

# Integrated Assessment of Teaching Efficacy: A Natural Language Processing Approach

Lalitha Manasa Chandrapati<sup>1</sup>, Dr. Ch.Koteswara Rao<sup>2</sup>  
School of Computer Science and Engineering, VIT-AP University  
Beside Secretariat, Amaravathi 522237  
Andhra Pradesh, India

**Abstract**—The most significant component in the education domain is evaluation. Apart from student evaluation, teacher evaluation plays a vital role in the colleges or universities. The implementation of a scientific and appropriate assessment method for enhancing teaching standards in educational institutions is absolutely essential. Conventional teacher assessment techniques have always been bounded to bias and injustice for single dimensional assessment criteria, biased scoring, and ineffective integration. In this regard, it is crucial to develop a specialized teacher evaluation assistant (TEA) system that integrates with some computational intelligence algorithms. This research concentrates on using Natural language processing (NLP) based techniques for empirically analysing teaching effectiveness. We develop a model in which a teacher is evaluated based on the content he delivers during a lecture. Two techniques are employed to evaluate teacher effectiveness using topic modelling and text clustering. By the application of topic modelling, an accuracy of 75% is achieved and text clustering achieved an accuracy of 80%. Thus, the method can effectively be deployed to assess and predict the effectiveness of a teacher's teaching.

**Keywords**—Teacher evaluation; topic modeling; clustering; Latent Dirichlet Allocation (LDA); K-means

## I. INTRODUCTION

The acute necessity for experts is an aspect of present socio-economic upheaval. It contributes significantly to the level of potential nurturing at the pedagogical stage. Universities and colleges should improve the monitoring of teacher efficacy in classroom and develop an appropriate evaluation system since the competence of teachers directly determines the teaching standards and the understanding of performance of pupils [1].

Teaching and teacher activity are intrinsically diverse and complicated entities. Teaching entails a multitude of activities and interactions both inside and beyond the classroom. Although the concept of evaluating quality of teacher seems simple, in reality it includes identifying, describing, gathering data on, and inferring from hundreds of complicated component variables. Teacher's competencies significantly influence quality of teaching and learner's knowledge reception; accordingly, universities or colleges should improve classroom supervision and develop a realistic scoring scheme [2]. Teachers can receive feedback on the implementation outcomes to reinforce or improve particular areas of their training and make sure that they finish their assignments within the allotted amount of time by monitoring and evaluating the quality of their student's learning [19]. To continually increase student's learning abilities and teacher's teaching abilities, a faultless,

complete, and appropriate teaching performance evaluation system must be established [3].

At this moment, the implementation of a quantitative assessment method is necessary. The conventional form of teaching assessment continues to be used in the majority of universities and colleges [4]. The conventional teaching assessment methodology has proven inadequate to match the instructional criteria in the effective instructional procedure [5]. However, the conventional assessment methods do have some drawbacks, such as the single dimension of the indexes and the absence of a fairly neutral evaluation criteria [6].

Deep-learning is used in various aspects such as object detection, deep auto-encoders [21], image-captioning [24], sentiment analysis [22,25], etc. With the continual advancement in the areas of Deep learning and NLP, it is relatively possible to monitor teaching efficacy of a teacher in real time [7,20]. Teacher evaluation assistance system plays a crucial part in enhancing the effectiveness of potential nurturing at the teaching level [2,23]. Besides that, the teacher will get an opportunity of being awarded by the organization relying on the TEA system's results. Several factors can be taken into consideration while evaluating a teacher. Sentiment [17], presentation abilities, speaking skills, and student reviews are some of the common features. In this paper, we propose a TEA system that presents the teacher effectiveness score which solely depends on the content that the teacher delivers irrespective of his body postures, voice, etc.

$$TUS \propto \text{bodypostures} + \text{gestures} + \text{voice} + \text{accent} + \text{content} \quad (1)$$

In the eq. (1), TUS represents teacher uniqueness score

The teacher uniqueness score (TUS) is a measure that describes how distinct a teacher is from the others. As shown in the eq. (1), TUS depends on the body postures, gestures, voice, language and content. We consider 'content' as the main feature.

TABLE I. FEATURE INTERPRETATION IN TEA

Features/Aspects	TEA	[8]	[9]
Text extraction	yes	no	no
Student reviews	no	yes	yes
Context based	yes	no	no
Clustering	yes	no	no
Topic Heterogeneity	yes	no	no

The Table I displays the main discrepancies between our findings and that of others. As previously stated, this approach

incorporates text extraction, whereas the other two do not. Student reviews are required for assessing the teacher for the analysis done in [8] and [9], however they are not considered in this study. This research is solely context based, which has not been presented in prior studies. In this work, the clustering technique is applied for teacher assessment, whereas [8] and [9] use deep neural networks, deep denoising autoencoders, and support vector regression models. Topic heterogeneity, that is, different topics are considered when acquiring the dataset, unlike in the prior works.

The contributions of this article includes the following:

- This research developed a novel teaching quality assessment technique for universities and colleges based on topic modelling and clustering techniques.
- The proposed system solely concentrates on the content delivered by the teacher while instructing a class unlike other research works.
- This system can be used to assess a teacher as well as determine 'how effective a speaker is' in any given video.

The rest of this paper is organised in the following way. Section II addresses significant contribution in the field. Section III provides the proposed methodology of this system, Section IV describes experimental details followed by the results and finally, the conclusion is given in Section V.

