Journal of Information Systems Engineering and Management
2025, 10(22s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Developing an Innovative NLP based Model to enhance Search Accuracy for Microlearning Videos on YouTube

Raghad Alharbi, Manal Abdullah, Shaimaa Salama
Faculty of Computing and Information, Technology, King Abdulaziz University, Jeddah, Kingdom of Saudi Arabia Jeddah, rhalharbi@stu.kau.edu.sa
Faculty of Computing and Information, Technology,King Abdulaziz University, Jeddah, Saudi Arabia Jeddah, maaabdullah@kau.edu.sa
Faculty of Computing and Information, Technology, King Abdulaziz University, Jeddah, Kingdom of Saudi Arabia
Faculty of Computers and AI, Helwan University, Cairo, Egypt

ARTICLE INFO

Received: 18 Dec 2024

Revised: 04 Feb 2025

Accepted: 20 Feb 2025

ABSTRACT

Technology has recently developed to support all fields and increase efficiency, especially in education. Microlearning has emerged as a powerful educational approach, allowing learners to view short, focused content that aligns with their needs and schedules. Micro-learning videos have recently become widespread, but the retrieved videos may not be relevant to the keywords used for search. This research aims to create an educational model specialized in micro-learning using Natural language processing. However, developing an effective microlearning model involves delivering highly relevant content, which requires developing a model that compares two NLP algorithms. Using two advanced NLP algorithms and comparing them after development to choose the best can significantly enhance the framework's accuracy. In this study, we develop and compare the NLTK and Gensim algorithms, measuring their cosine similarity to determine the best. The study's findings confirmed that NLTK outperformed Gensim regarding relevance, clarity, and alignment with the intended learning objectives. To ensure the reliability of the results and achieve high accuracy, we conducted a survey among teachers to select the videos that would be ranked based on their relevance. The survey results are aligned with NLTK's results, underscoring NLTK's potential as a more dependable tool for processing video content in microlearning applications.

Keywords: Microlearning, NLP, TF_IDF, YouTube

I. Introduction

Reliance on technology in education has increased, especially during the Covid-19 pandemic, and techniques have been made available to deliver information to students[1]. During the COVID-19 pandemic, the shift to distance education became a necessity following the decisions to close all educational institutions[2]. Distance learning is an educational system in which internet media is used to support the educational process. Learners tend to use many platforms to find information, and among those platforms is YouTube. YouTube contains a massive amount of information that covers all topics in various fields. The duration of the video sometimes exceeds two hours, and this affects the academic achievements of learners[3]. Also, a study showed that the level of attention of students decreases after 12 seconds, which confirms that increasing the duration of the video leads to a lack of concentration[4]. Dividing the educational content into small parts can motivate students and increase their focus, which is called micro-learning[5].

Micro-learning is a short-term self-learning strategy via the Internet that increases students' motivation to comprehend and retain information more effectively[6]. It responds to students' growing need for knowledge and reveals new ways to access information[7]. It overcomes the limitations and methods imposed by traditional learning in time and place as micro-learning can be anytime and

anywhere[8]. It provides small learning units, and each unit teaches a specific concept so that it does not exceed 15 minutes[9]. Micro-learning provides more fun and excitement for learners than traditional learning, as it contains essential topics quickly, and the learner does not make an effort to obtain information[10].

Most students tend to go to self-learning videos, especially during crises such as Covid, due to their large availability. Despite the importance of mini-learning videos in self-learning, a few studies discussed them on YouTube. Due to their extensive availability, most students attend self-learning videos, especially during crises such as COVID-19. Despite the importance of microlearning videos in self-learning, a few studies discussed them on YouTube. Also, given the massive number of videos provided by the YouTube platform, the Microlearning videos have not been studied before, whether the retrieved videos are relevant to the keyword used for search. Hence, there is a need to look at the microlearning videos on the YouTube platform to improve the educational process. This paper filters the videos on the YouTube platform to get the highest viewers, and their duration does not exceed 12 minutes.

However, developing an effective microlearning platform requires addressing the challenge of delivering highly relevant content. Using advanced natural language processing tools like NLTK and Gensim can significantly enhance the development of such a platform. NLTK provides robust capabilities for text preprocessing, keyword extraction, and sentiment analysis, while Gensim enables topic modeling and semantic understanding of content. Together, these tools provide the foundation for creating a dynamic and personalized microlearning website that meets diverse learning needs, ensures content relevance, and maximizes user engagement. Both algorithms are compared after development to choose best. The motivation of this study is to create a microlearning model that enhances learning efficiency, accessibility, and engagement by transforming YouTube's vast educational content into structured, digestible lessons tailored to time-constrained learners, improving knowledge retention, and making education more accessible through NLP. We present a model that compares both NLP algorithms and chooses the best after measuring the cosine similarity of both.

The remainder of the paper is organized as follows:

Section 2 contains the literature review; Section 3 presents the proposed approach; Section 4 discusses the results and provides a comprehensive analysis; and Section 5 concludes.

II. Literature review

A. *Background*

1) *Microlearning*

The main objective of micro-learning is to divide the educational content into small parts, or so-called micro-content, to increase the use of knowledge and focus on specific parts[11]. E-learning, which includes digital or online learning, differs from micro-learning, as e-learning offers macro-learning, such as substantial open courses and lessons. Micro-learning is characterized by its focus on relevant information content that is presented concisely and accurately[12]. E-learning allows learners access to information anytime, according to their schedule, and anywhere because portable devices are present[13]. Microlearning is short-term learning, which ranges from a few seconds to 15 minutes at most[14]. According to the study, the concentration rate of Internet users decreases after 8 seconds from the start of browsing[15]. According to previous research, video clips are becoming more popular with learners, and video viewership decreases after exceeding 7 minutes. In contrast, another study indicates that viewers are likelier to watch short videos from start to finish[16]. However, most of the educational materials presented via online platforms exceed 15 minutes[17]. This indicates a reduced human ability to focus and not be distracted for an extended period. Practical teaching that requires applying specific skills is one of the most influential in the educational field, such as programming and medicine[18].

