Text Classification of National Anthem using Agglomerative Hierarchical Clustering

Prajwal Rai [1], Nirdosh Bista [2], Kumar Prasun [3], Gajendra Sharma[4],

developer.prajwal007@gmail.com [1], bistanirdosh007@gmail.com [2] , erkprasun@gmail.com [3],gajendrasharma@kcc.edu.np [4]
Kantipur City College [1] [4], Nepal Engineering College [2], Padmakanya Campus. Kathmandu, 44600, Kathmandu, Nepal

ABSTRACT

Text clustering allows users to categorize different documents based on their similarities. Over the course of several years, this research topic has attracted significant attention from scholars, resulting in the emergence of many approaches and procedures. Nevertheless, the study primarily focuses on English and other languages that have ample resources. This paper presents a comprehensive assessment of clustering methods in the context of national anthems across 190 countries worldwide. The task of conceptually categorizing Anthem is difficult because of its restricted duration. The present study involved the extraction of various features from the anthem, such as stop-words, stemming, corpus tokenization, noise removal, and TF-IDF features. The Agglomerative Hierarchical Clustering technique is utilized for the clustering process. The results indicate that the utilization of a clustering technique in combination with an Agglomerative Hierarchical Clustering algorithm, which incorporates TF-IDF properties, is highly beneficial.
Keywords: Text Clustering, national anthem. Agglomerative Hierarchical Clustering

1. INTRODUCTION

A) BACKGROUND

The results of this text clustering study provide valuable insights into the procedure of categorizing similar publications according to their shared characteristics. Unlike previous research that mostly examines English and languages with ample resources, this study adopts a unique approach by assessing clustering algorithms specifically for the national anthems of 190 countries worldwide. The national anthem functions as a powerful emblem of national identity, reflecting historical records, fundamental values, and shared memories of the citizens of the country. Anthem writings are commonly characterized by their concise nature, posing difficulties in attaining a comprehensive classification of their content. One more element that contributes to the complexity of the classification procedure is the extensive range of literary forms and languages.

Text clustering, a key component of natural language processing (NLP), has drawn scholars who aim to group text by commonalities. Many clustering methods have been developed and studied in academia. However, English and other languages with rich linguistic resources have received much attention. There is little research on text clustering to national anthems in 190 nations. National anthems are usually limited to a set number of stanzas or verses, making clustering problematic. Despite these challenges, national anthems can reveal cultural, historical, and social aspects of distinct nations.

To remedy the research gap, this paper evaluates clustering algorithms expressly used in national anthems. It involves extracting various song aspects by eliminating stop-words, stemming, tokenizing the corpus, and removing noise. Quantitatively representing the anthems using TF-IDF (Term Frequency-Inverse Document Frequency) characteristics captures the meaning of words in the corpus. Agglomerative Hierarchical Clustering (AHC) was employed for this investigation since it can handle TF-IDF features and capture the anthem data's structure. This research uses AHC on TF-IDF representations to conceptually classify national anthems across countries. The study shows that grouping national anthems with TF-IDF traits is beneficial.

B) STATEMENT OF PROBLEM

The challenge of effectively clustering and categorizing national anthems, which serve as potent symbols of cultural identity and historical narratives for more than 190 nations globally, arises from their limited length and the diverse range of languages and literary forms they encompass. The current body of literature in the field of text clustering primarily concentrates on English and languages that possess ample resources, so overlooking the distinctive attributes of national anthems. Hence, there exists a knowledge deficit about the effectiveness of clustering algorithms that are specifically designed for national anthems. The objective of this study is to fill the existing research gap by assessing clustering algorithms, specifically focusing on the Agglomerative Hierarchical Clustering algorithm, within the domain of national anthems. This evaluation will incorporate many elements like stop-words, stemming, corpus tokenization, noise removal, and TF-IDF characteristics.

