BENGALI NEWS CLUSTERING USING K-MEANS CLUSTERING BASED ON LSA

BY

Rashedul Alam Zilani
ID: 192-15-13131

This Report Presented in Partial Fulfillment of the Requirements for the Degree of Bachelor of Science in Computer Science and Engineering

Supervised By

Md. Aynul Hasan Nahid
Lecture
Department of CSE
Daffodil International University

Co-Supervised By

Md. Ferdouse Ahmed Foysal
Lecturer
Department of CSE
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY
DHAKA, BANGLADESH

JANUARY 2024

APPROVAL

This Project/internship titled **"Bengali News Clustering Using K-Means Clustering based on LSA"**, submitted by Rashedul Alam Zilani, ID No: 192-15-13131 to the Department of Computer Science and Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation was held on *21 January 2024.*

BOARD OF EXAMINERS

Chairman

Narayan Ranjan Chakraborty (NRC)
Associate Professor & Associate Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner

Sadek

Md. Sadekur Rahman (SR)
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner

Shayla Sharmin

Shayla Sharmin (SS)
Senior Lecturer
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

External Examiner

Dr. Md. Zulfiker Mahmud (ZM)
Associate Professor
Department of Computer Science and Engineering
Jagannath University

DECLARATION

I hereby declare that, this project has been done by me under the supervision of **Md. Aynul Hasan Nahid, Lecturer, Department of CSE** Daffodil International University. I also declare that neither this project nor any part of this project has been submitted elsewhere for award of any degree or diploma.

Supervised by:

Md. Aynul Hasan Nahid
Lecturer
Department of CSE
Daffodil International University

Co-Supervised by:

Md. Ferdouse Ahmed Foysal
Lecturer
Department of CSE
Daffodil International University

Submitted by:

Zilani

Rashedul Alam Zilani
ID: -192-15-13131
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

First I express my heartiest thanks and gratefulness to almighty God for His divine blessing makes me possible to complete the final year project/internship successfully.

I am really grateful and wish profoundly my indebtedness to **Md. Aynul Hasan Nahid, Lecturer**, Department of CSE Daffodil International University, Dhaka. Deep Knowledge & keen interest of our supervisor in the field of "Neural Networks" to carry out this project. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior drafts and correcting them at all stages have made it possible to complete this project.

I would like to express our heartiest gratitude to **Dr. Sheak Rashed Haider Noori, Professor**, and Head, Department of CSE, for his kind help to finish our project and also to other faculty members and the staff of the CSE department of Daffodil International University.

I would like to thank our entire course mates at Daffodil International University, who took part in this discussion while completing the coursework.

Finally, I must acknowledge with due respect the constant support and patience of our parents.

ABSTRACT

Effective information retrieval and organization have become increasingly important, especially in contexts involving diverse cultural backgrounds, as the continued growth of digital content has demonstrated. The subject matter of this paper is the clustering of Bengali news using the K-means algorithm, which integrates LSA. Because of being uncommon, clustering news based on latent semantic analysis poses a tricky problem. Document clustering is also known as textual document clustering. It is one form of cluster analysis. Recent research in this technological age has focused on the implementation of text clustering techniques in diverse domains, including text extraction for extracting vast quantities of valuable content from the Internet and automated document organization [15] and [16]. This article introduces a more advantageous K-means clustering news clustering framework for the purpose of clustering text or news documents. A self-taught learning model is employed to cluster a given set of data into distinct groups, obviating the need for external labels or identifiers. We analyzed a dataset consisting of approximately 0.5 (504266) million portal news texts retrieved from several Bengali newspapers, as well as seven distinct kinds of news content. To categorize the dataset using clustering and semantic analysis, we first set the dataset up. Following that, the punctuation and keywords are converted into codes so that deep learning techniques may be applied to them for the training process. Once we have the learned groups, we cluster them using K-means. However, there are certain things to work on, like data processing and the separation of sentences and punctuation. We recommend a strategy neural network-based deep learning that can solve such issues. Since no groundbreaking work has been done on news text or document clustering yet, this is an effective method. Additionally, we have conducted a few experiments to show how the approach is specifically implemented, confirming the proposed method's efficacy.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Introduction

News clustering is a fundamental method in pattern recognition, image analysis, and data mining that divides news documents into distinct categories. News clustering integrates various perspectives of deep learning into the categorization process. The earlier studies discovered the application of text clustering techniques to languages other than Arabic, English, and Chinese [11], [12], and [10]. In newspapers or texts, innovative work has never been done before. Currently, a few particular languages have been utilized in the text clustering procedure when it comes to the technique. We conducted an analysis on a dataset comprising 5,04,266 (around 0.5 million) news articles obtained from various Bengali newspapers. We implemented latent semantic analysis and modified K-means methods to cluster the dataset. The traditional clustering method, K-Means News Clustering, which relies on Latent Semantic Analysis (LSA), partition data points into clusters connected with their equality. The model uses latent semantic analysis (LSA) and modified K-means algorithms to group datasets together. This is a common method that works well for exploring data because it organizes patterns. The text analysis method, LSA utilizes word relationships to convert a text into a numerical representation, enhancing machine learning tasks. In the training phase, words and punctuation are converted into codes for latent semantic analysis (LSA), preparing the dataset for both LSA and K-Means algorithms. After applying K-means clustering to the data, the classification of each group is based on a similar cluster set. The system underscores ongoing adjustments to maintain an intellectual foundation for news documents, facilitating automatic text organization. K-means clustering shows itself to be a successful technique for news clustering, employing latent semantic analysis for the unsupervised categorization of related news articles. This approach allows for the grouping of news articles based on their semantic similarity, enabling efficient organization and retrieval of relevant information. By leveraging the power of LSA and K-means clustering, the system a robust solution for automatic text categorization in the field of news analysis. The use of automated clustering assists editors and journalists in quickly arranging and grouping news articles based on their patterns. By utilizing K-means clustering, the system can identify patterns and relationships within the dataset, allowing for efficient organization and retrieval of relevant information. This automated approach saves time and effort for editors and journalists, enabling them to focus on analyzing and presenting the news rather than manually sorting through numerous articles. Analyzing stories, therefore improving the process of subject identification. The method effectively retrieves relevant information from huge document collections and has been utilized for tasks such as automatic subject identification and summarization. The study presents a novel clustering technique for news articles, specifically focused on addressing the issue of scarcity. The framework emphasizes the application of deep learning techniques based on latent semantic analysis to achieve continuous document grouping for news documents across different languages. The paper looks at what text clustering means, how it's done, and what studies have been done on it. It focuses on how deep neural networks (DNN) and convolutional filters are used to make text representations. This research aims to enhance the grouping of news texts using advanced deep learning techniques based on LSA, in contrast to the most previous research that has predominantly focused on English and Chinese.

The text highlights the lack of novel studies in the clustering of newspapers, specifically in Bengali, and highlights the uniqueness of the proposed approach, which aims to overcome the limitations of previous studies.

1.2 Motivation

This study seeks to provide valuable perspectives in the realm of Bengali news clustering, tackling both the immediate hurdles of information organization and the broader discussion on refining clustering approaches for languages with distinct characteristics. This research aims to contribute meaningful insights into the domain of Bengali news clustering, addressing both the
immediate challenges of information organization and the broader discourse on optimizing clustering methodologies for languages with unique characteristics.

1) Design and implement a clustering framework for Bengali news content using the K-means algorithm [14].

2) Establishing an efficient news text dataset and the best-performing news document clustering system are our objectives.

3) The proposed clustering method's effectiveness and efficiency will be assessed by evaluating the quality of the clusters and their relevance to the inherent semantic structure of Bengali news.

4) For news document clustering, provide a news document dataset and determine the most suitable model for the new dataset, granting permission for future research utilization.