During the epidemic, a micro-learning study was applied to students to teach them biochemical materials through interactive exercises and video clips of no more than 10 minutes. Before applying the study students' desire to obtain educational material was 15%. The student's desire to obtain educational material increased by 52%. The study was implemented for a month[19] .Another study used test-based

micro-learning during COVID-19, where the platform offered daily multiple-choice testing. The learner receives e-mail or text messages at the time specified by the learner. Learners reported that this format was effective and preferred over traditional education methods, indicating more opportunities for innovation[20]. To obtain better educational results and the ability to learn at a distance, educational materials must be divided logically into small parts or create new resources that achieve the desired.

2) Natural Language Processing (NLP)

NLP is considered a solution allowing machines to process, understand, and generate human language[21]. Rooted in linguistics, computer science, and artificial intelligence, NLP has significantly advanced, finding applications in education to improve teaching methods and enhance student performance[22]. The ability of NLP to analyze and generate text at scale provides a new dimension to educational tools, making them more interactive, personalized, and outperformed.

In teaching, NLP has been used to automate routine administrative tasks such as grading, provide personalized feedback, and even adapt teaching materials to meet the specific needs of individual learners. Beyond automation, it has facilitated the assessment of competencies in complex subjects like STEM (Science, Technology, Engineering, and Mathematics), enabling educators to focus more on pedagogical innovation[23]. Furthermore, studies indicate that NLP techniques, such as Neuro-Linguistic Programming (NLP), are gaining traction in language teaching classrooms. Research shows that NLP-trained teachers can effectively enhance students' emotional intelligence, motivation, and learning outcomes by employing strategies like rapport-building, pacing, and reframing[24]. These findings are particularly relevant for English as a Foreign Language (EFL) contexts, where language instructors benefit from NLP workshops to improve teaching practices through structured classroom observations and feedback[25].

a) NLTK

The Natural Language Toolkit (NLTK) is the oldest and most comprehensive Python library proposed for text processing and analysis[26]. It was developed as a platform to provide easy access to standard NLP algorithms, linguistic data, and educational resources. NLTK provides tools for tokenization, stemming, and sentiment analysis, enabling the evaluation of linguistic features like tone, grammar, and coherence[27]. Using parsers and syntax checkers, NLTK helps evaluate grammatical accuracy. For instance, it can identify subject-verb agreement errors or misplaced modifiers, offering targeted corrective feedback[28]. NLTK provides various tokenization techniques, including word and sentence tokenization, which are essential for breaking down text into manageable pieces[29]. NLTK includes a vast collection of text corpora, and lexical resources like WordNet are essential for various NLP tasks, such as training models or finding word synonyms[30].

b) Gensim

Gensim is a specialized library designed primarily for word embedding, document similarity, and topic modeling with large corpora[31]. Its focus on semantic analysis makes it especially useful for uncovering thematic trends in extensive textual datasets. Unlike NLTK, a general-purpose NLP library, Gensim focuses on unsupervised learning tasks and is optimized for processing large-scale text data. Gensim provides efficient methods for calculating document similarity, a crucial task in information retrieval[32]. It also provides topic modeling algorithms like Latent Dirichlet Allocation (LDA) and Latent Semantic Analysis (LSA). One of Gensim's strengths is its ability to handle large text datasets efficiently[33]. It also supports distributed computing, allowing users to scale their models across multiple processors or machines, making it ideal for big data applications[34].

B. Releted Works

Several studies have been conducted on NLP algorithms, and most of them used the NLTK and Gensim techniques.

Identify applicable funding agency here. If none, delete this text box.

Caratozzolo et al. [35]focused on assessing the benefits of incorporating NLP tools into evaluation procedures for advanced STEM. They

argued that NLP tools are adequate for assessing higher-order functions and gauging cognitive understanding of concepts. The NLP technique used is NLTK to text analysis. Also, they found that NLP tools can assist instructors in conducting more effective review and feedback sessions and providing personalized reports on students' oral and written communication skills.

Agarwal et al. [36]studied the mental state of students during the COVID-19 pandemic. The study was based on sentiment analysis using NLTK. Data was gathered from a survey conducted for students aged 18-22. The results showed that most students had negative thoughts and expressions during COVID-19, and only 15.7% had positive sentiments.

Haider et al. [37]suggested a model that integrates Gensim Word2Vec with the K-Means clustering algorithm and an innovative sentence scoring procedure. They tested the model using BBC news articles. They found the best performance of the model on business articles due to their higher concentration of numerical values, which the sentence scoring algorithm prioritizes.

Kulkarni et al. [38]developed a Python module using Natural Language Processing (NLP) to summarize online class videos. They employed algorithms like Term Frequency-Inverse Document Frequency (TF-IDF) and Gensim for summarization. The process incorporates two methods: cosine similarity, which doesn't require a reference summary, and ROUGE score, which does. They showed that the cosine similarity method achieves over 90% efficiency with both TF-IDF and Gensim, while the ROUGE score ranges from 40-50% efficiency.

Ponmalar et al. [39] addresses the challenge learners encounter when trying to obtain real-time answers to their questions during online video tutorials. They suggested a deep learning video streaming web application featuring AI bots that utilize natural language processing (NLP) to generate relevant keywords from uploaded videos, allowing users to contribute questions and answers to the database. The application was developed using Agile and Scrum methodologies. Additionally, they utilized natural language processing (NLP) algorithms to generate relevant keywords and match queries to answers in the database.

III. Methodology

The proposed micro video search model aims to retrieve the most relevant short videos related to the user keywords. The model is implemented in Python, and the model steps are illustrated in Fig.1. The user enters keywords, and then YouTube searches for videos related to these keywords. The model will find and download videos to be less than 12 minutes. Open AI whisper is used to convert downloaded videos into text[40]. After converting, we use two types of natural language processing NLTK and Gensim to preprocess the converted text then TF-IDF is used for text vectorizers. Using two NLP algorithms and comparing them after development to choose the best can significantly enhance accuracy. Finally, cosine similarity is used to rank videos and determine the best results of the two NLP libraries.

