

**K-MEANS CLUSTERING FOR NEWS CLUSTERING BASED ON
LATENT SEMANTIC ANALYSIS**

BY

MANISHA DAS JAYA

ID: 191-15-12340

MD. RABIUL ALAM

ID: 171-15-8766

SOMAIYA AKTER KAKOLI

191-15-12476

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

Supervised By

Md. Ferdouse Ahmed Foysal

Lecturer

Department of Computer Science and Engineering Faculty of Science
and Information Technology Daffodil International University

Co-Supervised By

Nishat Sultana

Lecturer

Department of Computer Science and Engineering Faculty of Science
and Information Technology Daffodil International University



DAFFODIL INTERNATIONAL UNIVERSITY

Dhaka, Bangladesh

January 2023

APPROVAL

This Thesis titled “K-Means Clustering for News Clustering Based on Latent Semantic Analysis”, submitted by Manisha Das Jaya, ID: 191-15-12340, Md. Rabiul Alam ID: 171-15-8766, Somaiya Akter Kakoli ID: 191-15-12476 to the Department of Computer Science and Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfilment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on 29 January 2023.

BOARD OF EXAMINERS

Chairman

Dr. Touhid Bhuiyan
Professor and Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University



Internal Examiner

Md. Abbas Ali Khan
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University



Internal Examiner

Ms. Aliza Ahmed Khan
Senior Lecturer
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University



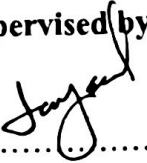
External Examiner

Dr. Md. Sazzadur Rahman
Associate Professor
Institute of Information Technology
Jahangirnagar University

DECLARATION

We hereby declare that, we completed this research project under the supervision of **Md. Ferdouse Ahmed Foysal**, Lecturer, Department of CSE, Daffodil International University and co-supervision of **Nishat Sultana**, Lecturer, Department of CSE, Daffodil International University. We also declare that neither this project nor any portion of it has been submitted to any institution for the award of a degree or diploma.

Supervised by:

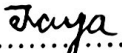


.....
Md. Ferdouse Ahmed Foysal
Lecturer, Department of CSE
Daffodil International University

Co-Supervised by:

.....
Nishat Sultana
Lecturer, Department of CSE
Daffodil International University

Submitted by:



.....
Manisha Das Jaya
ID: 191-15-12340
Department of CSE
Daffodil International University



.....
Md. Rabiul Alam
ID: 171-15-8766
Department of CSE
Daffodil International University



.....
Somaiya Akter Kakoli
ID: 191-15-12476
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

Firstly, we express our heartiest thanks and gratefulness to Almighty God for His divine blessing that made us possible to complete the final year thesis successfully.

We are really grateful and wish our profound indebtedness to our supervisor **Md. Ferdouse Ahmed Foysal**, Lecturer, Department of CSE, Daffodil International University and **Co-supervisor Nishat Sultana**, Lecturer, Department of CSE, Daffodil International University. Our supervisor has extensive knowledge and a deep interest in the field of "Deep Learning & Natural Language Processing" which helped us carry out this research. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior drafts and correcting them at all stage have made it possible to complete this thesis.

We would like to express our heartiest gratitude to **Dr. Touhid Bhuiyan**, Professor and Head, Department of CSE, for his kind help to finish our thesis and also to other faculty members and the staff of CSE department of Daffodil International University.

We would like to thank our entire course mate in Daffodil International University, who took part in this discuss while completing the course work. We also appreciate the encouragement and inspiration provided by all of our well-wishers, friends, family, and elders. This research is the result of a great deal of effort and the encouragement and cooperation of all those people.

Finally, we must acknowledge with due respect the constant support and passion of our parents.

ABSTRACT

K-means Clustering for News Clustering based on Latent Semantic Analysis is a challenging problem because of its rarity. Document clustering or text clustering is a cluster analysis process of textual documents. In this technological era, the use of text clustering processes in different domains is most happening work in research fields recently, including automatic document organization and text extraction for mining significant sets of valuable information on the Internet. In this article, we propose a K-means Clustering News Clustering using Latent Semantic Analysis-based framework, which is more beneficial for clustering news documents or text. The clustering process works by considering a set of data to cluster them in groups with the help of a self-taught learning model, which need not use any external labels or tags. We used the BBC news classification dataset and implemented our proposed model on the dataset. Firstly, we set the dataset to prepare for the clustering and semantic analysis method to categorize the dataset. Then the keywords and punctuation marks are processed into codes to apply the Deep Learning methods on it for the training procedure. Then we use K-means to cluster them after getting the learned delegations. But there are some issues to operate such as sentences and punctuation marks separation and data processing. To conquer these issues, we propound a Deep Learning framework using neural networks. This is an effective approach for news text or document clustering because no revolutionary work has been done about it yet. We also have done some experiments to demonstrate the specific implementation of the approach which confirms the effectiveness of the proposed method.

TABLE OF CONTENTS

CONTENTS	PAGE
CHAPTER	
CHAPTER 1: INTRODUCTION.....	1-9
1.1 Introduction.....	1-3
1.2 Objective.....	4
1.3 Motivation.....	5
1.4 Rationale of the Study.....	5-6
1.5 Research Questions.....	6-7
1.6 Expected Outcome.....	7-8
1.7 Layout of the Report.....	8-9
CHAPTER 2: BACKGROUND STUDY	10-14
2.1 Introduction.....	10
2.2 Related Works.....	10-12
2.3 Comparative Analysis and Summary.....	12-13
2.4 Challenges.....	13-14
CHAPTER 3: RESEARCH METHODOLOGY	15-26
3.1 Introduction.....	15
3.2 Workflow.....	15-17
3.3 Experiment Dataset.....	18
3.4 Proposed Methodology.....	19

3.4.1 Evaluating The Effectiveness.....	19
3.4.2 K-means Clustering on Text Features.....	19-20
3.4.3 Utilizing TfidfVectorizer.....	20-21
3.5 K-means clustering of sparse data.....	21-22
3.6 Performing dimensionality using LSA.....	22-23
3.7 Top Terms Per Cluster.....	23-24
3.8 HashingVectorizer.....	24-26
CHAPTER 4: EXPERIMENTAL RESULTS AND DISCUSSION	27-28
4.1 Evaluation Summary.....	27
4.2 Result Discussion.....	27-28
CHAPTER 5: CONCLUSION AND FUTURE WORK	29-30
5.1 Conclusion.....	29
5.2 Future Works.....	29-30
REFERENCES	31-33
PLAGIARISM REPORT	34

LIST OF FIGURES

FIGURES	PAGE NO
Fig 3.2.1: The Flowchart of the Workflow	17
Fig 3.3.1: The Categorized Dataset	18
Fig 3.4.3.1: Utilizing TfidfVectorizer	21
Fig 3.5.1: K-means Clustering of Sparse Data	21
Fig 3.5.2: K-means Clustering with Highest Inertia	22
Fig 3.6.1: Performing Dimensionality Reduction using LSA	23
Fig 3.6.2: Processing times for K-Means and MiniBatch K-Means	23
Fig 3.6.3: Conducting MiniBatch K-Means	23
Fig 3.7.1: Top Terms Per Cluster	24
Fig 3.8.1: HashingVectorizer	25
Fig 3.8.2: HashingVectorizer (2)	25
Fig 4.1.1: Clustering Evaluation Graph	27

