difficulty of attempting short answer questions against the level of difficulty of attempting essay question	0.891

More tests were done to validate the feedback regarding the correctness of the marking as reported by the students. Some studies have reported anxiety by students when attempting online examinations and it was important to see if this could also impact the perception of the scores that they receive. (Adenuga, Mbarika, and Omogbadegun 2019). There was a need to see if there was a relationship between how the students who reported the system as being usable also reported marking to be correct for both short answers and essay-type questions. If only those who reported the system to be usable saw the marking to be correct then the feedback would not be a true presentation since those who struggled to answer the exams did not trust the results returned. Two hypotheses were set for short-answer and essay-type questions as shown in table 4.

Table 4: Comparison between the level of difficulty and the score

Hypothesis	Chi-square Score Value	Decision
H1o: The score returned for short answer questions is dependent on the level of difficulty for attempting short answer questions	1.281	Reject H10
H4o: The score returned for essay-type questions is dependent on the level of difficulty for attempting essay-type questions	0.0172	Reject H20

Chi-Square was used to test if the feedback returned on the score was dependent on the usability of the system. Chi-square is used to check the relationship between a nominal variable and since the questionnaire was set using the Likert scale of nominal data then chi-square provided the best test (De Raadt et al., 2021). Since there was only one dependent and one independent variable for each text then the degree of freedom is 1 and the threshold (minimum) Chi-Square value for 0.05 to accept the hypothesis is 3.85. As seen on the table both hypotheses were rejected since the obtained P value in both cases (1.281 for Short Answers and 0.0172 for Essay type questions) were below the threshold. It was therefore not conclusive that there is a relationship between the level of difficulty of examining the score returned. It was then safe to say that the perception of the score returned by the automatic marker was independent of the difficulty of doing the exams and can be accepted the way it is.

In summary, the findings from the study have shown two main contributions in the area of online examination in Higher Learning Institutions. First, it has shown that NLP can be used to mark textual questions with a lot of degree of accuracy. Second, it has shown the trust the students have in the automatic marking of online examinations by NLP and revealed the readiness of students to use online examinations even for textual questions.

5. Conclusion

This study presented a way to mark descriptive questions using NLP. The study explored the way the text can be converted into mathematical vectors using TF-IDF for them to be interpreted well with cosine similarity. The prototype developed indicated that we can achieve high accuracy when marking using NLP similar to the results obtained from the Moodle marking tool. The findings indicated that students are ready to be assessed using automatic marking and institutions like OUT ought to try to improve the NLP systems so that they can benefit from them. Even with the pre-processing the of descriptive answer before conversion into a mathematical vector, a minimal possibility remains where the vector created is far different from the vector of the marking scheme because of the different words used by the learners and the marking scheme. The current NLP technologies still lack the semantic meaning when transforming text into vectors and more research is needed to improve the performance. Future work would involve enhancing semantic enhancement of TF-IDF using lemmatization and synonyms of lemmatized words to get more variations of both the marking scheme and student answers to boost up similarity score between them