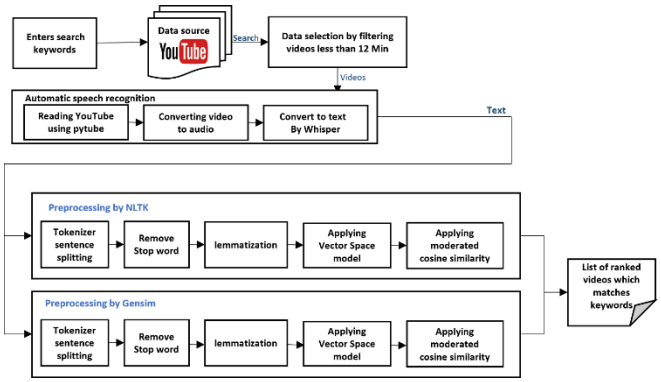

Fig. 1. Proposed Methodology

A. *Retrieve results from YouTube*

As a first step, users are required to enter search keywords into YouTube. The specified keywords are used to extract related videos from YouTube. We use the youtube-search-python library to search and

import videos only related to a specific keyword on YouTube[41]. In the research, the youtube.search().list function is used to import all the needed information, such as titles, links, and duration of videos.

B. Filter out videos

After searching for videos we need to filter out 12-minute videos or less from playlists or whole channels. The YouTube API provides video duration in this format: 5H12M30S; therefore, we need to convert it into seconds. The parse_duration() converts the provided time format (by the YouTube API) into seconds.

C. Download and convert videos into audio by PyTube

In this step, download and convert these videos to audio using another library called PyTube[42]. To use the PyTube library, we need to call the package yt-dlp (youtube-dl), which allows downloading videos from thousands of sites. Import YouTube from PyTube, pass the video URL and use the filter method to specify filters like output format and duration to download a video. Each video is downloaded in .Mp3 format.

D. Convert Videos into Text

In this step, the videos converted to audio are converted to text using Whisper deep learning model[40]. Whisper's deep learning model is one of the translators that convert audio into text. In our model, we only apply the English language. After installing Whisper into our model, we call the load_model function. It's a pre-trained function for which we have specified the English language only and passed the path file. Whisper's word error rate in English does not exceed 4.5%[43].

E. Text Preprocessing

Text preprocessing is one of the basic steps in NLP to obtain results that enable the machine to read the data. Data preprocessing improves model performance and produces more accurate results[44]. Basic steps in preprocessing text include stop word removal, tokenization, and lemmatization. Tokenization is a crucial first step in preprocessing NLP, essential for breaking down text into manageable pieces. Tokenization will help to understand each word individually and be able to count the number of times each word is repeated[45].

Example

Input: Databases are structured to facilitate the storage

output: "Databases","are","structured","to","facilitate","the","storage"

Stop words are a list of common words that do not provide helpful information on a specific topic in the document, such as "the,"" of," and "an,". It is often removed to improve model performance.

Example

Input: Databases are structured to facilitate the storage

output: "Databases","structured","facilitate","storage"

Preprocessing also needs to convert the words into their roots for meaningful words known as lemmatization. This technique is used to reduce words to their base forms through morphological analysis of the words. Unlike stemming, lemmatization considers the word's context and returns a valid lemma. It looks for the context and meaning of the sentence and the part of speech in a sentence [46].

In the proposed model two NLP libraries which are NLTK and Gensim are used for preprocessing. These two Libraries have some similar characteristics, but some differences distinguish each from the other.

F. Convert Video Text into Text Vector

After converting videos to text, we have converted video text into text vectors using TF-IDF. Text vectorization is a process that converts words in a text document to importance numbers. It is a Statistical method used to measure the importance of the word within the document relative to a collection of documents[47]. TF-IDF is the most common method for calculating text vectorization.

G. Calculate Cosine Similarity with respect to Searched Keyword

Cosine similarity brings similar documents or products closer. We can measure the similarity between two nonzero vectors. We can get cosine similarity by computing the angle between them. The smaller the angle, the higher the cosine similarity between the videos we are searching for concerning keywords. Cosine similarity is advantageous because, despite two similar data objects being different in terms of Euclidean distance due to their size, they can still have a minor angle between them. The similarity is higher when the angle is smaller[48].

We can find cosine similarity between videos in Python using Numpy, Matlab, or Sklearn libraries. Then, we found the cosine similarity among these videos concerning our searched keywords.

H. Display Final Ranked Videos

For the final step, we are concatenating the scores with the data frame videos with high cosine similarity scores are shown at the top of the results, followed by videos with lower cosine similarity.

IV. Methodology Results

The model was utilized for two NLP algorithms in Python: NLTK and Gensim. High school teachers were contacted to select topics to apply and test the model. They selected five keywords in HTML.

A. NLTK and Gensim

NLTK and Gensim algorithms were used, and the videos were ranked using cosine similarity for both algorithms.

1) Using NLTK

a) TF-IDF Calculation:

NLTK does not have built-in functions for TF-IDF. Instead, typically other libraries like scikit-learn are used for this purpose. However, TF-IDF with NLTK can be calculated by using its tokenization and frequency counting tools and then applying the TF-IDF formula.

