Clustering and Retrieval of News articles using Natural Language Processing

Dubey Dhanraj Shevendrakumar[1]

[1]Student, Department of Mechanical Engineering, AIML Specialization, Symbiosis Institute of Technology, SIU, Pune, Maharashtra, India.

Abstract - This study focused on organizing and finding news articles from a large dataset using Natural Language Processing (NLP). In journalism, it's important to manage information effectively due to the increasing amount of data available. Clustering groups similar articles together, while retrieval helps find specific articles based on certain criteria. These techniques help journalists and researchers identify patterns and extract useful information. The study used the Reuters dataset, which contains a lot of news stories. Different clustering algorithms were tested, and one called Agglomerative Clustering performed the best. For retrieval, a system was created that finds articles like user queries using cosine similarity. The study had some limitations, such as unclear categorization and a complex dataset. Future work could involve trying other clustering algorithms, considering additional features like author and publication date, and using a more diverse dataset.

***Key Words*:** Machine Learning, Retrieval, Reuters Dataset, News Articles

1.INTRODUCTION

In the current era, where information is overwhelming, the clustering and retrieval of news articles using Natural Language Processing (NLP) techniques have become essential for effectively managing and accessing large volumes of data. These techniques play a critical role in organizing information by identifying similarities and patterns, which allows us to extract valuable insights and trends that might not be readily apparent. Additionally, NLP-based retrieval systems offer customizable search criteria, enabling users to refine their searches based on specific parameters like keywords, dates, authors, or topics. This functionality saves significant time and effort that would otherwise be spent sifting through extensive datasets manually. Compared to traditional manual methods, NLP algorithms offer faster processing and improved accuracy. They can analyze massive amounts of data more efficiently, which is particularly advantageous for journalists and researchers striving to keep up with the competition. Furthermore, NLP algorithms excel at removing irrelevant data from the analysis, resulting in enhanced accuracy when clustering and retrieving news articles. The objective of this research paper is to emphasize the importance of NLP-based clustering and retrieval in transforming information organization and extracting valuable insights. By utilizing these techniques, researchers and journalists can gain a competitive edge in their respective fields. The paper aims to explore how NLP facilitates efficient data management, identifies trends, and provides comprehensive analysis, contributing to a deeper understanding of the subject matter. Ultimately, the utilization of NLP-based clustering and retrieval can significantly benefit the fields of journalism and research by streamlining processes and enabling better decision-making.

2. REVIEW ON LITERATURE

Dewan and Farzana provide a comprehensive review of clustering techniques for news articles. They discuss various methods such as k-means, hierarchical clustering, and topic modeling, highlighting their strengths and weaknesses. The paper also addresses challenges specific to news article clustering, including text preprocessing, feature extraction, and evaluation metrics. Overall, this review serves as a valuable resource for researchers interested in clustering news articles.[1]

Lee, Y. H., Hu, P. J. H., Zhu, H., & Chen, H. W. propose a novel clustering method for discovering event episodes from sequences of online news articles. Their approach leverages time-adjoining frequent item sets to capture temporal relationships between news articles and identifies event episodes based on the discovered patterns. The paper provides a detailed explanation of the method and presents experimental results demonstrating its effectiveness in extracting event episodes from news article sequences.[2]

Uma Priya and Santhi Thilagam propose an innovative approach that combines two techniques: incremental clustering and indexing. By employing incremental clustering, their method adapts dynamically to changes occurring in the dataset, ensuring that the data is organized into clusters effectively. Additionally, they develop an indexing mechanism that enhances the retrieval process by enabling rapid access to relevant clusters.[3]

Yang, P., Li, W., & Zhao, G. propose a language model-driven approach for topic clustering and summarization of news articles. Their method combines a language model with clustering algorithms to group articles based on similar topics. Additionally, it generates representative summaries for each cluster. The paper describes the methodology in detail and presents experimental results demonstrating the effectiveness of

their approach in organizing news articles into coherent clusters.[4]

Wang, J., Jatowt, A., Färber, M., & Yoshikawa, M. focus on improving question answering for event-focused questions using temporal collections of news articles. They propose a method that identifies relevant articles related to the query and extracts salient information to generate concise answers. The paper presents a detailed framework for question answering and discusses techniques for event extraction and answer generation. Experimental evaluations demonstrate the effectiveness of their approach in answering event-focused questions using news article collections.[5]

3. DATASET DESCRIPTION

The Reuters-21578 dataset includes news articles published by the Reuters news agency in 1987. It contains a diverse range of 21,578 news stories covering various categories such as business, politics, and sports. Over time, this dataset has gained significant popularity and has become a standard reference for studying text categorization and information retrieval. Researchers and professionals utilize it extensively for a wide range of purposes, including training and testing machine learning models, evaluating natural language processing algorithms, and initiating studies in information extraction and text mining.