## II. PRIOR RESEARCH AND NOVELTY OF TEA

The key approach for cultivating talents for a new generation is the enhancement of teaching excellence in academia, where the most significant aspect being the improvement of the foundation of teaching quality administrative systems. An efficient teacher assessment system in academia can depict teaching accomplishments in real times. The research on teacher assessment techniques in advanced nations seems to be more established owing to the early emergence of the education sector. In their paper [8], a comprehensive network for assessing teacher performance in universities and colleges was developed. The assemble evaluation data sets for the network, were collected from three different groups: students, peers, and leaders through survey questionnaire. The network of Student-NET, Peer-NET and Leader-NET was individually trained upon every dataset to build the hypothetical correlations between various evaluation indexes and outcomes from three perspectives. An integrated network Integrated-NET was implemented to merge the evaluations from the preceding networks in order to familiarise multi-dimensional analyses. This research eventually developed and implemented an online teaching assessment system, relying on the SSM framework, with a user-friendly functionality.

Yu Liu [9] in their study, analysed the traits and existent concerns of teacher's teaching assessment along with the critical aspects and approaches of university teacher teaching assessment. They designed an efficient methodology, a distinctive blend of deep denoising autoencoder and support vector regression model to assess the teaching effectiveness of college and university teachers. The model was built on multiple hidden layers and executed multiple feature transformations throughout the unsupervised training stage to achieve the

reconstruction between the output data and the input data. The developed approach was assessed in order to highlight the method's exceptional efficacy in enhancing decision-making about teacher effectiveness and its ability to accurately evaluate and anticipate the excellence of university education.

A research framework for analysing Physical Education teaching was proposed by Yuansheng Zeng [10], in which the authors developed a model for assessing both the teaching ability of the teacher and student's learning impact relying on a hybrid technology of data mining and the hidden Markov model respectively. Eventually, the cumulative assessment score is obtained.

In academic contexts, teachers, students, graduates and employers are often surveyed and this survey questionnaire includes open ended questions whose analyzation requires large amount of time and workload. To overcome this challenge, Buenano-Fernandez et al. in their study [11] suggested the use of a comprehensive technique based on topic modelling and text network modelling, which enables researcher to extract significant data from surveys containing open-ended questions. By use of these assessments vital data is gathered with the intention of determining the level of satisfaction that the aforementioned entities have with institutional education procedures.

A unique approach for evaluating the impact of practical knowledge on educational accomplishments was indicated by M. M. Rahman et al., in their study [12]. A remarkable approach for retrieving latent features and corresponding association rules from a legitimate dataset is presented in this research. For data clustering, an unsupervised k-means clustering technique is utilised, followed by the frequent pattern-growth approach for association rule mining. Using this framework many significant and relevant features were derived that are strongly linked to the learner's activities. To analyse the association among pragmatic (e.g. programming, logical implementations, etc.) abilities and entire scholastic accomplishment, statistical aspects of students are assessed, and the associated findings are presented. Relying on the determined latent attributes, a range of significant recommendations are presented for students for each cluster. Furthermore, the empirical outcomes of this study can assist teachers in developing productive instructional strategies, assessing programmes with precise arrangements, and pinpointing student's academic inadequacies.

By merging DBSCAN and k-means algorithms [13], the authors developed an ensemble unsupervised clustering paradigm for assessing student's behavioural traits. The efficiency of the suggested methodology is assessed by performing research on six categories of behavioural data generated by students at a Beijing university and assess the associations among diverse behavioural traits and student's grade point averages (GPAs). Besides detecting aberrant behavioural trends, the conclusions drawn from experiment also detect conventional behavioural trends more reliably.

As shown in Table II the prior work had some limitations. As it is seen, the data collected primarily was based on only reviews by different groups in their research [8]. Apart from those expressions, content, etc. could have been considered to evaluate the teacher. In their research [9], the data was collected only from teaching procedure with students, would

TABLE II. SUMMARY OF RELATED WORK

Reference #	Methodology	Findings	Limitations
[8]	Implemented an online teaching assessment system, relying on the datasets gathered from three distinct groups of people. To achieve this, they used ANNs.	An accuracy of 98.59% was achieved on verification dataset. The results of the testing reveal that the system acts effectively and fits the criteria to encourage pedagogical improvement in universities and colleges.	The data obtained for evaluating a teacher is primarily based only on the reviews from students, peers and leaders. But they could have considered other sources like expressions, content, etc., for evaluating the teacher.
[9]	A novel framework was designed by combining deep denoising encoder and support vector regression model to evaluate quality of teaching.	In comparison to other models, this model attained an accuracy of 85.23%	It would have been better if they could have considered other parameters too for evaluation, rather than depending only on the data taken from teaching procedure with students.
[10]	Combination of data mining and hidden markov models to evaluate PE teaching effectiveness with regard to both teachers and students, in universities was proposed.	Attained better accuracy and also high computational efficiency compared to other models.	The training data should be large in size for the model to perform better and the computation time is high in this case.
[11]	LDA for topic modelling and text network modelling was used to glean relevant insight from questionnaires containing open-ended questions.	The deployment of this approach allows the optimization of effort and time required to adhere with the assessment of the text data produced by the open questions.	Equitably finite amount of data was considered in the proposed research and it is a challenging task to retrieve topics from concise text.
[12]	K-means clustering and FP-growth techniques were used to identify the correlation and affinity among practical skills and academic performance.	According to the study, it is concluded that stronger practical abilities have a good influence on academic success.	K value of the enhanced k-means clustering method may vary depending on the dataset and might result in yielding stronger or even worse outcomes for varied datasets.
[13]	Proposed the application of DBSCAN and K-means clustering to analyse the behavioural patterns of the students.	The outcomes of this research help in providing students, effective services and governance, like psychological assistance and educational counselling.	The issue of this research arises when applying the suggested method to multisource behavioural characteristics with high dimensions.

be better if other parameters are considered. In their study [10], large dataset should be taken for the model to predict better results. It is a challenging task to retrieve topics from concise text [11] considering finite amount of data. The limitation in the study [12] is that the k-value of the enhanced k-means method varies depending on the dataset resulting in yielding stronger or even worse outcomes for varied datasets. The issue of the research [13] arises when applying the suggested method to multisource behavioural characteristics with high dimensions.

