Final Project Report

On

Course Review on Coursera

Asmita Deshpande, Mostafa Cham, Parthiv Gandhi

Department of Information Systems, University of Maryland, Baltimore County

IS603: Decision Making Support Systems

December 11, 2022

INDEX

1.Abstract

Educational videos have emerged as a dominant medium for educational purposes in many Massive Online Course (MOOC) platforms. One such online platform is Coursera. Before enrolling in an online course, students always look at the ratings and reviews. However, reading all the information, particularly the course reviews, might take a lot of time. Ratings and reviews are always a primary consideration for anyone enrolling in a new course. It takes a lot of time to go through each and every review before selecting the course. The purpose of this research paper is to conduct a data analysis of a Coursera course review to evaluate the effectiveness of online learning platforms. Using a dataset of over 1.4 Million reviews from various Coursera courses, the study will analyze the feedback provided by students on various aspects of the courses including content quality, instructor expertise, and overall satisfaction. The results of the data analysis will provide insights into the effectiveness of online learning platforms and highlight any potential areas for improvement. This research will be valuable for educators and institutions looking to incorporate online learning into their curriculums and for learners seeking to make informed decisions about their educational choices. By conducting data analysis of Coursera course reviews, we will be able to gain valuable insights into the effectiveness of online learning platforms and identify any potential areas for improvement. This information could be used by educators and institutions to enhance the quality of their online courses and improve the overall learning experience for students. Additionally, it could be helpful for students who are looking to make informed decisions about their educational choices by providing them with information on the quality of different online courses. Overall, this research could have a positive impact on the field of online education.

2.Introduction

In the last few years, the way of education and its delivery has changed. With the easy use of the internet and software technologies, most of the knowledge can be gained online. Apart from that online information can be gathered and learned from any place in the world due to advancements in the technological field. These advancements help us bring access to quality education irrespective of location and timing as long we have access to a computer and the internet.

Data analysis is the process of examining, cleaning, transforming, and modeling data with the goal of discovering useful information, informing conclusions, and supporting decision-making. In the context of a course review, data analysis could involve examining students' ratings and feedback on the course, looking for patterns and trends in the data, and using statistical techniques to analyze the results. This could help identify areas where the course is performing well and areas where it could be improved, providing valuable insights for the course instructor and educational institution.

Data analysis has become an increasingly important tool for understanding and improving various aspects of our lives, including education. There are multiple online learning platforms like Udemy,

Edx, Skillshare and Codecademy. Users can access these online courses through a web browser. Ratings and reviews are considered to be a major factor by students at the time of enrolling. It can be time-consuming to read all the reviews and ratings of the course review.

In this research paper, we will explore the use of data analysis to gain insights into the effectiveness of a Coursera course. By analyzing the course reviews left by students, we will attempt to answer questions such as: What are the most common themes mentioned in the reviews? Do students who leave positive reviews have different characteristics than those who leave negative reviews? What factors are most strongly correlated with a positive review? By providing answers to these questions, we hope to gain a better understanding of the factors that contribute to a successful online learning experience.

The data collected from the course reviews were cleaned and pre-processed to remove any irrelevant or missing information. The cleaning and preprocessing of the data involved several steps. Firstly, the reviews were checked for typos or errors, and these were corrected. Secondly, any reviews that were incomplete or did not provide relevant information were removed from the dataset. Thirdly, the reviews were standardized to ensure that they were in a consistent format and contained the necessary information. This included removing any unnecessary punctuation, capitalization, and special characters.

A clustering algorithm was used to determine the relationship between the students' overall satisfaction and their ratings for the course content and instructor. Sentiment analysis was not used in this research, as we focused on analyzing the content and ratings provided by the students in their reviews, rather than their emotional responses. In this scenario, clustering algorithms could be used to group the students' ratings and feedback into different clusters based on their overall satisfaction with the course. This could help identify any patterns or trends in the data, and potentially uncover any relationship between the student's satisfaction with the course and their ratings for the course content and instructor. By using clustering algorithms, the researchers could gain valuable insights into how different factors contribute to overall student satisfaction with the course.

3. Data and Method:

3.1. Data Collection:

In this research study, we have used the dataset of course reviews on Coursera from the Kaggle website. This dataset is consist of 1.4 million samples of course reviews posted by participants in more than 600 courses on Coursera. This Dataset includes two datasets; the first dataset is related to courses and consists of course name, institution name, course URL, and course ID. The second one is for course reviews and has five variables which are review text, name of the reviewer who wrote the review, date of the review, rating score, and course ID as a foreign key.

