Volume 67, Issue 2, 2023

Journal of Scientific Research

of

The Banaras Hindu University

Evaluation of Descriptive Answer by using Probability Approach, Cosine Similarity and Pretrained model

Bhagat Gayval[*1] and Vanita Mhaske[2]

[*1]Department of Statistics, SP College, Pune,Maharashtra, India.
gayvalbk@gmail.com
[2]Department of Computer Science, PVG's College of Science and Commerce, Pune, Maharashtra, India.
vanitamhaske04@gmail.com

***Abstract:*Evaluation of descriptive answers is important for analyzing the growth of students. It may be helpful for a job interview, for academic purposes, and in many more fields.In this research we discussed the importance of evaluating descriptive answers for analyzing student growth and how it is useful in various fields. With the increase in online exams due to the pandemic, objective-type questions are evaluated through different software, but there is a lack of system for evaluating descriptive answers.As manual evaluation is time-consuming, the probability approach is used in this research, which is compared with a pre-trained model and cosine similarity approach.In this research, we have used a probability approach, a pre-trained model, a cosine similarity approach, and compared it with a manually assigned score by a subject expert. The analysis concludes that the probability approach provides efficient results compared to other methods.**

***Index Terms:*Cosine Similarity, Descriptive answer, NLTK, Probability Approach, Similarity Score.**

I. INTRODUCTION

During the COVID pandemic situation, we have new experiences to familiarize ourselves with online exams.In the education sector, there is a lot of online student data. It may be their Google forms, assignments for examination purposes. The examination part plays a vital role in the student's academic phase. Because of the huge amount of data, it is important to handle it with a proper system. In the pandemic situation, many institutions shifted their examinations online too. Objective type questions are easy to evaluate, and they can be evaluated automatically with correct results. But the main purpose of the exam is knowledge understood by students. Descriptive answers may be helpful in checking overall student growth, progress, and positive change. But evaluation of descriptive answers is difficult through online mode. It is a lengthy textual answer given by students, and it will become difficult for the examiner to evaluate dozens of student submissions. It may get biased while checking the numbers on the paper or towards some students at the same time as comparing the solutions. To formulate scores obtained by students, we have used Natural Language Processing (NLP), a certain existing tool, and a probability approach.

II. OBJECTIVE

The main objective of this research is to use the concept of text analysis through a probabilistic approach, a pre-trained model, and a cosine similarity approach to accurately evaluate descriptive answers in an online mode. Here we use three techniques to score the student answer. Further, we compare those scores with scores given by a teacher or subject expert. Our purpose is to find out which method gives better results.

III. DATA PRE-PROCESSING

For this study, we collected responses from students who had basic ideas about the experiment. To gather answers from students, we ask the question, "What is the deterministic experiment?" We have collected 129 samples of data through a Google Form. Online-collected data is not structured. There is a need for structured data to apply the additional tools (see fig. 1). For comparison purposes, we need the ideal score, which we find out manually through subject experts. Using NLTK, the conversion of primary data into structured format is done.

A. Lower

If the textual content is in the same case, it is easy for a device to interpret the phrases because the lower case and upper case are handled differently through the machine. As an example, words like Exam and exam are treated differently via machine. So, we need to make the text within the identical case and the most desired case is a lower case to keep away from such issues.

B. Tokenization

Tokenization is the system of dividing textual content into a set of significant pieces. Those pieces are referred to as tokens. For instance, we will divide a bit of textual content into words, or we can divide it into sentences. Depending on the task at hand, we will outline our personal situations to divide the entered text into significant tokens.

word as run. Essentially stemming is to get rid of the prefix or suffix from phrases like ing, s, es, etc. NLTK library is used to stem the wordsdone.

E. Lemmatization

Lemmatization is similar to stemming, used to stem the words into root words however differs in running. actually, Lemmatization is a systematic way to reduce the words into their lemma by way of matching them with a language dictionary.