Initially stored in SGML format, an older markup language no longer in widespread use, the dataset has been transformed into different formats like plain text and XML to make it more accessible and usable for researchers and practitioners.

Table -1: Important Features of Reuters Dataset

FIELD	DESCRIPTION
Name	Reuters-21578 Text Categorization Collection
Source	Reuters News Agency
Type	Text classification dataset
Size	21,578 documents
Categories	90 categories
Language	English
Format	XML
Features	Document ID, Title, Body, and Category
Task	Multi-class classification
Split	Predefined train-test split
Metadata	Contains metadata for each document, including author, date, and topic code
Preprocessing	Includes tokenization, stop-word removal, and stemming
Availability	Available for download from various sources, including the NLTK corpus
Usage	Commonly used as a benchmark dataset for text classification and information retrieval research, and has been used in various academic and commercial applications

4. PROPOSED METHODOLOGY

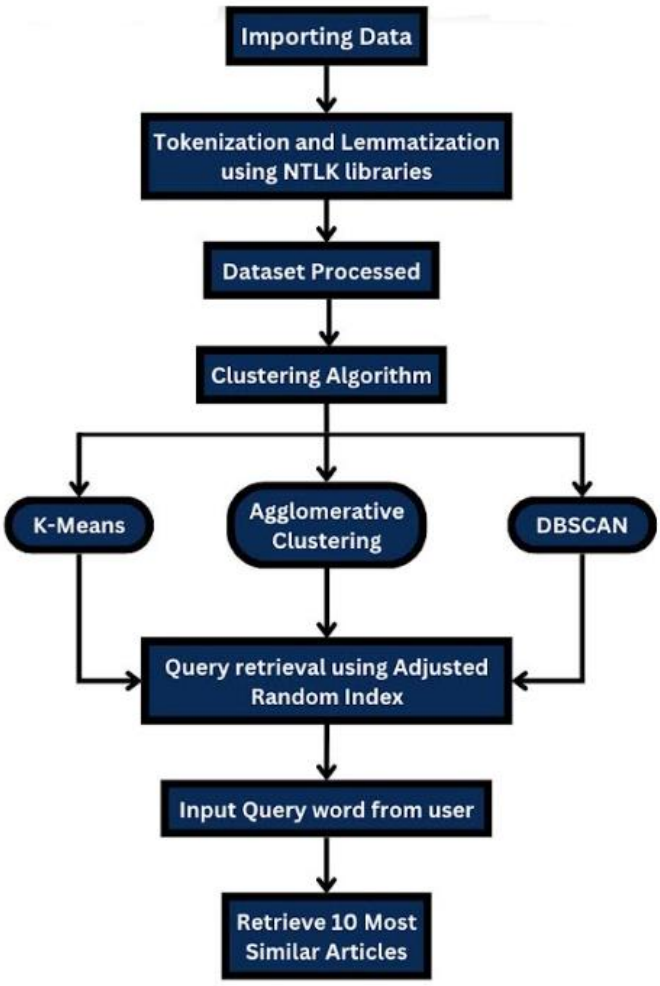

Fig-1: Flow chart of proposed work

STEP 1: Dataset Importing and Preprocessing

The project begins by introducing the Reuters dataset from the nltk.corpus package. This dataset contains many news stories categorized under different labels. A Pandas DataFrame is created with two columns, namely 'text' and 'categories', to store the text data and relevant

categories of each article. Prior to analysis, the text data is preprocessed. Essential nltk libraries for text preprocessing, such as punkt, wordnet, and omw-1.4, are downloaded and installed. The text input is tokenized using the nltk.tokenize.word_tokenize() method and lemmatized using the nltk.stem.WordNetLemmatizer() method. Lemmatization is the process of reducing words to their base or root form, which aids in normalizing the text data.

