ANKARA SCIENCE UNIVERSITY, RESEARCHER
Vol. 4, No. 2, Decmber 2024
e-ISSN: 2717-9494
Research Article/Araştırma Makalesi

# An Investigation on the Execution of the Document Clustering Process on Internet News

**Metin Oktay BOZ[1], Jale BEKTAŞ[2]**
[1]Mersin üniversitesi; 0009-0006-8620-7775
[2] Mersin üniversitesi; 0000-0002-8793-1486
* Corresponding Author: 22023010006@mersin.edu.tr

Received: 27 June 2024 ; Accepted: 18 June 2024

**Reference/Atıf:** M. O. Boz and J. Bektaş, "An Investigation on the Execution of the Document Clustering Process on Internet News", Researcher, vol. 04, no. 02, pp. 113–119, 2024.

## Abstract

Numerous investigations have focused on recognizing Internet news as valid documents. This study encompasses the application of text mining techniques to generate a TF-IDF matrix and the subsequent automatic identification and categorization of an optimal number of clusters. The research examines the impact of K-Means document clustering on internet news articles, integrating the User Engagement dataset which includes articles from various esteemed publishers. Prior to implementing the K-Means algorithm, several preprocessing steps were undertaken to prepare the TF-IDF matrix. Due to the absence of the content attribute data, the description attribute was selected for document clustering. During preprocessing, extraneous ASCII symbols, punctuation marks, line breaks, emails, mentions, internet extensions, stopwords, and words outside the 2 to 21 character range were removed. Words were stemmed to consolidate different forms of the same root. The Elbow method was employed on the TF-IDF matrix to determine the optimal number of clusters, followed by an analysis of results using prominent words and word clouds. Ultimately, five clusters of document counts 797, 408, 89, 364, and 8755 were identified.

**Keywords:** K-Means, TF-IDF, Clustering, Document Clustering

## 1. Introduction

In this study, we embark on a journey through the intricate domain of document clustering, with a particular focus on the "Internet news data with readers engagement" dataset. Our mission unfolds in several pivotal stages, each contributing to the overarching goal of unveiling insightful cluster analysis. It all begins with meticulous data preprocessing, where we meticulously cleanse the dataset of empty rows and parse texts into individual words, setting the stage for effective clustering. Leveraging the TF-IDF (Term Frequency-Inverse Document Frequency) method, we extract features to gauge the significance of words within documents, laying a robust foundation for subsequent analysis. The heart of our study lies in the application of the K-Means algorithm, a cornerstone of unsupervised machine learning, which diligently partitions the dataset into K clusters based on similarity. Through iterative refinement and the judicious use of the Elbow Method to determine the optimal number of clusters, we unveil clusters ripe for analysis. Visualizations further enhance our understanding by showcasing the most prominent terms within each cluster, revealing distinct thematic threads lurking within the corpus, ranging from geopolitics to journalism. This systematic approach not only refines the dataset but also unveils nuanced insights, illuminating the latent patterns and narratives embedded within the textual data.

## 2. Data Preprocessing

Before processing the dataset, it must be prepared depending on the algorithm or operations to be applied. To perform preprocessing, understanding the structure of the dataset is essential. Within the dataset, there are two headers suitable for clustering: 'description' and 'content'. While data from the 'content' header is preferred for its representation of the articles' content, the 'description' header, containing summaries of the news, is also utilized due to the unavailability of complete content in the dataset. Data in this header was cleaned of empty rows before processing, resulting in a dataset

comprising 10,413 rows. Following the removal of empty spaces, text within the dataset was segmented into individual words. Furthermore, stop words in English, words with less than two characters, ASCII symbols, commas, special characters, emails and mentions, and internet extensions were cleansed using appropriate functions. Additionally, for more meaningful clustering results, words within the dataset were organized to retain only their roots.

```
Latest figures suggest Governmet is on course to hit its tax and spending targets
latest figur suggest governmet cours hit tax spend target
```

Figure 1: Comparison Before and After Data Preprocessing

## 3. Feature Extraction with TF-IDF

At this stage, determining the importance of a word for a document is crucial for utilizing the preprocessed dataset in the K-Means algorithm. TF-IDF (Term Frequency-Inverse Document Frequency) is a numerical statistic aimed at reflecting the importance of a term for the document it is deemed relevant to. This method operates by increasing the weight of a term when it appears multiple times in a document, while decreasing the weight when it is common across many documents. This weight typically ranges between 0 and 1, signifying the term's significance within the context of the document.

| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | ... | 10403 | 10404 | 10405 | 10406 | 10407 | 10408 | 10409 | 10410 | 10411 | 10412 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| tesla | 0.398867 | 0.0 | 0.0 | 0.000000 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | ... | 0.0 | 0.000000 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.00000 |
| board | 0.365565 | 0.0 | 0.0 | 0.000000 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | ... | 0.0 | 0.000000 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.00000 |
| culver | 0.261741 | 0.0 | 0.0 | 0.000000 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | ... | 0.0 | 0.000000 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.00000 |
| autopilot | 0.245217 | 0.0 | 0.0 | 0.000000 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | ... | 0.0 | 0.000000 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.00000 |
| mode | 0.224751 | 0.0 | 0.0 | 0.000000 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | ... | 0.0 | 0.000000 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.00000 |
| struck | 0.220808 | 0.0 | 0.0 | 0.000000 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | ... | 0.0 | 0.000000 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.00000 |
| truck | 0.206473 | 0.0 | 0.0 | 0.000000 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | ... | 0.0 | 0.000000 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.00000 |
| assist | 0.198552 | 0.0 | 0.0 | 0.000000 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | ... | 0.0 | 0.000000 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.00000 |
| transport | 0.194501 | 0.0 | 0.0 | 0.000000 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | ... | 0.0 | 0.000000 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.00000 |
| model | 0.191618 | 0.0 | 0.0 | 0.000000 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | ... | 0.0 | 0.000000 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.00000 |
| safeti | 0.187761 | 0.0 | 0.0 | 0.000000 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | ... | 0.0 | 0.000000 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.00000 |
| driver | 0.186580 | 0.0 | 0.0 | 0.000000 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | ... | 0.0 | 0.000000 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.00000 |
| crash | 0.174416 | 0.0 | 0.0 | 0.000000 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | ... | 0.0 | 0.000000 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.00000 |
| involv | 0.170811 | 0.0 | 0.0 | 0.000000 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | ... | 0.0 | 0.000000 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.00000 |
| system | 0.170473 | 0.0 | 0.0 | 0.000000 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | ... | 0.0 | 0.000000 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.00000 |
| seri | 0.163884 | 0.0 | 0.0 | 0.264569 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | ... | 0.0 | 0.000000 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.00000 |
| california | 0.157189 | 0.0 | 0.0 | 0.000000 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | ... | 0.0 | 0.000000 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.00000 |
| fire | 0.149508 | 0.0 | 0.0 | 0.000000 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | ... | 0.0 | 0.000000 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.00000 |
| citi | 0.142234 | 0.0 | 0.0 | 0.000000 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | ... | 0.0 | 0.123276 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.00000 |
| investig | 0.136722 | 0.0 | 0.0 | 0.000000 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | ... | 0.0 | 0.000000 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.15508 |

Figure 2: The Top 15 Terms with Highest Weight

## 4. Application of K-Means Algorithm and Cluster Analysis

K-Means stands as one of the simplest and most popular machine learning algorithms to date. Operating without the utilization of labeled data, it epitomizes an unsupervised algorithm, signifying that in this context, no single text belongs to a specific class or group. It functions as a clustering algorithm that categorizes a dataset into K clusters. The underlying principle of this algorithm revolves around defining the clusters by K centroids. Each centroid represents a center of a cluster. The algorithm iteratively operates by initially randomly placing each centroid into the dataset's vector space and then shifting them towards points closer to themselves. With each iteration, the distances between each centroid and

points are recalculated, and centroids are relocated to the centers of the nearest points. The algorithm concludes when either the positions or groups no longer change, or when the distance by which centroids change falls below a predefined threshold. Given the ambiguity regarding the ideal number of clusters initially, the K-Means algorithm was executed with cluster numbers ranging from 1 to 10, and the results were scrutinized to determine the optimal K value. The Elbow Method was employed for this determination, wherein the sum of squared distances of points to the cluster centroids is calculated for each K value, and a graphical representation is generated. The inflection point on the graph, where the rate of decrease in the sums begins to diminish, denotes the most suitable K value.