IV. Research Methodology

1) By Probability Method

The probability formula defines the likelihood of an event

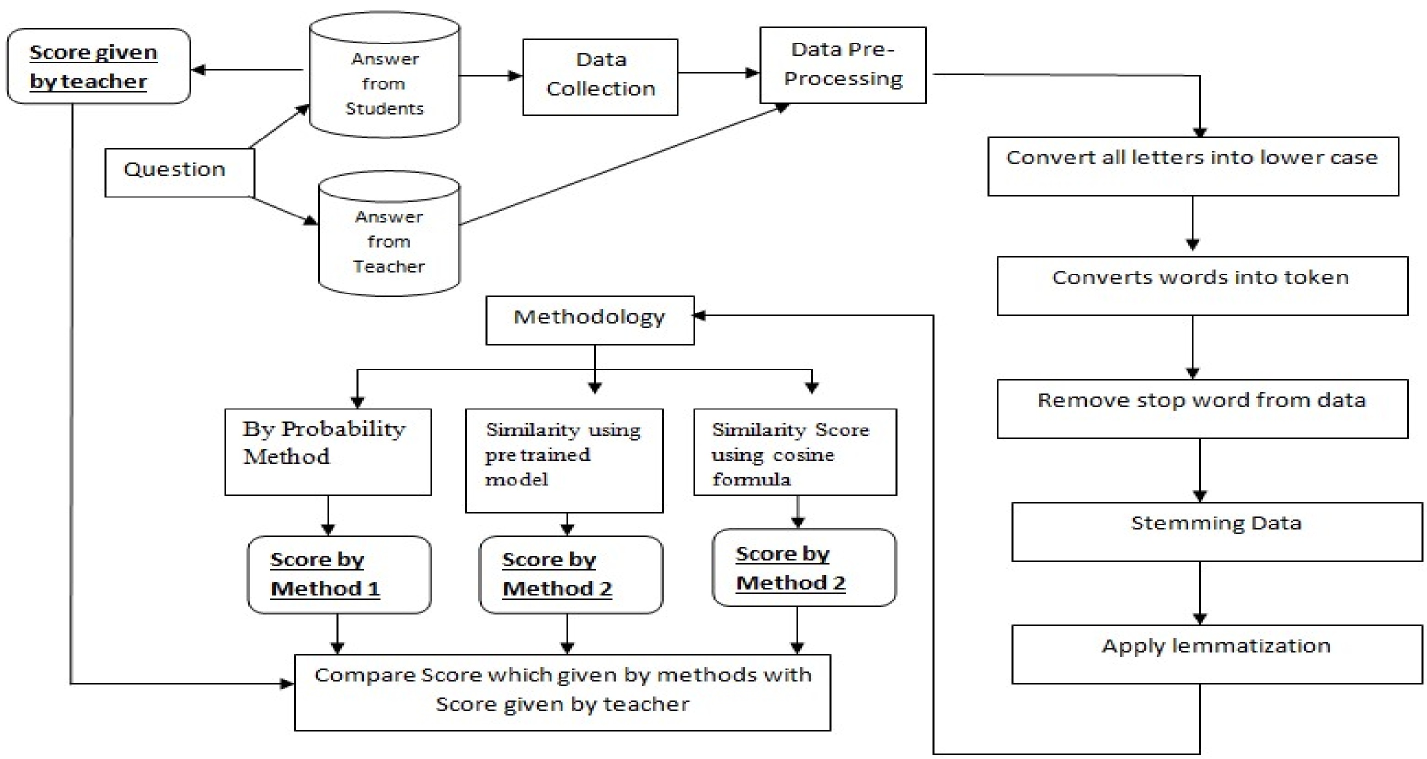

Fig. 1. Flowchart of the Process

C. Remove Stopwords

Stopwords are the most typically taking place words in a text Fig which do not provide any valuable data. stopwords like they, there, this, where, and many others are a number of the stopwords. NLTK library is a common library that is used to remove stopwords and consists of approximately 180 stopwords which it removes.