CHAPTER 1

INTRODUCTION

1.1 Introduction

News clustering is the process of the cluster news documents accurately. Clustering is the basic process of analysis pattern recognition methods, image analysis, and data mining which converts the data into clusters. News clustering also embeds the clustering process turning to the advantage of different perspectives of Deep Learning. The antecedent research works unraveled the utilization of text clustering methods for other languages like English and Chinese. No revolutionary work has not been done on news documents or text before. At present, in terms of the text clustering approach, there are mainly some specific languages that have been used in the text clustering process. But news text clustering in all languages and its rarity is a challenging problem nowadays. It is a method that is a unique solution for the clustering process of news text or document. The previous studies worked for some specific fields and limited sentence expressions of news documents to show manifest clustering procedures in the sentence structure. So the news text clustering method by latent semantic analysis for all languages is accepted. Our proposed method is not a limited field or sentence expression, it works with all types of news textual documents and expressions. Basically, we used the BBC news classification dataset which has five categories and implemented our proposed model on the dataset. The BBC news dataset has both training and testing records to enhance the desired results. We used modified K-Means algorithms and Latent Semantic Analysis to cluster the dataset. The K-Means algorithm is a classic clustering algorithm used for grouping data points into clusters based on their similarities. It is a popular choice for tasks of data exploration and analysis as it is relatively easy to implement and interpret. The algorithm works by first randomly selecting k-number of points from the dataset to serve as the initial cluster centers. Then, the data points are assigned to the closest cluster center in terms of Euclidean distance. Latent semantic analysis (LSA) is an algorithm used for text

analysis that uses the relationships between terms to produce a numerical representation of the text. It is often used in natural language processing (NLP) and computational linguistics, as well as information retrieval. This numerical representation of the text can be used in tasks such as machine learning, as it provides a way to capture the contextual information of the text in a computer-readable format. Furthermore, LSA can be used to identify relationships between words that may not be obvious from the text alone, allowing for improved searching and retrieval of documents. First, we process the dataset to prepare for the latent semantic analysis and K-means clustering method. Then the keywords and punctuation marks are processed into codes to apply the latent semantic analysis methods on it for the training procedure. Then we use K-means to cluster them after getting the learned delegations. After the clustering process, the recognition of the same cluster set is done for each group. Besides, it is important to set up a progressive update cluster and gradually structure the text automatically and maintain the news document in an intellectual system. K-means clustering is a popular method for news clustering that takes advantage of the underlying latent semantic analysis. It uses an unsupervised learning approach to group similar items of news together for better categorization and organization. By utilizing a process of iterative refinement, k-means clustering is able to identify similarities among different types of news articles and group them together automatically. This automated news clustering method helps editors and journalists quickly organize and analyze their stories, making it easier to find interesting and related topics. K-means clustering is a popular clustering algorithm which is used to partition data into k distinct clusters. This algorithm is particularly useful for clustering based on latent semantic analysis. Through this process, one can find a better way of structuring their data and can discover latent patterns within the data. Furthermore, it allows for the exploration of various topics and themes in the data set. K-means clustering is an unsupervised learning algorithm which is used to group similar data points into clusters. It is commonly used in News clustering and relies on latent semantic analysis to identify similarities between documents. This algorithm has been successfully used in various news applications, such as automatic topic identification and summarization. K-means clustering has proven to be an effective tool for extracting the most valuable information from large document collections. K-means clustering is a popular supervised learning algorithm used to grouping similar data points together. This form of

analysis was utilized in a recent study that focused on the application of latent semantic analysis (LSA) for news articles clustering. The use of LSA enables the algorithm to identify the semantic meaning of news documents which makes it easier to cluster them into more meaningful groups. The researchers utilized k-means for their data clustering and their results showed that LSA was successful in clustering news articles into more meaningful groups as compared to traditional clustering approaches.

This article proposes an excellent clustering method for news documents which is a very important task and a challenging problem of its sparseness. The study provides an unprecedented research framework for the progressive document clustering method which is mainly for news documents for all languages and is based on deep learning approaches based on latent semantic analysis. In this research, text clustering is studied, some fundamental definitions and discussions of methods and some experiments of clustering are given and the news document clustering is discussed. With Deep Neural Networks (DNN), which is the recently revived interest in the research field, Deep Learning is being used to learn features by many researchers. Nowadays, in creating text representation, neural networks are giving their best performance with the help of word embeddings, such as Recursive Neural Networks and Recurrent Neural Networks (RNN). However, Recursive Neural Networks represent the complexity of high time to create the textual tree and RNN, using the computed layer at the final word representing the text as a preferential model. Applying the convolutional filters to recognize the local expressions of sentences, which are included in Convolution Neural Networks (CNN), has acquired a better performance in many NLP works, like different types of traditional NLP tasks, relation classification, and sentence modeling. Maximum preceding works focus on solving the clustering problem of other languages mainly English and Chinese languages, while in this paper we aim to achieve a better solution for news text clustering by using Deep Learning approaches based on latent semantic analysis.

1.2 Objective

In this global era of technology, there is a new invention nowadays, and a new addition is being made to our life and technology. To move with this global development and modern life, we have to keep ourselves updated and technology enthusiasts to keep pace with this global development. Differences in language shouldn't be the reason for falling behind anymore. News text clustering in English and Chinese and other languages has been so close to the accuracy that they have started implementing text clustering in different devices and software. But news text clustering for all languages has not been there yet. The goal of this document is to discuss how k-means clustering can be used to efficiently cluster news articles using latent semantic analysis. By clustering news articles in this way, we can create a more organized and efficient way of sorting through and understanding large amounts of news data. This can help to improve news reader comprehension, ensuring that readers are receiving the most important and relevant news stories in a timely manner. Additionally, it can help to reduce the amount of noise present in news media, allowing readers to focus more on the stories that matter and less on stories that are irrelevant to their interests. We invested ourselves in news document clustering through Deep Learning based Model so that we can get a more accurate and functional system to generate news headlines and use it for further improvement.

We can describe these goals in a list like the following:

- Our goal is to create a news document clustering system with the best performance alongwith a resourceful news text dataset.
- We have to cluster the dataset or documents of news text.
- As the documents of the dataset are given randomly, we have to generate the system of clustering in such a way that is easier to cluster news documents.
- Set the news document dataset for News Document Clustering and find out the bestworking model for the improved dataset so that it can be used for furthermore research.

1.3 Motivation

Nowadays some inventions appear in this technological era constantly. As a human being, we have to keep updated ourselves with these new inventions and closer to the present global era of technology. As a student, we can help others to do that through different types of work. Throughout the courses, there were very few opportunities to get connected with people and do something for them. News content analysis has recently seen a surge in use in various industries. With the increased availability of data, methods like k-means clustering are becoming more popular for grouping news stories by their latent semantic features. This approach is often used to find related stories, identify trends, and understand how news is being received. To make use of these clustering algorithms, data must be pre-processed in ways that can be used to accurately and efficiently cluster the available stories.

News Text Clustering was something that actually will have a great impact on serving people with disabilities, security, and surveillance. When we started reading about this, the work seemed interesting and positive to people. But most of the advancements in this sector were recorded or performed in English and Chinese and other languages. For this, we think that if want to get the full benefits of text clustering, we need to develop News Text Clustering models and datasets for all languages. Following the findings of earlier text clustering research, we were inspired to make a standard News Clustering model that would eliminate the discrimination between our proposed method and other news clusterings as the sparsity of research about news document clustering is a challenging problem.