C) SCOPE OF STUDY

- The primary objective of this study is to conduct an analysis of national anthems with the purpose of gaining a deeper understanding of shared historical narratives, cultural values, and national identities. This will be achieved through the identification of common themes, attitudes, and language patterns.
- Evaluates the effectiveness and limitations of the Agglomerative Hierarchical Clustering algorithm in grouping anthems.
- Potential avenues for future research encompass the investigation of alternative clustering algorithms, the enhancement of text preprocessing techniques, and the examination of ethical implications.

D) OBJECTIVES

- This study aims to fill the current research void by examining text clustering approaches that are specifically designed for national anthems. These anthems hold significant cultural and historical value as powerful emblems of identity.
- To identify common themes, sentiments, and linguistic patterns across national anthems from diverse linguistic and cultural backgrounds, shedding light on shared historical narratives, cultural values, and national identities.

2. LITERATURE REVIEW

In the paper Text Clustering using K-Mean , Document clustering is used which allows users group comparable documents. Various methodologies and procedures have been created research issue for many years. The study primarily examines English and high-resource languages. And provides an experimental estimate of clustering approaches for Pakistani national anthems. Due to its brief length, thematically clustering Anthem is challenging. The paper extracted stop-words, stemming, corpus tokenization, noise removal, and TF-IDF features from the anthem, then clustered them using the K-Means algorithm. Results indicate a used clustering technique using a K-mean algorithm and TF-IDF features[1].

In Trends in the texts of national anthems: A comparative study identifies previous research on national anthems identified preferred topics and biases like identification, battling, well-being subjectively. If an objective, automated comparison of national anthem texts worldwide can uncover systematic tendencies and their magnitude. Tropes and Semantria software packages are used to analyze chosen subjects like state, feeling, body, time, land, religion, family, combat. Latin anthems emphasize "liberty" more than Asian ones, Germanic anthems mention "feelings" less, and African anthems lack the first-person singular "I". Anthem sentiment scores range from neutral to highly positive[2].

The Nigeria's National Anthem: A text linguistic Linguists study examines the anthem's language arrangement to improve comprehension. The study uses the Integrated Theory of Text Linguistics (ITTL) to understand the anthem's nationhood, concise language, and expressive qualities. The research focuses on how each linguistic feature enhances understanding, enabling language use in everyday life and solving social concerns. The study highlights the anthem's organization and objectives within a complex communication framework, highlighting its expressive and emotional nature[3].

The study of Comparing automated text classification method, this study compares ten methodologies for automated categorization of unstructured text data from 41 social media datasets. Random forest (RF) and naive Bayes (NB) algorithms show superior performance in accurately capturing human intuition, particularly for sentiment analysis. SVM does not consistently outperform other approaches, and lexicon-based approaches, including LIWC, show suboptimal performance compared to machine learning techniques. The findings suggest incorporating NB and RF in marketing research for advantageous outcomes [4] .

The research paper Opinion Mining on Twitter Data Using Supervised Machine Learning Algorithms. The digital age has generated a vast amount of computerized data, with people often using digital media platforms to express opinions. Opinion Mining, or Sentiment Analysis, helps classify reviews and determine their orientation. This study examines tweets related to the "The Necessity of the National Anthem in Cinema Theatres" news thread using various classifiers, with Random Forest and Multinomial Naive Bayes showing the highest accuracy values [5].

In the research Deep Learning--based Text Classification: A Comprehensive Review, this article examines over 150 deep learning models for text categorization, highlighting their technical contributions, commonalities, and strengths. It provides a brief overview of 40 datasets used, and evaluates their efficacy using benchmarks. The article also discusses potential future research avenues [6].

In document clustering based on text mining K- Means algorithm using Euclidean distance similarity examines Text mining, a specialised field within data mining, is employed to effectively classify large volumes of semi-structured data through the process of clustering. The process of maximum text documents include the efficient retrieval of information, the organisation of papers, and the exploration of information contained within the documents. The process of declaring text

input data and classifying documents is a multifaceted undertaking. The primary aim of this study is to develop a dedicated open source system for categorising clusters of identical documents within interconnected folders, with the intention of reducing the complexity associated with document retrieval [7].