1.3 Objective

In today's technological age, certain inventions are made every day. As humans, we must stay on top of these new developments and move toward the current global technological era. As students, our capacity to contribute to this progress is highlighted, albeit with limited opportunities for impactful engagement throughout courses. The increased availability of data is what is driving the surge in news content analysis across various industries. Significantly, techniques such as k-means clustering are increasingly favored for categorizing news articles according to latent semantic features. This method plays a crucial role in identifying interconnected stories, recognizing trends, and assessing the audience's response to news. Nevertheless, the successful application of clustering algorithms requires thorough data preprocessing to ensure the precise and efficient clustering of the available stories. News text clustering emerges as a potent force with significant implications for serving individuals with disabilities, bolstering security measures, and enhancing surveillance systems. While the initial exploration of this field garnered interest and positivity, it is noteworthy that most advancement has been documented in English, Chinese, and other languages. It becomes essential to create essential to create news text clustering models and datasets that are language-neutral to fully realize the benefits of text clustering. Previous research on text clustering techniques and algorithms for news publications highlighted the need for a standard model to efficiently arrange and categorize articles based on content and subjects. This study aims to eliminate discrimination and develop a comprehensive approach to news clustering across languages, a crucial and challenging aspect of contemporary studies.

1.4 Rationale Investigation

The study implements deep learning and machine learning strategies to cluster 5,00,000 Bengali news stories using LSA and the K-means algorithm, filling a research gap. A popular approach for analyzing data and machine learning, K-means clustering splits a dataset into k clusters and gives points to each one according to how similar it is to the cluster center [17]. On the other hand, LSA is an algorithm that extracts and evaluates data from the text by discerning relationships between words and concepts. What sets our research apart is its focus on news text documents and the utilization of deep learning techniques for clustering. This approach deviates from prior studies, delivering more effective outcomes in text clustering for news documents. It represents a significant advancement in the realm of grouping text, particularly for news text documents.

1.5 Research Questionnaire

To find fresh insights, resolve issues, and find answers to concerns, research is an essential tool. It is the methodical study of a subject to learn more or make judgments. The goal of research is to shed light on the world around us and find answers to questions. Many research questions remain unresolved, such as the clustering approaches that work best for news clustering. Additionally, it enables us to investigate the impact of cluster size on the accuracy of clustering outcomes and determine if increasing cluster size can enhance performance. We had a difficult time conducting research because of the specific grouping of news. In order to ensure a thorough understanding of the desired knowledge and explanation, the research question is an essential first step in the research process.

To share their ideas and accomplishments, the researchers would like to set out the following queries:

1) How can the LSA-based approach be effectively used to cluster Bengali news articles?
2) Can we cluster the four categories of the news text dataset properly?
3) Is it possible to reduce cluster time in the dataset?
4) Is it possible to enhance text clustering through this tactic?
5) How have individuals took advantage from such a strategy?

1.6 Predicted Outcomes

To better the productivity and performance of information document classification, our study makes use of the "Bengali News Clustering using K-means Clustering based on LSA" model. Our work is anticipated to yield the following outcomes:

1) Automatic clustering of news texts into distinct groups.
2) Accurate classification of unseen news articles into relevant categories through semantic analysis.
3) Clustering of both specific and unspecific news data using the proposed K-means clustering approach.
4) Grouping for news documents in the relevant categories is automatically generated.
5) Benefiting news readers by providing a more organized and structured approach to accessing news articles.
6) The objective of our research is to address real-world issues, and we make every effort to ensure that our findings are trustworthy and reproducible and advance news clustering techniques as a whole. Our research works to address real-world difficulties utilizing the machine learning algorithms like K-means clustering. We aim to provide a trustworthy tool for wide and skillful analysis of news data, focusing on value distribution, and ensure reproducible findings.

1.7 Research Layout

Chapter 1: Provides summary of a research, outlining its introduction, goals, motivation, research inquiries, and framework.

Chapter 2: Explains the project's fundamental framework, discussing research difficulties, related works, and comparative analysis.

Chapter 3: Show the research's theoretical discussions. It first outlines the entire research project's workflow before addressing the process of obtaining and preparing the data. Next, it displays the algorithms needed for the suggested model and their correct implementation.

Chapter 4: Demonstrates the outcome of execution and analyzes the suggested model's effectiveness. The project's proper explanation and clarification are provided in this section.

Chapter 5: The research project concludes with a summary of its achievements and findings, recommending additional investigation in the relevant field.

CHAPTER 2

BACKGROUND

2.1 Preliminaries and Terminologies

During the era in technological abundance, the effective organization and retrieval of news content have become paramount. As linguistic diversity continues to shape the global digital landscape, the demand for tailored approaches to information categorization in specific languages intensifies. This research embarks on a journey into the realm of Bengali news clustering, leveraging the power of K-means clustering augmented by Latent Semantic Analysis (LSA). The objective is to address the unique linguistic nuances inherent in Bengali news content and provide a robust framework for efficient categorization. K-means clustering, a widely utilized algorithm in data clustering is chosen as the primary tool for organizing Bengali news articles. It’s adaptability and efficiency make it an appealing choice for this task. To further enhance the clustering process, we integrate Latent Semantic Analysis, a technique that captures the underlying semantic relationships within a corpus, providing a more nuanced understanding of the content [13].

This paper begins by providing a MiniBatch K-means approach, which is a modified version of K-means clustering that offers faster computation time and Latent Semantic Analysis, establishing the basis for their integration into the Bengali news clustering framework. A critical review of existing literature illuminates the landscape of news clustering methodologies, underscoring the scarcity of studies tailored to the linguistic and contextual intricacies of Bengali content.

2.2 Related Articles

Over the years, numerous studies have explored news text clustering employing deep learning techniques. Unsupervised text clustering research has long relied on the K-means clustering method. These studies have utilized K-means clustering and latent semantic analysis to group text into data points, and identify articles for clustering based on document similarity scores. D.D. Lee and H.S. Seung's "Document Clustering based on Non-negative Matrix Factorization" explores the use of NMF as an alternative to traditional methods, demonstrating its ability to produce meaningful and interpretable clusters [8]. "Statistical classification methods for Arabic news articles" evaluates the efficacy of different statistical classification methods in Arabic news articles, revealing the challenges and opportunities of automated classification in this linguistic field [2]. "Improving News Article Recommendations via User Clustering" focuses on improving news article recommendations through user clustering, aiming to provide more accurate and tailored suggestions based on individual user preferences and behaviors [5].

"A text clustering approach to Chinese news based on a neural network language model" presents a novel text clustering method for Chinese news articles, utilizing a neural network language model to enhance efficiency and accuracy in grouping content based on semantic relationships [7]. The study "A Comparative Study of Clustering English News Articles Using Clustering Algorithms" evaluates and compares the effectiveness of various clustering algorithms in organizing and categorizing English news articles, providing valuable insights for researchers and practitioners in natural language processing and information retrieval [3]. "Arabic Words Clustering by Using K-means Algorithm" explores the use of the K-means algorithm for Arabic word clustering, aiming to improve natural language processing and enhance applications like information retrieval and language understanding by organizing words based on semantic similarities [6]. "A Comparison of Document Clustering Techniques" by Steinbach, Karypis, and Kumar evaluates and compares document clustering techniques, assessing their strengths and weaknesses, and providing insights into their performance, applicability, and suitability for specific data types, offering valuable guidance for researchers and practitioners [9]. "A Survey of Text Clustering Algorithms" provides a thorough analysis of various text clustering techniques, comparing their methodologies, strengths, and applications in grouping textual data [1]. "A novel document clustering model based on latent semantic analysis" propose a novel document clustering model based on latent semantic analysis, aiming to improve the efficiency and effectiveness of document clustering by utilizing latent semantic analysis to group related documents [4]. "LSA & LDA Topic Modeling Classification: Comparison Study on E-books" compares Latent Semantic Analysis (LSA) and Latent Dirichlet Allocation (LDA) topic modeling techniques in e-books, assessing their performance and effectiveness in classifying and organizing topics, providing insights into their strengths and limitations in this specific domain [14]. "Bengali Text Document Categorization Based on Very Deep Convolution Neural Network" explores the use of a very deep convolutional neural network (CNN) for Bengali text document categorization, aiming to improve accuracy and efficiency in this field by leveraging the deep learning model's capabilities [18]. "Bengali News Headline Categorization Using Optimized Machine Learning Pipeline" explores the use of an optimized machine learning pipeline for categorizing Bengali news headlines, aiming to improve efficiency and accuracy, contributing to the field of natural language processing and machine learning in Bengali news content [21]. To the best of our knowledge, this research is the first to look at the use of the K-means algorithm for latent semantic analysis-based news grouping.

