ISSN 2278-3091
Volume 10, No.4, July - August 2021
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse371042021.pdf
https://doi.org/10.30534/ijatcse/2021/371042021

Text Clustering using K-MEAN

Chaman Lal[1,2], Awais Ahmed[1],Reshman Siyal[3], Suresh Kumar Beejal[3],Shagufta Aftab[3],Arshad Hussain[2]
[1]Mohammad Ali Jinnah University Karachi Pakistan
[2]Faculty of Engineering, Science and Technology, Indus University Karachi Pakistan, [3]FCIT, Indus University Karachi Pakistan.
chaman.lal@indus.edu.pk , awais.ahmed@jinnah.edu,reshman.siyal@indus.edu.pk, suresh.poorani@indus.edu.pk, shagufta.aftab@indus.edu.pk, arshad15me04@gmail.com

ABSTRACT

Document clustering allows the user to add similar documents to a group. For many years, it has been a fascinating research topic, developed various methods and techniques. However, the study focuses mostly on English and high-resource languages. About Pakistan national anthems, this research gives an experimental estimation of clustering techniques. Because of its short length, thematically clustering Anthem is a difficult task. This paper extracted various characteristics, including stop-words, stemming, corpus tokenization, noise removal, and TF-IDF features from the anthem, and the clustering was conducted using the K-Means algorithm. The results show that a clustering strategy paired with a K-mean clustering algorithm with TF-IDF features has already been used. The dataset is available on GitHub (https://www.kaggle.com/lucasturtle/national-anthems-of-the-world).

1. INTRODUCTION

The goal of this research is to learn how to encode text data (in this case, the national anthem) so, such that the K-Means algorithm can use it as input. Since we are dealing with countries, it would be good to imagine that this is the result of an interactive map with Folium.

The entire procedure can be broken down in the following steps:

1. Dataset Pre-processing
2. Attribute to extraction through TF-IDF
3. K-mean & cluster analysis running
4. Visualization of clusters in a map with folium.

2. RELATED WORK

In this section is a brief overview of the literature. For predicting the emotion of Urdu tweets, Zarmeen Nasim and Sayeed Ghani [1] suggested an emotion analysis approach according to the Markov chains. The recovery of valuable information from a vast volume of data is one of the fields of text mining. In which classification is subsumed by the intricate process of grouping interconnected documents. Document clustering is a technique for gathering documents with a high degree of similarity into a single cluster. One way to cluster a document is k-means clustering, which clusters of documents according to the centroid & cluster similarities [2]. Ammar Ismael Kadhim, Yu-N Cheah, Nurul Hashimah Ahamed, et.al. [3] Introducing document clustering is a way to retrieve process text before the relevant information form. Knowledge of pre-processing text has a significant effect. The complexity of discovering a document is directly proportional to the number of documents. As a result, clustering algorithms are unable to handle data with a high dimensionality. Techniques such as DR are used to solve this problem. The authors demonstrate how SVD with k-means clustering and pre-processing can be used to reduce the difficulty of a task. The accuracy in outcome is 95 percent and 94.6667 percent, respectively, based on the BBC news and BBC sports datasets. This demonstrates that the proposed strategy improves the grouping of English texts in text documents. Large data dimensions are a barrier to extracting relevant information, according to RutujaKumbhar and SnehalMhamane [4], hence the dimensions of data matrices are reduced using the dimensionality reduction (DR) technique. TF-IDF, SVD, k-means clustering and NMF are used to further partition data into clusters using the k-means clustering algorithm. On Urdu tweets, Zarmeen Nasim and Sajjad Haider [5] provided an experimental evaluation of clustering algorithms. Due to their short length, thematically clustering tweets is a difficult operation. Tweets are analyzed for several features such as phrase, phrase level embedding, TF-IDF properties. clustering is done using three different methods: K-Means, Affinity Propagation and Bisecting K-mean techniques. According to the findings, The K-mean clustering algorithm paired with T-IDF features outperforms the previously used clustering technique.