Taking into account the shortcomings of previous studies, a novel approach for teacher evaluation is proposed. The study's novelty includes the teacher being evaluated solely on the content of his lectures, regardless of feedback or ratings.

### III. PROPOSED METHODOLOGY FOR TEA

For a college or university to develop and prosper, teaching standards must be of the highest calibre. So we try to implement a model that identify good teacher on their domain without human support. The detailed methodology is further discussed in this session.

#### A. A System level overview of TEA to evaluate teacher effectiveness

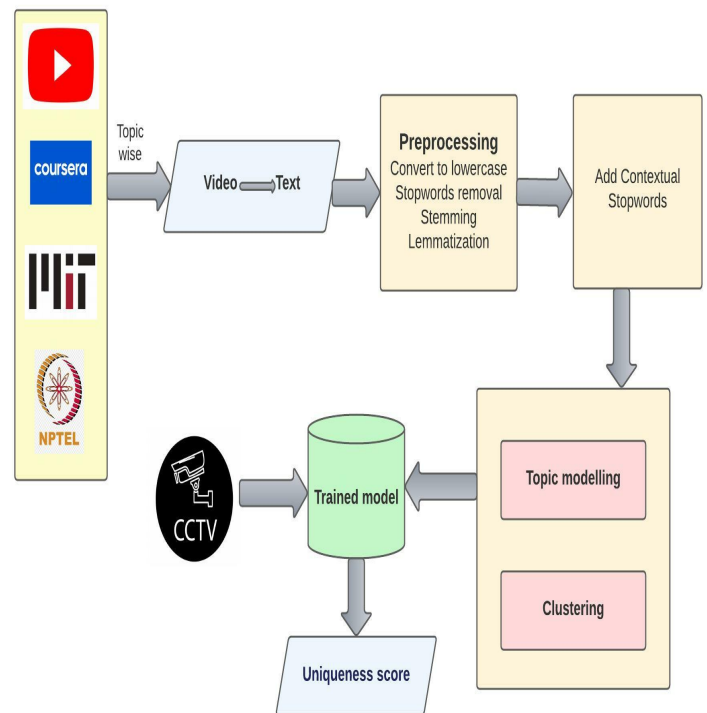


Fig. 1. Framework of the model to evaluate teacher

Fig. 1 depicts a schematic representation of the model for the TEA system. As illustrated in the Fig. 1, the steps involved in teacher evaluation are as follows:

- Videos spanning multiple sources such as YouTube, MIT courses, NPTEL, Coursera, and so on are considered.
- Using the IBM Watson and FFmpeg libraries, the acquired videos are transcribed to text and this text is treated as input.
- Pre-processing operations including converting to lowercase, stemming, lemmatization, stop word elimination, and so on are applied to the input text.
- Contextual stop words are added to the list.
- Topic modelling and Clustering approaches are employed on the pre-processed text and the trained model is stored.
- The lecture videos are acquired from the CCTVs installed in the classrooms, and are undergone through the trained model to provide a uniqueness score which describes “How effective a lecture of the teacher is?”

### B. Flowchart of the Proposed Model for Teacher Assessment

This research, which is based on topic modelling and clustering strategies, proposes a novel framework for teacher evaluation in universities and colleges, the flowchart of which is illustrated in Fig. 2.

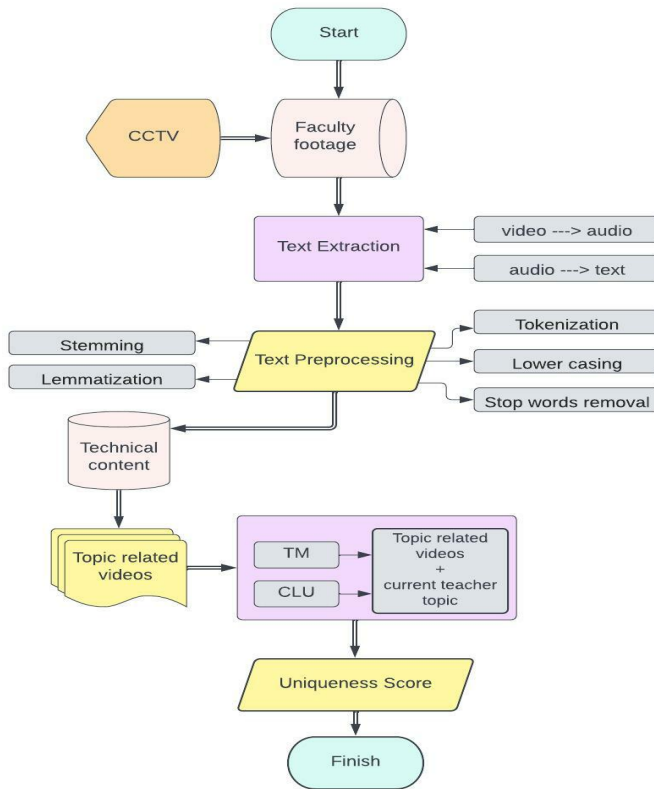


Fig. 2. Flow chart of the evaluation model for TEA based on topic modelling and clustering techniques

It can be seen from Fig. 2 that includes data required for the evaluation model is generated from Faculty footage

acquired of installed CCTVs at classrooms, then the generated data is undergone text extraction and text pre-processing steps. Finally, using both topic modelling and clustering techniques a uniqueness score is produced which evaluates the teacher effectiveness.

### C. Preparation of Data

As data acquired from CCTVs are videos, there is a need to transcript these videos to text format and also the transcribed text should be preprocessed.