The execution of an effective machine learning pipeline for the categorizing Bengali news headlines, seeking to improve efficiency and accuracy, contributing to the field of machine learning and NLP in Bengali news content. For the best of what we know, the present study is the first to appear at how it's done of the method known as K for LSA based news categorization.

2.3 Comparative Studies and review

Research on news text grouping has primarily utilized deep learning and neural network techniques, with deep learning showing promise in news article clustering and neural networks being key components. K-means clustering, the widely used data clustering approach, finds applications in text classification, facial recognition, and image segmentation. In news clustering, K-means is employed to categorize related items into appropriate groups. Despite its simplicity, K-means clustering outcomes may not always be reliable. In our comparative analysis of prior studies, we note both similarities and differences, with deep learning approaches being commonly employed. However, our proposed method introduces significant advancements in text clustering. The goal is to improve the news document clustering performance and accuracy by utilizing a new dataset that includes training and test records.
This study introduces a novel news clustering method that combines latent semantic analysis (LSA) with K-means. Our method has consistently yielded reliable results across diverse datasets, showcasing its effectiveness. We aim to improve clustering performance by introducing a modified K-means algorithm incorporating a local search technique. We also explore the possibility of using LSA to more efficiently extract semantic material from articles.

2.4 Challenges

Despite its effectiveness in finding latent concepts within extensive document collections, K-means clustering is not without its drawbacks. Document topics may not be accurately represented if latent semantic analysis which is based on word frequencies and similarities—is relied upon. Euclidean distance measures in K-means clustering could distort the true data structure. Addressing this, a combination of methods becomes crucial for accuracy. The research tackles the initial challenge of dataset processing and the implementation of a modified method for news document clustering. Utilizing the portal news classification dataset, thorough cleaning and normalization enhance data readiness. Deep Learning algorithms are applied, overcoming difficulties in extracting underlying topics using K-means clustering. This proves valuable for journalists and researchers, revealing relevant topics and aiding in article recommendations. Despite dataset complexities, including training and testing sets, the proposed model outperforms previous approaches showing improved results and accuracy.

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Research Subject and Instrumentation

The first step involves collecting and preprocessing the dataset by performing data cleaning and normalization procedures. This ensures that the format of the data is appropriate for further analysis. Word embedding techniques are then applied to represent the textual data in a numerical format that captures the semantic meaning of words. This enhances the model's understanding of the context of the news articles. The embedded representations of the news articles are grouped according to similarity using a k-means clustering algorithm. This clustering technique aims to uncover latent patterns and themes within the dataset, enabling the categorization of similar news articles into clusters.

The algorithm uses LSA to determine the semantic meaning of news documents, enabling more meaningful categorization. The relevance and effectiveness of the model have been enhanced with MiniBatch, modified K-means clustering, and metrics to evaluate precision. Heat maps of the confusion matrix are used to visualize classification results and identify misclassifications or patterns.

3.2 Workflow

The research methodology involves collecting and preprocessing a dataset of 0.5 million portal news texts, applying word embedding techniques, utilizing k-means clustering, incorporating an LSTM model, and evaluating the performance using accuracy metrics and visualizations.

Step 1 - Data Collection: Collect 0.5 million portal news texts from reliable sources.

Step 2- Data Preprocessing: Perform preprocessing and data cleaning procedures, such as text normalization, missing value handling, and noise removal.

Step 3- Feature Extraction: The data is clustered based on its similarities, potentially uncovering significant characteristics that would otherwise be unobserved.

Step 4- Model Accomplishment: The documents' dimensionality is decreased while maintaining their informational content through the application of LSA (Latent Semantic Analysis). It provides this by first creating a matrix out of the documents, which is then reduced in dimension using a technique called singular value decomposition.

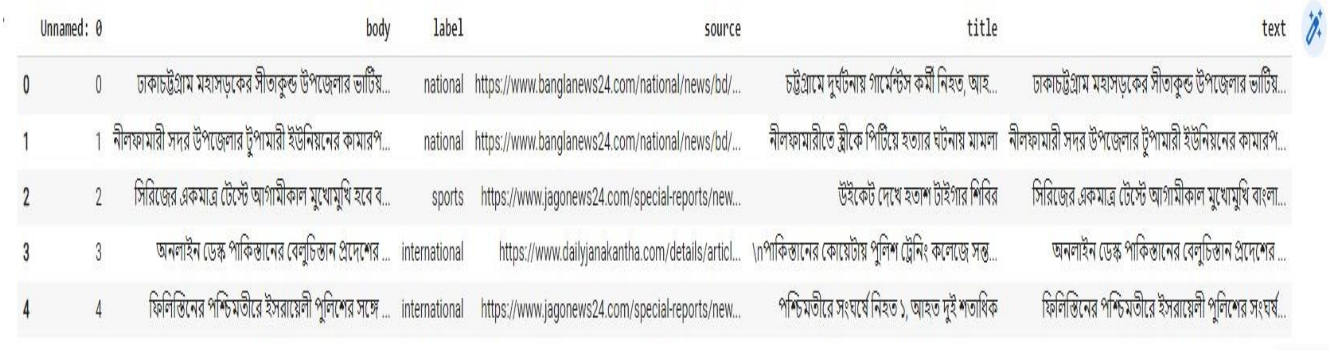
Step 5- Apply K-means Clustering: Using the best K value from the previous step, the technique known as K-means can be utilized for dividing texts based on to their LSA properties.

Step 6- Performance Evaluation: This section contains an evaluation of the clustering result displayed graphically. We show that k-means clustering can be improved by utilizing various data sets. In addition, we demonstrate that applying latent semantic analysis in a supervised. In addition, the study demonstrates that utilizing latent semantic analysis in a supervised learning environment leads to improved outcomes.

Step 7- Next investigations and summary: Researchers found that the clustering method with k-means performed more accurately than latent semantic analysis. A subsequent study aims to investigate its effectiveness in latent semantic analysis-based news clustering, potentially improving the algorithms' accuracy.

3.3 Experiment Dataset

We used a dataset made up of seven different sorts of news categories and roughly 0.5(504266) million portal news texts pulled from different Bengali newspapers. The dataset was collected from various online news sources, covering a diverse range of topics and domains. The news texts received preprocessing to manage missing values, reduce noise, and group the textual data. The resulting dataset serves as the input for the word embedding techniques, neural network model training, and clustering algorithm.



	Unnamed: 0	body	label	source	title	text
0	0	ঢাকাচট্টগ্রাম মহাসড়কের সীতাকুন্ড উপজেলার ভাটিয়...	national	https://www.banglanews24.com/national/news/bd/...	চট্টগ্রামে দুর্ঘটনায় গার্মেন্টস কর্মী নিহত, আহ...	ঢাকাচট্টগ্রাম মহাসড়কের সীতাকুন্ড উপজেলার ভাটিয়...
1	1	নীলফামারী সদর উপজেলার টুপামারী ইউনিয়নের কামারপ...	national	https://www.banglanews24.com/national/news/bd/...	নীলফামারীতে স্ত্রীকে পিটিয়ে হত্যার ঘটনায় মামলা	নীলফামারী সদর উপজেলার টুপামারী ইউনিয়নের কামারপ...
2	2	সিরিজের একমাত্র টেস্টে আগামীকাল মুখোমুখি হবে ব...	sports	https://www.jagonews24.com/special-reports/new...	উইকেট দেখে হতাশ টাইগার শিবির	সিরিজের একমাত্র টেস্টে আগামীকাল মুখোমুখি বাংলা...
3	3	অনলাইন ডেস্ক পাকিস্তানের বেলুচিস্তান প্রদেশের ...	international	https://www.dailyjanakantha.com/details/articl...	\nপাকিস্তানের কোয়েটায় পুলিশ ট্রেনিং কলেজে সন্ত...	অনলাইন ডেস্ক পাকিস্তানের বেলুচিস্তান প্রদেশের ...
4	4	ফিলিস্তিনের পশ্চিমতীরে ইসরায়েলী পুলিশের সঙ্গে ...	international	https://www.jagonews24.com/special-reports/new...	পশ্চিমতীরে সংঘর্ষে নিহত ১, আহত দুই শতাধিক	ফিলিস্তিনের পশ্চিমতীরে ইসরায়েলী পুলিশের সংঘর্ষ...