Text mining is a multifaceted discipline that utilises information retrieval, data mining, machine learning, parameter statistics, and computational linguistics to extract important features and knowledge from unstructured text texts. Traditional approaches depend on comparable categories, however, grouping documents might enhance the extraction of knowledge. This research paper presents a novel approach that utilises the k-means algorithm to enhance the effectiveness of document clustering in the context of English text texts [8].

Text mining is a technique used to extract features and knowledge from unstructured text documents. It uses document clustering to categorize similar papers into groups. However, data dimensions can hinder extraction. This work uses TF-IDF, SVM, NMF, and k-means clustering techniques to categorize data. Results are compared using 20 newsgroup datasets, demonstrating the effectiveness of these techniques [9].

Document clustering is a classification process that operates without the need for supervision, wherein documents are grouped together into separate groups based on their similarities and differences. This research examines the application of K-means, heuristic K-means, and fuzzy C-means algorithms in the classification of textual content. The utilisation of tf.idf representation and stemming approaches in experiments has been found to enhance the results of clustering. Fuzzy clustering exhibits superior performance compared to K-means and heuristic K-means on several datasets, showcasing improved stability[10].

3. RESEARCH METHODOLOGY

This chapter details the research methodology used in the dissertation, emphasizing the importance of selecting a suitable technique for productivity and efficacy. The acquisition of qualitative data is crucial for achieving desired results.

A) Primary data

The Primary data is taken from Kaggle.

B) Secondary data

The secondary data sources were obtained via journals, books and papers.

C) Research process

The Research process is a organized procedure that encompasses gathering, evaluating and comprehend facts in order to address specific issue.

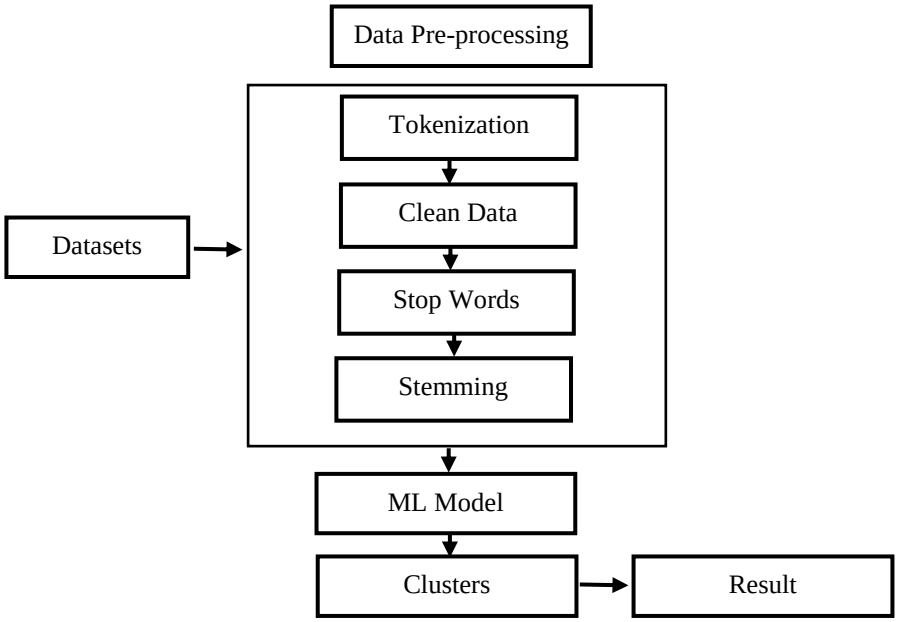

Figure 1: Research Process

Dataset

A total of 190 country national anthem were collected in English language.

Data Pre-processing

Machine learning encompasses the process of preparing and structuring raw data for both training and development purposes. This entails employing various techniques such as tokenization, data cleaning, stop words, and stemming to facilitate analysis.