1.4 Rationale of the Study

There so many works that have been done in the field of text clustering that is maximum for English and Chinese and other languages. But the work in news text clustering using deep learning and machine learning algorithms is infrequent. We used the BBC news classification data for clustering by K-means clustering method based on latent semantic analysis. K-means clustering is a common and well-known algorithm used for data analysis and machine learning.

The algorithm works by partitioning a dataset into a set of k clusters, where the number of clusters 'K' is chosen by the user. It then calculates the sum of squared errors for each of the points in each cluster, which is used to measure the accuracy of the clusters. To assign points to the clusters, it uses an iterative process that assigns points to clusters based on their similarity to the cluster's center point. Latent Semantic Analysis (LSA) is an algorithm used to extract and analyze information from text. It is a form of natural language processing that uses mathematical and statistical techniques to identify the relationships between words and concepts. LSA looks at the context of each word in a document to better understand the meaning of the text. This algorithm can be used to identify relationships between documents, classify them based on topics, and help discover hidden topics and relationships. Despite all the text clustering works completed, our research is unique and different because of the news text and its clustering methods of Deep Learning for news text documents. It is the most significant resolution in the field of text clustering in news text documents, which implement different approaches than the preceding research works. We applied Deep Learning methods to the dataset to get the best clustering results in this work, which is more effective than other research works in text clustering and also a unique resolution for news text clustering.

1.5 Research Questions

Research is an important tool to answer questions, solve problems, and uncover new insights. It is the systematic investigation into a subject in order to discover knowledge or draw conclusions. For starters, it is essential to formulate a specific research question or hypothesis. This will guide the research process and help to keep it focused. Research seeks to answer questions and provide insight into the world around us. By asking questions about our environment, behavior, and other phenomena, we can find out more about the world that we inhabit. Many research questions such as what clustering techniques are best suited for news clustering based on latent semantic analysis. Moreover, it also allows us to investigate how the size of the clusters affect the accuracy of the clustering results and whether or not the size of the clusters can be optimized for better performance. Finally, k-means clustering can be used

to effectively classify data points into distinct groups, thus providing a better understanding of the underlying structure of the data. The research work was quite challenging for us because of its sparseness in news clustering. The research question is an important part to start research work to be sure about that what we actually want to find out and give a clear concept about it. To find a pragmatic, effective, and accurate resolution to the problem, the researchers would like to proffer the following questions to express their thinking and accomplishment.

- Can we process the dataset and extract the raw data properly?
- Can we get the desired result of news text clustering?
- Will this work enhance the field of news text clustering by utilizing this approach? Can this work help the readers to read the news?
- How the news readers will be benefited from the proposed model?

1.6 Expected Outcome

Most of the points of this section are based on the research questions of our proposed model. As our research progresses, we continue to uncover new ways to make our project faster, more reliable, and more efficient. We are confident that our data and results will show a significant improvement in the overall performance of our product. Our goal is to develop a product that can solve real-world issues while being user-friendly and cost-effective. Additionally, we intend to make sure our research is repeatable and reliable by taking careful steps to ensure accuracy and validity. The anticipated goal of our research work is to cluster news documents successfully by applying the ‘K-means Clustering for News Clustering based on Latent Semantic Analysis’ model to classify unseen news articles into the right category. After the documents have been clustered, the algorithm then uses the average of the metrics to assign a score to each cluster. This score is used to identify the most relevant articles for a given topic and can be used to generate summaries of news stories. K-means clustering has also been used in natural language processing, text mining, and other types of machine learning. With its ability to identify the most relevant articles, this algorithm has become a valuable tool for quickly summarizing large amounts of news data. It is a valuable tool for quickly summarizing large

amounts of news data. It provides a powerful and efficient way to analyze news articles by grouping them into meaningful categories. This allows researchers to quickly identify similarities between articles and find trends or topics within a collection of documents. By leveraging the power of machine learning algorithms, news sources can be analyzed more quickly and accurately than with manual methods. Furthermore, K-means clustering can be used to help identify topics of importance within a data set, allowing for more accurate and efficient analysis.

The following expected outcome of the proposed work would be proposed by the researchers.

- The news text will be automatically clustered into different types of groups.
- Accurately classify unseen data or news articles into the best category by semantic analysis.
- Not only the specific data but also the unspecific data or news text will be clustered by this proposed model using K-means clustering methods.
- Automatically generate the clusters by K-means approaches of news documents in the right category.
- The news readers will be benefitted from this proposed model.

1.7 Layout of the Report

Chapter 1 shows the overall layout of the research which includes the introduction of the research with objective, motivation, research questions, and rationale of the study research's total formation.

Chapter 2 demonstrates the framework of the research work. This section speculates the related works, comparative analysis and summary of this study, and the challenges of the research.

Chapter 3 demonstrates the theoretical discussions of the research. Firstly, it shows the workflow of the whole research project and then discusses the data procurement and data

preparation procedure. Then it shows the required algorithms for the proposed model and the execution of the algorithms properly.

Chapter 4 shows the result of implementation and evaluates the performance of the proposed model. This section gives the proper explanation and clarification about the project.

Chapter 5 concludes the research work with the conclusion and future works. This section represents the overall work and outcomes of the project and also discusses some future recommendations for this study.

CHAPTER 2

BACKGROUND STUDY

2.1 Introduction

K-means clustering is a powerful tool for analyzing the content of news articles. With the ability to group documents by similarity, it provides a way to quickly identify articles that are most relevant to a particular topic. Although the concept of clustering has been known for many years, the development of the K-means algorithm made it possible to automate the feature extraction process. K-means clustering is an unsupervised learning algorithm that takes a set of data points, divides them into clusters, and assigns each point to a cluster based on its distance from the cluster's centroid. This algorithm is highly efficient and its results can be used to identify patterns and trends in data that may have otherwise gone unnoticed. The algorithm also provides a way to assess the quality of a given dataset by measuring the cohesion between points in each cluster. Additionally, this method of clustering is much faster than traditional methods, making it a useful tool for quickly analyzing large sets of data. By combining the power of machine learning with the speed of clustering algorithms, K-means clustering can provide insights into the content of news articles that would not be possible with other method. In this section, we will discuss related previous works, research summary, and some challenges of the research. In the related works section, there will be discussed the related research works and with that their working procedure and results of accuracy. In the comparative analysis and research summary section, there is a research summary of our associated works and some comparison between our work and previous works which we will present. We will also discuss some research challenges and how we enhanced the research results of accuracy.