- - Tokenize and preprocess text with NLTK.
- - Compute term frequency (TF) and inverse document frequency (IDF) manually.
- - Combine TF and IDF to get TF-IDF scores.

b) Cosine Similarity:

Like TF-IDF, NLTK does not have built-in support for cosine similarity. Vector space representations (possibly created with other libraries like scikit-learn) is needed to compute cosine similarity using numerical libraries such as NumPy.

```
In [17]: # Calculate similarity
         cosine_similarities = linear_kernel(doc_vectors[0:1], doc_vectors).flatten()
         document_scores = [item.item() for item in cosine_similarities[1:]]
         document_scores

Out[17]: [0.49506588927712425,
          0.46242726359776 15,
          0.32299277298602425,
          0.2534728380500141,
          0.12689606243326543,
          0.16866201595964836]
```

Fig. 2 .NLTK Score Calculation

2) Using Gensim

a) TF-IDF Calculation:

Gensim has built-in support for TF-IDF through its TF-IDF Model. You can easily compute TF-IDF for a corpus of documents.

b) Cosine Similarity:

Gensim also provides tools to compute cosine similarity using its similarities module.

```
In [25]: # Calculate similarity
         cosine_similarities = linear_kernel(doc_vectors[-1], doc_vectors).flatten()
         document_scores = [item.item() for item in cosine_similarities[0:-1]]
         document_scores

Out[25]: [0.16866201595964836,
          0.18257828193919293,
          0.1805062636108074,
          0.310358433073355,
          0.18626474340626958,
          0.1726707038372099]
```

Fig. 3. Gensim Score Calculation

V. Result verififcation

To ensure the reliability of the results and achieve high accuracy, we surveyed teachers to select the videos that would be ranked based on their relevance. The survey was sent to fifteen secondary school teachers with videos in random order to determine which algorithm performed better. Results of Gensim and NLTK were measured, and random videos were sent to teachers to analyze the videos from 5 main aspects: relevance of video with title, relevance of video with keywords, Pronunciation of video, Sound affecting video, and whether the video achieved its goal.

As shown in Fig.4 , cosine similarity for both algorithms are added. A shown Video 8 was lowest in NLTK and highest in Gensim. Video 5 was highest in NLTK, and Video 9 was lowest in Gensim. These results are hidden from teachers.

```
[31] df['doc_sim_gensim'] = pd.Series(document_scores)
     df.sort_values('doc_sim_gensim', ascending=False)
```

	links	title	doc_sim	doc_sim_gensim
8	https://www.youtube.com/watch?v=hpz34Pcu5I0	Colorful table with HTML and CSS	0.000000	0.483727
7	https://www.youtube.com/watch?v=bil9OFH6Nmg	Styling HTML tables with CSS - Web Design/UX ...	0.189874	0.405192
2	https://www.youtube.com/watch?v=CILWGhUnMbo	How To Create Table In HTML And CSS I HTML Web...	0.132255	0.282665
1	https://www.youtube.com/watch?v=A3NIKULhZIs	HTML tables I Web Technology I Lec-8 I Bhanu P...	0.194219	0.277355
5	https://www.youtube.com/watch?v=-kXZvKxs9oA	Learn HTML lists in 4 minutes	0.554290	0.256544
3	https://www.youtube.com/watch?v=K21yqCmo6Qg	HTML Table Using Rowspan & Colspan I Html Tuto...	0.141911	0.233771
4	https://www.youtube.com/watch?v=fGNwmSwHksE	How to create table, using html only? for Begi...	0.127213	0.222390
0	https://www.youtube.com/watch?v=iDA0kF5IrVk	Learn HTML tables in 3 minutes	0.153870	0.136687
6	https://www.youtube.com/watch?v=EKSPK_Zk9X8	Create a Data Table in Bootstrap 5 (2023)	0.075634	0.084536
9	https://www.youtube.com/watch?v=dK27jWKlJxs	25: Table In HTML and CSS I How To Create Tabl...	0.136687	0.000000

NLTK Gensim

Fig. 4 .HTML Videos Measurements NLTK/Gensim

Videos 8, 0, 5, 9, and 2 were chosen randomly and sent to 15 teachers in the secondary school. Keywords were provided by teachers; these videos were sent to teachers for evaluation to know which algorithm performed better. and teachers were asked 5 questions on each video.

- How relevant are keywords to the video?
- How relevant are keywords to the video title?
- Is there a sound that affects the clarity of speech?
- Are the pronunciations clear?
- Did the video achieve the goal?

1) Video: Learn HTML lists in 4 Minutes Highest Rank in NLTK

As shown from Figure 4-3 this video was highly ranked in NLTK and its rank in Gensim was 0.25. When teachers were asked about this video relevance of keywords towards title and video content, the result shown in Figure 4-4 showed that 14 teachers confirmed that the keywords were both relevant to title and video content.1 teacher said that it was probably relevant.

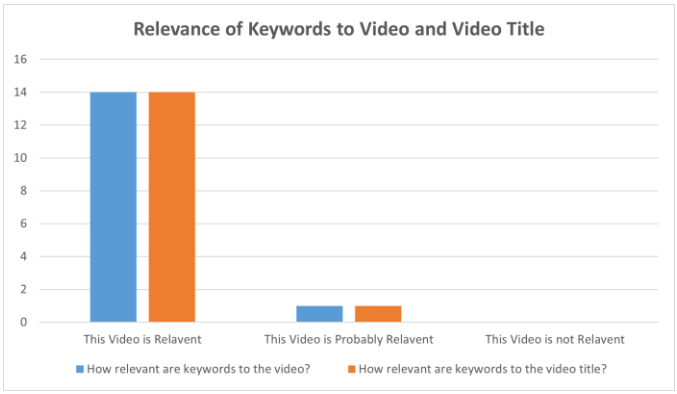

Fig. 5. Results on Relevance of Keywords to Video and Video Title on Learn HTML lists in 4 Minutes

Figure 5 shows the results of the 3 aspects regarding sound affecting clarity of speech, clearness of pronunciation and goal achievement of video. As shown in Figure 4-2 14 teachers confirmed that there was no sound affecting the clarity of speech, while only one teacher said there was sound affecting clarity. Regarding pronunciation and achievement of goal all 15 teachers confirmed that the pronunciation was clear, and goal was achieved.