ML model

Agglomerative Hierarchical Clustering (AHG) is a widely employed technique in the field of cluster analysis, wherein objects with similar characteristics are grouped together to create clusters. The technique commences by considering each data point as an individual cluster and subsequently combines the nearest pairs of clusters through an iterative process, utilising a selected distance metric, such as Euclidean distance or cosine similarity. When clusters combine, a hierarchical structure is created, which visually represents the order of merges and the connections between clusters. The utilisation of a hierarchical structure facilitates a versatile examination of the data, since it offers valuable insights into both intricate and broad clusters. The method persists in the process of merging clusters until a predetermined stopping criterion is satisfied, which may involve attaining a predetermined number of clusters or surpassing a specific distance threshold. Agglomerative Hierarchical Clustering is a method that is both intuitive and effective for comprehending the data's structure. However, it can be computationally burdensome for big datasets because of the computations involved in determining pairwise distances.

4. RESULTS

The Agglomerative Hierarchical Clustering (AHC) algorithm is a hierarchical technique that begins by assigning each data point to its own cluster. Subsequently, clusters are merged iteratively according to a selected criterion until a predetermined stopping condition is satisfied. In the AHC algorithm, the initial step involves treating each data point as an individual cluster. Subsequently, at each step, the closest pair of clusters is combined using a distance metric.

AHC is capable of adapting to noise and outliers present in the data. At higher levels of the hierarchy in AHC, noise points can either exist as isolated clusters or be assimilated into bigger clusters. In the context of AHC, it is common for the user to establish a predetermined number of clusters or employ a stopping condition that is determined by the distance between clusters.

The efficacy of AHC can be impacted by several factors, such as the selection of a distance metric (e.g., Euclidean distance or cosine similarity), the linkage criterion employed to ascertain the distance between clusters (e.g., single linkage, complete linkage, or average linkage), and the approach utilised to construct the hierarchical structure (e.g., dendrogram visualisation).