2.2 Related Works

Many research works have been completed in News Text Clustering for years using Deep Learning approaches. K-means clustering has long been a staple of work in the field of

unsupervised text clustering. Related works have seen k-means applied to news clustering based on latent semantic analysis. These works have typically leveraged k-means to cluster text into data points, while also using a measure of similarity between documents to determine which documents to cluster together. This has been a powerful tool for providing more meaningful insights into news articles without manual categorization by a human. In 2001 a research paper entitled "Statistical classification methods for Arabic news articles." was published, which worked for experimental results on document clustering that was acquired on the Arabic news text applying statistical approaches [8]. Another research work published in 2016 designated "A text clustering approach of Chinese news based on neural network language model." has applied a new method where neural network language model is used foremost in text clustering [2]. In 2012 a research work titled "A clustering technique for news articles using WordNet." was published that worked with various techniques of document clustering and evaluated their application on news articles from the web [5]. In 2014 another research paper was published entitled "Unsupervised feature selection for multi-view clustering on text-image web news data." which proposed a new multi-view feature selection algorithm in news clustering that is unsupervised [7]. In 2017 a published research paper designated "News clustering based on similarity analysis." where the author proposed a two-step algorithm to reach the goal of news text clustering [3]. In 2017 another research work was published entitled "Improving news articles recommendations via user clustering." which proposed a personalized recommendation algorithm that implements both user and text clustering algorithms based on deep learning [6]. In 2010 "A Two-layer text clustering approach for retrospective news event detection." titled research work has published that proposed a two-layer text clustering algorithm to achieve the goal of clustering to propagate the final news events [9]. The latest research work designated "A news text clustering method based on similarity of text labels." has been published in 2019. This research paper proposed a new clustering approach to cluster news text as a series of text labels that work with data latitude [1]. Another recent research work published in 2019 entitled "Arabic text clustering using improved clustering algorithms with dimensionality reduction." has worked in three approaches which are unsupervised, semi-supervised, and supervised techniques to fabricate a classifier for Arabic news text based on clustering algorithms [4]. Some articles whose

references are given in the reference section also studied news clustering by the K-means clustering algorithm and other deep learning approaches [12] [18] [21]. The goal of this paper is to study the efficacy of the k-means clustering algorithm for news clustering based on latent semantic analysis. This type of clustering is an important task in the field of natural language processing, as it helps to organize and group related content into distinct categories. In particular, the use of k-means clustering has shown to be effective in clustering text documents. To the best of our knowledge, no previous study has investigated the use of the k-means algorithm for news clustering based on latent semantic analysis.

2.3 Comparative Analysis and Summary

Most of the previous research work on classification of news text clustering which have applied Deep Learning and neural network algorithms for clustering news documents. Deep learning is a subset of Machine Learning raised on artificial neural networks. Deep learning has three categories- supervised learning, semi-supervised learning, and unsupervised learning. The neural network is an algorithm that follows the operating system of neurons of the human brain. In text clustering, deep learning based on neural networks is the most exotic model. K-means clustering is a popular algorithm for data clustering, used in a variety of applications such as image segmentation, text classification, and facial detection. In news clustering, it is used to group similar articles together into one category. It is advantageous to compare k-means clustering to latent semantic analysis (LSA) for news clustering, as both algorithms have their strengths and weaknesses. While k-means clustering is simple to use, its results may not always be accurate. In the comparative analysis, we will see that there are some similarities and dissimilarities between our work and previous works. The similarities are the uses of deep learning approaches that are implemented in the remaining works. In the analysis of the preceding works, we can see that there are some types of methods that are used repeatedly and we proposed a new modified method that is very significant to do the text clusterings. We used the BBC news dataset which has both training and testing records and cluster the dataset which has not been done before. So this is very useful to reach our goal to cluster the news documents. In this paper, we present a method for news clustering using k-

means and latent semantic analysis. This approach has been successfully applied to various datasets and is capable of producing accurate results. We aim to improve the performance of the clustering process by introducing an adapted version of k-means that utilizes a local search strategy. Additionally, we investigate the use of latent semantic analysis in order to better capture the semantic meaning of news articles.

2.4 Challenges

Although K-means clustering is a powerful tool for uncovering hidden topics in large collections of documents, it is not without its challenges. For example, the latent semantic analysis used to classify documents is based on word frequencies and similarities, which may not accurately reflect the underlying topics of the documents. Additionally, the results of K-means clustering may not accurately reflect the true structure of the underlying data due to its reliance on Euclidean distance measures. Therefore, it is important to use a combination of methods to ensure that the results are as accurate as possible. The initial challenge of this research work is to process the dataset and implement the modified new method to cluster the news document. We collect the dataset of BBC news classification and process the dataset for our proposed methods. We had to make ready the data to be a perfect dataset. By cleaning and normalizing the data we prepared the dataset perfectly to work on our model. Then we applied different methods and algorithms of Deep Learning based on neural networks over the dataset. Once the documents have been grouped together, K-means clustering can be used to extract the underlying topics of the articles. By analyzing the words and phrases used in each group, it is possible to uncover topics that are relevant to all of the documents. This can be a useful tool for journalists or researchers looking to uncover topics that might be hidden under the surface of a large collection of news articles. Additionally, it can be used to recommend articles similar to a particular article, or even to identify breaking news about certain topics. It was quite difficult to implement the modified methods and algorithms on the dataset and get the desired result and accuracy. The dataset was containing records of training and testing sets which were also used to improve the result of the implementation of our modified approaches. As a result, we

overcome the research challenges and comparatively get better results from previous research works. Our proposed model gives better performances than the other remaining models and approaches.

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Introduction

K-means clustering is a popular supervised learning algorithm used to grouping similar data points together. This form of analysis was utilized in a recent study that focused on the application of latent semantic analysis (LSA) for news articles clustering. The use of LSA enables the algorithm to identify the semantic meaning of news documents which makes it easier to cluster them into more meaningful groups. The researchers utilized k-means for their data clustering and their results showed that LSA was successful in clustering news articles into more meaningful groups as compared to traditional clustering approaches. In this part, we discussed the workflow of our proposed model with the implementation. This section includes the data collection and processing, the proposed model with the implementation processes and methods. Applying a modified method of K-means clustering which is MiniBatch K-means clustering in the dataset, has made the model more relevant and effective. Demonstrating the implementation process and outputs, this section concludes with a discussion of the modified K-means clustering approaches.

3.2 Workflow

Firstly, we start our proposed model and read the dataset as input. Then we process the dataset and implement the K-means method for feature extractions of the dataset and categorize them in different groups. Then we implement the LSA (Latent Semantic Analysis) for dimensionality reduction with both K-means and K-means MiniBatch K-means clustering methods. After the implementation of the modified methods, we evaluate our model and demonstrate the result and then finish the process.

Phase 1 – Dataset Processing: We accumulated the BBC News dataset from google and processed it. K-means clustering to news clustering based on latent semantic analysis requires extensive dataset processing.

Phase 2 – The data must first be pre-processed to remove noise and outliers, as well as normalized in order to ensure that all data points are of equal importance. After pre-processing, the data must be transformed into a matrix representation, which can then be clustered using k-means.

Phase 3 – Feature Extraction: It groups the data points into clusters based on similarities in the data. By doing this, it can allow us to extract key features from the data that we would not have otherwise been able to see.

Phase 4 – Model Accomplishment: LSA (Latent Semantic Analysis) is used to reduce the dimensionality of the documents, while still preserving their information content. It does this by constructing a matrix of the documents and then using a process known as singular value decomposition to reduce the number of dimensions.

Phase 5 – Performance Evaluation: In this part, the clustering result has been evaluated in a graphical representation. We demonstrate that by using different sets of data, k-means clustering can be improved. In addition, we show that latent semantic analysis produces better results when used in a supervised learning setting.

Phase 6 – Future work and Conclusion: The results showed that k-means clustering outperformed the latent semantic analysis in terms of accuracy. In terms of future work, we plan to investigate the effectiveness of using k-means clustering for news clustering based on latent semantic analysis. This could potentially improve the accuracy of the algorithms and would allow us to have better insights into the news.

This is the workflow representation of our proposed method.