D. Stemming

Stemming is a method to reduce the word to its root stem as an example run, running, runs, runed derived from the same

happening. The formula to calculate the probability of an event is equivalent to the ratio of favourable outcomes to the total number of outcomes. Probabilities always range between 0 and 1. For an experiment having 'n' number of outcomes, the number of favorable outcomes can be denoted by x. The formula to calculate the probability of an event is as follows:

$x : No\_of\_Favourale\_outcome$

$n : No\_of\_all\_Possible\_outcome(\Omega)$

$$\Pr obability = \frac{x}{n} \ldots\ldots\ldots Eq(1)$$

In this case, using probability, we can find out whether the answer given by the student is likely to be the correct answer given by the experts or not. After multiplying the scores by the probability, we get the scores obtained by students. This gives us a value that represents to what extent the ideal response and the student's response are similar. In this case, we may consider the favourable outcome to be the presence of common words in both sentences, i.e., the response of the student and the ideal response. And the number of all possible outcomes is given by the number of all words present in the student's answer and the ideal answer. In this case, using probability, we can find out whether the answer given by the student is likely to be the correct answer given by the experts or not. After multiplying the scores by the probability, we get the scores obtained by students. This gives us a value that represents to what extent the ideal response and the student's response are similar. In this case, we may consider the favourable outcome to be the presence of common words in both sentences, i.e., the response of the student and the ideal response. And the number of all possible outcomes is given by the number of all words present in the student's answer and the ideal answer.

$$\Pr obability = \frac{\begin{array}{c} Number\_of\_common\_words\_present\_in \\ Both\_sentences \end{array}}{\begin{array}{c} Number\_of\_all\_words\_present\_in \\ Both\_sentences \end{array}} \ldots Eq(2)$$

2) Pre-trained Model

The Hugging Face is a community and data science platform that provides tools that enable users to build, train, and deploy ML models based on open source (OS) code and technologies. This can be useful for semantic textual similarity, semantic search, or paraphrase mining. Hugging Face is a large open-source community that quickly became an enticing hub for pre-trained deep learning models, mainly aimed at NLP. Their core mode of operation for natural language processing revolves around the use of transformers. You can use this framework to compute sentence / text embeddings for more than 100 languages.

Using Hugging Face pre-trained models, calculate students' scores. This is a sentence-transformer model: It maps sentences and paragraphs to a 768-dimensional dense vector space and can be used for tasks like clustering or semantic search. It has been trained on 215M (question-answer) pairs from diverse sources. Next, multiply the result of the score by its similarity using the pre-trained model to get the score obtained by the student.

3) Similarity Score using cosine formula

The probability Cosine similarity measures the similarity between two vectors in an inner product space. It is measured by the cosine of the angle between two vectors and determines whether two vectors are pointing roughly in the same direction. Cosine similarity is a measure of similarity, often used to measure document similarity in text analysis. Cosine similarity is a measure of similarity that can be used to compare documents or, say, give a ranking of documents with respect to a given vector of query words. Let x and y be two vectors for comparison. We use the equation to compute the cosine similarity.

$$Similarity = \frac{(A \bullet B)}{\|A\| * \|B\|} \ldots\ldots\ldots Eq.(3)$$

where A and B are vectors:

A.B is the dot product of A and B. It is computed as the sum of the element-wise products of A and B.

||A|| is the Euclidean norm of A. It is computed as the square root of the sum of squares of the elements of the vector A.

In this case, to discover the cosine similarity among phrases, we regard A as a student's answer and B as an ideal answer. Then multiply the result of the score with cosine similarity to get the score obtained by the student.

V. Result

For Comparing different methods' scores with the ideal score using ANOVA (this analysis is given by SPSS), we checked the assumption of homogeneity of variance. Levene's test is used to determine whether two or more groups have equal variances or not.

Problem-1:

H_0= Variance is equal across methodology

H_1=Variane is not equal across methodology

Table I. Test of Homogeneity of variance

Score

Levene Statistic	df1	df2	Sig.
13.030	3	512	.000

From Table 1, in problem 1, the p-value is less than .05, so we reject the null hypothesis. This means we have sufficient evidence to say that the variance between the three methods is significantly different. In other words, the three groups do not have equal variances.

The variance between the scores is not the same for all methodologies. Therefore, we use a robust test of equality of means to check the variation of the mean between the methodologies.