Figure 3.3.1: The Classified Set of data

3.4 Proposed Methodology

The study used standard k-means for smaller datasets and minibatch k-means for bigger datasets due to their flexibility and efficiency. TfidfVectorizer and HashingVectorizer were employed for text feature extraction, and latent semantic analysis reduced the dataset's dimensionality. Clustering results were created using several methods.

3.4.1 Evaluation the Effectiveness

This section presents a collection of metrics designed for evaluating various clustering strategies:

Homogeneity: Calculates the proportion of samples of a single class that are within the cluster.
Completeness: This shows the degree to which every sample in a class is a member of the same cluster.
V-measure: The item represents the harmonious equivalent of completeness as well as homogeneity and offers a comprehensive assessment of the quality of clustering.
Rand-Index: Assesses how frequently both data points are allocated to the specific group and class.
Adjusted Rand-Index: A modified version of the Rand-Index that considers chance agreement, with an expected value of 0.0 for randomly assigned clusters.

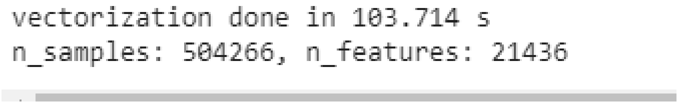
3.4.2 Clustering by K-means on Text Features

The benefits and drawbacks of using K-means clustering to analyze large-scale text data are covered in length in this work. The use of unsupervised machine learning is considered, particularly K-means clustering, for the identification and categorization of text feature clusters within large datasets [19]. The study evaluates the effectiveness of K-means clustering, an unsupervised machine-learning technique, in text data analysis by examining its advantages, shortcomings, potential applications, and limitations. Semantic analysis and text classification are two of its most widely used applications. The program constructs centroids, divides data points into 'k' clusters based on similarity, and then averages the data points inside each cluster. It is often used for semantic analysis and text categorization, creating centroids and separating data points into clusters.

In summary, this paper thoroughly explores the positive aspects and challenges associated with employing K-means clustering for text data analysis. It discusses the methodology, advantages, disadvantages, and feature extraction techniques utilized in the process. The final goal is to evaluate K-means clustering's suitability for textual data analysis and to draw relevant conclusions from the study's findings.

3.4.3 Utilizing TfidVectorizer for Feature Extraction

The work presents a novel feature extraction method that is utilizing the K-means algorithm, categorizing data points according to how close they are to every cluster's centroid as well. With the help of this technique, features from a dataset can be effectively extracted and used for further processing or analysis. To evaluate the effectiveness of the approach, the paper utilizes TfidfVectorizer's dictionary vectorizer and IDF normalization as benchmarks for comparison.

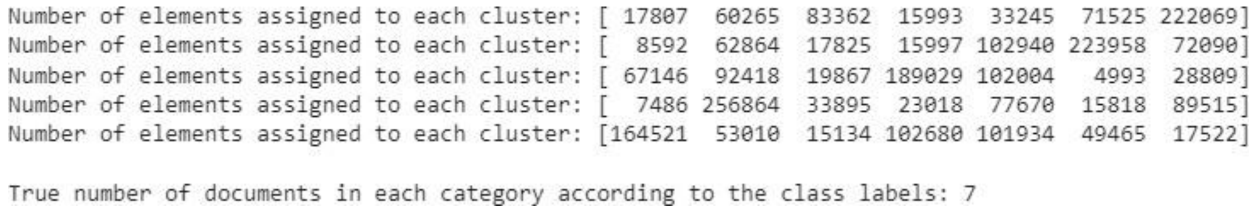


```
vectorization done in 103.714 s
n_samples: 504266, n_features: 21436
```

Figure 3.4.3.1: TfidfVectorizer Utilization

3.5 K-menas Clustering of Sparse Data

A technique called K and MiniBatch K-Means algorithms for clustering might not produce the best outcomes for unsupervised desired functions because to their inherent sensitivityThe process becomes more visible as handling complex, limited data, such as data vectorized with bag-of-words approaches.In such scenarios, K-Means may initiate centroids at isolated data points, steering the clustering process toward instability. Consequently, this can result in imbalanced clusters, as demonstrated in the given output.

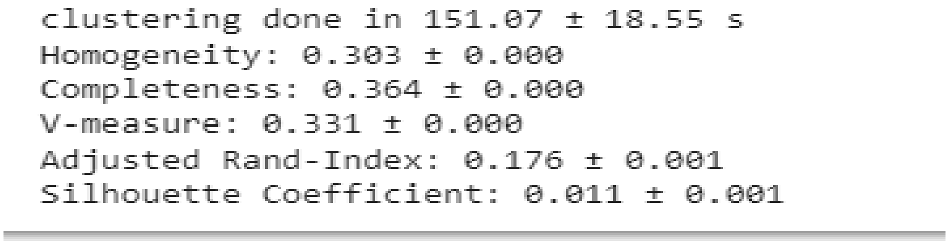


```
Number of elements assigned to each cluster: [ 17807  60265  83362  15993  33245  71525 222069]
Number of elements assigned to each cluster: [  8592  62864  17825  15997 102940 223958  72090]
Number of elements assigned to each cluster: [ 67146  92418  19867 189029 102004   4993  28809]
Number of elements assigned to each cluster: [  7486 256864  33895  23018  77670  15818  89515]
Number of elements assigned to each cluster: [164521  53010  15134 102680 101934  49465  17522]

True number of documents in each category according to the class labels: 7
```

Figure 3.5.1: Clustering with K-means of Weak Records

As a means of resolving the issue, enhance the overall number associated with runs through unique independent initialization (n_init) and resume the procedure known as K-Means multiple times, selecting the final clustering with maximum cohesion and rigidity. This technique aids in reducing the influence of suboptimal initialization, resulting in enhanced overall quality of the clustering results.

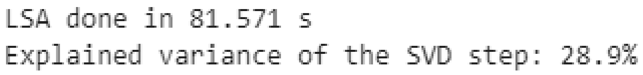


```
clustering done in 151.07 ± 18.55 s
Homogeneity: 0.303 ± 0.000
Completeness: 0.364 ± 0.000
V-measure: 0.331 ± 0.000
Adjusted Rand-Index: 0.176 ± 0.001
Silhouette Coefficient: 0.011 ± 0.001
```

Figure 3.5.2: Clustering by K-mean with Maximum Friction

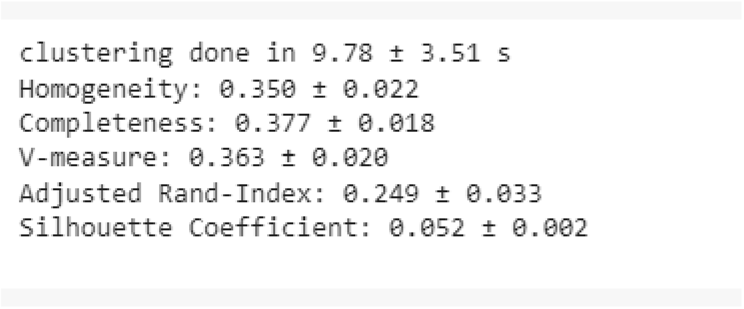
3.6 Performing dimensionality using LSA

We explore applying LSA in data mining and dimensionality reduction. LSA is a method that determines the underlying semantic meaning of a dataset by examining the relationships between words. By utilizing singular value decomposition (SVD), LSA reduces the dimensionality of the dataset, resulting in a lower-dimensional representation that can be used for various analysis tasks. Our proposed approach aims to leverage LSA to enhance the data mining process by reducing the dimensions and providing a more manageable representation of the data. However, we further reduce the dimensionality of the vectorized space using truncated SVD to standardize and the aim appears to be primarily to enhance the durability that results from the ensuing clustering procedure using K-means. Application of Singular Value Decomposition (SVD) to TF-IDF document vectors is a helpful technique for dimensionality reduction; this technique is also referred to as Latent Semantic Analysis (LSA) and has been extensively researched in the literature. The amount of data and patterns in the text corpus is efficiently decreased using this technique.



```
LSA done in 81.571 s
Explained variance of the SVD step: 28.9%
```