1) *Text extraction*:: Converting videos to text can be accomplished in a variety of approaches, including utilising python, existing web APIs and soon. We transcript the videos using python FFmpeg and IBM Watson libraries here. The first phase includes conversion of video to audio using FFmpeg library. The second phase involves using IBM Watson library, which is SaaS provider for AI applications. The IBM Cloud offers a variety of solutions such as Text to Speech, Speech to Text, Natural Language Classifier, Language Translator, Visual Recognition, and so on. The Speech to Text service converts audio to text so that applications could use voice transcription features.

2) *Text preprocessing*: Text data is extensively available and is utilised to assess and solve business challenges. However, prior to actually using the data for research or prediction, it must be processed. Text preprocessing is used to prepare text data for model formulation. It is the first stage in any NLP project. Some of the preprocessing steps are: Removing punctuations like . , ! \$( ) \* % @, Removing Stop words, Removing URLs, Tokenization, Lower casing, Lemmatization, Stemming. Based on the dataset, we need to perform the appropriate preprocessing steps. For our dataset, we considered using Tokenization, Lowercasing, stop words removal, Lemmatization.

- **Tokenization**: The text is fragmented into small components in this stage. Based on our task specification, we can employ either sentence tokenization or word tokenization.
- **Lowercasing**: One of the most common preprocessing tasks is to convert the text to the same case, ideally lower case.
- **Stop words removal**: Stopwords are frequently used words that are eliminated from the text because they provide no relevance to the assessment. Those terms have little or zero significance. Apart from the existing list of stop words in NLTK library, we can modify the list by adding or eliminating terms depending on the scenario.
- **Stemming**: It involves stemming or reducing the words to it's core form. For instance, programmer, programming, gets reduced to the word “program”. However, the drawback of stemming is that it separates the phrases, that the base form lacks meaning or is not reduced to a valid English word.
- **Lemmatization**: The difference between stemming and lemmatization is that it stems the term yet ensures that it retains its meaning.

#### D. Methodologies

For our research we have deployed two existing technologies to evaluate the teacher. The techniques used here are: Topic modelling and Clustering.

1) *Topic modelling-latent Dirichlet allocation*: Topic modelling is a type of unsupervised NLP that is used to depict a text document combining numerous topics that help describe the inherent content in a specific document. Analysis of text data is done using a model and clusters of words are generated on that dataset. Latent Dirichlet Allocation [14] is the most effectively used model for topic modeling approach [15].

In LDA, latent represents the underlying concepts in the data, while Dirichlet is a kind of distribution. The Dirichlet distribution is not same as normal distribution. In LDA [18], the topics for every document are allocated in the following manner:

- 1) For K number of predefined topics, in every document each word to a topic is arbitrarily configured.
- 2) For each document d:  
Determine the following for every word w in the document:
  - P(topic t / document d): The percentage of words in document d allocated to subject t.
  - P(word w / topic t): The percentage of assignments to topic t from words in w among all documents.
- 3) Given all previous words and their topic assignments, reassign topic T' to word w with probability  $p(t'/d)*p(w/t')$ .

The final phase is performed several times until we reach a stable condition in which the topic allocations do not vary any further. These topic allocations are then used to establish the percentage of subjects for each document.

2) *Clustering-K means*: The process of dividing a dataset into clusters is called Clustering. The idea is to divide the data so that elements within a single group are relatively same and those in different clusters are dissimilar. It defines how unlabelled data is categorized. In the context of using text data, it is Text Clustering, which involves different phases including Text pre-processing, feature extraction, Clustering. Text clustering uses machine learning and NLP to recognize and analyse textual data. The most well researched clustering algorithm is K Means [16], and it persistently tends to produce good results.

The main goal is to categorise given data collection into prespecified k number of distinct groups. The algorithm is divided into two phases: the initial step is to establish k centroids for each cluster, the following stage is to choose each point from the provided dataset and connect it to the closest centroid. To compute the distance among data points and centroids, Euclidean distance method is used extensively. Once all the elements are included in certain clusters, the first phase is accomplished. The revised centroids must be reassessed at this stage since the addition of new points could cause a variation in the centroids of the clusters. Once k new centroids are identified, a new association among the same data points and the closest new centroid is established, resulting in a

loop. As a consequence of this loop, the k centroid's locations may vary in a step-by-step process and this leads to a condition called clustering convergence which means the centroids stop drifting.

#### E. Our Proposed Methodology

1) *Evaluation parameters*: We consider normalized value of likes/views of a video as a parameter for evaluating a teacher. It is known that for a real time lecture footage, neither likes nor views exist. For this purpose, we deploy a model where the real time videos are compared to you tube videos taking NL/V into consideration. As a result, if the video needs to be uploaded to a web source, then the effectiveness score can be known prior to uploading thus making it easy to identify whether the video be success or not.

##### 2) Users of our system:

- Public domain(YouTube, MIT, Coursera, etc.)  
As mentioned earlier it would be easy to depict a video to be success or not using our proposed model. This can help the trainer to improve his presentation and provide better videos. Also, the effectiveness of the speaker in a video can also be depicted.
- College/University Management  
They can be benefited in varied ways like to assess a teacher without direct classroom intervention. A trainer can also be recognised and awarded by the management for his teaching abilities by application of this model.

## IV. VALIDATION OF TEA

### A. Dataset Collection

The major source of dataset for evaluating a teacher is the lecture footage acquired from classrooms. It is a challenging task to accumulate a significant amount of data based just on footage in a short span of time. As a result, we used YouTube videos to train the model, and once the model gives constructive results, it can be applied on real-time lecture footage.

Videos are converted to text format and the data is undergone through pre-processing techniques thus resulting in providing clean data. Topic modelling and clustering approaches, LDA and K-Means, respectively are deployed on the data to assess the teacher.

