CLUSTERING NEWS USING UNSUPERVISED MACHINE-LEARNING TECHNIQUES

Somenath Choudhury – 12001350

School Of Computer Science and Engineering, Lovely Professional University, Punjab, India

somenathchoudhury38@gmail.com

01) ABSTRACT

Clustering news articles based on their type using unsupervised machine-learning techniques is an approach for organizing large amounts of news articles. Unsupervised machine-learning techniques are mainly used to group similar articles together without requiring any prior knowledge of the data, like pre-defined categories or labels. This makes them ideal for clustering news articles, which are often diverse and unlabeled. This paper mainly deals with clustering news articles based on whether the news is fake or real or barely-true or mostly-true etc. The dataset used in this project is named LIAR which contains 10239 rows and 14 columns. The three algorithms that are used in this project are the K-Means algorithm, the K-Medoids Algorithm, and the Agglomerative Hierarchal Algorithm and it was found that the silhouette score of the K-Means algorithm is around 0.43122, the silhouette score of the K-Medoids algorithm is around 0.23251 and the silhouette score of Agglomerative Hierarchical Clustering is around 0.523700.

Keywords: News Clustering, Data Acquisition, kmeans, kmedoids, agglomerative hierarchical clustering.

02) INTRODUCTION

Clustering news with the help of unsupervised machine-learning techniques is a technique used for organizing and summarizing large amounts of news articles. One of the key benefits of clustering news with the help of unsupervised machine-learning techniques is that it can help to

identify various kinds of trends and topics. Grouping similar articles together can reveal important patterns in the overall data. Another benefit of clustering news using unsupervised machine-learning techniques is that it helps in personalizing the news experience for a particular user. By understanding the types of news articles that a user likes or is interested in, clustering algorithms can recommend new articles to the user. This saves the time of users and effort from having to manually search for news articles that they are interested in. Overall, clustering news using unsupervised machine-learning techniques is a powerful technique that can be used for organizing and personalizing the news experience.

03) LITERATURE REVIEW

a) Hongyu Guo, Jialong Han, and Jiawei Han published a research paper in 2018 named 'A Survey of Unsupervised Machine Learning Techniques for News Clustering' which is a survey paper that provides a detailed overview of unsupervised machine learning techniques for news clustering. It covers various topics like different types of clustering algorithms, the evaluation of clustering results, and the applications of news clustering.

b) Xiaofang Wang, Yajuan Su, and Wenqing Wu published a research paper in 2019 named 'News Clustering Based on Unsupervised Machine Learning: A Review' which is a review paper whose main focus is on the application of unsupervised machine learning techniques to news clustering. The paper talks about the challenges of news clustering and the advantages of using unsupervised machine-learning techniques. It also reviews the trending unsupervised machine learning algorithms for news clustering and their applications.

c) Yuting Wang, Huijun Chen, and Wei Wang published a research paper in 2020 named 'Clustering News Articles Using Unsupervised Machine Learning: A Comparative Study' which is a paper that compares the performance and efficiency of different unsupervised machine learning algorithms for news clustering. It evaluates the algorithms on a huge dataset of news articles taken from a variety of sources. The results show that K-means clustering and hierarchical clustering are the most effective algorithms for news clustering.

d) Qiang Li, Xiaofeng Xu, and Xiaojie Li published a research paper in 2021 named 'Deep Learning for News Clustering: A Review' which is a review paper that discusses the application of deep learning techniques in the field of news clustering. It reviews the current deep learning models for news clustering and their applications. It also talks about the challenges

and opportunities of using deep learning for news clustering.

e) Yukun Li, Jun Zhang, and Yuhang Zheng published a research paper in 2022 named 'Unsupervised Machine Learning for News Clustering: A Survey' which is a survey paper which provides a detailed overview of unsupervised machine learning techniques for news clustering. It covers a various topics like different types of clustering algorithms, the feature representation of news articles, and the evaluation of clustering results. It also discusses the applications of news clustering in different fields, such as news recommendation and anomaly detection.