6. Funding

This research paper received no internal or external funding

References

1. Abimanyu, C.G., & Sanjaya, A.N. E. (2020). Automatic Essay Answer Rating Using the Text Similarity Method. *Jurnal Elektronik Ilmu Komputer Udayana P-ISSN*, *8*(4), 463. Retrieved from https://ojs.unud.ac.id/index.php/JLK/article/download/53153/33557
2. Adenuga, K. I., Mbarika, V. W., & Omogbadegun, Z. O. (2019). Technical Support: Towards Mitigating Effects of Computer Anxiety on Acceptance of E-Assessment Amongst University Students in Sub Saharan African Countries. *IFIP Advances in Information and Communication Technology*, *558*, 48–72. https://doi.org/10.1007/978-3-030-20671-0_5/COVER
3. Aninditya, A., & MA Hasibuan. (2019). Text mining approach using TF-IDF and naive Bayes for classification of exam questions based on cognitive level of bloom's taxonomy. *IEEE International*. https://doi.org/10.1109/IoTaIS47347.2019.8980428
4. Asaju, K., & Ogar, O. B. (2022). The use of Information and Communication Technology (ICT) and its implications on academic excellence in Federal University Wukari, Taraba State. *Journal of Emerging Technologies*, 2(2), 85–94.
5. Barb, A. S., & Kilicay-Ergin, N. (2020). Applications of natural language techniques to enhance curricular coherence. *Procedia Computer Science*, *168*, 88–96.
6. De Raadt, A., Warrens, M. J., Bosker, R. J., & Kiers, H. A. L. (2021). A Comparison of Reliability Coefficients for Ordinal Rating Scales. *Journal of Classification*, *38*(3), 519–543. https://doi.org/10.1007/S00357-021-09386-5/TABLES/11
7. Ferreira-Mello, R., André, M., Pinheiro, A., Costa, E., & Romero, C. (2019). Text mining in education. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *9*(6), e1332.
8. Genelza, G. G. (2024). Unlocking the opportunities and challenges of using ChatGPT tools for educational services: *A narrative literature review. Journal of Emerging Technologies (JET)*, 4(1).
9. Ghouali, K., Benmoussat, S., & Cecilia, R. R. (2020). *E-assessment on the spotlight: Present and future prospects*. Retrieved from https://digibug.ugr.es/handle/10481/59151
10. Godfrey, D., Johns, C., Meyer, C., & Race, S. (2014). A case study in text mining: Interpreting twitter data from world cup tweets. *ArXiv Preprint ArXiv (Arxiv.Org)*. https://arxiv.org/abs/1408.5427
11. Gregor, S., & Hevner, A. R. (2013). Positioning and presenting design science research for maximum impact. *MIS Quarterly*, 37(2), 337-355. DOI: 10.25300/MISQ/2013/37.2.02
12. Huang, G., Fan, C., Sun, Z., & Zhu, H. (2019). Key word extraction for short text via word2vec, doc2vec, and textrank. *Turkish Journal of Electrical Engineering and Computer*. https://doi.org/10.3906/elk-1806-38
13. Jain, H., Sherali Shaikh, M., Shankar, R., & Mishra, V. (2022). Automatic Grading of Handwritten Answers. *International Research Journal of Engineering and Technology*. Retrieved from www.irjet.net
14. Jiang, K., & Lu, X. (2020). Natural language processing and its applications in machine translation: A diachronic review. *2020 IEEE 3rd International Conference of Safe Production and Informatization (IICSPI)*, 210–214.
15. Kim, M., Jung, Y., Jung, D., & Hur, C. (2014). Investigating the congruence of crowdsourced information with official government data: the case of pediatric clinics. *Journal of Medical Internet Research*, *16*(2), e3078.
16. Kitasuka, T., Aritsugi, M., & Rahutomo, F. (2012). *Semantic Cosine Similarity*. Retrieved from https://www.researchgate.net/publication/262525676
17. Lavanya, A., Gaurav, L., Sindhuja, S., Seam, H., Joydeep, M., Uppalapati, V., Ali, W., & SD, V. S. (2023). Assessing the performance of python data visualization libraries: a review. *Int J Comput Eng Res Trends*, 10(1), 29–39.
18. Muzaffar, A. W., Tahir, M., Anwar, M. W., Chaudry, Q., Mir, S. R., & Rasheed, Y. (2021). A Systematic Review of Online Exams Solutions in E-Learning: Techniques, Tools, and Global Adoption. *IEEE Access*, *9*, 32689–32712.
19. Namabira, J., Kamanzi, A., & Chawene A. (2022). Adoption of e-assessment technologies to enhance the happiness of academics. Insights from higher learning institutions in Tanzania. *African Journal of Education and Practice*. Retrieved from https://iprjb.org/journals/index.php/AJEP/article/view/1556

20. Nandini, V., & Uma Maheswari, P. (2020). Automatic assessment of descriptive answers in online examination system using semantic relational features. *Journal of Supercomputing*, *76*(6), 4430–4448. https://doi.org/10.1007/S11227-018-2381-Y/METRICS
21. Ozden, M. Y. (2004). Students' perceptions of online assessment: A case study. *International Journal of E-Learning & Distance Education/ Revue Internationale Du e-Learning et La Formation `a Distance*, *19*(2), 77–92.
22. Qasrawi, R., & BeniAbdelrahman, A. (2020). The Higher and Lower-Order Thinking Skills (HOTS and LOTS) in Unlock English Textbooks (1st and 2nd Editions) Based on Bloom's Taxonomy: An Analysis Study. *International Online Journal of Education and Teaching*, *7*(3), 744–758.
23. Semlambo, A., Almasi, K., of, Y. L.-I. J., & 2022, undefined. (2022). Facilitators' Perceptions on Online Assessment in Public Higher Learning Institutions in Tanzania: A Case Study of the Institute of Accountancy Arusha (IAA). *Researchgate.Net*. https://doi.org/10.18535/ijsrm/v10i6.lis02
24. Shaik, T., Tao, X., Li, Y., Dann, C., McDonald, J., Redmond, P., & Galligan, L. (2022). A review of the trends and challenges in adopting natural language processing methods for education feedback analysis. *Ieee Access*, *10*, 56720–56739.
25. SoftmaxAI. (2023). *Cosine Similarity – Text Similarity Metric – Study Machine Learning*. SoftmaxAI. Retrieved from https://studymachinelearning.com/cosine-similarity-text-similarity-metric/
26. Stowell, J. R., & Bennett, D. (2010). Effects of Online Testing on Student Exam Performance and Test Anxiety. *Journal of Educational Computing Research*, *42*(2), 161–171. https://doi.org/10.2190/EC.42.2.b
27. Wang, J., & Dong, Y.. (2020). Measurement of text similarity: a survey. *Mdpi.Com*. Retrieved from https://www.mdpi.com/813058
28. Xia, P., Zhang, L., & Li, F. (2015). Learning similarity with cosine similarity ensemble. *Information Sciences*, *307*, 39–52. https://doi.org/10.1016/J.INS.2015.02.024
29. Yan, D., Li, K., Gu, S., & Yang, L. (2020). *Network-Based Bag-of-Words Model for Text Classification | IEEE Journals & Magazine | IEEE Xplore*. IEEE Access. Retrieved from https://ieeexplore.ieee.org/abstract/document/9079815