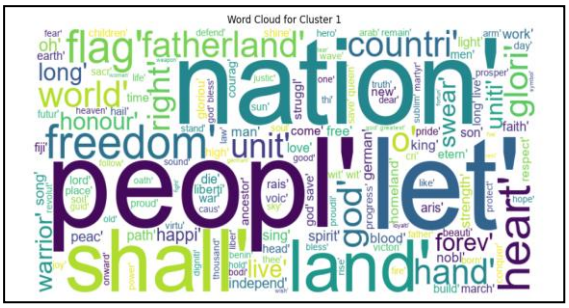

Figure 2: Cluster 1

Figure 3: Cluster 2

Figure 4: Cluster 3

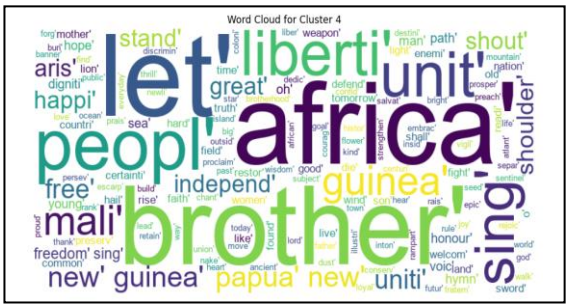

Figure 5: Cluster 4

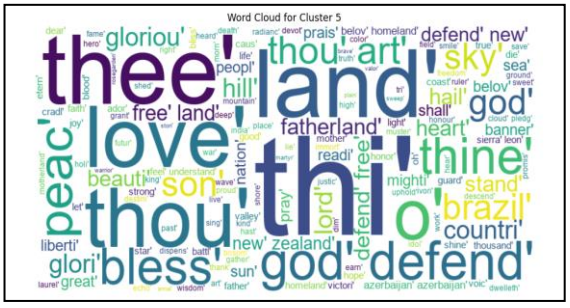

Figure 6: Cluster 5

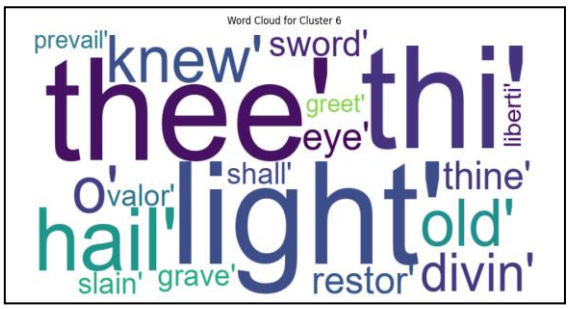

Figure 7: Cluster 6

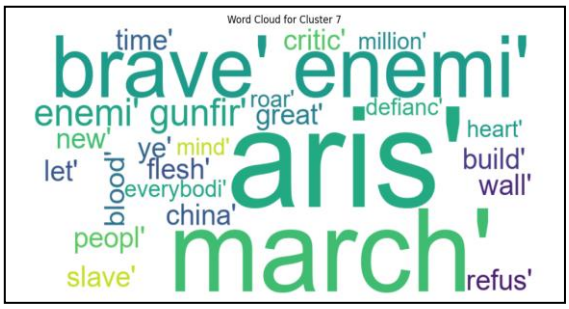

Figure 8: Cluster 8

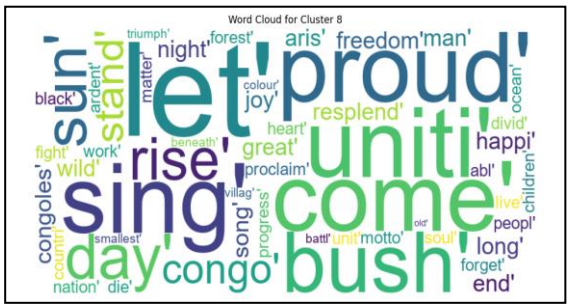

Figure 9: Cluster 8

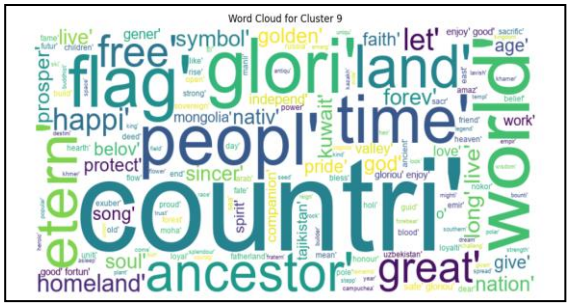

Figure 10: Cluster 9

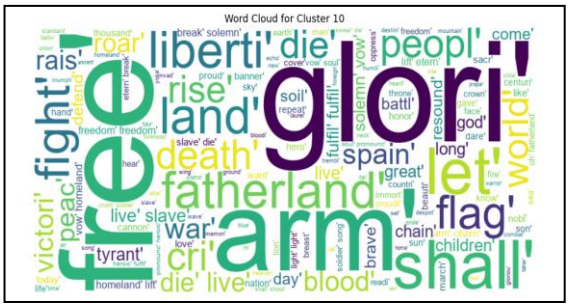

Figure 11: Cluster 10

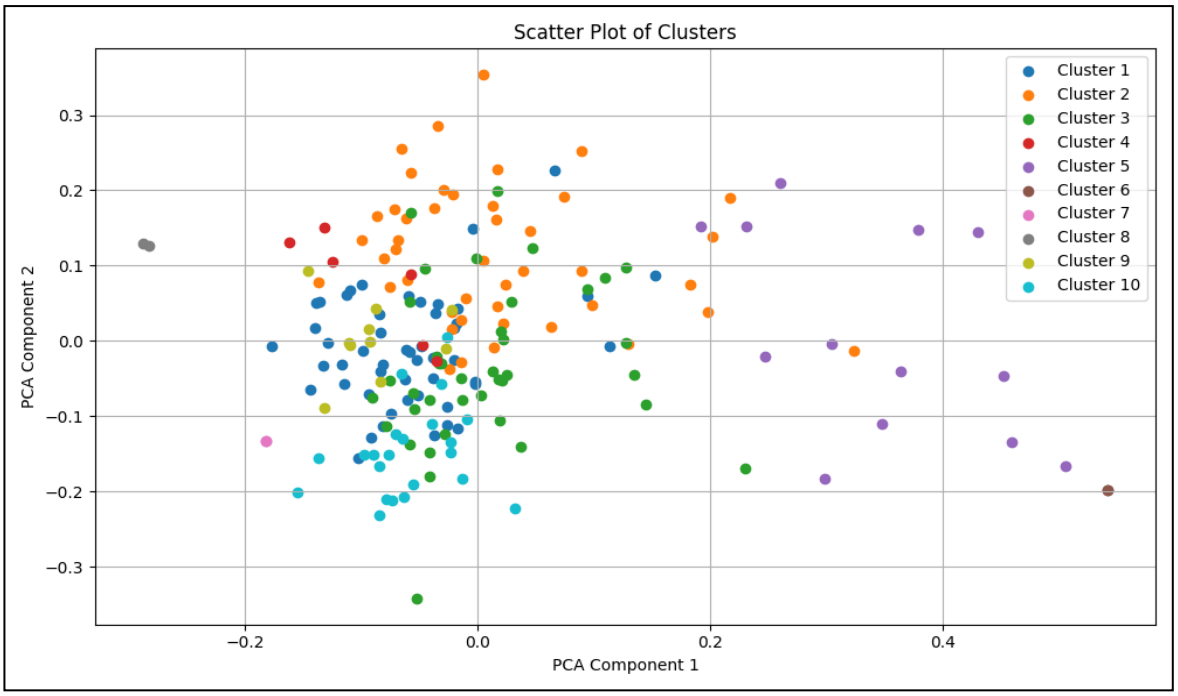

Figure 12: Cluster using Agglomerative Hierarchical Clustering

Table 1: Summary of Cluster

Clusters	Country
Cluster 1	'Albania' 'Armenia' 'Belgium' 'Germany' 'Liechtenstein' 'Lithuania' 'Netherlands (the)' 'Poland' 'Republic of North Macedonia' 'Slovenia' 'United Kingdom of Great Britain and Northern Ireland (the)' 'Venezuela' 'Haiti' 'El Salvador' 'Panama' 'Barbados' 'Grenada' 'Bermuda' 'Greenland' 'Fiji' 'Bahrain' 'Bhutan' 'Brunei' 'Jordan' 'Laos' 'North Korea' 'Palestine' 'Qatar' 'Saudi Arabia' 'Singapore' 'Thailand' 'Turkey' 'Vietnam' 'Yemen' 'Algeria' 'Angola' 'Benin' 'Burkina Faso' 'Burundi' 'Central African Republic' 'Chad' 'Eritrea' 'Gabon' 'Libya' 'Morocco' 'Mozambique' 'Niger' 'Sao Tome and Principe' 'Togo' 'Western Sahara'