After preprocessing the text data, it is vectorized using the TfidfVectorizer from the sklearn.feature_extraction.text library. Vectorization changes the text input into numerical representation, allowing the implementation of machine learning strategies. The TfidfVectorizer methodology is employed, which computes the importance of each word in a document based on its frequency and inverse document frequency. It then creates a vector visual for each text.

STEP 2: Use of Clustering Algorithms

Following the composition of the Reuters dataset, clustering methodologies are applied to group similar articles from the preprocessed data. Three notable clustering algorithms, namely KMeans, Agglomerative Clustering, and DBSCAN, are implemented in this step.

STEP 3: Evaluation using Adjusted Rand Index

The working of the clustering methods is evaluated in Step 3 of the project. The Adjusted Rand Index (ARI) is utilized as an evaluation metric. The ARI measures the resemblance between awaited and true clustering labels while accounting for chance. A score of 1 signifies perfect agreement, while 0 indicates random clustering.

STEP 4: Retrieval of Results

In this step, articles that closely resemble a given query are retrieved. This is accomplished by converting the query into a vector representation using the same TfidfVectorizer employed in the previous step. The cosine resemblance between the query vector and the vectorized data is computed for each article. By sorting the cosine similarity scores in descending order, the top ten articles most like the query are selected. These articles are considered the most relevant to the query based on their resemblance to the query vector.

5. RESULTS

According to the evaluation conducted, the Agglomerative Clustering algorithm demonstrated the highest Adjusted Rand Index (ARI), indicating superior performance compared to KMeans and DBSCAN. Nevertheless, all three algorithms exhibited relatively average ARI scores, implying that further research is required in this domain. However, it is important to note that this does not affect the retrieval performance, as a single input query can be associated with multiple categories of articles.