As shown in Table III, videos for each sample topic are acquired and processed according to our research. Videos are converted to text format and the data is undergone through pre-processing techniques thus resulting in providing clean data. Topic modelling and clustering approaches, LDA and K Means respectively are deployed on the data to assess the teacher. Likes and views of each video is taken and percentage of likes to views is calculated. As the videos are of varied durations, there is a need for normalizing the likes/views value. The normalization technique used here is minmax normalization. We contemplate NL/V as evaluation parameter, it is equal to 1 indicates that likes of the video are identical as the views of video. This depicts that the video is best as "who all viewed it, liked it". Hence, it can be deduced from the above said logic that  $NL/V = 1$  means the video is best of all.

TABLE III. DATASET

S.No.	Sample topic
1	What is an activation function?
2	Distinguish between AI vs ML
3	Explain Generations of computers
4	Explain GSM architecture
5	What is MQTT protocol?
6	What is a neural network?
7	What is virtualization?
8	What is cloud computing?
9	What is database?
10	What is data mining?
11	Differentiate between classification and regression
12	What is an embedded system?
13	Define gradient descent
14	Define internet of things
15	Introduction to machine learning
16	Explain OOPs concepts
17	What is an operating system?
18	Differentiate between supervised and unsupervised learning
19	What are the types of operating systems?
20	What are the types of neural networks?

**B. Experimental Procedure**

1) *Topic modelling:* Data pre-processing is performed first. After pre-processing the data, topic modelling is implemented on the data to extract significant topics. Word frequency count is calculated on the collection of words obtained for each topic. Frequent words with a threshold value of 4 or less are removed as more frequently occurred words in the corpus determines that the speaker tried to deliver the significant content by insisting those words. Average word count and average normalized count of the words are computed.

$$WC = \sum \text{frequency of words in topics} \quad (2)$$

$$AWC = \frac{\sum WC}{\text{Total no. of topics}} \quad (3)$$

2) *Clustering:* For clustering approach, the data is summarized using python code. K-Means clustering was deployed on the summarized data thus resulting in clusters. The optimal number of clusters is predicted by the usage of elbow method. Once the clusters were formed the participation ratio of every video across each cluster is calculated. Average clustering score for each video is also calculated. High average clustering score depicts the corresponding video being the best of all.

$$PR = \frac{\text{No. of sentences of a video in the cluster}}{\text{Total no. of sentences across the cluster}} \quad (4)$$

Where PR = Participation Ratio of video

$$ACS = \frac{\sum_{i=1}^k PR}{k} \quad (5)$$

Where ACS = Average clustering score, K = Total number of clusters

**C. Results and Analysis**

After applying topic modelling on each video of the dataset, topics are generated which are a collection of words. Word count and Average Word Count are calculated using the eq. (2) and (3), respectively across each topic for all the videos. Average word count is normalized using minmax normalization to obtain average normalized count values.

*Intertopic distance plot* The intertopic distance plot is also created using the topics that were generated. It presents a broad perspective of the topics and their difference from one another, whilst providing for a detailed analysis of the words most intrinsically correlated with each topic. The areas of the circles encode the relative prevalence of each topic. The horizontal bar chart is illustrated on the right whose bars indicate the distinct words which are most beneficial for comprehending the present topic of interest on the left.