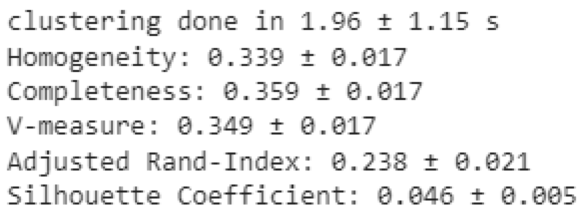
Figure 3.6.1: LSA is applied to perform dimension reduction



```
clustering done in 9.78 ± 3.51 s
Homogeneity: 0.350 ± 0.022
Completeness: 0.377 ± 0.018
V-measure: 0.363 ± 0.020
Adjusted Rand-Index: 0.249 ± 0.033
Silhouette Coefficient: 0.052 ± 0.002
```

Figure 3.6.2: Execution times(K-means and MiniBatch K-Means)

The formation of clusters applying the document's LSA representation is strongly faster and improved in all statistical tests, as established by the diminished LSA feature space and its implementation of modified k-means.

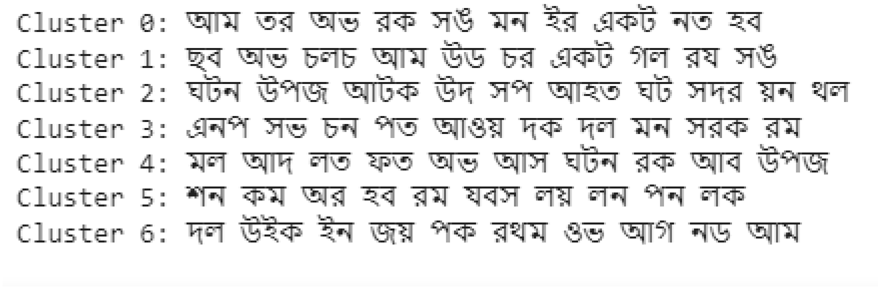


```
clustering done in 1.96 ± 1.15 s
Homogeneity: 0.339 ± 0.017
Completeness: 0.359 ± 0.017
V-measure: 0.349 ± 0.017
Adjusted Rand-Index: 0.238 ± 0.021
Silhouette Coefficient: 0.046 ± 0.005
```

Figure 3.6.3: MiniBatch K–Means execution

3.7 Greatest Phrases According to Group

The TfidfVectorizer's convertibility enables us that determine the cluster centers, providing insights into the most influential words in each cluster. By utilizing minimal features, we can perform text document classification and compare the predictive terms for each target class.

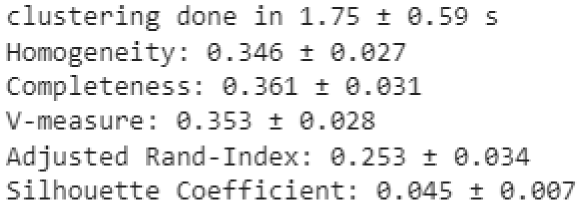


```
Cluster 0: আম তর অভ রক সঙ মন ইর একট নত হব
Cluster 1: ছব অভ চলচ আম উড চর একট গল রয সঙ
Cluster 2: ঘটন উপজ আটক উদ সপ আহত ঘট সদর য়ন থল
Cluster 3: এনপ সভ চন পত আওয় দক দল মন সরক রম
Cluster 4: মল আদ লত ফত অভ আস ঘটন রক আব উপজ
Cluster 5: শন কম অর হব রম যবস লয় লন পন লক
Cluster 6: দল উইক ইন জয় পক রথম ওভ আগ নড আম
```

Figure 3.7.1: Greatest Phrases According to Group

3.8 Hashing Vectorizer

This study addresses the use of a hash vectorizer for processing unstructured data and its application in discovering word patterns and associations. The approach involves assigning numerical values to words, enabling numerical representation for analysis. The study also investigates the possibility of reducing the dimensionality of the hashed vectors by using latent semantic analysis (LSA). The results demonstrate results that are comparable to conventional LSA vectors. The K-Means technique is used to cluster data points based on their similarity. The study demonstrates improved results and accuracy using the proposed model and suggests its applicability to other news document clustering tasks. Future work includes implementing the model on additional news datasets to enhance its performance.



```
clustering done in 1.75 ± 0.59 s
Homogeneity: 0.346 ± 0.027
Completeness: 0.361 ± 0.031
V-measure: 0.353 ± 0.028
Adjusted Rand-Index: 0.253 ± 0.034
Silhouette Coefficient: 0.045 ± 0.007
```

Figure 3.8.1: HashingVectorizer

CHAPTER 4

RESULTS AND DISCUSSION

4.1 Evaluation Summary

The 'curse of dimensionality' is a well-known problem that significantly affects the effectiveness of clustering algorithms such as K-Means and MiniBatch K-Means when dealing with large text datasets. This issue arises due to the high-dimensional nature of text data, which makes it challenging to identify relevant features and patterns. This, in turn, can lead to poor clustering results and reduced accuracy.

However, latent semantic analysis (LSA) can be used to alleviate this problem. LSA employs a shortened SVD (Singular Value Decomposition) to reduce the dimensionality of the data, thereby enhancing the clustering results. By reducing the data's dimensionality, LSA can identify the underlying semantic structure of the text, making it easier to cluster similar documents together.

Although leveraging LSA-reduced data can be time-consuming, particularly when applied to hashed vectors, it increases stability and reduces the amount of time needed for clustering. This is particularly beneficial when dealing with extensive textual studies that involve a large number of documents. By using LSA, researchers can improve the accuracy of their clustering results and obtain meaningful insights from their data.

4.2 Result Discussion

When evaluating the quality of clustering in a dataset, one commonly used metric is the Silhouette Coefficient. It measures how similar an object is to its own cluster compared to other clusters, with values ranging from 0 to 1. However, it is important to note that the Silhouette Coefficient's performance is sensitive to the dataset's dimensionality. As such, it is not advisable to compare Silhouette Coefficient values between different dimensional spaces directly.

Moreover, random labeling is not a suitable baseline for assessing metrics such as homogeneity, completeness, and v-measure, especially when dealing with datasets containing multiple clusters. Instead, adjusted indices like the Adjusted Rand Index (ARI) are more appropriate in these cases.

They are particularly useful for smaller sample sizes or situations that involve a large number of clusters.