Chart-1: Comparison of proficiency of algorithms

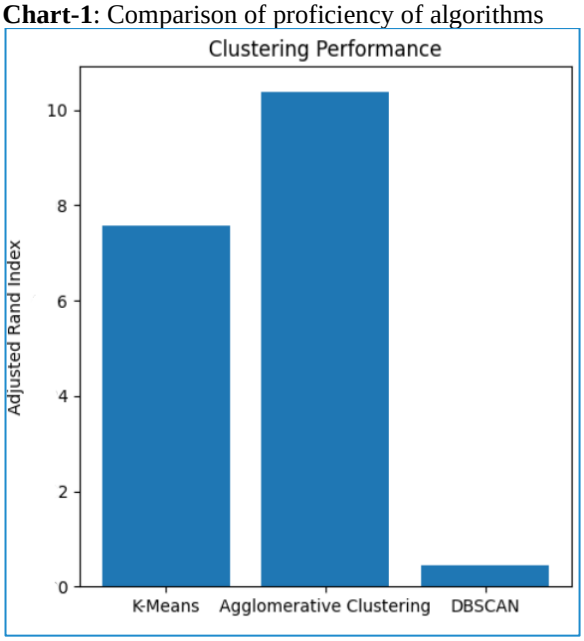

Chart-2: Clusters using K Means method

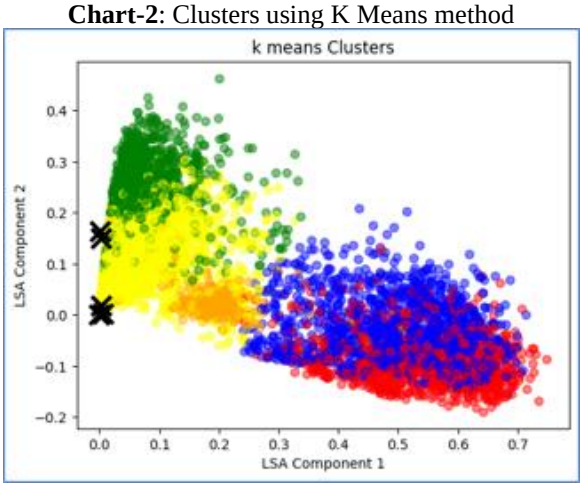

Chart-3: Clusters using Agglomerative method

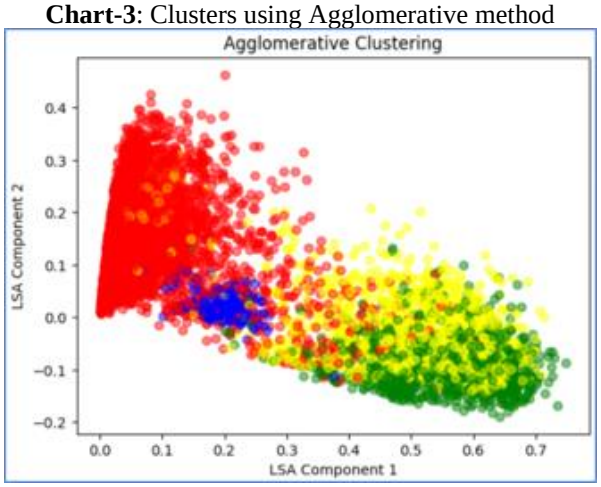

Chart-4: Clusters using DBSCAN method

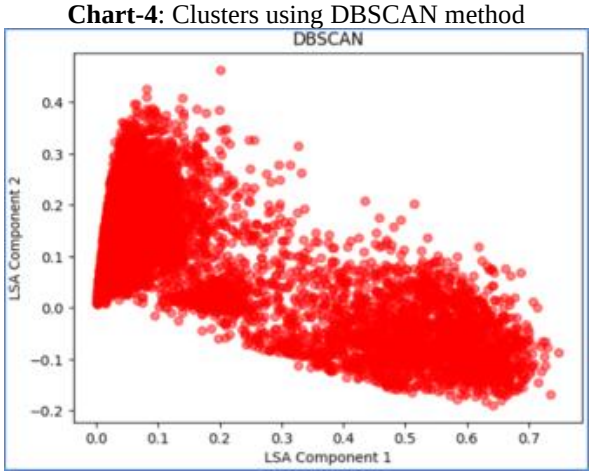

Chart-5: Overall Clusters

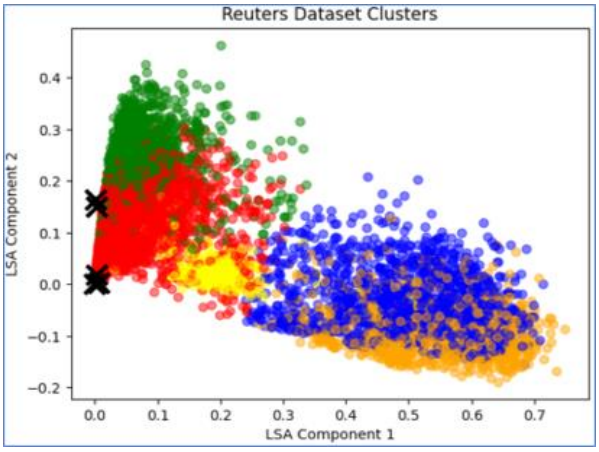

Chart 2, 3, 4 represent the clusters formed by three different clustering algorithms. Chart 5 represents the clusters in Reuters dataset combined, LSA stands for Latent Semantic Analysis. LSA 1 and LSA 2 refer to the first and second principal components obtained through dimensionality reduction.

The query retrieval stage is an important aspect of the project since it allows users to search for publications depending on their interests and requirements. It can be used to create a search engine or a recommendation system that provides users with appropriate content depending on their searches.

```
query = "oil prices"
query_vec = vectorizer.transform([query])

cosine_similarities = cosine_similarity(query_vec, X).flatten()
most_similar_indices = cosine_similarities.argsort()[::-1][:10]
for i, idx in enumerate(most_similar_indices):
    print('Rank {}:'.format(i+1))
    print('Category:', df.iloc[idx]['categories'])
    print('Text:', df.iloc[idx]['text'])
    print('-' * 50)
```

Fig-2: Input query from user, here oil prices

```
Rank 1:
Category: heat
Text: u.s. court upholds apex decision favoring nymex the u.s. court of appeal for the second
--------------------------------------------------
Rank 2:
Category: crude, nat-gas
Text: energy/foreign investor lured by the weakening dollar and the conviction that oil price are
--------------------------------------------------
Rank 3:
Category: crude
Text: u.s. warns of dependence on foreign oil a white house-ordered report said that growing u.s.
--------------------------------------------------
Rank 4:
Category: crude
Text: u.s. oil dependency seen rising to record level the united state ' dependency on foreign oil
```

Fig-3: Output in form of articles having oil prices repeated most times

In Figure 2, the input from the user is taken, here the ‘oil prices. When this code is run, the model returns top 10 most relevant articles from the dataset based on the query provided as shown in Figure 3.