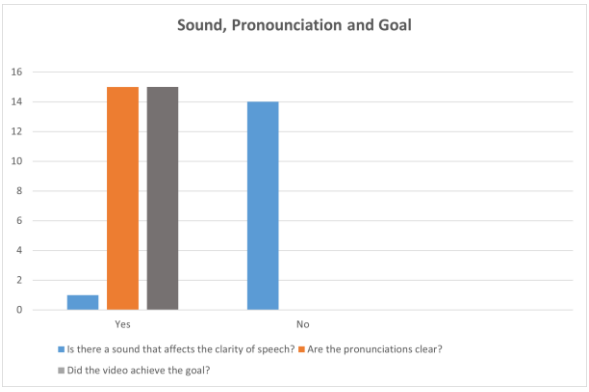

Fig. 6. Results on Clarity of Speech Pronounciation and Goal on Learn HTML lists in 4 Minutes

2) Video: Colorful Tables with HTML and CSS Highest Rank in Gensim and Lowest in NLTK

This video is ranked high in Gensim and lowest in NLTK, asking teachers about this video the results were as shown in Figure 4-6 regarding the relevance of the keyword to the video content and video title. As shown in Figure 4-6 11 teachers found that the keywords are relevant to content while 4 found it "Probably relevant". 9 teachers found keywords relevant to the title while 6 found it "Probably Relevant"

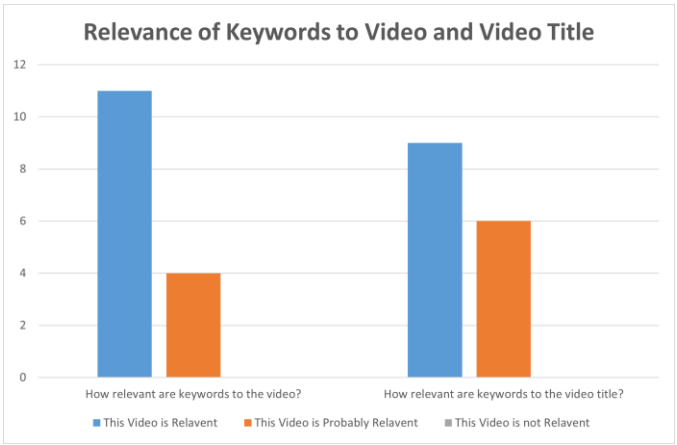

Fig. 7. Results on Relevance of Keywords to Video and Video Title on Colorful Tables with HTML and CSS

All respondents agreed as shown in Figure 7 that sound issues were present, significantly affecting the clarity of speech. This unanimous feedback indicates a pressing issue that must be addressed to improve the video's overall quality and effectiveness in communication. 14 participants noted that pronunciations were unclear as shown in Figure 4-7, with only one respondent indicating clarity. This result underscores a critical barrier to comprehension and suggests the need for targeted improvements. Despite issues with sound and pronunciation, all 15 respondents confirmed that the video successfully achieved its intended this indicates that while technical issues exist, the content and purpose of the video resonated well with the audience.

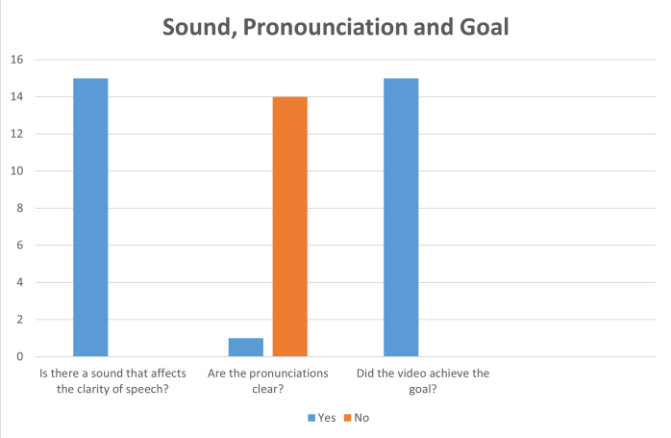

Fig. 8. Results on Clarity of Speech Pronounciation and Goal on Colorful Tables with HTML and CSS

3) Video: 25 Table in HTML and CSS | How to Create Tables | Learn HTML and CSS | HTML Tutorial | CSS Tutorial Lowest Rank in Gensim

Figure 8 shows that 14 respondents found the keywords "relevant, "to the video content with one opting for "probably relevant." This consistency across multiple evaluations underscores the effectiveness of keyword selection. Also Figure shows that 14 participants rated the keywords as "relevant," to the video title with one marking them "probably relevant." This confirms the sustained alignment between keywords and the video title.

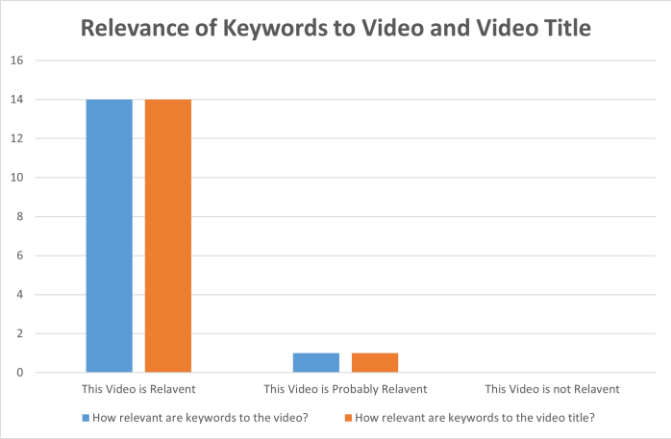

Fig. 9. Results on Relevance of Keywords to Video and Video Title 25 Table in HTML and CSS | How to Create Tables

Figure 9 shows that 14 respondents did not observe sound issues, with one indicating otherwise. This suggests varying experiences with audio clarity across different segments. Figure also shows that most respondents (13) confirmed clarity in pronunciations, while two noted issues. This slight deviation points to isolated challenges in articulation, all 15 participants agreed that the video met its intended goal, maintaining consistent feedback on its overall effectiveness.