However, the size of the data is too much more than 270 MB and it requires a huge computational power. Therefore, we made a sample of 300,000 data from this data set for our case study.

3.2. Data preprocessing:

After collecting data we could not simply take the raw data and run it through machine-learning algorithms. We first need to preprocess the data that we need to pass to the clustering algorithms. Preprocessing means making the data more understandable for the algorithm. In order to do that, all of the unnecessary letters and words should be removed from the review data to lower the complexity of our data and avoid bias in the result of the algorithm. Accordingly, first, we remove the null values from the data set. The second step is to remove numbers and special characters which include emojis from reviews by replacing them with empty spaces.

The next step is dropping stop words, which are the words used commonly in text and do not have a huge impact on the meaning of the text. As a result, they do not have a positive impact on the clustering process and we could remove them from the text. For this purpose, we provide a list of stop words in the English language and applied a for loop on the reviews column, and replace the stop words with empty spaces. After removing the stop words we should apply Stemming to review texts.

Stemming is the process of reducing inflected words to their word root. Although inflected words give meaning to sentences, they might confuse the clustering algorithms. In order to apply to stem to the text, we need to tokenize the text. To clarify, each text is a combination of characters and words and could be converted to a list of words. In the concept of text mining, this process is called Tokenizing.

After applying all the aforementioned steps, we have a clean list of words for each review. However, there is one more step before applying the data to the clustering algorithm. Instead of passing a list of words, we should pass a list of numbers as the reviews to the clustering algorithms. For this purpose, we use TF-IDF vectorization. The terms "Tf" and "tf-idf" stand for term frequency and inverse document frequency, respectively. This word-weighting technique is widely used in information retrieval and is very effective in document classification. The goal of using tf-idf is to scale down the impact of tokens that occur very frequently in a group of texts, which, empirically speaking, are less informative than features that only appear in a tiny portion of the training corpus. Check Figure 1.

3.3. Clustering Algorithms:

For clustering, we have used two algorithms for benchmarking. The goal of using two algorithms is to compare the results of each of them in order to choose which algorithm works better with our data.

The first algorithm we used in this research study is Kmeans. This algorithm is one of the most popular and frequently used clustering algorithms. The iterative Kmeans algorithm attempts to divide the dataset into K unique, non-overlapping subgroups (clusters), each of which contains a single data point. While keeping the clusters as distinct (far) apart as possible, it aims to make the intra-cluster data points as comparable as possible. It allocates data points to clusters in a way that minimizes the sum of the squared distances between the data points and the cluster centroid. The homogeneity (similarity) of the data points within a cluster increases as the amount of variance within the cluster decreases.

The second algorithm is the Feature Agglomeration algorithm which is one of the Hierarchical clustering algorithms. This family of algorithms creates nested clusters by gradually merging or breaking clusters. A tree is used to depict the clusters' hierarchical structure (or dendrogram). The unique cluster at the tree's base has all of the samples, while the clusters at the tree's leaves each contain a single sample. Each observation begins in its own cluster, which is then gradually combined by the Agglomeration Clustering object to create a hierarchical clustering. The metric employed for the merging approach is determined by the linkage criteria. In other words, linkage determines which distance to use between sets of features and the algorithm merges the pairs of the cluster that minimize this distance. There are for types of linkage for this algorithm:

- “ward” minimizes the variance of the clusters being merged
- “complete”: maximum distances between all features of the two sets.
- “average” uses the average of the distances of each feature of the two sets.
- “single” uses the minimum of the distances between all features of the two sets.

3.4. Evaluation Metrics:

To evaluate and compare the algorithms we have used two evaluation metrics; Davies-Bouldin Index and Silhouette Coefficient.