6. FUTURE WORK

Exploring other clustering algorithms: While KMeans, Agglomerative Clustering, and DBSCAN were examined, there are numerous other clustering methodologies that could be explored to enhance the performance of clustering and retrieval.

Using different text preprocessing techniques: In addition to tokenization and lemmatization, other techniques like stemming or stop word removal could be employed to improve the performance of the features utilized in clustering and retrieval.

Incorporating additional features: Currently, our clustering and retrieval approaches rely solely on the text data from the articles. Incorporating additional features such as author, publication date, or sentiment analysis could enhance the accuracy and relevance of the results.

Using a broader and more diverse dataset: The Reuters dataset utilized in this project is primarily focused on financial news stories. Expanding the dataset to include articles from various disciplines or sources may enhance the generalizability of the findings.

Utilizing transfer learning techniques: Transfer learning techniques can be employed to pretrain a language model on a larger dataset, which can then be fine-tuned on the Reuters dataset. This approach has the potential to improve the working of the features incorporated in clustering and retrieval.

7. CONCLUSIONS

The study successfully used clustering methods to the Reuters dataset to group like articles in group based on their categories, enabling for more efficient information retrieval.

The TfidfVectorizer was used to effectively preprocess and vectorize text data, enabling clustering methods and retrieval methodologies to be applied.

The project built different clustering algorithms and evaluated them using the Adjusted Rand Index and Silhouette score, allowing the best approach for the given dataset to be chosen.

The cosine similarity-based query retrieval system was able to efficiently retrieve the most comparable articles to a user's input query, providing a handy tool for quick and great information retrieval.

This study provides a foundation for future work in the field of textual grouping and retrieval, namely in the areas of adding domain-specific knowledge and enhancing feature sparsity and high dimensionality.

REFERENCES

1. Nidhi Dewan, & Shagufta Farzana. (2022, March 30). Clustering News Articles: A Review. International Journal of Advanced Research in Science, Communication and Technology, 767–771. https://doi.org/10.48175/ijarsct-5499
2. Lee, Y. H., Hu, P. J. H., Zhu, H., & Chen, H. W. (2020, November). Discovering event episodes from sequences of online news articles: A time-adjoining frequent itemset-based clustering method. Information & Management, 57(7), 103348. https://doi.org/10.1016/j.im.2020.103348
3. Uma Priya D, & Santhi Thilagam P. (2020, July 1). Dynamic Data Retrieval Using Incremental Clustering and Indexing. International Journal of Information Retrieval Research, 10(3), 74–91. https://doi.org/10.4018/ijirr.2020070105
4. Yang, P., Li, W., & Zhao, G. (2019). Language Model-Driven Topic Clustering and Summarization for News Articles. IEEE Access, 7, 185506–185519. https://doi.org/10.1109/access.2019.2960538
5. Wang, J., Jatowt, A., Färber, M., & Yoshikawa, M. (2021, January 2). Improving question answering for event-focused questions in temporal collections of news articles. Information Retrieval Journal, 24(1), 29–54. https://doi.org/10.1007/s10791-020-09387-9

BIOGRAPHIES

Dubey Dhanraj, a passionate and committed student at the esteemed Symbiosis Institute of Technology in Pune, is currently pursuing a Bachelor of Technology in Mechanical Engineering. In addition to his core studies, Dubey has also chosen to specialize in Artificial Intelligence and Machine Learning (AIML), reflecting his keen interest in this rapidly advancing field. To complement his theoretical knowledge, Dubey has actively sought practical experience by working as a developer trainee at prestigious companies like Salesforce and Persistent. Furthermore, he has had the privilege of internships at GE Aerospace and TATA, allowing him to gain valuable insights and hands-on exposure to real-world engineering challenges. Dubey's curiosity and dedication extend beyond the confines of his educational endeavors, as he actively participates in national and international conferences, eagerly exploring diverse facets of engineering. Notably, his remarkable research papers have been recognized and chosen for presentation at some of these conferences, and he has proudly published his valuable contributions to the field.