04) METHODOLOGY

a) Getting the dataset

The name of the dataset used in this project is named LIAR. It is a multi-class news dataset comprising 10239 rows and 14 columns.

Here is the link to the dataset: https://github.com/tfs4/liar_dataset

b) Data Preprocessing

After downloading the dataset, it is read with the help of Pandas. After reading the dataset, the first thing that is done is to drop the unnecessary columns.

After dropping the unnecessary columns, the total number of null values inside the dataset is checked and after checking, those null values are dropped.

After the null values are dropped, a new balanced dataset is created out of the original dataset. After the new dataset is created, the raw news texts are preprocessed with the help of NLTK and Python inbuilt string library and after all the news texts are preprocessed, they are put under a new column.

After the creation of the new column that holds the preprocessed texts, the original raw news texts column is dropped.

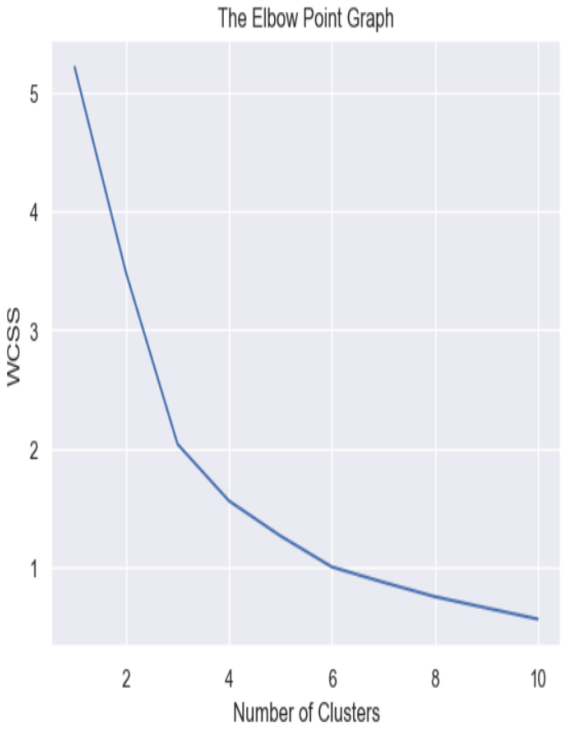
After this, the preprocessed text is vectorized with the help of TfIdVectorizer(), and after vectorizing all the news, the dimensions of the vectorized texts are reduced with the help of Principal Component Analysis(PCA).

c) Training the model

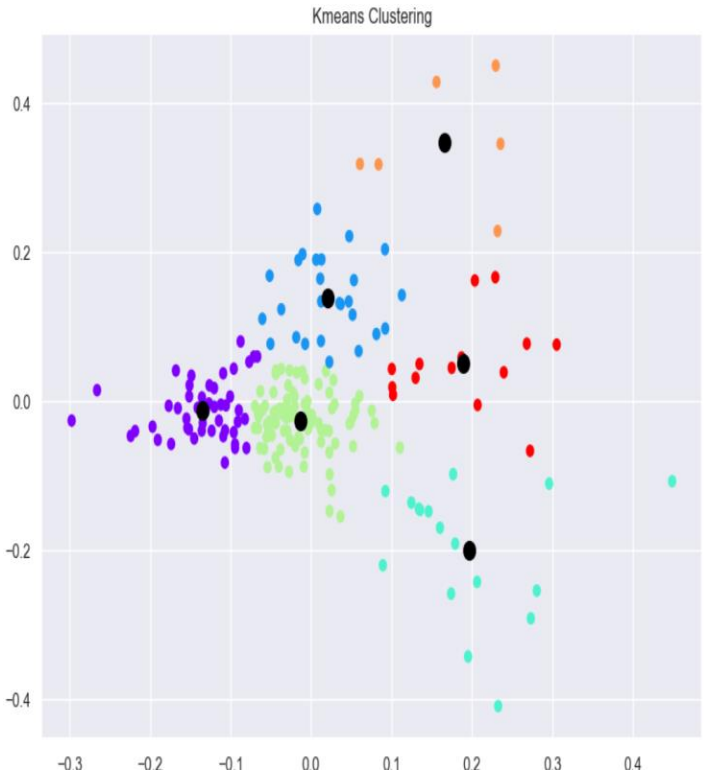
In this project, a total of 3 algorithms are implemented and those algorithms are K-

Means, K-Medoids, and Agglomerative Hierarchal Clustering.