Davies-Bouldin Index score is defined as the average similarity measure of each cluster with its most similar cluster, where similarity is the ratio of within-cluster distances to between-cluster

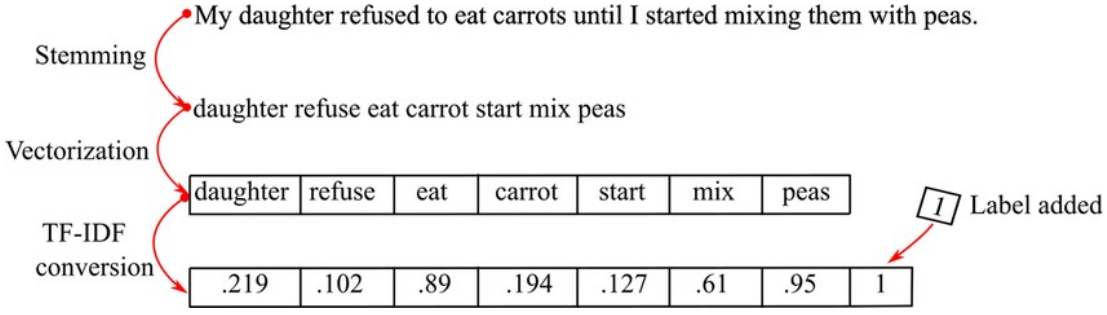

Figure 1: Stemming and TF-IDF Vectorization

distances. Therefore, groups that are more evenly spaced apart will score higher. In this metric minimum score is zero and the lower score indicates better clustering.

Silhouette Coefficient metric takes account of the mean distance of each data point with intra-cluster samples and nearest cluster samples, where -1<= score <=1 and 1 is the best score, -1 is the worst score and 0 indicates overlapping clusters.

4. Results:

In this case study because of the huge amount of data and lack of computational power we used half of the sample data, which is almost 150,000 samples, in order to tune our model and find the best value for the hyperparameters. The first hyperparameter we need to tune is the best type of linkage for feature agglomeration clustering in the case of our data. For this purpose, we applied