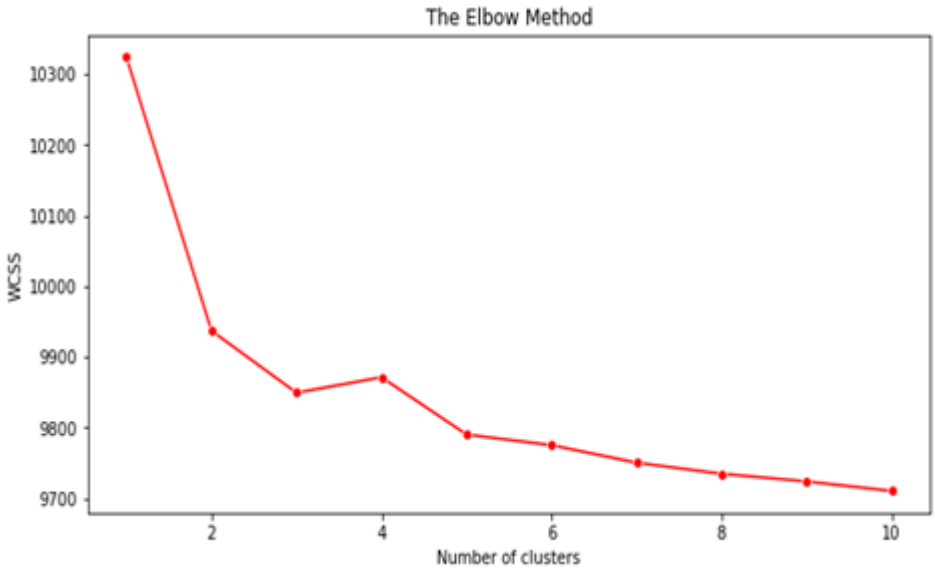


Figure 3: The Graphs Drawn for Each K Value

Based on the trend observed in the graph, where the difference starts diminishing from the point where K equals 5, this value is chosen as the optimal one.

## 5. Results and Evaluation

Following the execution of the K-Means algorithm with a K value of 5, the most popular terms in each cluster were visualized, and word clouds were generated.