3. DATASET PRE-PROCESSING

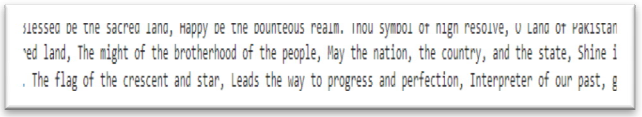
It is crucial to understand the structure of the dataset before we begin preprocessing it:

	country	alpha-2	alpha-3	continent	anthem
80	Australia	AU	AUS	Oceania	Australians all let us rejoice For we are youn...
81	Papua New Guinea	PG	PNG	Oceania	O arise all you sons of this land, Let us sing...
82	New Zealand	NZ	NZL	Oceania	God of Nations at Thy feet, In the bonds of lo...
83	Fiji	FJ	FJI	Oceania	Blessing grant oh God of nations on the isles ...
84	Solomon Islands	SB	SLB	Oceania	God Save our Solomon Islands from shore to sho...
85	Vanuatu	VU	VUT	Oceania	We are, we are, we are happy to proclaim We ar...

Figure. 1 data pre-processing

It is worth noting that the columns 'alpha 2' and 'alpha-3' indicate the ISO-Codes for various countries show in (fig. 1). It will be useful in the future.

It is also interesting to know how the anthems are composed. The dataset contains noise, as evidenced by Pakistan's national anthem show in (fig. 2). Anthems often incorporate references to places and individuals from many cultures, which are frequently not constituted by UTF-8 characters.



blessed be the sacred land, Happy be the bounteous realm. Thou symbol of high resolve, O Land of Pakistan
red land, The might of the brotherhood of the people, May the nation, the country, and the state, Shine i
. The flag of the crescent and star, Leads the way to progress and perfection, Interpreter of our past, g

Figure. 2with noise, stop words of National anthem**.**

Our dataset will now be referred to as a corpus. A set of procedures that must be applied to the corpus before it can be well constituted by a statistical framework is referred to as the pre-processing stage. The K-Means algorithm is then applied to the data to analyze it and make a decision about what to do with it[6]. The pre-processing routine is divided into the sections listed below:

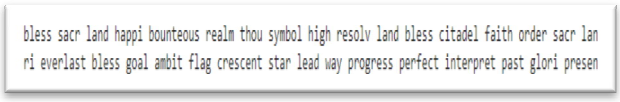
1. **Corpus tokenization**, which is the process of separating different texts into individual words.

2. S**top words removal**, these are ordinary words (not, the, a, et cetera) that do almost nothing in the meaning of a text.

3. **Noise removal,** which include anything from the text that cannot be recognized such as words with non-ASCII symbols, an English word, such as words coupled with digits, fall into this category.

4. **Stemming**, is the process of reducing a word to its root. For example [History, Historical, Histories] ->histra.

Auxiliary operates that eliminate a list of terms (such as pause words) from the text, add steaming and delete words of 2 characters or fewer, and words with 21 characters or more are listed below show in (fig. 3)(the longest word which has 21 letters in the English alphabet is "Incomprehensibilities"). We will now build the central processing operate, which uses regular feedback to eliminate noise and calls the said process.After the operation on our corps, the Pakistan national anthem sounded like this.



bless sacr land happi bounteous realm thou symbol high resolv land bless citadel faith order sacr lan
ri everlast bless goal ambit flag crescent star lead way progress perfect interpret past glori presen

Figure. 3without noise, stop words of National anthem**.**

4. ATTRIBUTE TO EXTRACTION THROUGH TF-IDF

The pre-processing phase was completed to produce the best possible results for this phase. We desire to define the value of a word in the collection of documents so that we can use the encoded data through the algorithm. The essence of the feature refers to the process of converting text-related data into real vectors. Term Frequency - Reverse Document Frequency (TF-IDF) is a numerical statistic that aims to show the significance of a word in the corpus, in the corpus, which is considered more relevant. When a term appears multiple times in a document, this approach operates by increasing its weight and decreasing it as well, when commonly used in many documents.