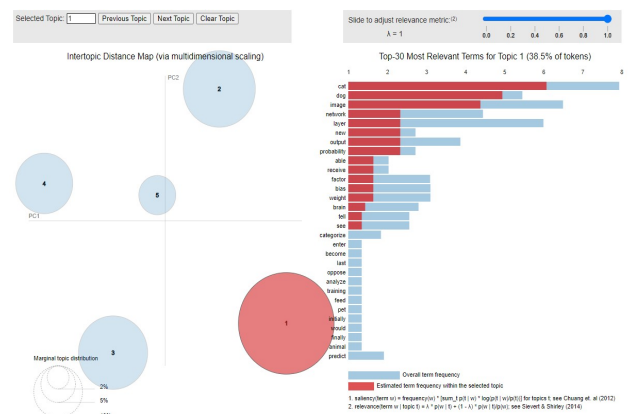


Fig. 3. Intertopic distance plot for neural network

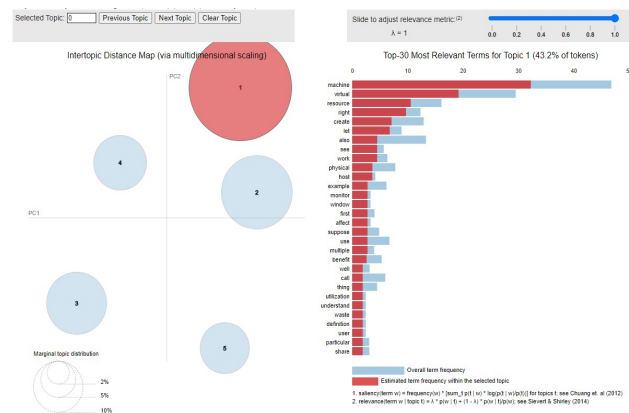


Fig. 4. Intertopic distance plot for virtualization



TABLE IV. STATISTICS OF TOPICS AFTER TOPIC MODELLING

TITLE	VIDEO-1	VIDEO-2	VIDEO-3	ANC(HIGH)
What is neural network?	NL/V=0 AC=31.4 ANC=0.44	NL/V=1 AC=41 ANC=0.47	NL/V=0.408 AC=38.6 ANC=0.45	YES
What is virtualization?	NL/V=0.57 AC=71 ANC=0.54	NL/V=1 AC=80.2 ANC=0.6	NL/V=0 AC=74.4 ANC=0.52	YES
What is database?	NL/V=1 AC=31.4 ANC=0.557	NL/V=0.329 AC=36.6 ANC=0.425	NL/V=0 AC=74.4 ANC=0.568	YES
What is cloud computing?	NL/V=0.67 AC=22 ANC=0.58	NL/V=1 AC=16.6 ANC=0.65	NL/V=0 AC=28.4 ANC=0.64	YES
What is data mining?	NL/V=0 AC=22.2 ANC=0.58	NL/V=0.414 AC=25.2 ANC=0.21	NL/V=1 AC=61.6 ANC=0.6	YES
Explain GSM architecture	NL/V=0 AC=24.4 ANC=0.408	NL/V=1 AC=36.8 ANC=0.48	NL/V=0.103 AC=62.6 ANC=0.38	YES
Distinguish between AI vs ML	NL/V=0 AC=22.2 ANC=0.26	NL/V=0.63 AC=36.4 ANC=0.6	NL/V=1 AC=38.6 ANC=0.51	NO
what is an activation function?	NL/V=0.85 AC=45.2 ANC=0.41	NL/V=1 AC=52.2 ANC=0.67	NL/V=0 AC=62 ANC=0.51	YES
What is MQTT protocol?	NL/V=1 AC=74.6 ANC=0.57	NL/V=0 AC=143 ANC=0.36	NL/V=0.06 AC=50.2 ANC=0.38	YES
Explain Generations of computers	NL/V=0.59 AC=42.4 ANC=0.44	NL/V=0 AC=28.6 ANC=0.487	NL/V=1 AC=35.8 ANC=0.54	YES
Differentiate between classification and regression	NL/V=0 AC=41.6 ANC=0.53	NL/V=1 AC=83.2 ANC=0.64	NL/V=0.74 AC=64.4 ANC=0.43	YES
What is an embedded system?	NL/V=0.75 AC=57.2 ANC=0.52	NL/V=0 AC=90.6 ANC=0.5	NL/V=1 AC=64.6 ANC=0.56	YES
Define gradient descent	NL/V=1 AC=95.6 ANC=0.45	NL/V=0.083 AC=130.6 ANC=0.55	NL/V=0 AC=108 ANC=0.65	NO
Define internet of things	NL/V=0 AC=28 ANC=0.46	NL/V=1 AC=34.8 ANC=0.48	NL/V=0.7 AC=27.8 ANC=0.38	YES
Introduction to machine learning	NL/V=0 AC=46.2 ANC=0.57	NL/V=0.33 AC=25.4 ANC=0.49	NL/V=1 AC=34.2 ANC=0.28	NO
Explain OOPs concepts	NL/V=0 AC=65 ANC=0.43	NL/V=0.89 AC=51.4 ANC=0.65	NL/V=1 AC=103.4 ANC=0.51	NO
What is an operating system?	NL/V=1 AC=41.6 ANC=0.43	NL/V=0.26 AC=48 ANC=0.53	NL/V=0 AC=29.6 ANC=0.475	NO
Differentiate between supervised and unsupervised learning	NL/V=0 AC=93.2 ANC=0.5	NL/V=1 AC=122.6 ANC=0.67	NL/V=0.46 AC=115.2 ANC=0.62	YES
What are the types of operating systems?	NL/V=0.83 AC=47.4 ANC=0.59	NL/V=0 AC=116.4 ANC=0.41	NL/V=1 AC=218.6 ANC=0.65	YES
What are the types of neural networks?	NL/V=0 AC=128.2 ANC=0.48	NL/V=1 AC=73.8 ANC=0.55	NL/V=0.088 AC=66 ANC=0.42	YES