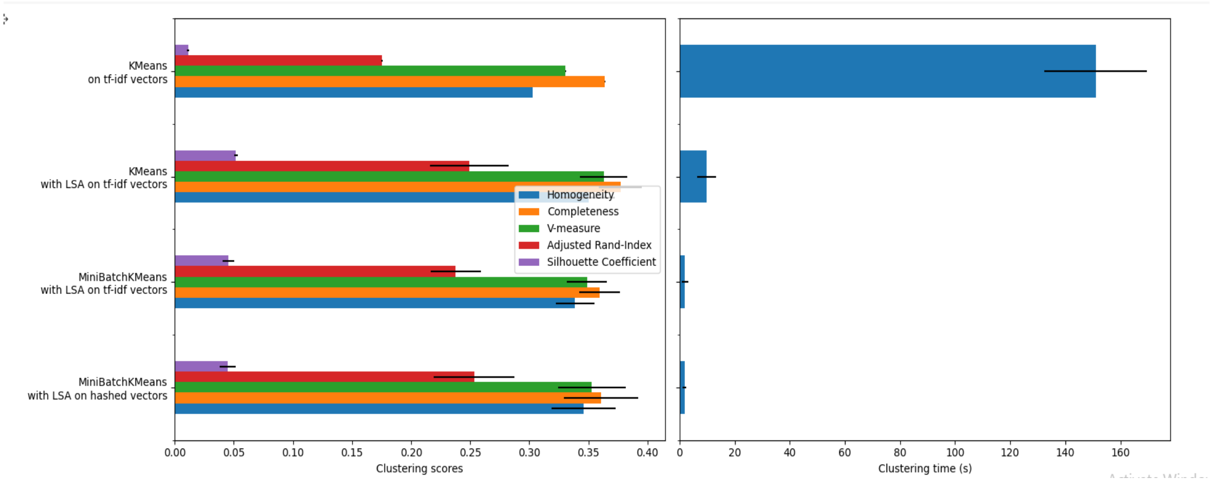


Figure 4.1.1: Clustering Evaluation Graph

It is also worth noting that the MiniBatch K-Means algorithm may have lower reliability than the traditional K-Means algorithm in some datasets. However, it may prove more advantageous for larger sample sizes where processing time is a concern. Therefore, it is important to consider the characteristics of the dataset and the specific goals of the analysis when selecting an appropriate clustering algorithm.

CHAPTER 5

CONCLUSION AND FUTURE WORK

5.1 Impact on Society

The clustering of Bengali news using the K-means algorithm integrated with Latent Semantic Analysis (LSA) significantly enhances information accessibility for society. With the exponential growth of digital content, efficient organization and retrieval of information have become imperative. By categorizing news content into distinct groups based on semantic similarities, this research enables individuals, policymakers, researchers, and the general public to access relevant information more effectively. Improved access to organized news content empowers individuals and stakeholders to make informed decisions. Whether it's citizens staying abreast of current events, policymakers researching policy implications, or researchers analyzing trends, the ability to quickly retrieve and comprehend news articles fosters a more informed and engaged society. A well-organized and accessible news ecosystem promotes transparency and accountability in society. By facilitating access to diverse perspectives and a wide range of news sources, the clustering framework helps counteract misinformation and promotes critical thinking. This, in turn, strengthens democratic principles by fostering an informed citizenry capable of holding institutions accountable. Clustering Bengali news content also contributes to the preservation and promotion of cultural heritage. By organizing news articles according to semantic similarities, the clustering framework can highlight key themes, events, and narratives relevant to Bengali culture, fostering a deeper appreciation and understanding of cultural nuances.

5.2 Impact on Environment

The clustering framework plays a vital role in promoting environmental sustainability by optimizing the utilization of resources. Although the environmental impact of this framework may not be immediately noticeable, it indirectly contributes to reducing carbon emissions associated with digital infrastructure by minimizing redundant storage and processing resources in data centers.

The framework organizes and categorizes news content efficiently, leading to significant reductions in the digital footprint associated with information retrieval processes. By minimizing redundant searches and processing, the clustering framework substantially reduces the overall energy consumption and environmental impact of digital operations.

Additionally, the framework facilitates easier access to news articles related to environmental issues, promoting awareness and advocacy for environmental sustainability. It presents organized and accessible information, making individuals more inclined to engage with environmental topics. This increased engagement leads to support for environmental initiatives and behavioral changes that benefit the environment, such as reducing the use of single-use plastics, adopting sustainable transportation methods, and conserving energy resources.

Furthermore, the clustering framework can promote sustainable business practices by providing companies with valuable insights into their environmental impact. By analyzing patterns in news articles related to environmental sustainability, businesses can identify areas where they can improve their sustainability practices and reduce their carbon footprint.

Overall, the clustering framework is a powerful tool for promoting environmental sustainability in the digital age. Its impact goes beyond reducing energy consumption and carbon emissions; it creates a culture of environmental awareness and advocacy, leading to positive changes that benefit the environment and society as a whole.

5.3 Ethical Aspects

When it comes to the clustering of news articles, ethical considerations play a crucial role in ensuring that the rights of individuals mentioned in the articles are respected. This means that sensitive information must be handled with the utmost care, and data protection regulations must be followed to the letter. It is also essential to take steps to mitigate any biases that may arise from the clustering process. This includes identifying and addressing algorithmic biases and ensuring that there is fair representation of diverse perspectives within the clustered news content.

To uphold ethical standards, it is imperative to transparently document the clustering methodology and validate the results. This means making it clear how the clustering algorithm works, what data is used, and how the results are generated. By providing insights into the clustering process and its limitations, we can foster trust and confidence in the research outcomes. It is also essential to consider potential unintended consequences of the clustering algorithm, such as reinforcing stereotypes or promoting misinformation. Therefore, ethical oversight and continuous evaluation of the clustering framework are necessary to identify and address any adverse impacts on individuals or society.

Overall, upholding ethical standards in news article clustering involves a comprehensive approach that considers various factors, including privacy, bias mitigation, transparency, and unintended consequences. By taking a multifaceted approach to ethics, we can ensure that our research outcomes are not only reliable and trustworthy but also ethical and beneficial to society.

CHAPTER 6

SUMMARY, CONCLUSION, RECOMMENDATION, AND IMPLICATION FOR FUTURE RESEARCH

6.1 Summary of the study

The primary focus of this study is to explore the use of the K-means algorithm with Latent Semantic Analysis (LSA) for the purpose of clustering Bengali news content in a more efficient manner. The main objective of this research is to enable stakeholders to access and retrieve specific information from a vast amount of news content retrieved from Bengali newspapers.

The study employs a self-taught learning model to cluster the news dataset into distinct groups, without relying on external labels or identifiers. The dataset analyzed in this research comprises approximately 0.5 million news texts from multiple Bengali newspapers, covering seven distinct categories of news content.

The study is designed to optimize the clustering process and identifies key steps involved in the process. The dataset is preprocessed, and punctuation and keywords are converted into codes for deep learning techniques. The learned groups are then clustered using the K-means algorithm. The research also identifies areas for improvement, such as data processing and sentence separation, and suggests the exploration of neural network-based deep learning approaches to address these challenges.

Through experiments and analysis, the study demonstrates the effectiveness of the proposed K-means clustering framework in organizing Bengali news content. The research highlights the potential of this framework for practical implementation in automated document organization and information retrieval systems. The study emphasizes the need for further research to optimize the clustering process and to explore the potential of other deep learning techniques for this task. Overall, this study is a valuable contribution to the field of automated document organization and information retrieval.

6.2 Conclusions

The study has reached a significant conclusion that holds great promise for the efficient organization and categorization of Bengali news content. Specifically, the researchers have proposed the integration of the K-means algorithm and Latent Semantic Analysis (LSA) to cluster Bengali news content. This approach has demonstrated the ability to streamline the process by effectively categorizing the news content and highlighting its adaptability to complex linguistic structures typical of Bengali text. It is expected that the proposed integration will enhance the categorization of Bengali news content and simplify the process of organizing it.

This study's remarkable accomplishment is attributed to its utilization of self-taught learning models and advanced deep learning techniques. The proposed framework's innovative and practical approach is evident in its ability to function independently without requiring external annotations. The framework's autonomy is a testament to its robustness and potential for practical applications in the real world. By autonomously clustering news articles into relevant and meaningful groups, the framework demonstrates its capacity to provide valuable insights that can benefit various industries.

The study showcases the impressive effectiveness of the K-means clustering approach despite the challenges of dealing with complex data processing and the intricacies of natural language processing. The research team meticulously experimented and analyzed the framework, highlighting its ability to accurately categorize Bengali news content, even in the presence of linguistic complexities and ambiguities. The results of the study are a testament to the robustness and reliability of the K-means clustering approach in handling sophisticated natural language processing tasks.