This approach is broken down into two parts:

a. *Term-Frequency (TF)*

The term-frequency tf (t, d) is a measure of how often a term t occurs in a document d, which weighs more frequently than the terms.

$$tf(t,d) = log(1 + freq(t,d))$$

b. *Inverse-Document-Frequency (IDF)*

Inverse document frequency idf (t, D) measures the importance of a term t in the repository of documents D. This statistic reduces the importance of post-terminology and increases the weight of rare terms that make the text more meaningful.

The production of TF (t, d) by IDF (T, D) achieves a TF-IDF score for each term.

$$tfidf(t,d,D) = tf(t,d) \cdot idf(t,D)$$

Although we have understood all the math, it is very simple to apply TF-IDF show in (fig. 4) using the on-screen library!

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
advanc	0.368967	0.000000	0.000000	0.000000	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.000000
fair	0.368967	0.000000	0.000000	0.000000	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.000000
ve	0.368836	0.000000	0.000000	0.000000	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.000000
let	0.247619	0.097719	0.059606	0.059606	0.0	0.0	0.0	0.0	0.092678	0.0	0.0	0.0	0.0	0.0	0.0	0.124088
strain	0.245891	0.000000	0.000000	0.000000	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.000000

rows × 110 columns

Figure. 4 Overweight words

5. RESULT AND DISCUSION

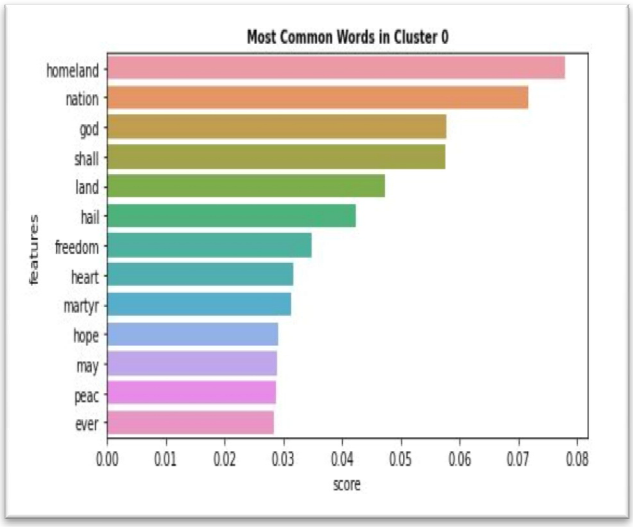
K-MEAN AND CLUSTER ANALYSIS RUNING

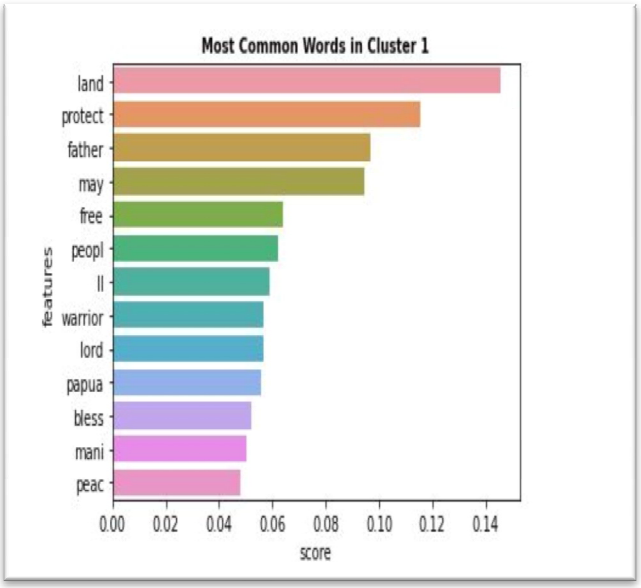
K-Means is a machine learning algorithm that is both simple and popular. Because it does not use data, it is an unsupervised algorithm; in our situation, this means that no one text be relevant to a class or clusters. It is also a clustering algorithm that classifies datasets into clusters of K numbers.