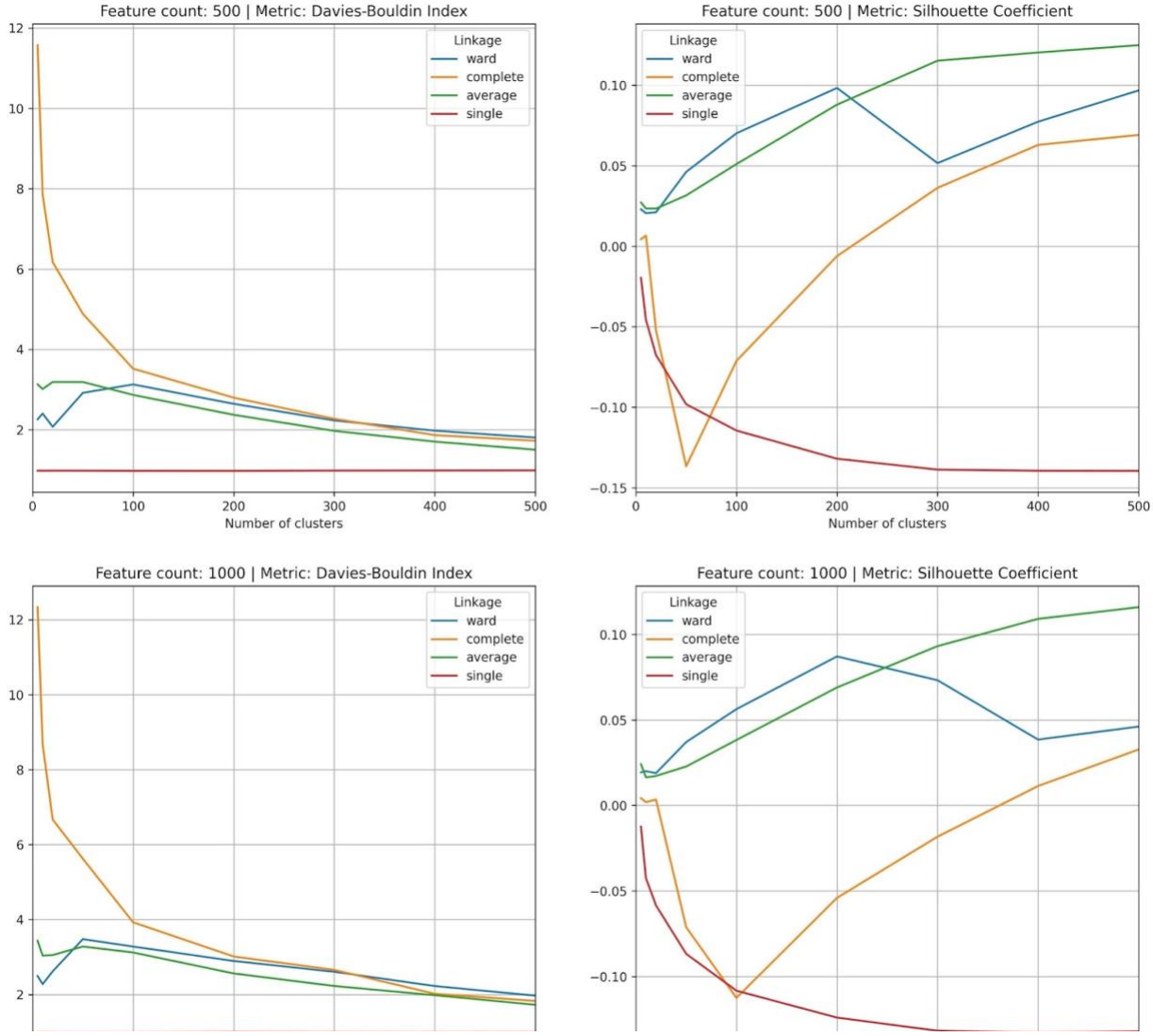

Figure 2: Plot of different linkages in feature agglomeration clustering. The Y axis in each graph indicates the score of metric mentioned at the top of the graph.

the algorithm to different metrics and the number of top feature counts. It is worth mentioning that the number of top feature counts is the number of top impactful words selected to be passed to the clustering algorithm. As shown in Figure 2 best Linkage for our case study is the “Average” linkage.

After selecting the proper linkage for feature agglomeration clustering, we need to find the best number of top features for both clustering algorithms. Simultaneously, we should find the proper number of clusters to apply the algorithms based on that to our data. Accordingly, we applied two algorithms coupled with two evaluation metrics to 150,000 samples of course review. Based on