The study has revealed potential opportunities for further improvement and streamlining of the clustering framework. One of the most promising avenues for enhancing the framework's efficiency and scalability is to delve into neural network-based approaches. The application of deep learning techniques represents a particularly exciting prospect for unlocking new insights and capabilities that could elevate the clustering framework to unprecedented levels of accuracy and performance. By exploring these methodologies, we may be able to achieve significant advancements in clustering techniques and gain a deeper understanding of complex data structures.

The conclusion of the study presents a promising outlook for progress and growth. It not only affirms the practicality and effectiveness of the suggested framework, but also sets the stage for further innovations in the domains of news clustering and information structuring. The framework has the potential to transform the way we interact with and comprehend large amounts of textual data in the digital era, by leveraging the latest technologies and refining them for optimal output. As the field evolves and new techniques are explored, the framework can offer a revolutionary approach to managing and making sense of information overload.

6.3 Implication for Further Study

As this research unfurls new dimensions in predicting developer categories on GitHub through Graph Neural Networks (GNN), it opens a vista of intriguing avenues for future exploration. The following considerations beckon researchers to delve deeper into the nuances and potential extensions of this study:

6.3.1 Exploration of Advanced Deep Learning Techniques: The proposed task is to conduct a thorough investigation into the application of advanced deep learning techniques, specifically convolutional neural networks (CNNs) or recurrent neural networks (RNNs), to enhance the performance of the clustering framework. The objective is to explore the potential of these deep learning models in capturing intricate semantic relationships and contextual nuances present in Bengali news articles, thereby leading to more accurate and insightful clustering results.

To achieve this, the study will involve experimenting with different deep learning architectures and analyzing their effectiveness in optimizing the clustering process. This will also involve conducting comprehensive comparative studies to evaluate the performance of various deep learning models, taking into consideration factors such as computational efficiency and scalability. The ultimate goal is to establish a framework that can accurately cluster Bengali news articles while being both efficient and scalable.

6.3.2 Investigation into Alternative Clustering Algorithms: The task at hand is to explore various clustering algorithms beyond the commonly used K-means, such as hierarchical clustering, density-based clustering, or spectral clustering, to discern their suitability for organizing Bengali news content. This involves assessing how these alternative algorithms handle the inherent complexities of Bengali text, which includes syntactic variations, semantic ambiguities, and cultural nuances. The aim is to compare their performance with the K-means approach to identify which algorithm suits best for clustering Bengali news content.

Moreover, we aim to investigate hybrid approaches that combine multiple clustering methods to leverage their respective strengths and mitigate their limitations, aiming for improved clustering accuracy and robustness. This will help us identify the most optimal clustering method for Bengali news content and provide better insights into organizing the vast amount of news data available.

6.3.3 Cross-Cultural Applications and Linguistic Diversity: The proposed study's framework can be extended to analyze news content in various languages and cultural contexts. This expansion can broaden its applicability beyond Bengali news and make it more relevant to other cultural contexts. An investigation into how the clustering framework performs when applied to diverse linguistic structures and cultural narratives can provide valuable insights. This exploration can help identify any necessary adaptations and optimizations required for different language families and writing systems.

Moreover, the impact of cultural biases and linguistic variations on the clustering process should be examined. This examination should consider the factors such as regional dialects, idiomatic expressions, and socio-political contexts that can impact the clustering process. By doing so, we can ensure that the clustering framework performs effectively and efficiently, regardless of the language or cultural context it is applied to.

6.3.4 Ethical Considerations and Algorithmic Transparency: The task at hand involves performing a comprehensive analysis of the ethical implications associated with the use of automated news clustering systems. This analysis will cover a range of topics, including but not limited to bias mitigation, privacy preservation, and algorithmic transparency.

To address the issue of bias, we will explore various methodologies that can be used to ensure that the clustering process is fair and unbiased. This will include examining fairness-aware clustering algorithms, as well as post-processing techniques that can be used to correct or mitigate any biases that are found.

Maintaining algorithmic transparency and accountability is also of utmost importance in this context. As such, we will develop frameworks that ensure that news clustering systems are transparent and accountable. This will include establishing documentation standards, audit mechanisms, and stakeholder engagement strategies that promote trust and transparency. By doing so, we can ensure that these systems are used in a responsible and ethical manner, while still providing all the benefits that they offer.

6.3.5 Real-World Applications and Industry Integration: In order to effectively implement the clustering framework, it is essential to collaborate with industry partners and stakeholders. This will enable the deployment and evaluation of the framework in real-world applications, such as news recommendation systems, content discovery platforms, and media analytics tools. This will provide the opportunity to assess the framework's performance, scalability, and user satisfaction in practical settings. By gathering real-world usage feedback, areas for optimization and refinement can be identified and addressed.

Furthermore, exploring potential avenues for commercialization and technology transfer is crucial. By leveraging the clustering framework, companies can address pressing challenges in media and information management industries. It is important to ensure that responsible and ethical deployment practices are followed in order to optimize the framework's potential for positive outcomes. This includes considering the potential impact on users, privacy concerns, and other ethical considerations.

6.3.6 Real-world Application Testing: The first step in real-world application testing involves deploying the Bengali news clustering framework in a production environment. This environment should mimic the conditions under which the clustering framework will operate once deployed for actual use.

Ensure that all components of the clustering framework, including data ingestion pipelines, preprocessing modules, clustering algorithms, and user interfaces, are properly configured and integrated into the production environment.

Collect a diverse and representative dataset of Bengali news articles from various sources, including newspapers, online portals, and news agencies. The dataset should cover a wide range of topics, genres, and writing styles to adequately test the clustering framework's versatility.
Integrate the collected dataset into the production environment, ensuring seamless data ingestion and preprocessing to prepare the news articles for clustering analysis.

Define evaluation metrics and performance benchmarks to assess the clustering framework's effectiveness and efficiency in organizing Bengali news content. Common metrics include clustering accuracy, silhouette score, and computational performance metrics such as runtime and memory usage.
Establish baseline performance benchmarks based on initial testing results or existing literature to gauge the clustering framework's performance improvements over time.

6.3.7 Feedback Incorporation and Iterative Improvement: To ensure the security and privacy of data, it is essential to conduct thorough testing to identify any vulnerabilities or compliance issues related to data privacy, confidentiality, and integrity. This testing should include various techniques such as penetration testing, vulnerability scanning, and compliance audits. Once vulnerabilities or compliance issues are identified, it is crucial to implement appropriate measures such as encryption, access controls, and data anonymization techniques to protect sensitive information. These measures should be in line with regulatory requirements such as GDPR or HIPAA, which govern data privacy and security.

Gathering feedback from stakeholders, end-users, and testing results can help identify areas for improvement and prioritize feature enhancements. This feedback can be obtained through surveys, interviews, or user testing sessions. It is important to note that feedback should be carefully analyzed and prioritized based on the impact on the overall system's functionality, performance, and usability. Finally, the iterative improvement process should be followed to enhance the clustering framework's functionality, performance, and usability continuously. This process should be based on feedback and testing results, following an agile development approach. The agile approach allows for flexibility and adaptability in response to new requirements or changes in the system's environment.

REFERENCES

[1] Aggarwal, C. C., & Zhai, C. (2012). A survey of text clustering algorithms. Mining text data, 77-128.

[2] Sawaf, H., Zaplo, J., & Ney, H. (2001). Statistical classification methods for Arabic news articles. Natural Language Processing in ACL2001, Toulouse, France.

[3] Disayiram, N., & Rupasingha, R. A. H. M. (2022, September). A Comparative Study of Clustering English News Articles Using Clustering Algorithms. In 2022 International Research Conference on Smart Computing and Systems Engineering (SCSE) (Vol. 5, pp. 108-113). IEEE.

[4] Song, W., & Park, S. C. (2007, October). A novel document clustering model based on latent semantic analysis. In Third International Conference on Semantics, Knowledge and Grid (SKG 2007) (pp. 539-542). IEEE.