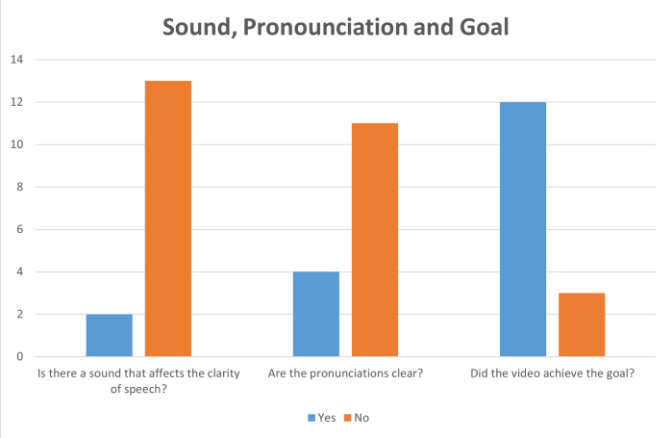

Fig. 10. Results on Clarity of Speech Pronounciation and Goal on 25 Table in HTML and CSS | How to Create Tables

VI. Discussion

The survey results reveal important insights into the effectiveness of the analyzed video content and the algorithms employed NLTK and Gensim. NLTK was better in preprocessing text data, extracting keywords, and assessing their relevance to the video content and titles. The algorithm's precision in identifying key terms is more accurate in understanding of the video's alignment with its intended purpose. Gensim, on the other hand, offered significant value in topic modeling and semantic analysis.

The feedback from teachers and instructors showed the consistent relevance of keywords, suggesting that the natural language processing methods were effective in maintaining alignment. However, the survey also identified some issues with sound clarity and pronunciation, which, while unrelated to the NLP algorithms, require further attention for improvement. By addressing these technical challenges and refining keyword selection based on the survey insights, the platform can achieve a higher level of content quality and engagement. Combination of NLTK and Gensim provided a comprehensive analytical model that not only validated the relevance of the content but also highlighted areas for enhancement, ensuring that the microlearning platform meets the needs of its users.

From previous teacher's and instructors survey result it was proven that NLTK algorithm ranking was better than Gensim. For TF-IDF and cosine similarity, additional libraries like scikit-learn or manually implement these concepts.

Other categories rather than education could be tested and compared. This thesis focused on educational videos.

Gensim: Provides high-level functions and is more efficient for working with large corpora and vector space models. It has built-in support for both TF-IDF and cosine similarity, making it more convenient for these tasks.

To sum it up, NLTK's TFIDF Vectorizer directly computes similarity using the query terms, which tends to be more effective for keyword-based matching. Gensim relies on BoW and a topic model with TFIDF Model, which may not align closely with keyword-specific searches, leading to less precise results for this task.

VII. Conclusion and future work

The microlearning mobile application was developed to deliver concise, engaging, and personalized learning experiences by leveraging cutting-edge natural language processing algorithms. The project implemented and evaluated two prominent algorithms NLTK and Gensim to recommend YouTube videos for microlearning. The effectiveness of these algorithms was assessed through a survey conducted among educators, where they were randomly sent video recommendations generated by both methods. The feedback consistently indicated that NLTK outperformed Gensim in terms of relevance, clarity, and alignment with the intended learning objectives. These findings highlight the potential of NLTK as a

more reliable tool for video content curation in microlearning applications. Overall, the website demonstrates a significant step forward in utilizing NLP techniques to enhance learner satisfaction and engagement. Based on this study's findings and limitations, future work should focus on conducting broader evaluations involving diverse users, including students, to ensure the recommendations meet varied learning needs and preferences. Future work will expand the model to include interactive quizzes, gamification elements, and progress tracking, which can complement the video recommendations and boost learner engagement. We will also combine the two algorithms and use different analytical methods to get more accurate results.

VIII. Acknowledgment

We would like to express our gratitude to all those who assisted us throughout this endeavor, particularly the teachers who contributed to the data collection process

IX. References

[1] P. Blessinger and T. J. Bliss, *Open Education : International Perspectives in Higher Education*. 2016.

[2] D. L. Edy, Widiyanti and Basuki, "Revisiting the impact of project-based learning on online learning in vocational education: Analysis of learning in pandemic covid-19," in *2020 4th International Conference on Vocational Education and Training (ICOVET),* 2020, . DOI: 10.1109/ICOVET50258.2020.9230137.

[3] M. Slimani *et al*, "The Effect of Mental Fatigue on Cognitive and Aerobic Performance in Adolescent Active Endurance Athletes:Insights from a Randomized Counterbalanced, Cross-Over Trial," 2018. Available: https://research.ebsco.com/linkprocessor/plink?id=e01ad304-cd06-34d5-901a-b05988f7a236.

[4] Alyson Gausby, "Microsoft attention spans research report," 2015.

[5] M. S. Shail, "Using Micro-learning on Mobile Applications to Increase Knowledge Retention and Work Performance: A Review of Literature," *Cureus,* vol. 11, *(8),* pp. e5307, 2019. Available: https://research.ebsco.com/linkprocessor/plink?id=1958a23d-2242-3882-8ee3-02efa4afc048. DOI: 10.7759/cureus.5307.

[6] J. C. De Gagne *et al*, "Microlearning in Health Professions Education: Scoping Review," *JMIR Medical Education,* vol. 5, *(2),* pp. e13997, 2019. Available: https://research.ebsco.com/linkprocessor/plink?id=8beade15-f379-3164-8fbf-9728eda10256. DOI: 10.2196/13997.

[7] O. Jomah *et al*, "Micro Learning: A Modernized Education System," *Brain: Broad Research in Artificial Intelligence and Neuroscience,* vol. 7, *(1),* pp. 103–110, 2016. Available: https://research.ebsco.com/linkprocessor/plink?id=49f9fa26-79b9-3831-a664-528cfc3759ed.