Cluster 2	'Belarus' 'Estonia' 'Iceland' 'Latvia' 'Malta' 'Moldova (the Republic of)' 'Norway' 'Serbia' 'Canada' 'Jamaica' 'Trinidad and Tobago' 'Belize' 'Saint Lucia' 'Antigua and Barbuda' 'Solomon Islands' 'Vanuatu' 'Kiribati' 'Federated States of Micronesia' 'Indonesia' 'Israel' 'Japan' 'Kyrgyzstan' 'Malaysia' 'Myanmar' 'Nepal' 'Oman' 'Pakistan' 'South Korea' 'Botswana' 'Comoros' 'Ethiopia' 'Gambia' 'Ghana' 'Kenya' 'Madagascar' 'Mauritania' 'Namibia' 'Nigeria' 'Rwanda' 'Saint Helena' 'Seychelles' 'South Africa' 'Tanzania' 'Zimbabwe'
Cluster 3	'Austria' 'Bosnia and Herzegovina' 'Bulgaria' 'Croatia' 'Czechia' 'Denmark' 'Finland' 'Georgia' 'Hungary' 'Luxembourg' 'Montenegro' 'Slovakia' 'Ukraine' 'Colombia' 'Guyana' 'United States of America' 'Costa Rica' 'Puerto Rico' 'Bahamas' 'Dominica' 'Australia' 'Samoa' 'Afghanistan' 'Bangladesh' 'Iran' 'Iraq' 'Maldives' 'Sri Lanka' 'Syria' 'Djibouti' 'Egypt' 'Guinea-Bissau' 'Lesotho' 'Liberia' 'Malawi' 'Somalia' 'South Sudan' 'Uganda' 'Zambia'
Cluster 4	'Papua New Guinea' 'Cape Verde' 'Equatorial Guinea' 'Guinea' 'Mali' 'Senegal'
Cluster 5	'Azerbaijan' 'Switzerland' 'Brazil' 'Nicaragua' 'New Zealand' 'Tonga' 'India' 'Philippines' 'Cameroon' 'Ivory Coast' 'Mauritius' 'Sierra Leone' 'Swaziland'
Cluster 6	'Cyprus' 'Greece'
Cluster 7	'China' 'Macau'
Cluster 8	'Democratic Republic of Congo' 'Republic of the Congo'
Cluster 9	'Russian Federation (the)' 'Cambodia' 'Kazakhstan' 'Kuwait' 'Lebanon' 'Mongolia' 'Tajikistan' 'Turkmenistan' 'United Arab Emirates' 'Uzbekistan'
Cluster 10	'France' 'Ireland' 'Italy' 'Portugal' 'Romania' 'Spain' 'Sweden' 'Argentina' 'Bolivia' 'Chile' 'Ecuador' 'Paraguay' 'Peru' 'Suriname' 'Uruguay' 'Mexico' 'Guatemala' 'Cuba' 'Dominican Republic' 'Honduras' 'Sudan' 'Tunisia'