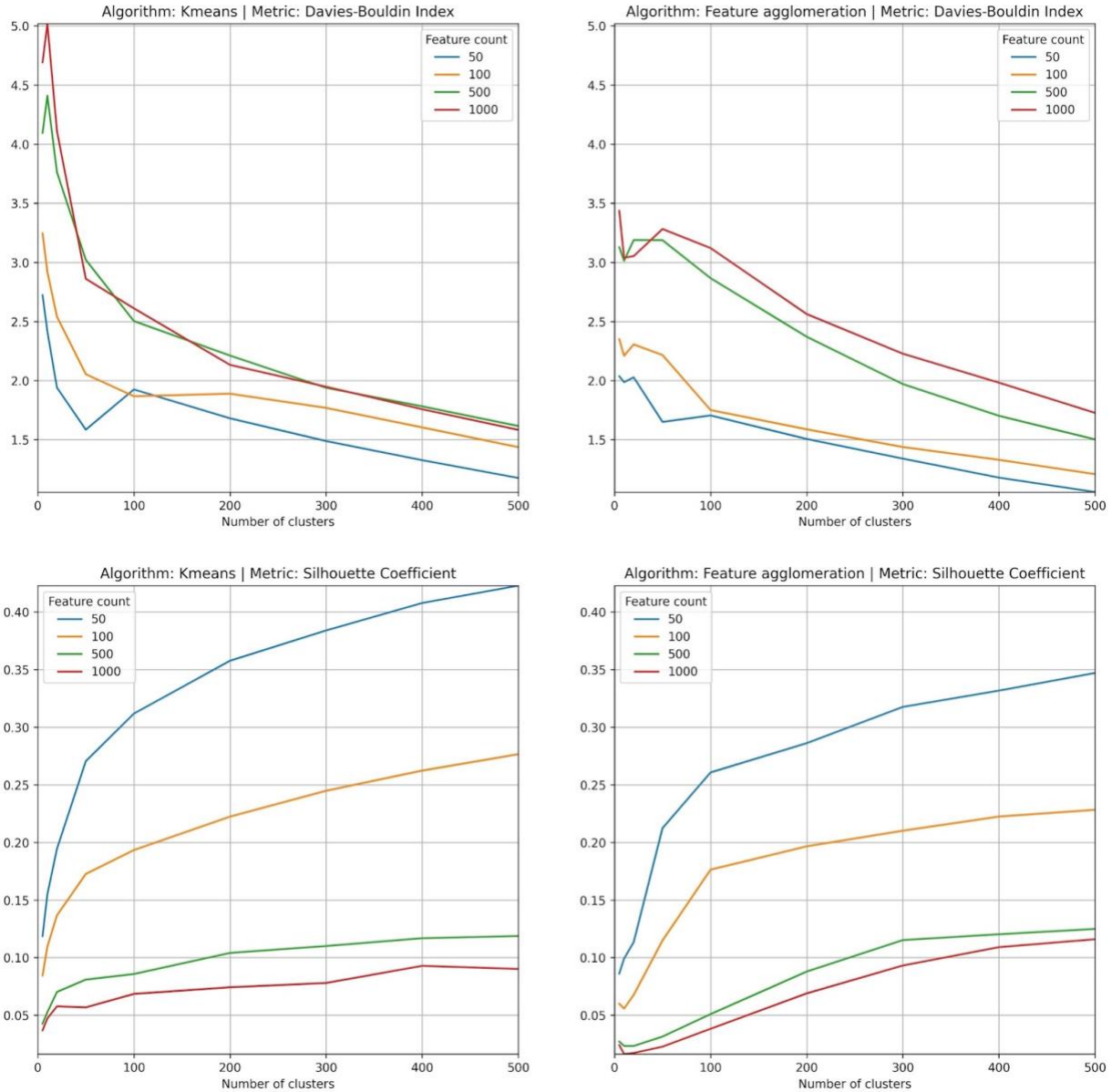

Figure 3: Result for top feature count and number of clusters. The Y axis in each graph indicates the score of metric mentioned at the top of the graph.

the results from Figure 3 we can figure out that the best number of top features and the best number of clusters are, 50 and 500 respectively.

By finding the best values for the hyperparameters in our algorithms we applied the algorithms to a larger set of course reviews data with a number of top features equals to 50 and number of cluster equal to 500. We used the Average linkage for feature agglomeration clustering. The evaluation results are mentioned as benchmarking in Table 1.

	Silhouette Coefficient	Davies-Bouldin Index
Kmeans	0.4226	1.1765
Feature Agglomeration	0.3468	1.0578

Table 1: Final Evaluation Result.

5. Conclusion

Overall, the results of the data analysis showed that online learning platforms, such as Coursera, can be effective in providing high-quality education to learners. However, the study also highlighted the need for continuous improvement and addressing any challenges faced by students in order to maximize the effectiveness of online learning.

This process sets a benchmark. We can use the clustered data in the near future. We can send this information to the instructors or the Coursera team to improve the course or modify the course and make any necessary changes.

Coursera Review helps course-taking seekers to gain quick insights including positive and negative reviews and also enables to make a quick comparison between multiple courses. This process can help save a lot of time for the course seeker in the course selection process. The Reviews which are put up below every course can also help the instructors or the course provider to understand the strengths and weaknesses and modify the course in the near future.

6. Future work

1. We can use all the positive reviews and keywords as an opportunity to promote and encourage more students to join the course.
2. We can use the information on which users are taking which courses, as an opportunity to recommend them more similar or advanced courses, recommending them books.
3. We can take this opportunity to also recommend jobs related to those specific courses. Inform and give a report to Coursera/Instructors about how they can improve a specific course or modify it to make it more attractive and informative. We can promote and market the courses which have great reviews.

7. REFERENCES

1. H. Y. Chan, R. Rajamohan, K. H. Gan and N. -H. Samsudin,(2021) "Text Analytics on Course Reviews from Coursera Platform," IEEE International Conference on Artificial Intelligence in Engineering and Technology (IICAIET), 2021.

2. B. Wu and Y. Zhou,(2020) "The Impact of MOOC Instructor Group Diversity on Review Volume and Rating— Coursera Specialization as an Example," in IEEE Access, vol. 8.

3. Donelson, Curtis & Sutter, Carolyn & Pham, Giang & Narang, Kanika & Wang, Chen & Yun, Joseph. (2021). Using a Machine Learning Methodology to Analyze Reddit Posts regarding Child Feeding Information. Journal of Child and Family Studies.

4. Cerro Martínez, J.P., Guitert Catasús, M. & Romeu Fontanillas, T. (2020). Impact of using learning analytics in asynchronous online discussions in higher education. Int J Educ Technol High Educ 17, 39 .