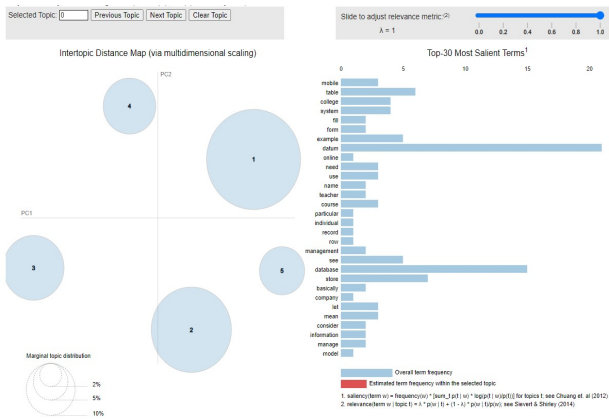


Fig. 5. Intertopic distance plot for database

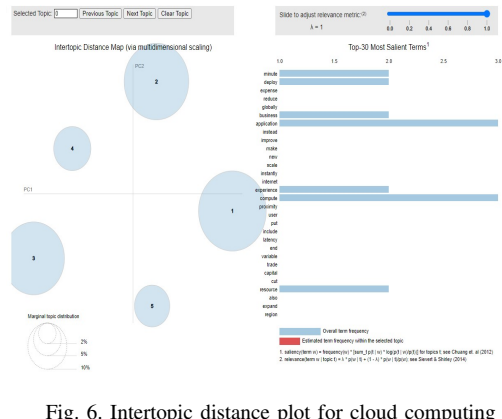


Fig. 6. Intertopic distance plot for cloud computing

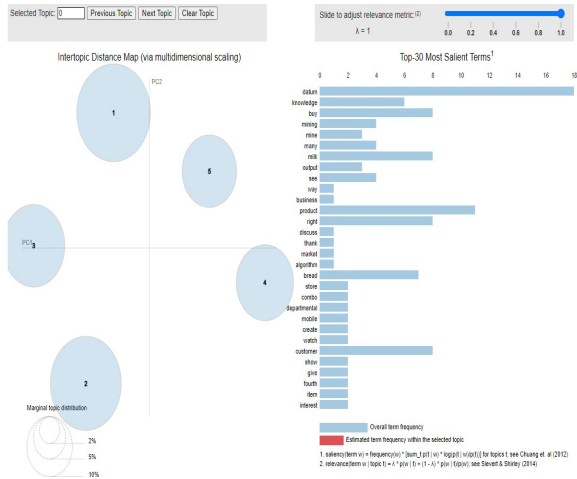


Fig. 7. Intertopic distance plot for datamining

The images Fig. 3 through Fig. 7 depicts the intertopic distance plots of the five sample topics. It can be depicted that in distance plot the size of the circle represents each topic's population in the corpora. Each bubble depicts a distinct topic. The greater size of the bubble indicates the greater proportion of terms related to that topic. The greater the distance between the bubbles, the greater the degree to which they differ from one another. When there is more space between the bubbles, there is a huge differentiation among each of the individual bubbles. On the right side, the box plot represents the top thirty salient terms of each topic and their respective frequencies. The term frequency across the whole document is represented by blue bars and the red bars are categorized by the significance of the words within every topic.

As it is shown in Fig. 3 and Fig. 4, the bubble 1 is selected, thus red bars indicate the particular term's frequency inside the topic 1 for both respectively. For example, it can be seen in Fig. 4 that the term "machine" is the most frequently used for 48 times. The red bar depicts that the word is used for 32 times in that particular topic 1. Similarly in the plots Fig. 5, Fig. 6, Fig. 7, the frequency of each term in the corpus is shown by blue bars. The blue bars indicate that no topic is selected and hence, most widely spoken terms are presented. In Fig. 5, the most frequently used words are datum and database. As the plot is obtained from the sample topic that discusses about database, it is clear that the teacher is trying to induce the topic by concentrating on those words.

The Table IV represents the statistical measures calculated after acquiring topics by implementing LDA on the dataset. We calculated the term frequencies across the whole corpus for five topics over all the videos in the dataset and it is represented as WC, by using eq. (2). The average word count is obtained by usage of eq. (3) and can be identified as AC. ANC is the normalized value of AC. As seen in the Table IV ANC score is high for 15 sample topics out of 20 sample topics. Those fifteen instances whose ANC is high, it can be observed that their respective video's NL/V value is 1. Hence, a conclusion can be drawn from the observed pattern that ANC score is high for the videos whose NL/V value is equal to 1 which comprehends that the particular video is the best. The Fig. 8

represents a graph for the distribution of the sampled topics whose average normal count is high. The videos with high ANC are considered to be the best videos and it can be deduced that the teacher delivered the lecture in an effective manner compared to other two videos.

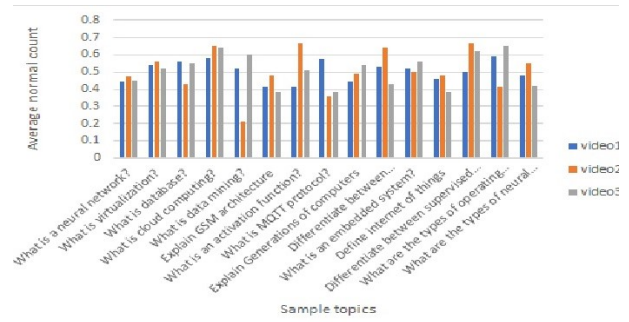


Fig. 8. Distribution plot of videos after topic modelling

TABLE V. TABLE OF CLUSTERING SCORE

Sample Videos	K	ACS(HIGH)	NL/V=1
what is an activation function	5	v2=0.401	v2
what is neural network	5	v6=0.128	v6
what is virtualization?	6	v3=0.18	v2
what is database?	8	v1=0.12	v1
what is cloud computing?	4	v4=0.2	v4
what is data mining?	8	v3=0.162	v3
Explain generations of computers	7	v9=0.121	v9
Explain OOPs concepts	7	v4=0.178	v10
Explain GSM architecture	6	v2=0.38	v2
what is MQTT protocol?	3	v2=0.37	v2

After applying clustering to the dataset, the clusters are formed and number of optimal clusters is identified by using elbow method. Using the eq. (4), participation ratio of the videos across the clusters for all the videos is calculated. ACS, average clustering score of all the sample topics is calculated using the eq. (5). The Table V depicts the number of optimum clusters for sample topics and their highest ACS score among the videos. It can be seen in Table V that 8 out of 10 sample topics have highest average clustering score of the corresponding video are same as the video having NL/V is equal to 1. Thus it can be concluded that the best videos of the sample topics after clustering have the highest average clustering score.



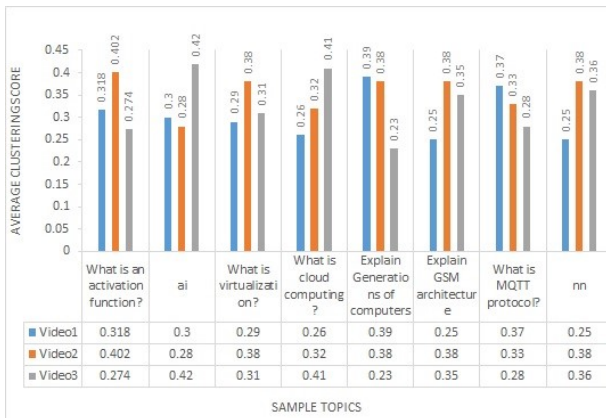


Fig. 9. Distribution plot of videos after clustering

The Fig. 9 shows the distribution plot of the sample topics after clustering technique been applied. It portrays the occurrences that have the highest average clustering score. Finally, We have drawn a conclusion from the patterns observed that by applying topic modelling and clustering methods, teacher evaluation can be performed based on the content and it got successful results with an accuracy of 75% and 80%, respectively.