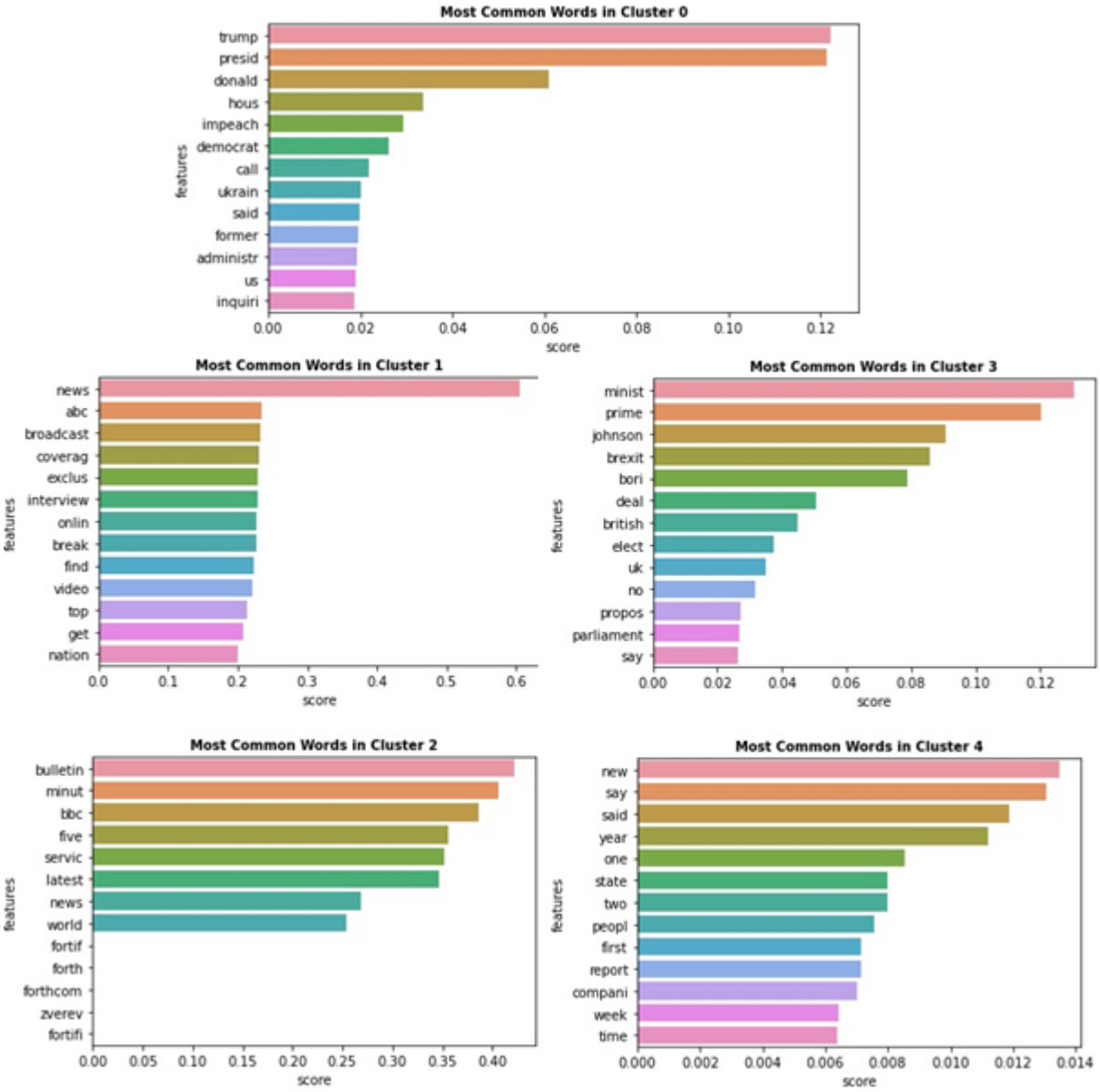


Figure 4: The Most Popular Terms for Each Cluster

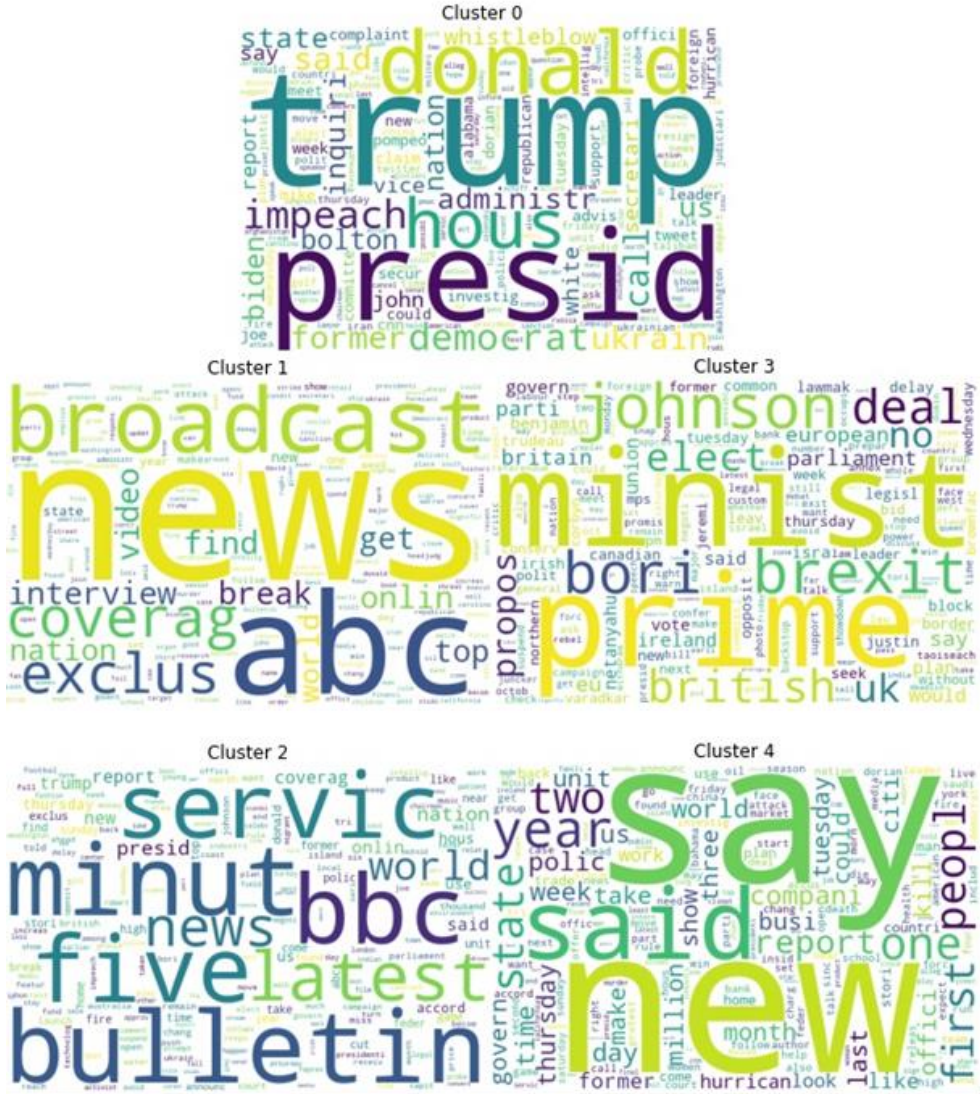


Figure 5: F Word Clouds Created for Each Cluster

Upon examining the clusters generated by the K-Means algorithm, distinct thematic threads and patterns emerged, shedding light on the underlying structures within the dataset. Notably, the first cluster prominently featured terms related to the United States, suggesting a concentration of news articles pertaining to American affairs. In contrast, the second cluster showcased terms associated with publishing, indicative of articles discussing the journalism industry and its practices. The third cluster revealed a focus on journalism-related terms, suggesting a thematic emphasis on the profession itself. Similarly, the fourth cluster exhibited a concentration of terms related to the United Kingdom, indicating a distinct regional focus within the dataset. Finally, the fifth cluster presented a diverse array of fundamental terms, reflecting a broad spectrum of topics encompassing various aspects of news and information dissemination.

Furthermore, the respective sizes of each cluster provided valuable insights into the distribution of articles across different thematic categories. The largest cluster, comprising 8755 documents, represented a comprehensive collection of articles covering diverse subjects. In contrast, smaller clusters, such as the third and fourth clusters with 89 and 364 documents respectively, highlighted more specialized topics within the dataset. These findings underscore the effectiveness of the K-Means algorithm in discerning meaningful patterns and clustering documents based on their thematic content.

Through this systematic approach, our study has successfully elucidated the inherent structure of the dataset, offering valuable insights into the composition and distribution of internet news articles.