[8] P. Galiatsatos *et al*, "The use of social media to supplement resident medical education – the SMART-ME initiative," *Medical Education Online,* vol. 21, *(0),* pp. 1–5, 2016. Available: https://research.ebsco.com/linkprocessor/plink?id=6aa94c5d-53b2-3158-9607-06494e48d9e7. DOI: 10.3402/meo.v21.29332.

[9] S. A. Nikou and A. A. Economides, "Mobile-based assessment: A literature review of publications in major referred journals from 2009 to 2018," *Computers and Education,* vol. 125, pp. 101–119, 2018. Available: https://research.ebsco.com/linkprocessor/plink?id=e2079f06-1770-35a4-ba0a-366142651556. DOI: 10.1016/j.compedu.2018.06.006.

[10] E. Surahman *et al*, "The Effect of Blended Training Model to Improving Learning Outcomes: A Case in Micro Learning Object Training," *2019 5th International Conference on Education and Technology (ICET), Education and Technology (ICET), 2019 5th International Conference On,* pp. 33–38, 2019. Available: https://research.ebsco.com/linkprocessor/plink?id=8a52d78c-8d0c-3158-b292-566a25bd0a23. DOI: 10.1109/ICET48172.2019.8987210.

[11] J. Lin, "Hybrid Translation and Language Model for Micro Learning Material Recommendation," *2020 IEEE 20th International Conference on Advanced Learning Technologies (ICALT),* 2020. Available: https://research.ebsco.com/linkprocessor/plink?id=73470b97-1f75-37d4-95d1-8cf9db62e95e. DOI: 10.1109/icalt49669.2020.00121.

[12] J. Smolle *et al*, "Lecture recording, microlearning, video conferences and LT-platform - medical education during COVID-19 crisis at the Medical University of Graz," *GMS Journal for Medical Education; VOL: 38; DOC11 /20210128/,* 2021. Available: https://research.ebsco.com/linkprocessor/plink?id=db6857c4-8cfb-34b1-824f-73b07b90e408.

[13] A. Erradi, H. Almerekhi and S. Nahia, "Game-Based Micro-learning Approach for Language Vocabulary Acquisition Using LingoSnacks," *2013 IEEE 13th International Conference on Advanced Learning Technologies, Advanced Learning Technologies (ICALT), 2013 IEEE 13th International Conference On,* pp. 235–237, 2013. Available: https://research.ebsco.com/linkprocessor/plink?id=2010e5dd-0758-39b1-926e-58e0344323bf. DOI: 10.1109/ICALT.2013.73.

[14] G. Sun *et al*, "Towards Bringing Adaptive Micro Learning into MOOC Courses," *2015 IEEE 15th International Conference on Advanced Learning Technologies, Advanced Learning Technologies (ICALT), 2015 IEEE 15th International Conference On,* pp. 462–463, 2015. Available: https://research.ebsco.com/linkprocessor/plink?id=1943e324-6675-34ac-b29b-6068d3dd8da7. DOI: 10.1109/ICALT.2015.26.

[15] N. K. Hayles, "Hyper and Deep Attention: The Generational Divide in Cognitive Modes," *Profession,* pp. 187–199, 2007. Available: http://www.jstor.org/stable/25595866.

[16] P. J. Guo, J. Kim and R. Rubin, "How video production affects student engagement: An empirical study of MOOC videos," in *Proceedings of the First ACM Conference on Learning@ Scale Conference,* 2014, .

[17] G. Sun *et al*, "MLaaS: A Cloud-Based System for Delivering Adaptive Micro Learning in Mobile MOOC Learning," *IEEE Transactions on Services Computing, Services Computing, IEEE Transactions on, IEEE Trans.Serv.Comput.,* vol. 11, *(2),* pp. 292–305, 2018. Available: https://research.ebsco.com/linkprocessor/plink?id=e969b8a5-1aa9-3082-97f0-978c6cc46785. DOI: 10.1109/TSC.2015.2473854.

[18] A. Gill *et al*, "Reacting to the coronavirus: A case study of science and engineering education switching to online learning in a sino-foreign higher education institution," in *2020 International Conference on Open and Innovative Education (ICOIE 2020), Hong Kong, China,* 2020, .

[19] E. Y. Sözmen, O. Karaca and A. H. Bati, "The Effectiveness of Interactive Training and Microlearning Approaches on Motivation and Independent Learning of Medical Students during the COVID-19 Pandemic," *Innovations in Education and Teaching International,* vol. 60, *(1),* pp. 70–79, 2023. Available: https://research.ebsco.com/linkprocessor/plink?id=fb58b2a0-c744-3342-a46b-cb0e66c7ffc0.

[20] B. M. Miller *et al*, "Quiz-Based Microlearning at Scale: a Rapid Educational Response to COVID-19," *Medical Science Educator,* vol. 31, pp. 1731–1733, 2021. Available: https://research.ebsco.com/linkprocessor/plink?id=b61cb695-a67d-3ad2-938d-1f0265b5e38f. DOI: 10.1007/s40670-021-01406-8.

[21] D. Khurana *et al*, "Natural language processing: state of the art, current trends and challenges," *Multimedia Tools Appl,* vol. 82, *(3),* pp. 3713–3744, 2023. .

[22] X. Chen *et al*, "Application and theory gaps during the rise of artificial intelligence in education," *Computers and Education: Artificial Intelligence,* vol. 1, pp. 100002, 2020. .

[23] C. V. McDonald, "STEM Education: A review of the contribution of the disciplines of science, technology, engineering and mathematics." *Science Education International,* vol. 27, *(4),* pp. 530–569, 2016. .

[24] Dr. Dhara Ashish Darji and Dr. Sachinkumar Anandpal Goswami, "The Comparative study of Python Libraries for Natural Language Processing (NLP)," *International Journal of Scientific Research in Computer Science, Engineering and Information Technology,* vol. 10, *(2),* pp. 499–512, 2024. . DOI: 10.32628/CSEIT2410242.