[5] Bouras, C., & Tsogkas, V. (2017). Improving news articles recommendations via user clustering. International Journal of Machine Learning and Cybernetics, 8, 223-237.

[6] Al-Azzawy, D. S., & Al-Rufaye, F. M. L. (2017, March). Arabic words clustering by using K-means algorithm. In 2017 Annual Conference on New Trends in Information & Communications Technology Applications (NTICT) (pp. 263-267). IEEE.

[7] Fan, Z., Chen, S., Zha, L., & Yang, J. (2016). A text clustering approach of Chinese news based on neural network language model. International Journal of Parallel Programming, 44, 198-206.

[8] Xu, W., Liu, X., & Gong, Y. (2003, July). Document clustering based on non-negative matrix factorization. In Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval (pp. 267-273).

[9] Steinbach, M., Karypis, G., & Kumar, V. (2000). A comparison of document clustering techniques.

[10] Yao, M., Pi, D., & Cong, X. (2012). Chinese text clustering algorithm based k-means. Physics Procedia, 33, 301-307.

[11] Aliwy, A. H., Aljanabi, K., & Alameen, H. A. (2022, January). Arabic text clustering technique to improve information retrieval. In AIP Conference Proceedings (Vol. 2386, No. 1). AIP Publishing.

[12] Naeem, S., & Wumaier, A. (2018). Study and implementing K-mean clustering algorithm on English text and techniques to find the optimal value of K. International Journal of Computer Applications, 182(31), 7-14.

[13] Yang, J., & Wang, J. (2017). Tag clustering algorithm LMMSK: improved K-means algorithm based on latent semantic analysis. Journal of Systems Engineering and Electronics, 28(2), 374-384.

[14] Mohammed, S. H., & Al-augby, S. (2020). Lsa & lda topic modeling classification: Comparison study on e-books. Indonesian Journal of Electrical Engineering and Computer Science, 19(1), 353-362.

[15] Fujisawa, H., Shima, Y., Koga, M., & Murakami, T. (1992). Automatically organizing document bases using document understanding techniques. In Future Databases' 92 (pp. 244-253).

[16] Salloum, S. A., Al-Emran, M., Monem, A. A., & Shaalan, K. (2018). Using text mining techniques for extracting information from research articles. Intelligent natural language processing: Trends and Applications, 373-397.

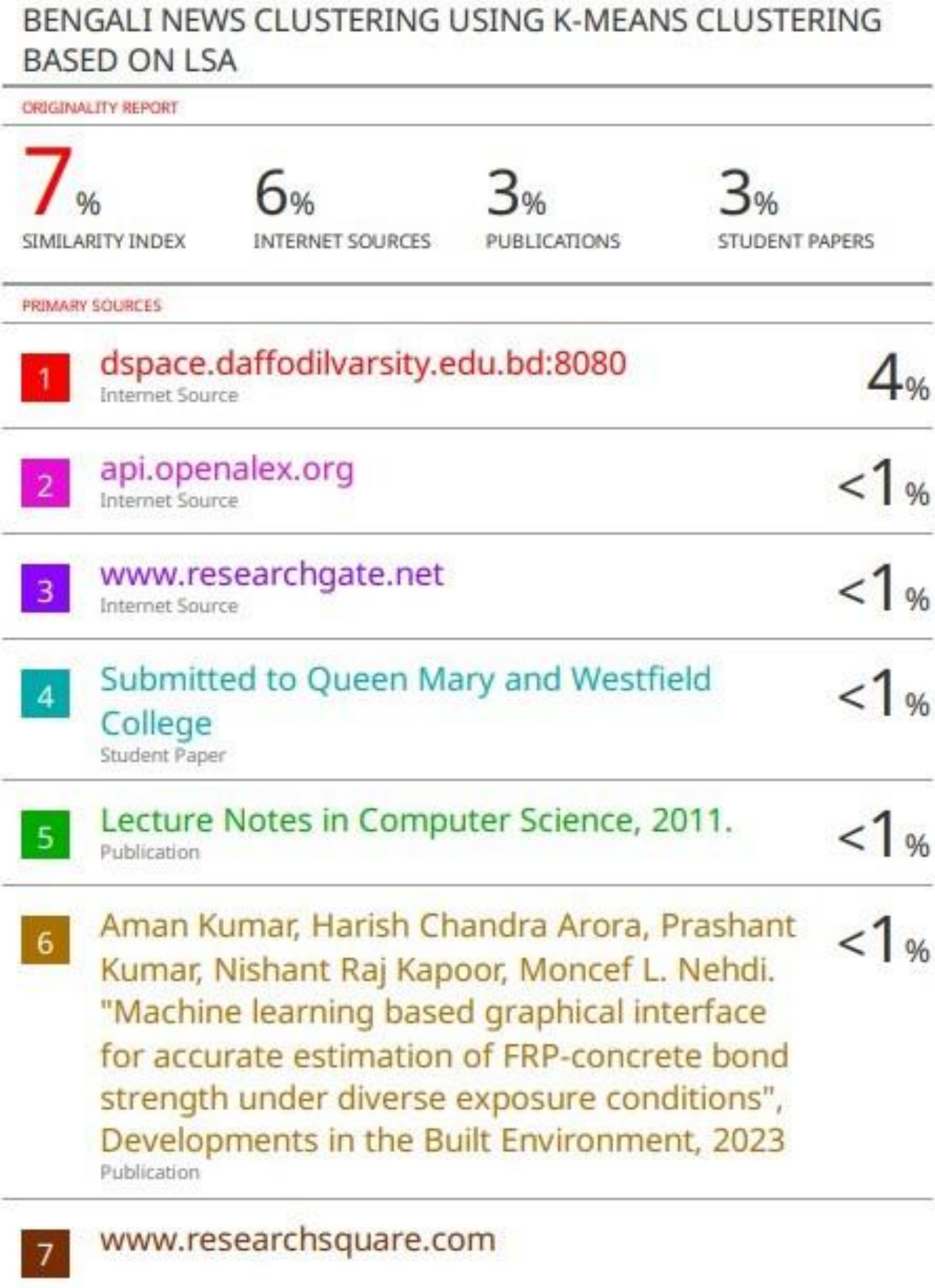
[17] Kanungo, T., Mount, D. M., Netanyahu, N. S., Piatko, C., Silverman, R., & Wu, A. Y. (2000, May). The analysis of a simple k-means clustering algorithm. In Proceedings of the sixteenth annual symposium on

[18] Hossain, M. R., Hoque, M. M., Siddique, N., & Sarker, I. H. (2021). Bengali text document categorization based on very deep convolution neural network. Expert Systems with Applications, 184, 115394.

[19] Li, Y., Luo, C., & Chung, S. M. (2008). Text clustering with feature selection by using statistical data. IEEE Transactions on knowledge and Data Engineering, 20(5), 641-652.

[20] Zhao, G., Liu, Y., Zhang, W., & Wang, Y. (2018, January). TFIDF based feature words extraction and topic modeling for short text. In Proceedings of the 2018 2nd international conference on management engineering, software engineering and service sciences (pp. 188-191).

[21] Dhar, P., & Abedin, M. Z. (2021). Bengali News Headline Categorization Using Optimized Machine Learning Pipeline. International Journal of Information Engineering & Electronic Business, 13(1)



BENGALI NEWS CLUSTERING USING K-MEANS CLUSTERING BASED ON LSA

ORIGINALITY REPORT

7%	6%	3%	3%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	dspace.daffodilvarsity.edu.bd:8080 Internet Source	4%
2	api.openalex.org Internet Source	<1%
3	www.researchgate.net Internet Source	<1%
4	Submitted to Queen Mary and Westfield College Student Paper	<1%
5	Lecture Notes in Computer Science, 2011. Publication	<1%
6	Aman Kumar, Harish Chandra Arora, Prashant Kumar, Nishant Raj Kapoor, Moncef L. Nehdi. "Machine learning based graphical interface for accurate estimation of FRP-concrete bond strength under diverse exposure conditions", Developments in the Built Environment, 2023 Publication	<1%
7	www.researchsquare.com	