## 6. General Evaluation and Conclusions

To facilitate the application of the K-Means algorithm and achieve cleaner results, preprocessing was conducted on the dataset. Once the dataset was refined to a state conducive for obtaining accurate outcomes, feature extraction was performed using the TF-IDF method. This methodology enabled the determination of the weight associated with each term for every document, providing valuable insights into the dataset's structure. Subsequently, the K-Means algorithm was applied with cluster numbers ranging from 1 to 10, and the optimal K value was determined using the Elbow Method, resulting in a choice of 5 clusters. Within these 5 clusters, the most popular terms were identified for each, and word clouds were generated accordingly. Upon examining the clusters, distinct thematic threads emerged, with the prominence of terms related to the United States in the first cluster, publishing in the second, journalism in the third, and the United Kingdom in the fourth. The fifth cluster exhibited a collection of fundamental terms. The respective term counts for the clusters were as follows: 797, 408, 89, 364, and 8755. Throughout the implementation process, insights from the article "Text Clustering with K-Means" were taken into consideration.

## Contribution of Researchers

Metin Oktay BOZ was responsible for the collection of a large dataset comprising internet news articles and the execution of various preprocessing steps. This involved the removal of extraneous characters, handling stop-words, and filtering out irrelevant information, which enhanced the quality of the dataset. Also performed the feature extraction and vectorization of the news texts. The TF-IDF (Term Frequency-Inverse Document Frequency) method was used to transform the news articles into vectorized representations, with each article represented by a vector based on the importance levels of its contained words. Furthermore, he applied the K-Means algorithm, a clustering technique that partitions the dataset into a predetermined number of clusters, using the vectorized representations of the news articles. This facilitated the grouping of news articles based on similar themes. Finally, he conducted the cluster analysis and interpreted the findings. This analysis revealed how news articles were grouped around specific topics and explored the contribution of clustering results to understanding and analyzing news content.