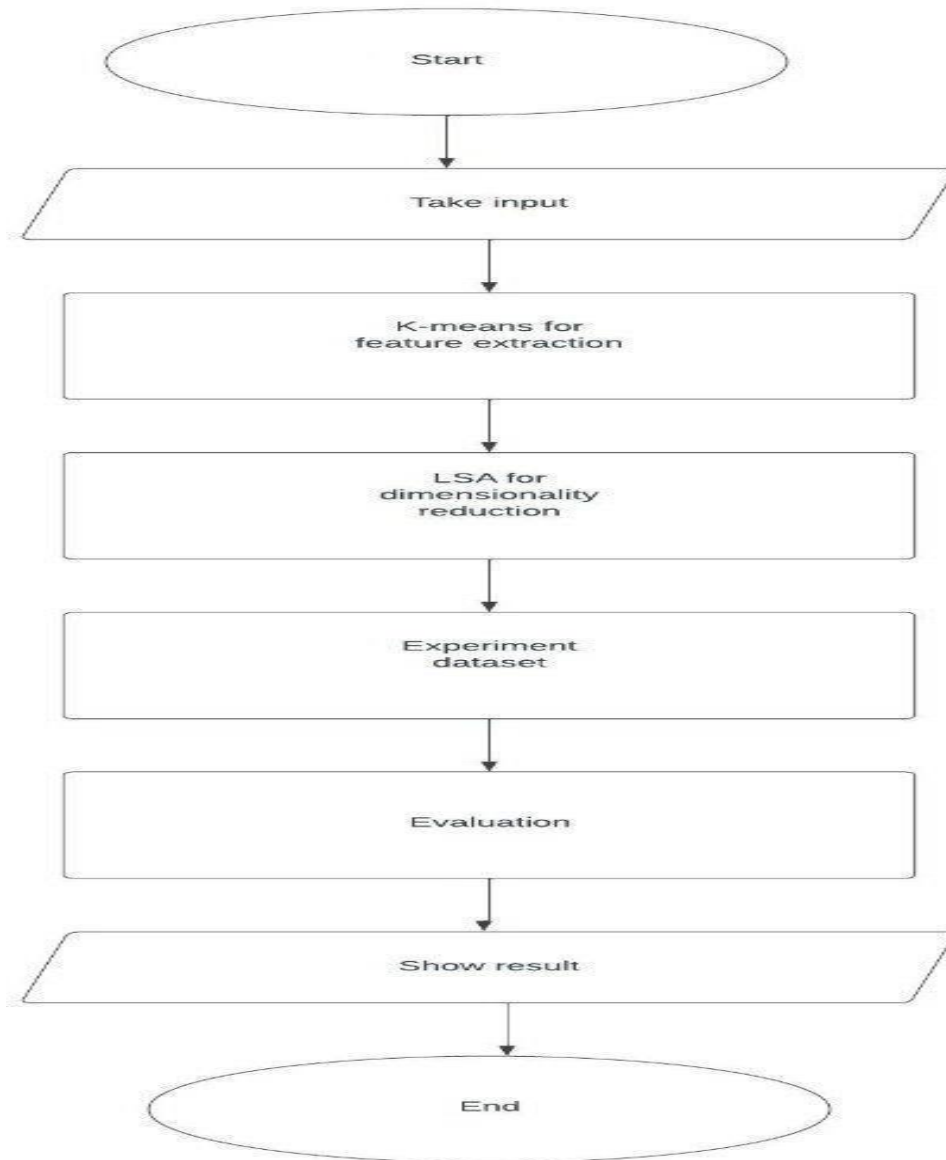
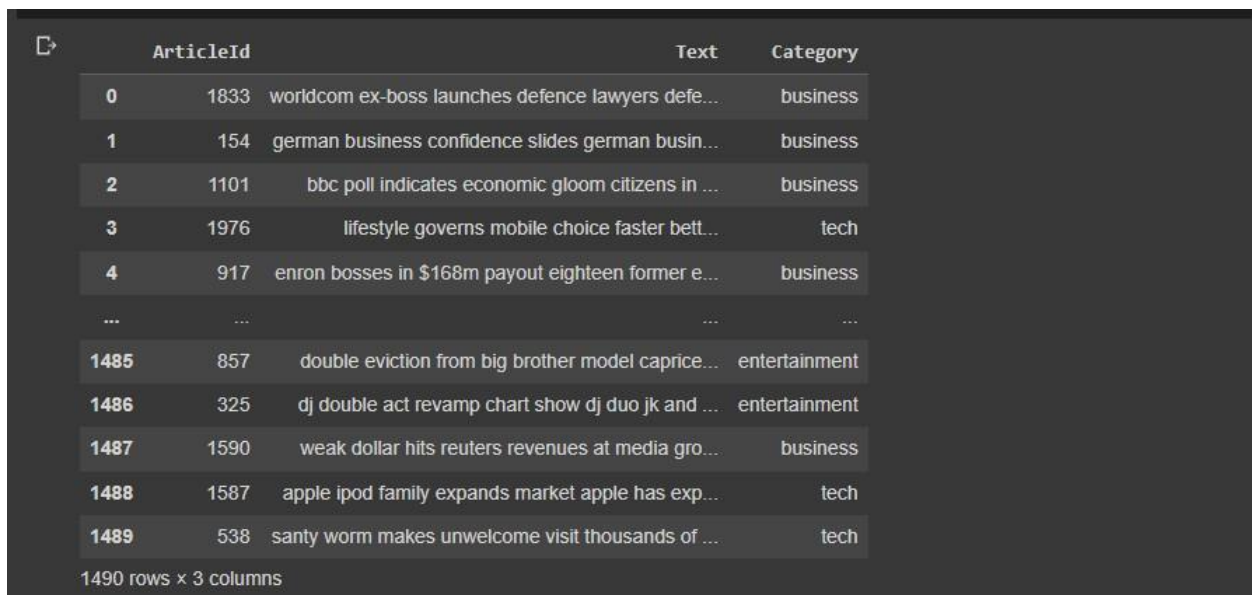


Fig 3.2.1: The Flowchart of the Workflow

3.3 Experiment Dataset

We import information from the 20 newsgroups text dataset, which includes over 18,000 postings across 20 subjects in newsgroups. To serve as an example and cut down on computational costs, we only choose a selection of 4 themes that together account for over 3,400 papers. see the illustration Text document classification with sparse characteristics to acquire insight into the overlap of such themes. Be aware that the text samples already include certain message metadata by default, including "headers," "footers" (signatures), and "quotes" from other posts. To remove those features and create a more logical clustering problem, we use the remove argument from fetch 20 newsgroups.



	ArticleId	Text	Category
0	1833	worldcom ex-boss launches defence lawyers defe...	business
1	154	german business confidence slides german busin...	business
2	1101	bbc poll indicates economic gloom citizens in ...	business
3	1976	lifestyle governs mobile choice faster bett...	tech
4	917	enron bosses in \$168m payout eighteen former e...	business
...
1485	857	double eviction from big brother model caprice...	entertainment
1486	325	dj double act revamp chart show dj duo jk and ...	entertainment
1487	1590	weak dollar hits reuters revenues at media gro...	business
1488	1587	apple ipod family expands market apple has exp...	tech
1489	538	santy worm makes unwelcome visit thousands of ...	tech

1490 rows x 3 columns

Fig 3.3.1: The Categorized Dataset

3.4 Proposed Methodology

We used the MiniBatch K-means approach, a K-means clustering method that has been adjusted to take less time and K-means for data stability. For small amounts of data, K-means clustering is employed, and for large amounts of data, MiniBatch K-means clustering. For feature extraction, we employed K-means clustering on text features. TfidfVectorizer and HashingVectorizer are two approaches used in this feature extraction. The dataset's dimensionality is reduced using the latent semantic analysis. Tf-idf and hashed vectors were combined with five different clustering techniques.