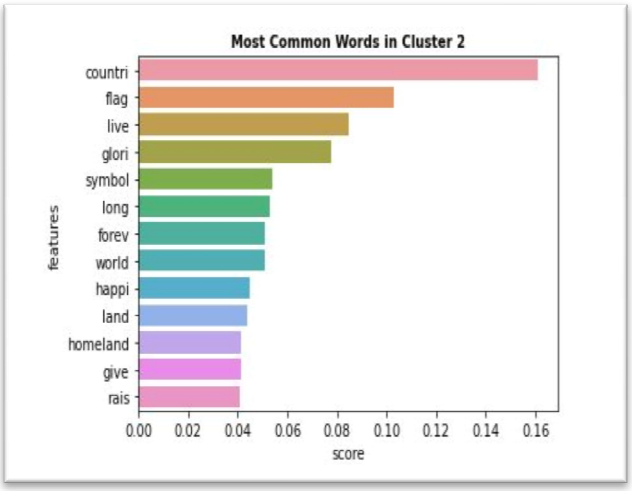
The principle subsequently this technique is that its clusters are defined by K centroids, with each centroid representing the cluster's center. This approach works interactively, with each centroid starting in a random location in the dataset's vector space and moving to the center of the points that are closest to them.The distance amid each centroid and the points is calculated again with each subsequent iteration, and the centroids are moved back to the center of the nearest points. When the location or groupings of centroids do not change any longer, or when the distance between the centroids changes does not exceed a pre-defined threshold, the method is complete.

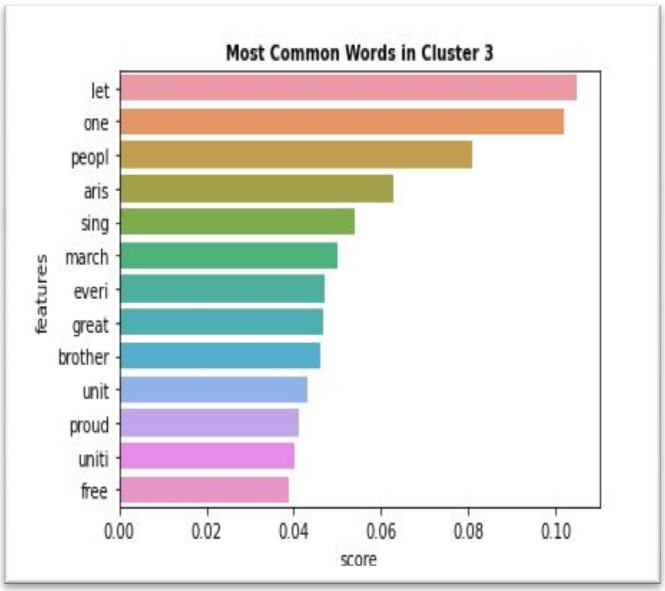
We shall run the algorithm with 2 to 5 clusters at first because we do not know the right number of groups.

We must decide which K number is the best based on the outcomes of all the other categories, which might be a very subjective study.The Elbow Method and the Silhouette Score are two methods for evaluating the algorithm's performance and providing insight into how well defined the groups were went through all the possible clusters in each grouping by look at graphs with the most prevalent term in groups to determine the best number of clusters show in (fig. 5& 6). After dividing the groupings into five different clusters, I came to opinion that they made more sense.