Problem-2:
H_0=There is no difference between the methodology and Ideal score
H_1=There is difference between the methodology and Ideal score

Table I. Robust Tests of Equality of Means

Score

	Statistic[a]	df1	df2	Sig.
Brown-Forsythe	98.739	3	440.095	.000

a. Asymptotically F distributed.

From Table 2, this p-value is less than .05, so reject the null hypothesis. This means we have sufficient evidence to say that there is a difference between the methodology and the ideal score. To pairwise compare, we use the Multiple Comparison Test (MCT). This test is performed when certain experimental conditions have a statistically significant mean difference or when there is a specific aspect between the group means. In some cases, the equal variance or homoscedasticity assumption is violent during the ANOVA process or pairwise comparisons. Here for multiple comparison, by considering problem 1, we take statistics of Tamhane's and Dunnett's T3. For that we define problem-3 , problem-4 and problem-5.

Problem-3:

$$H_0 : \mu_{Ideal\_score} = \mu_{score1}$$

$$H_1 : \mu_{Ideal\_score} \neq \mu_{score1}$$

Problem-4:

$$H_0 : \mu_{Ideal\_score} = \mu_{score2}$$

$$H_1 : \mu_{Ideal\_score} \neq \mu_{score2}$$

Problem-5:

$$H_0 : \mu_{Ideal\_score} = \mu_{score3}$$

$$H_1 : \mu_{Ideal\_score} \neq \mu_{score3}$$

Table II. Multiple Comparison

Dependent Variable: Score

	(I) Methodology	(J) Methodology marks_by_teacher				
		Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval Lower Bound	Upper Bound
Tamhane	marks_by_teacher					
	score1	-.07133	.06221	.826	-.2365	.0938
	score2	-.29935*	.06695	.000	-.4769	-.1218
	score3	-.88909*	.05926	.000	-1.0465	-.7316
Dunnett T3	marks_by_teacher					
	score1	-.07133	.06221	.824	-.2364	.0937
	score2	-.29935*	.06695	.000	-.4768	-.1219
	score3	-.88909*	.05926	.000	-1.0465	-.7317

*. The mean difference is significant at the 0.05 level.

From Table III, we can see that p-value is less than 0.05, therefore, we reject the null hypothesis for Problem-4 as well as for the Problem-5. For Problem-3, p-value is 0.826 (for Tamhane's method) and 0.824 (for Dunnett's T3 method) which is greater than 0.05. Hence, we failed to reject the null hypothesis. Therefore, we can conclude that there is no significant difference in Ideal Score and score _1 which we calculated from Probability approach.

VI. Conclusion

In the future, on-line coaching study approaches will be extensively used in many institutions. Descriptive solution checking techniques will assist in evaluating students' solutions. We have used a probability approach, a pre-trained model, the cosine similarity technique, and compared it with a manually assigned score by a subject expert. We have applied a robust test to check the equality mean and conclude that there is a difference between the methodologies. As per the statistical result (from problem 3), we can observe that the probability approach gives a better result as compared to the pre-trained model and cosine similarity approach. For the future, the probability approach will be more useful in descriptive answer checking as compared to a pre-trained model and the cosine similarity approach.

VII. References

A. References

1) Book with one author

Andy Field. Discovering Statistics Using IBM SPSS Statistics: SAGE Publication.

2) Book with two authors

Vijay K. Rohatgi and A. K. MD. Ehsanes Saleh, An Introduction to Probability and Statistics. (PP. 7-11): A Wiley-Interscience Publication.

Sabine Landau and Brian S. Everitt. (November 24, 2003). A Handbook of Statistical Analyses Using SPSS: Chapman and Hall/CRC; 1st edition.

3) Book with more than two authors

Bharadia, Sharad & Sinha, Prince & Kaul, Ayush. (2018). Answer Evaluation Using Machine Learning.

4) Journal article

A. R. Lahitani, A. E. Permanasari and N. A. Setiawan(2016).*Cosine similarity to determine similarity measure: Study case in online essay assessment*.(pp. 1-6). 4th International Conference on Cyber and IT Service Management, Bandung. Indonesia.

Tripathi, Vasu. (2020). *Automated Answer-Checker*. (pp. 152-155). International Journal for Modern Trends in Science and Technology.