3.4.1 Evaluating the Effectiveness of Clustering Findings

This section provides a tool to compare different clustering algorithms using a range of metrics.

- homogeneity, which measures the proportion of clusters that solely include individuals from a single class
- completeness, which measures the proportion of a class's members who are placed in the same clusters.
- The harmonic mean of completeness and homogeneity is the V-measure.
- when two data points are regularly grouped together as a result of the clustering method and the ground truth class assignment, the Rand-Index is used to measure this frequency.
- A chance-adjusted Rand-Index is one that has an ARI of 0.0 in expectation for random cluster assignment.

3.4.2 K-means Clustering on Text Features

This paper will discuss the various benefits and challenges of using K-means clustering for text data. It will provide a comprehensive overview of the method, which involves using unsupervised machine learning to identify and categorize clusters of text features from large

datasets. Furthermore, this paper will go into detail regarding the advantages and disadvantages of K-means clustering, as well as the potential applications and pitfalls. The end goal of this paper is to come to a conclusion on the suitability of K-means clustering as a reliable means of text data analysis. K-means clustering is an unsupervised machine learning algorithm used to group similar data points together. It is often used for various types of text analysis, such as in text classification and semantic analysis. K-means clustering works by taking a given set of data points and assigning each of them to one of 'K' different clusters. Then, the algorithm calculates the mean of all the data points in each cluster, effectively creating a centroid for each cluster.

Two feature extraction methods are used here.

- **TfidfVectorizer:** computes a sparse word occurrence frequency matrix by mapping the most popular terms to feature indices using an in-memory vocabulary (a Python dict) The Inverse Document Frequency (IDF) vector, which was gathered feature-wise over the corpus, is then used to reweight the word frequencies.
- **HashingVectorizer:** word occurrences are hashed into a set-dimensional space with potential collisions. It appears crucial for k-means to operate in high dimensional space that the word count vectors be normalized such that each has a l2-norm of one (projected to the euclidean unit-sphere).

3.4.3 Utilizing TfidfVectorizer for Feature Extraction

In this paper, we present a novel approach to feature extraction by using the K-means algorithm. K-means is a clustering algorithm that partitions data points into K clusters. The clusters are determined by minimizing the sum of distances between the data points and their cluster centroids. This approach provides an efficient way to extract features from a dataset, which can then be used for further analysis or processing. In order to benchmark the estimators, we first use TfidfVectorizer's dictionary vectorizer and IDF normalization.

```
vectorization done in 0.449 s
n_samples: 3387, n_features: 7929
```

Fig 3.4.3.1: Utilizing TfidfVectorizer

3.5 K-means Clustering of Sparse Data

The clustering produced by K-Means and MiniBatch K-Means isn't always the best for a given random unit because they both maximize a non-convex objective function. Furthermore, when dealing with sparse high-dimensional data, such as text that has been vectorized using the Bag of Words technique, k-means can initialize centroids on very isolated data points. The centroids of those data points can continue to exist. Depending on the random initialization, the prior phenomenon might occasionally result in severely unbalanced clusters, as shown by the output below:

```
↳ Number of elements assigned to each cluster: [ 1  1 3384  1]
Number of elements assigned to each cluster: [1688 725 238 736]
Number of elements assigned to each cluster: [2004 446 646 291]
Number of elements assigned to each cluster: [1695 649 446 597]
Number of elements assigned to each cluster: [ 338 2155 417 477]

True number of documents in each category according to the class labels: [799 973 987 628]
```

Fig 3.5.1: K-means Clustering of Sparse Data

One solution to this problem is to increase the quantity of runs with independent random initiations (n_init). The grouping with the greatest inertia is chosen in this case (objective

function of k-means).

```
└─ clustering done in 0.32 ± 0.16 s
Homogeneity: 0.343 ± 0.029
Completeness: 0.404 ± 0.009
V-measure: 0.370 ± 0.018
Adjusted Rand-Index: 0.213 ± 0.012
Silhouette Coefficient: 0.008 ± 0.001
```

Fig 3.5.2: K-means Clustering with Highest Inertia

3.6 Performing Dimensionality Reduction using LSA

Latent Semantic Analysis (LSA) is a way of reducing the dimensionality of a data set by analyzing the relationships among its words. It is a technique for extracting and representing the meaning of a document using semantic meaning. This is done by using latent semantic indexing, which uses singular value decomposition to reduce the dimensionality of the data set. The resulting model is a lower-dimensional representation of the data set that can be used for different analysis tasks. This paper seeks to apply latent semantic analysis (LSA) to the task of dimensionality reduction and data mining. LSA is a technique that falls under the umbrella of natural language processing, and it is used to analyze the relationship between documents and the terms they contain. Our proposed approach will utilize LSA to reduce dimensions of the data, allowing for a more robust data mining process. Additionally, this method will allow for an easier understanding of the data by representing it in a more manageable form. A $n_{init}=1$ can still be used as long as the dimension of the vectorized space is initially reduced to make k-means more stable. To do this, we use the phrase count/tf-idf matrix-operating Truncated SVD. SVD findings are not normalized, thus we redo the normalization to enhance the K-Means result. In the literature on information retrieval and text mining, Latent Semantic Analysis (LSA), which uses SVD to reduce the dimensionality of TF-IDF document vectors, is frequently mentioned.

```
↳ LSA done in 0.523 s
   Explained variance of the SVD step: 18.4%
```

Fig 3.6.1: Performing Dimensionality Reduction using LSA Processing times for K-Means and MiniBatch K-Means will both be slashed by using a single initialization.

```
✘ clustering done in 0.14 ± 0.00 s
   Homogeneity: 0.398 ± 0.012
   Completeness: 0.432 ± 0.023
   V-measure: 0.414 ± 0.016
   Adjusted Rand-Index: 0.319 ± 0.015
   Silhouette Coefficient: 0.030 ± 0.001
```

Fig 3.6.2: Processing times for K-Means and MiniBatch K-Means

We can see that clustering on the LSA representation of the document is substantially faster because $n_{init}=1$ and the size of the LSA feature space is much lower. All of the metrics used to evaluate clustering have also changed for the better. We conduct the test once more using MiniBatch K-Means.

```
clustering done in 0.13 ± 0.06 s
   Homogeneity: 0.395 ± 0.011
   Completeness: 0.402 ± 0.008
   V-measure: 0.398 ± 0.008
   Adjusted Rand-Index: 0.360 ± 0.036
   Silhouette Coefficient: 0.028 ± 0.001
```

Fig 3.6.3: Conducting MiniBatch K-Means

3.7 Top Terms Per Cluster

The TfidfVectorizer's ability to be inverted allows us to locate the cluster centers, which give us a general idea of the words that contribute the most to each cluster. View the following script. Text document classification utilizing minimal features for comparison with the most prognosticative terms for each target class.

```
Cluster 0: thanks graphics image know file program files looking does software
Cluster 1: people don think just like say know time did does
Cluster 2: space launch orbit earth shuttle nasa moon just like mission
Cluster 3: god jesus believe bible faith say people does christian belief
```

Fig 3.7.1: Top Terms Per Cluster

3.8 HashingVectorizer

This paper will also discuss the use of a hash vectorizer, which is a machine-learning technique used to help process unstructured data. This enables the system to identify, classify and compare large sets of data. The hash vectorizer works by assigning numerical values to each word, allowing the data to be represented numerically, which can then be used to identify patterns and relationships between words. This approach is particularly powerful in natural language processing, where it can be used to identify relationships between words in a sentence. Given that this model is stateless, a different vectorization can be carried out using a HashingVectorizer object that does not use IDF weighting (the fit method does nothing). When necessary, IDF weighting can be added by pipelineing the HashingVectorizer output to a TfidfTransformer instance. In this case, we also include in LSA in the pipeline to reduce the size and sparsity of the hashed vector space. One can see that the LSA stage, especially with hashed vectors, requires a considerable amount of time to fit. The reason is that a hashed space (set to `n_features=50 000` in our case) is normally big. As seen in the example notebook `FeatureHasher and DictVectorizer Comparison`, one can try reducing the number of features at the cost of having a higher percentage of features with hash collisions.