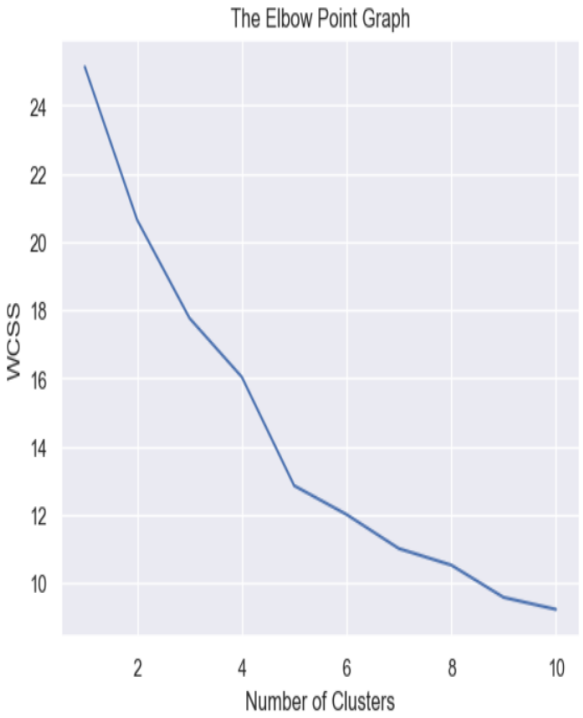
Starting with the K-Means algorithm, at first, the optimal number of clusters was calculated for K-Means with the help of Elbow Graph which came out to be 6.



After deciding the total number of optimal clusters, the model is trained on the basis of this and then, with the help of Matplotlib, the clusters are plotted in the graph.