Asst.Prof.Dr. Jale BEKTAŞ provided invaluable guidance and support throughout the study, contributing to the design of the study, interpretation of results, and review of the manuscript.

Throughout all these stages, Metin Oktay BOZ and Asst.Prof.Dr. Jale BEKTAŞ collaborated to ensure the quality and accuracy of the research.

## Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this research. The collection, analysis, and interpretation of data, as well as the writing of the manuscript, were conducted in an objective and unbiased manner. No financial or personal relationships with any individuals or organizations that could potentially bias the findings of this study exist. Furthermore, there are no competing interests related to employment, consultancy, patents, products in development, or marketed products that could influence the research process or outcomes. This research was conducted solely for academic and scientific purposes, with the aim of contributing to the field of document clustering and analysis.

## References

[1] Adolfsson, A., Ackerman, M., & Brownstein, N. C. (2019). To cluster, or not to cluster: An analysis of clusterability methods. Pattern Recognition, 88, 13-26.

[2] Al-Anazi, S., AlMahmoud, H., & Al-Turaiki, I. (2016). Finding Similar Documents Using Different Clustering Techniques. Procedia Computer Science, 82, 28-34.

[3] Bezdan, T., Stoean, C., Naamany, A. A., Bacanin, N., Rashid, T. A., Zivkovic, M., & Venkatachalam, K. (2021). Hybrid Fruit-Fly Optimization Algorithm with K-Means for Text Document Clustering. Mathematics, 9(16), 1929.

[4] Capó, M., Pérez, A., & Lozano, J. A. (2020). An efficient K-means clustering algorithm for tall data. Data Mining and Knowledge Discovery, 34, 776–811.

[5] Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A K-Means Clustering Algorithm. Journal of the Royal Statistical Society. Series C (Applied Statistics), 28(1), 100-108.

[6] Hu, G., Zhou, S., Guan, J., & Hu, X. (2008). Towards effective document clustering: A constrained K-means based approach. Information Processing & Management, 44(4), 1397-1409.

[7] Janowski, S. (2020). Internet News Data with Readers Engagement. https://www.kaggle.com/datasets/szymonjanowski/internet-articles-data-with-users-engagement. [Accessed: 19-Dec-2023].

[8] Jarvis, R. A., & Patrick, E. A. (1973). Clustering Using a Similarity Measure Based on Shared Near Neighbors. IEEE Transactions on Computers, C-22(11), 1025-1034.

[9] Liang, M., & Niu, T. (2022). Research on Text Classification Techniques Based on Improved TF-IDF Algorithm and LSTM Inputs. Procedia Computer Science, 208, 460-470.

[10] Luo, C., Li, Y., & Chung, S. M. (2009). Text document clustering based on neighbors. Data & Knowledge Engineering, 68(11), 1271-1288.

[11] Mahdavi, M., & Abolhassani, H. (2009). Harmony K-means algorithm for document clustering. Data Mining and Knowledge Discovery, 18, 370–391.

[12] Minaee, S., Gao, J., Kalchbrenner, N., Cambria, E., Nikzad, N., & Chenaghlu, M. (2021). Deep Learning--based Text Classification: A Comprehensive Review. ACM Computing Surveys, 54(3), Article 62, 1–40.

[13] Mussabayev, R., Mladenovic, N., Jarboui, B., & Mussabayev, R. (2023). How to Use K-means for Big Data Clustering? Pattern Recognition, 137, 109269.

[14] Ridzuan, F., & Zainon, W. M. N. (2019). A Review on Data Cleansing Methods for Big Data. Procedia Computer Science, 161, 731-738.

[15] Sa, L. (2019). Text Clustering with K-Means. https://medium.com/@lucasdesa/text-clustering-with-k-means-a039d84a941b. [Accessed: 05-Jan-2024].

[16] Saha, B., & Srivastava, D. (2014). Data quality: The other face of Big Data. In 2014 IEEE 30th International Conference on Data Engineering (pp. 31 March 2014 - 04 April 2014). IEEE. https://doi.org/10.1109/ICDE.2014.6816764

[17] Thangaraj, M., & Sivakami, M. (2018). Text Classification Techniques: A Literature Review. Interdisciplinary Journal of Information, Knowledge, and Management, 13, 117-135.