On this hashed-lsa-reduced data, now fit & assess the K-means and minibatch K-means instances:


```
clustering done in 0.23 ± 0.15 s
Homogeneity: 0.400 ± 0.003
Completeness: 0.451 ± 0.005
V-measure: 0.424 ± 0.004
Adjusted Rand-Index: 0.320 ± 0.010
Silhouette Coefficient: 0.029 ± 0.001
```

Fig 3.8.1: HashingVectorizer

```
↳ clustering done in 0.08 ± 0.06 s
Homogeneity: 0.349 ± 0.059
Completeness: 0.364 ± 0.060
V-measure: 0.357 ± 0.059
Adjusted Rand-Index: 0.317 ± 0.072
Silhouette Coefficient: 0.025 ± 0.004
```

Fig 3.8.2: HashingVectorizer

Both approaches produce good outcomes that are comparable to those of using the same models with conventional LSA vectors (without hashing).

The K-Means algorithm is a classic clustering algorithm used for grouping data points into clusters based on their similarities. It is a popular choice for tasks of data exploration and analysis as it is relatively easy to implement and interpret. The algorithm works by first randomly selecting

K-number of points from the dataset to serve as the initial cluster centers. Then, the data points are assigned to the closest cluster center in terms of Euclidean distance. Latent semantic analysis (LSA) is an algorithm used for text analysis that uses the relationships between terms to produce a numerical representation of the text. It is often used in natural language processing (NLP) and computational linguistics, as well as information retrieval. We found that our model has acquired better results and accuracy by using this dataset and it can be helpful for other sectors of text clustering. We wish to implement our proposed model on the other datasets of news documents for being a perfect model for clustering news documents.

CHAPTER 4

EXPERIMENTAL RESULTS AND DISCUSSION

4.1 Evaluation Summary

For large dimensional datasets like text data, K-Means and MiniBatch K-Means experience the "Curse of Dimensionality" a phenomenon. This explains why employing LSA results in higher total scores. Although the LSA phase itself takes a while, especially with hashed vectors, using LSA reduced data increases stability and necessitates less clustering time.

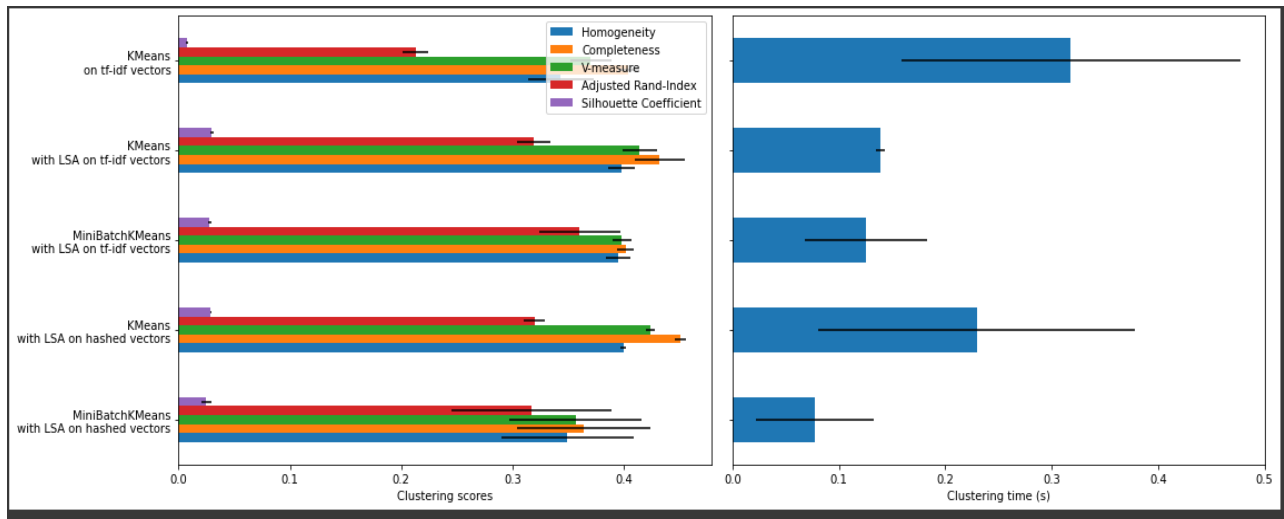


Fig 4.1.1: Clustering Evaluation Graph

4.2 Result Discussion

The range of the silhouette coefficient is 0 to 1. Its definition requires measuring distances, unlike other assessment measures like the V-measure and the Adjusted Rand Index which are

merely dependent on cluster assignments instead of distances, hence in every case we receive values near to 0 (even if they increase a little after employing LSA). Notably, due to the various ideas of distance they entail, The Silhouette Coefficient should not be directly compared between spaces with differing dimensions.

With regard to completely random labeling, the homogeneity, completeness, and consequently v-measure metrics do not produce a baseline; hence the outcomes will vary depending on the number of samples, clusters, and ground truth classes. Particularly when there are several clusters, random labeling won't result in zero scores. When there are more than a thousand samples and fewer than ten clusters, as there are in the current example, this issue can be safely disregarded. Using an adjusted index, like the Adjusted Rand Index, is safer for smaller sample sizes or more clusters (ARI). For a demonstration of the impact of random labeling, view the performance evaluation of clustering example of adjusting for chance. For this relatively tiny dataset, the magnitude of the error bars demonstrates that MiniBatch K-Means is less reliable than K-Means. Although compared to the traditional k-means approach, it can result in a little drop in clustering quality, it is more interesting to use when the number of samples is much higher.

CHAPTER 5

CONCLUSION AND FUTURE WORK

5.1 Conclusion

This exceptional work is based on submitting a modified version of the proposed model to provide the news text clustering approaches in a different way. We used the news text dataset which is collected from the BBC news classification dataset which is divided into five categories such as sports, entertainment, business, politics, and technology. The dataset has both training and testing records to get a better experimental result in unseen news articles clustering to classify accurately into the right category. We found that our model has acquired better results and accuracy by using this dataset and it can be helpful for other sectors of text clustering. The proposed deep-learning approaches generate clusters by using the clustering algorithm based on neural networks which are more relevant. In this paper, we propose a novel approach to news clustering based on the K-Means Clustering algorithm and Latent Semantic Analysis. We propose that the combination of K-Means Clustering and Latent Semantic Analysis can produce more accurate and meaningful clusters of news articles than traditional clustering algorithms. We evaluate our proposed approach using a real-world dataset and demonstrate superior performance against existing clustering techniques. In conclusion, our proposed method of clustering news articles is a viable and effective approach to obtain meaningful information from large volumes of news articles.