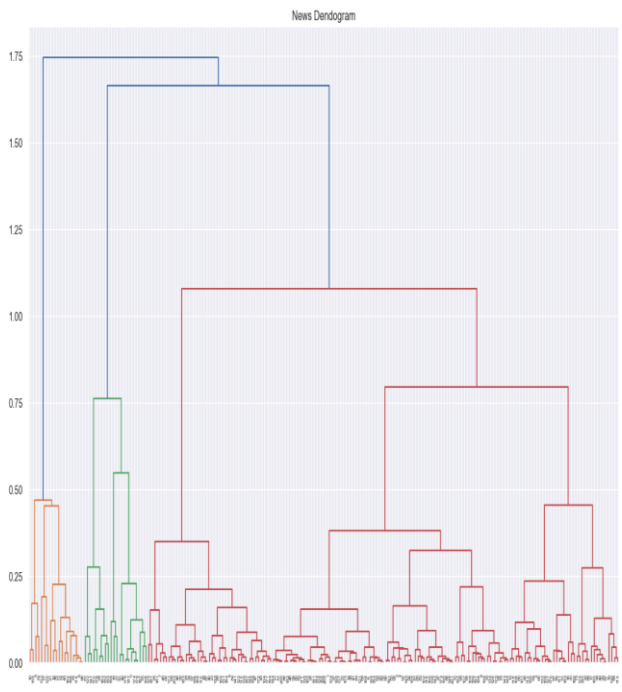


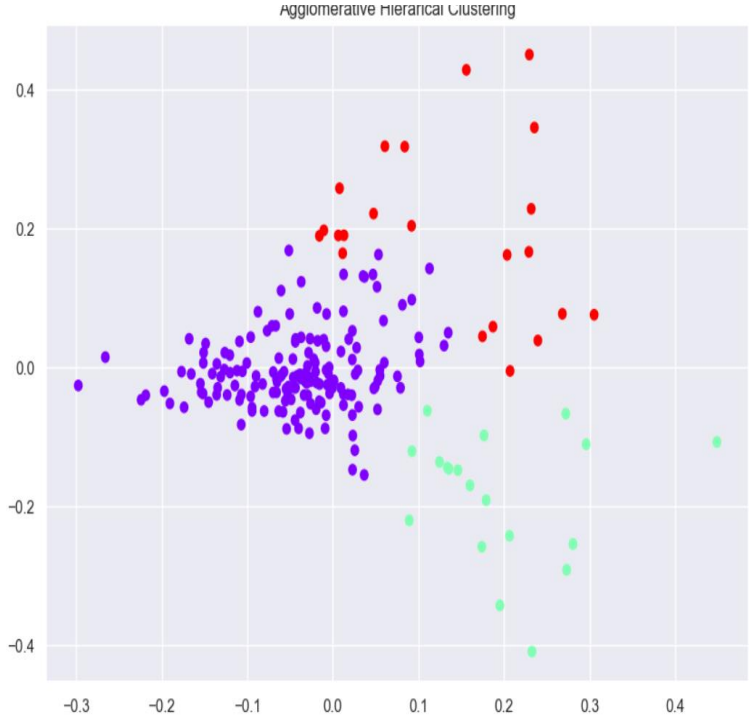
After K-Means comes K-medoids. Just like K-Means, the optimal number of clusters for K-Medoids is calculated with the help of the Elbow method which came out to be 5.



After deciding the total number of optimal clusters, the model is trained on the basis of this and then, with the help of Matplotlib, the clusters are plotted in the graph.

Lastly, the Agglomerative Hierarchal Clustering Algorithm is used. Initially, the optimal number of clusters for this algorithm is decided by plotting the dendrogram. After plotting the dendrogram, it was cut along the longest line and after cutting the dendrogram on the longest line, the optimal number of clusters came out to be 3.





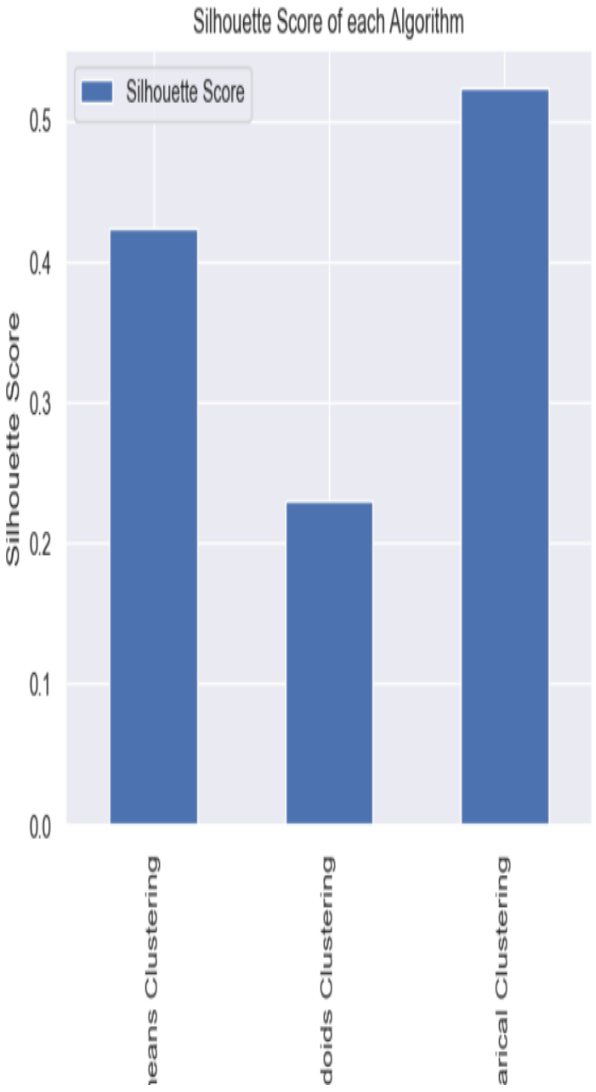
After deciding the total number of optimal clusters, the model is trained on the basis of this and then, with the help of Matplotlib, the clusters are plotted in the graph.