Reimers, Nils &Gurevych, Iryna. (2019). *Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks*. (pp. 3973-3983).

Lourdusamy, Ravi & Abraham, Stanislaus. (2018). *A Survey on Text Pre-processing Techniques and Tools*.(pp. 148-157). International Journal of Computer Sciences and Engineering.

Supriya Sanjay Sasane, Aishwarya Gulab Thorat and Sayali Kiran Joshi .(2019).*Automatic Evaluation of Descriptive*

Answers. (pp.40-42).6(4). International Journal of Emerging Technologies and Innovative Research.

Menaka (2014). *Text Classification using Keyword Extraction Technique*.

D. L. Lee, Huei Chuang and K. Seamons(1997). *Document ranking and the vector-space model*. 14(2). (pp. 67-75). in IEEE Software.

K. Meena and R. Lawrance. (2016).*Semantic similaritybased assessment of descriptive type answers*.(pp. 1-7). International Conference on Computing Technologies and Intelligent Data Engineering (ICCTIDE'16), Kovilpatti, India.

Haneem, Faizura& Ali, Rosmah & Kama, Nazri&Basri, Sufyan. (2017).*Descriptive analysis and text analysis in Systematic Literature Review: A review of Master Data Management*. (pp. 1-6). International Conference on Research and Innovation in Information Systems (ICRIIS).

Muhammad Umer, Saima Sadiq, Malik Muhammad Saad Missen, Zahid Hameed, Zahid Aslam, Muhammad Abubakar Siddique, Michele NAPPI.(2021). *Scientific papers citation analysis using textual features and SMOTE resampling techniques*. (150). (pp. 250-257).

Mohan, Vijayarani.(2015).*Preprocessing Techniques for Text Mining - An Overview*.

Hickman, L., Thapa, S., Tay, L., Cao, M., & Srinivasan, P. (2022). *Text Preprocessing for Text Mining in Organizational Research: Review and Recommendations*. 25(1). (pp. 114–146). Organizational Research Methods.

H. Drucker, D. Wu, V.N. Vapnik.(1999).*Support vector machines for spam categorization*. 10. (pp. 1048–1054). IEEE Transactions on Neural Networks.

Lee, G. Lee.(2006). *Information gain and divergence-based feature selection for machine learning-based text categorization*. 42(1). (pp. 155 -165). Information processing & management.

Y. Yang, J.O. Pedersen.(1997). *A comparative study on feature selection in text categorization*. (pp. 412–420). Proceedings of the 14th International Conference on Machine Learning.

L. Tan, J. Na, Y. Theng, K. Chang. (2011). *Sentence-level sentiment polarity classification using a linguistic approach*. (pp. 77 -87). Digital Libraries: For CulturalHeritage, Knowledge Dissemination, and Future Creation.

Rosy Salomi Victoria D, Viola Grace Vinitha P, & Sathya R..(2020).*Intelligent Short Answer Assessment using Machine Learning*. 9(4). (pp. 1111–1116). International Journal of Engineering and Advanced Technology (IJEAT).

P. Kartheek Rachabathuni.(2017). *A survey on abstractive summarization techniques*. (pp. 762-765). International Conference on Inventive Computing and Informatics (ICICI). Coimbatore. doi: 10.1109/ICICI.2017.8365239.

P. Selvi and A. K. Bnerjee.(2010). *Automatic Short – Answer Grading System (ASAGS)*.2(1). (pp.18-23).InterJRI Computer Science and Networking .

Thompson, Victor &Panchev, Christo & Oakes, Michael. (2015).). *Performance Evaluation of Similarity Measures on Similar and Dissimilar Text Retrieval*.(pp. 577-58410.5220/0005619105770584.

Nandini, V., Uma Maheswari, P.(2020).*Automatic assessment of descriptive answers in online examination system using semantic relational features*.(pp. 4430–4448). J Supercomput 76. https://doi.org/10.1007/s11227-018-2381-y.

Patwardhan S, Banerjee S, Pedersen T. (2003).*Using measures of semantic relatedness for word sense disambiguation*. (pp 241–257). In: International Conference on Intelligent Text Processing and Computational Linguistics. Springer.

***