## V. CONCLUSION

In this paper, LDA topic modelling and K-Means clustering methods are deployed on the dataset to evaluate a teacher's teaching effectiveness. The main feature for this research is considered as the content that the teacher delivers. Based on this context, videos from public domain are gathered for the dataset and applied the said models on the data resulting in obtaining an accuracy 75% and 80% of LDA, K-Means respectively. The utilization of two models is beneficial for acquiring effective results, like topic modelling consumes less computing time and clustering gives best accuracy. Finally, this paper provides a teacher evaluation system which is based on the content of the lecture distinct from traditional methods which are based on the reviews, emotion detection and so on.

## REFERENCES

- [1] Bain, Ken. What the best college teachers do. Harvard University Press, 2004.
- [2] Howe, Christine, et al. "Teacher-student dialogue during classroom teaching: Does it really impact on student outcomes?." *Journal of the Learning Sciences* 28.4-5 (2019): 462-512.
- [3] Chawinga, Winner Dominic. "Taking social media to a university classroom: teaching and learning using Twitter and blogs." *International Journal of Educational Technology in Higher Education* 14.1 (2017): 1-19.
- [4] Martinelli, Susan M., et al. "Results of a flipped classroom teaching approach in anesthesiology residents." *Journal of graduate medical education* 9.4 (2017): 485-490.
- [5] Kay, Robin, Thom MacDonald, and Maurice DiGiuseppe. "A comparison of lecture-based, active, and flipped classroom teaching approaches in higher education." *Journal of Computing in Higher Education* 31.3 (2019): 449-471.

- [6] Barrie, S., P. Ginns, and R. Symons. "Rewarding and recognising quality teaching and learning in higher education." (2007).
- [7] Joyce, Jeanette, Drew H. Gitomer, and Charles J. Iaconangelo. "Classroom assignments as measures of teaching quality." *Learning and Instruction* 54 (2018): 48-61.
- [8] Li, Guannan, et al. "Intelligent evaluation of teaching based on multi-networks integration." *International Journal of Cognitive Computing in Engineering* 1 (2020): 9-17.
- [9] Liu, Yu. "Evaluation Algorithm of Teaching Work Quality in Colleges and Universities Based on Deep Denoising Autoencoder Network." *Mobile Information Systems* 2021 (2021).
- [10] Zeng, Yuansheng. "Evaluation of physical education teaching quality in colleges based on the hybrid technology of data mining and hidden markov model." *International Journal of Emerging Technologies in Learning (iJET)* 15.1 (2020): 4-15.
- [11] Buenano-Fernandez, Diego, et al. "Text mining of open-ended questions in self-assessment of university teachers: An LDA topic modeling approach." *IEEE Access* 8 (2020): 35318-35330.
- [12] Rahman, Md Mostafizer, et al. "Impact of practical skills on academic performance: A data-driven analysis." *IEEE Access* 9 (2021): 139975-139993.
- [13] Li, Xiaoyong, et al. "An unsupervised ensemble clustering approach for the analysis of student behavioral patterns." *Ieee Access* 9 (2021): 7076-7091.
- [14] Guo, Chonghui, Menglin Lu, and Wei Wei. "An improved LDA topic modeling method based on partition for medium and long texts." *Annals of Data Science* 8.2 (2021): 331-344.
- [15] Muchene, Leacky, and Wende Safari. "Two-stage topic modelling of scientific publications: A case study of University of Nairobi, Kenya." *PloS one* 16.1 (2021): e0243208.
- [16] Nazeer, KA Abdul, and M. P. Sebastian. "Improving the Accuracy and Efficiency of the k-means Clustering Algorithm." *Proceedings of the world congress on engineering*. Vol. 1. London, UK: Association of Engineers London, 2009.
- [17] Onan, Aytug, Serdar Korukoglu, and Hasan Bulut. "LDA-based Topic Modelling in Text Sentiment Classification: An Empirical Analysis." *Int. J. Comput. Linguistics Appl.* 7.1 (2016): 101-119.
- [18] Mutanga, Murimo Bethel, and Abdultaofeek Abayomi. "Tweeting on COVID-19 pandemic in South Africa: LDA-based topic modelling approach." *African Journal of Science, Technology, Innovation and Development* 14.1 (2022): 163-172.
- [19] Liang, Liqun, Qian Yin, and Chunrang Shi. "Exploring proper names online and its application in English teaching in university." *ASP Transactions on Computers* 1.1 (2021): 24-29.
- [20] Nava, Imelda, et al. "Measuring teaching quality of secondary mathematics and science residents: A classroom observation framework." *Journal of Teacher Education* 70.2 (2019): 139-154.
- [21] Gomed, Everton, Rodolfo Miranda de Barros, and Leonardo de Souza Mendes. "Deep auto encoders to adaptive E-learning recommender system." *Computers and Education: Artificial Intelligence* 2 (2021): 100009.
- [22] Pong-Inwong, Chakrit, and Konpusit Kaewmak. "Improved sentiment analysis for teaching evaluation using feature selection and voting ensemble learning integration." *2016 2nd IEEE international conference on computer and communications (ICCC)*. IEEE, 2016.
- [23] Ge, Dongjun, Xiaoyue Wang, and Jingting Liu. "A teaching quality evaluation model for preschool teachers based on deep learning." *International Journal of Emerging Technologies in Learning (iJET)* 16.3 (2021): 127-143.
- [24] Sirisha, Uddagiri, and Boleam Sai Chandana. "Semantic interdisciplinary evaluation of image captioning models." *Cogent Engineering* 9.1 (2022): 2104333.
- [25] Sirisha, Uddagiri, and Sai Chandana Boleam. "Aspect based Sentiment & Emotion Analysis with ROBERTa, LSTM." *International Journal of Advanced Computer Science and Applications* 13.11 (2022).