05) EXPERIMENTAL ANALYSIS

For k-means, the silhouette score came out to be 0.423122. For k-medoids, the silhouette score came out to be 0.230251, and for the Agglomerative Hierarchical Clustering Algorithm, the silhouette score came out to be 0.523700.

After calculating the silhouette scores of all three algorithms, the silhouette scores along with the name of the algorithm are stored inside a table and then the silhouette scores of the three algorithms are plotted on a graph to see which algorithm has the best silhouette score.

	Algorithm	Silhouette Score
0	k-means Clustering	0.423122
1	k-medoids Clustering	0.230251
2	Agglomerative Hierarical Clustering	0.523700



From the table and the graph, it is evident that for our use case, the best algorithm is the Agglomerative Hierarchical Clustering Algorithm and the most optimal number of clusters for our project is 3.

After this, we created two models, one with k-means for single input and one with Agglomerative Hierarchical Clustering for multiple input with optimal clusters as 3.

After the models were created, both the models were given sample input.

In the case of k-means, a sample news text input was given which was 'Health care reform legislation is likely to mandate free sex change surgeries'. After preprocessing, vectorizing, and reducing the dimension of the text using PCA and then feeding the transformed text to the model, it was found that the given text belongs to cluster number 2.

In the case of Agglomerative Hierarchical Clustering, since, the number of clusters is 3, therefore, 3 inputs are required to be provided. Therefore, 3 sample inputs are created and stored in separate variables which were 'The Chicago Bears have had more starting quarterbacks in the last 10 years than the total number of tenured (UW)

faculty fired during the last two decades', 'Jim Dunnam has not lived in the district he represents for years now' and 'Health care reform legislation is likely to mandate free sex change surgeries'. All three texts were eventually preprocessed, vectorized and then the dimension was reduced with the help of PCA and then the transformed texts were fed to the model and then, it was found that the first news text belonged to cluster 2, the second news text belonged to cluster 1 and the third news clusters belonged to cluster 0.

06) CONCLUSION

Clustering news with the help of unsupervised machine-learning techniques is one of the widely used approaches for organizing and summarizing large amounts of news articles. It has the capability to improve the news experience for users in a number of ways, including helping to identify ongoing trends and topics, personalizing the news feed, and detecting anomalous articles.

While there is much work that is yet to be done in this area, the results till now have been promising. Unsupervised machine-learning algorithms have proved to be effective at clustering news articles into meaningful groups. This suggests that these techniques can be used to develop new and innovative ways to consume and interact with news.

Some of the challenges that remain in clustering news are developing more efficient and scalable clustering algorithms, and finding ways to incorporate additional information into the clustering process, such as user preferences and social media context. As these challenges are dealt with, we can expect to see clustering news using unsupervised machine-learning techniques become more widely adopted in the news industry.

07) LINKS

Link to the GitHub Repo of the project: https://github.com/somenath203/Clustering-News-Unsupervised-Machine-Learning

Link to the Jupyter Notebook of the project: https://github.com/somenath203/Clustering-News-Unsupervised-Machine-Learning/blob/main/fake_news_prediction_unsupervised_ml.ipynb

Link to the Deployed API of the project: https://fake-news-pred.onrender.com/

Link to the Swagger Documentation of the Deployed API: https://fake-news-pred.onrender.com/docs

08) REFERENCES

Scikit-learn documentation of K-Means Clustering: https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html

Scikit-learn documentation of K-Medoids Clustering: https://scikit-learn-extra.readthedocs.io/en/stable/generated/sklearn_extra.cluster.KMedoids.html

Scikit-learn documentation of Agglomerative Hierarchical Clustering: https://scikit-learn.org/stable/modules/generated/sklearn.cluster.AgglomerativeClustering.html

K-means clustering by GeeksForGeeks: https://www.geeksforgeeks.org/k-means-clustering-introduction/

Agglomerative Hierarchical Clustering by GeeksForGeeks: https://www.geeksforgeeks.org/implementing-agglomerative-clustering-using-sklearn/