[25] M. Rayati, "Neuro-Linguistic Programming and Its Applicability in EFL Classrooms: Perceptions of NLP-Trained English Teachers." *Language Teaching Research Quarterly,* vol. 24, pp. 44–64, 2021. .

[26] R. Egger and E. Gokce, "Natural language processing (NLP): An introduction: Making sense of textual data," in *Applied Data Science in Tourism: Interdisciplinary Approaches, Methodologies, and Applications*Anonymous 2022, .

[27] M. Kumar, L. Khan and H. Chang, "Evolving techniques in sentiment analysis: a comprehensive review," *PeerJ Computer Science,* vol. 11, pp. e2592, 2025. .

[28] S. Pal *et al*, "Learner question's correctness assessment and a guided correction method: enhancing the user experience in an interactive online learning system," *PeerJ Computer Science,* vol. 7, pp. e532, 2021. .

[29] A. K. Ohm and K. K. Singh, "Study of tokenization strategies for the santhali language," *SN Computer Science,* vol. 5, *(7),* pp. 807, 2024. .

[30] P. Lahoti, N. Mittal and G. Singh, "A survey on nlp resources, tools, and techniques for marathi language processing," *ACM Transactions on Asian and Low-Resource Language Information Processing,* vol. 22, *(2),* pp. 1–34, 2022. .

[31] M. Vu, "Building Topic Modelling on Theses Abstracts Data: Thesis Supervisors Finder for Students," 2021. .

[32] M. H. Teodorescu, "Natural language processing techniques in management research," in *Research Handbook on Artificial Intelligence and Decision Making in Organizations*Anonymous 2024, .

[33] M. Pandey *et al*, "Hybrid Technique of Topic Modelling and Text Summarization: A Case Study on Predicting Trends in Green Computing." *Int.J.Perform.Eng.,* vol. 20, *(3),* pp. 139–148, 2024. .

[34] Z. Wu *et al*, "Recent developments in parallel and distributed computing for remotely sensed big data processing," *Proc IEEE,* vol. 109, *(8),* pp. 1282–1305, 2021. .

[35] P. Caratozzolo, J. Rodriguez-Ruiz and A. Alvarez-Delgado, "Natural Language Processing for Learning Assessment in STEM," *2022 IEEE Global Engineering Education Conference (EDUCON), Global Engineering Education Conference (EDUCON), 2022 IEEE,* pp. 1549–1554, 2022. Available: https://research.ebsco.com/linkprocessor/plink?id=cfb3d63c-4c7b-39a3-b839-af095c51f04f. DOI: 10.1109/EDUCON52537.2022.9766717.

[36] N. Agarwal and Vijayalaxmi, "Covid-19 impact on mental health: Sentiment analysis using NLTK," in *2023 6th International Conference on Contemporary Computing and Informatics (IC3I),* 2023, . DOI: 10.1109/IC3I59117.2023.10398068.

[37] M. M. Haider *et al*, "Automatic Text Summarization Using Gensim Word2Vec and K-Means Clustering Algorithm," *2020 IEEE Region 10 Symposium (TENSYMP), Region 10 Symposium (TENSYMP), 2020 IEEE,* pp. 283–286, 2020. Available: https://research.ebsco.com/linkprocessor/plink?id=a535c732-8b1c-35d1-89c2-6b4a0c92baa2. DOI: 10.1109/TENSYMP50017.2020.9230670.

[38] K. Kulkarni and R. Padaki, "Video based transcript summarizer for online courses using natural language processing," in *2021 IEEE International Conference on Computation System and Information Technology for Sustainable Solutions (CSITSS),* 2021, . DOI: 10.1109/CSITSS54238.2021.9683609.

[39] P. S. Ponmalar, S. Singh and R. K. Agrahari, "Deep Learning Video Streaming Web Application with AI Botsusing NLP," *2023 International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICSES), Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICSES), 2023 International Conference On,* pp. 1–6, 2023. Available: https://research.ebsco.com/linkprocessor/plink?id=50da65b7-ec43-3458-a907-63da459dfd48. DOI: 10.1109/ICSES60034.2023.10465475.

[40] A. Kiefer, "Improving Automatic Transcription Using Natural Language Processing," 2024. .

[41] (October 5). *youtube-search*. Available: https://pypi.org/project/youtube-search/.

[42] . *pytube 15.0.0 documentation*. Available: https://pytube.io/en/latest/.

[43] C. Deuerlein *et al*, "Human-robot-interaction using cloud-based speech recognition systems," *Procedia CIRP,* vol. 97, pp. 130–135, 2021. Available: https://research.ebsco.com/linkprocessor/plink?id=96e16d45-2e55-39d7-a757-702ccf1568ad. DOI: 10.1016/j.procir.2020.05.214.

[44] J. Samuel, R. Kashyap and A. Kretinin, "When the Going Gets Tough, The Tweets Get Going! An Exploratory Analysis of Tweets Sentiments in the Stock Market," *American Journal of Management,* vol. 18, 2018. Available: https://research.ebsco.com/linkprocessor/plink?id=84582b2a-e9e4-396b-a867-78e63f225964. DOI: 10.33423/ajm.v18i5.251.

[45] S. J. Mielke *et al*, "Between words and characters: A brief history of open-vocabulary modeling and tokenization in NLP," *arXiv Preprint arXiv:2112.10508,* 2021. .

[46] K. D. Bafitlhile, "A context-aware lemmatization model for setswana language using machine learning," 2022. .

[47] L. Xiang, "Application of an Improved TF-IDF Method in Literary Text Classification," *Advances in Multimedia,* vol. 2022, *(1),* pp. 9285324, 2022. .

[48] M. Kirişci, "New cosine similarity and distance measures for Fermatean fuzzy sets and TOPSIS approach," *Knowledge and Information Systems,* vol. 65, *(2),* pp. 855–868, 2023. .