The above table defines that countries are divided into 10 different clusters which national anthem lies in same cluster.

5. DISCUSSION AND ANALYSIS

The purpose of this research is to tackle the difficulty of efficiently grouping and classifying national anthems from more than 190 countries worldwide. National anthems hold great cultural significance since they function as powerful emblems of historical narratives and cultural identity. Nevertheless, the restricted duration and the wide array of languages and literary genres they encompass pose difficulties in their classification by conventional text clustering methodologies.

The objective of this study is to address the current research void by explicitly concentrating on text clustering methods designed for national anthems. Through the assessment of clustering algorithms, with a specific focus on the Agglomerative Hierarchical Clustering algorithm, this research aims to offer significant insights into the shared characteristics, themes, attitudes, and language patterns found in national anthems originating from various linguistic and cultural contexts. The objective of this analysis is to provide insight into the common historical narratives, cultural values, and national identities that are evident in these songs.

Moreover, the objective of this study is to make a scholarly contribution by enhancing the approaches and procedures employed in text clustering within an area that has received limited attention in the current body of literature. The work seeks to improve the clustering process and showcase the efficacy of the selected algorithm in addressing the distinctive attributes of national anthems by integrating diverse text pre-processing approaches, including stop-words removal, stemming, noise removal, and TF-IDF features.

Limitation

- The analysis has a limited scope as it specifically examines the grouping of national anthems from more than 190 nations, perhaps disregarding alternative clustering methods that could produce different outcomes.
- Restricted text preprocessing methods such as stop-word elimination, stemming, noise elimination, and TF-IDF may not comprehensively encompass the intricacies of preprocessing procedures for diverse languages and literary structures.
- The quality and availability of data can differ among countries, which can impact the accuracy and reliability of clustering conclusions.

6. CONCLUSION

Ultimately, this study attempted to address the task of effectively grouping and classifying national anthems from more than 190 countries across the globe. The work offers vital insights into the efficacy of text clustering algorithms, specifically the Agglomerative Hierarchical Clustering algorithm, in categorizing national anthems according to shared characteristics. However, it is important to acknowledge numerous drawbacks.