5.2 Future Works

There are abundant chances for development in the news clustering sectors by using our proposed research model. We are conscious of our proposed model's enhancement and to work on it in the upcoming days. We wish to implement our proposed model on the other datasets

of news documents for being a perfect model for clustering news documents. Because if we want to make it a robust deep-learning model then we have to work with a larger dataset of news documents. The model is basically designed in a supervised learning algorithm. Developing an unsupervised learning algorithm can be a further solution to this certain problem and achieve a robust model. Traditional approaches in automated news clustering including K-Means Clustering and Latent Semantic Analysis (LSA) have demonstrated promise in providing some degree of topical classification of news articles. However, it is often difficult to obtain satisfactory results due to the presence of a high degree of variance in topics among different news sites. As a result, K-Means Clustering and LSA are often found to be inadequate for automated news clustering purposes. To address this issue, there is potential for future research to investigate alternative methods that may be more suitable for automated news clustering tasks.

REFERENCES

- [1] Tong, Yuqiang, and Lize Gu. "A news text clustering method based on similarity of text labels." *International Conference on Advanced Hybrid Information Processing*. Springer, Cham, 2019.
- [2] Fan, Zhaoxin, et al. "A text clustering approach of Chinese news based on neural network language model." *International Journal of Parallel Programming* 44.1 (2016): 198-206.
- [3] Blokh, Ilya, and Vassil Alexandrov. "News clustering based on similarity analysis." *Procedia computer science* 122 (2017): 715-719.
- [4] Sangaiah, Arun Kumar, et al. "Arabic text clustering using improved clustering algorithms with dimensionality reduction." *Cluster Computing* 22.2 (2019): 4535-4549.
- [5] Bouras, Christos, and Vassilis Tsogkas. "A clustering technique for news articles using WordNet." *Knowledge-Based Systems* 36 (2012): 115-128.
- [6] Bouras, Christos, and Vassilis Tsogkas. "Improving news articles recommendations via user clustering." *International Journal of Machine Learning and Cybernetics* 8.1 (2017): 223-237.
- [7] Qian, Mingjie, and Chengxiang Zhai. "Unsupervised feature selection for multi-view clustering on text-image web news data." *Proceedings of the 23rd ACM international conference on conference on information and knowledge management*. 2014.
- [8] Sawaf, Hassan, Jörg Zaplo, and Hermann Ney. "Statistical classification methods for Arabic news articles." *Natural Language Processing in ACL2001, Toulouse, France* (2001).
- [9] Dai, Xiangying, Yancheng He, and Yunlian Sun. "A Two-layer text clustering approach for retrospective news event detection." *2010 International Conference on Artificial Intelligence and Computational Intelligence*. Vol. 1. IEEE, 2010.

- [10] Wu, Xiao, Chong-Wah Ngo, and Alexander G. Hauptmann. "Multimodal news story clustering with pairwise visual near-duplicate constraint." *IEEE Transactions on Multimedia* 10.2 (2008): 188-199.
- [11] Marutho, Dhendra, Sunarna Hendra Handaka, and Ekaprana Wijaya. "The determination of cluster number at k-mean using elbow method and purity evaluation on headline news." *2018 international seminar on application for technology of information and communication*. IEEE, 2018.
- [12] <https://www.kaggle.com/competitions/learn-ai-bbc/overview/description>
- [13] Kaur, Sukhpal, and Er Mamoon Rashid. "Web news mining using Back Propagation Neural Network and clustering using K-Means algorithm in big data." *Indian Journal of Science and Technology* 9.41 (2016): 1-8.
- [14] Rahman, Zahid, et al. "Urdu News Clustering Using K-Mean Algorithm On The Basis Of Jaccard Coefficient And Dice Coefficient Similarity." *ADCAIJ: Advances in Distributed Computing and Artificial Intelligence Journal* 10.4 (2021): 381-399.
- [15] Hamzah, Amir, et al. "Concept-based text document clustering." *Proceedings of International Conference on Electrical Engineering and Informatics, Indonesia June*. 2007.
- [16] Al-Azzawy, Dhyaa Shaheed, and Faiez Musa Lahmood Al-Rufaye. "Arabic words clustering by using K-means algorithm." *2017 Annual Conference on New Trends in Information & Communications Technology Applications (NTICT)*. IEEE, 2017.
- [17] Bhole, Pankaj, and A. J. Agrawal. "Single Document Text Summarization Using Clustering Approach Implementing for News Article." *International Journal of Engineering Trends and Technology (IJETT)* 15.7 (2014): 364-368.
- [18] <https://medium.com/@krause60/news-article-clustering-using-unsupervised-learning-7647600a04fd> Marutho, Dhendra, Sunarna Hendra Handaka, and Ekaprana Wijaya. "The determination of cluster number at k-mean using elbow method and purity evaluation on headline

news." *2018 international seminar on application for technology of information and communication*. IEEE, 2018.

[19] Kandhro, Irfan Ali, et al. "Roman Urdu headline news text classification using RNN, LSTM and CNN." *Advances in Data Science and Adaptive Analysis* 12.02 (2020): 2050008.

[20] https://scikit-learn.org/stable/auto_examples/text/plot_document_clustering.html

[21] Duygulu, Pinar, Jia-Yu Pan, and David A. Forsyth. "Towards auto-documentary: Tracking the evolution of news stories." *Proceedings of the 12th annual ACM international conference on Multimedia*. 2004.

[22] Yongyi, L. I., and Gao Yin. "News Events Clustering Method Based on Staging Incremental Single-Pass.

[23] Bashir, E., and M. Luštrek. "Self Learning of News Category Using AI Techniques." *Intelligent Environments 2021: Workshop Proceedings of the 17th International Conference on Intelligent Environments*. Vol. 29. IOS Press, 2021.

[24] Wu, Xiao, Chong-Wah Ngo, and Alexander G. Hauptmann. "Multimodal news story clustering with pairwise visual near-duplicate constraint." *IEEE Transactions on Multimedia* 10.2 (2008): 188-199.

[25] Azzopardi, Joel, and Christopher Staff. "Incremental clustering of news reports." *Algorithms* 5.3 (2012): 364-378.

[26] Miranda, Sebastiao, et al. "Multilingual clustering of streaming news." *arXiv preprint arXiv:1809.00540* (2018).

[27] Siegler, Matthew A., et al. "Automatic segmentation, classification and clustering of broadcast news audio." *Proc. DARPA speech recognition workshop*. Vol. 1997. 1997.

K-MEANS CLUSTERING FOR NEWS CLUSTERING BASED ON LATENT SEMANTIC ANALYSIS

ORIGINALITY REPORT

19%

SIMILARITY INDEX

15%

INTERNET SOURCES

13%

PUBLICATIONS

9%

STUDENT PAPERS

PRIMARY SOURCES

1

dspace.daffodilvarsity.edu.bd:8080

Internet Source

3%

2

assets.researchsquare.com

Internet Source

1%

3

"Evolutionary Data Clustering: Algorithms and Applications", Springer Science and Business Media LLC, 2021

Publication

1%

4

Submitted to Daffodil International University

Student Paper

1%

5

Shengsheng Qian, Jinguang Wang, Jun Hu, Quan Fang, Changsheng Xu. "Hierarchical Multi-modal Contextual Attention Network for Fake News Detection", Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2021

Publication

1%

6

Submitted to Coventry University

Student Paper

1%