Initially, the analysis was restricted to the Agglomerative Hierarchical Clustering algorithm, thus disregarding alternative clustering methods that could have distinct outcomes. Furthermore, the efficacy of text pre-processing methods may exhibit variability contingent upon the linguistic and cultural attributes of the national anthems, hence giving rise to apprehensions regarding the applicability of the results.

Moreover, it is important to acknowledge that the interpretation of clustering outcomes might be influenced by subjective factors. Additionally, it is crucial to carefully address ethical problems pertaining to the utilization and interpretation of cultural artefacts, such as national anthems.

Notwithstanding these constraints, the study enhances the current body of knowledge by providing valuable perspectives on the grouping of national anthems and emphasizing the difficulties and possibilities in this specific field. In order to strengthen the robustness and usefulness of clustering methodology for national anthems, future research may investigate exploring other clustering algorithms, incorporating more complete text preparation techniques, and addressing ethical considerations.

In general, the results of this study offer significant foundation for future investigations in the domain of text clustering. However, it is important to exercise caution when making conclusive assertions, and future research endeavours should strive to overcome the identified constraints in order to uphold the credibility and dependability of their findings.

REFERENCES

[1] "Text Clustering using K-MEAN," *Int. J. Adv. Trends Comput. Sci. Eng.*, vol. 10, no. 4, pp. 2892–2897, Aug. 2021, doi: 10.30534/ijatcse/2021/371042021.
[2] R. Silaghi-Dumitrescu, "Trends in the texts of national anthems: A comparative study," *Heliyon*, vol. 9, no. 8, p. e19105, Aug. 2023, doi: 10.1016/j.heliyon.2023.e19105.
[3] A. S. Oyeyemi, "THE NIGERIA'S NATIONAL ANTHEM: A TEXT LINGUISTIC EXPLORATION," no. 1, 2018.
[4] J. Hartmann, J. Huppertz, C. Schamp, and M. Heitmann, "Comparing automated text classification methods," *Int. J. Res. Mark.*, vol. 36, no. 1, pp. 20–38, Mar. 2019, doi: 10.1016/j.ijresmar.2018.09.009.
[5] D. M. Mathews and S. Abraham, "Opinion Mining on Twitter Data Using Supervised Machine Learning Algorithms," *Int. J. Comput. Sci. Eng.*, vol. 06, no. 06, pp. 63–66, Jul. 2018, doi: 10.26438/ijcse/v6si6.6366.
[6] S. Minaee, N. Kalchbrenner, E. Cambria, N. Nikzad, M. Chenaghlu, and J. Gao, "Deep learning--based text classification: a comprehensive review," *ACM Comput. Surv. CSUR*, vol. 54, no. 3, pp. 1–40, 2021.
[7] E. L. Lydia, P. Govindaswamy, S. Lakshmanaprabu, and D. Ramya, "Document clustering based on text mining K-means algorithm using euclidean distance similarity," *J. Adv. Res. Dyn. Control Syst.*, vol. 10, 2018.
[8] A. I. Kadhim, Y.-N. Cheah, and N. H. Ahamed, "Text document preprocessing and dimension reduction techniques for text document clustering," presented at the 2014 4th international conference on artificial intelligence with applications in engineering and technology, IEEE, 2014, pp. 69–73.
[9] R. Kumbhar, S. Mhamane, H. Patil, S. Patil, and S. Kale, "Text document clustering using k-means algorithm with dimension reduction techniques," presented at the 2020 5th International Conference on Communication and Electronics Systems (ICCES), IEEE, 2020, pp. 1222–1228.
[10] V. K. Singh, N. Tiwari, and S. Garg, "Document clustering using k-means, heuristic k-means and fuzzy c-means," presented at the 2011 International conference on computational intelligence and communication networks, IEEE, 2011, pp. 297–301.