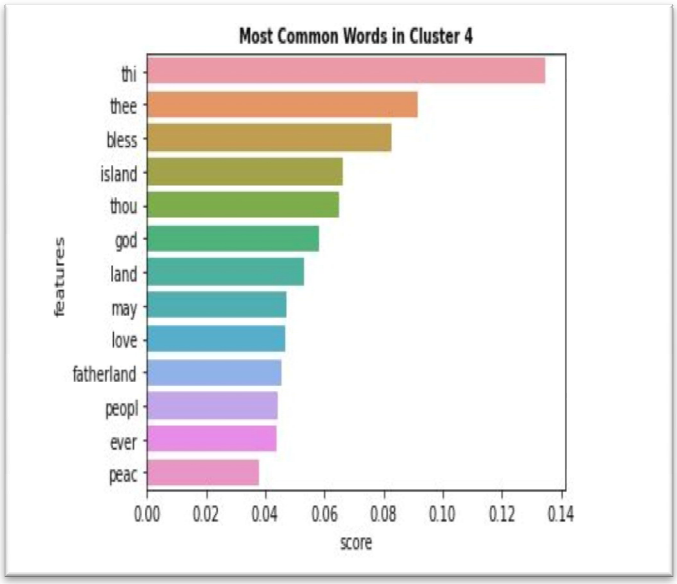
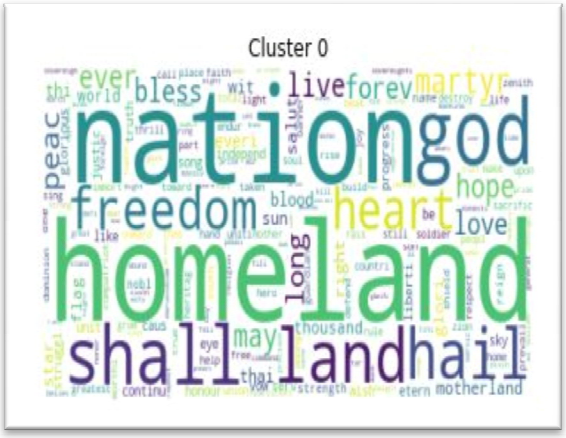
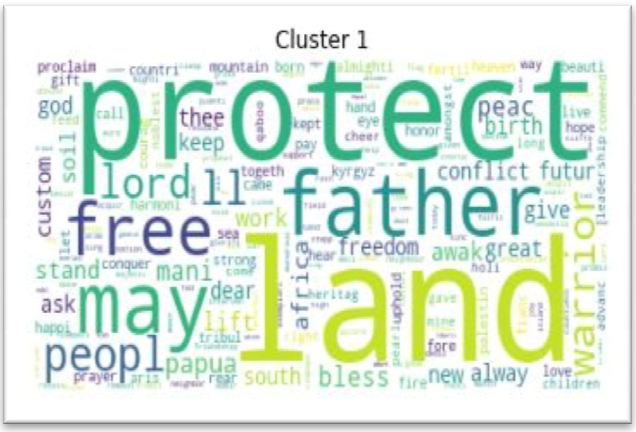
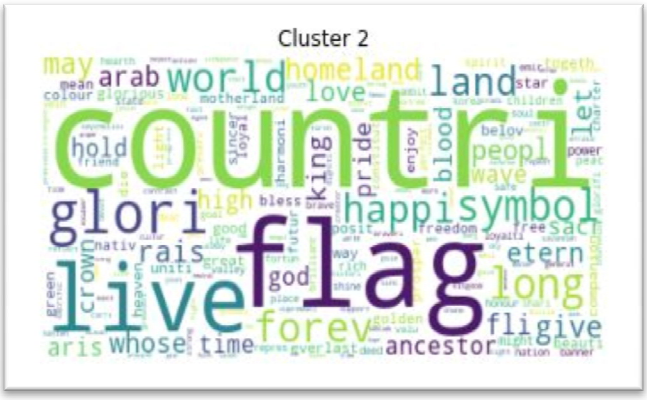


Figure. 5 Most dominant words by cluster







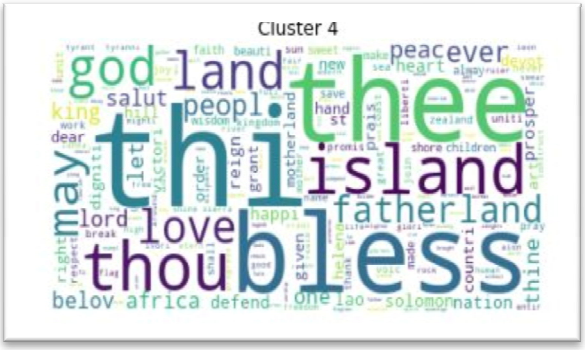


Figure. 6 Word Clouds

The words in each cluster have a theme, as can be seen by looking at the clusters. For example, positive terms like “homeland,” “heart,” “freedom,” and "peac" appear more frequently in Cluster 0.
Now we will categories the labels from the K-Means results by nation show in (fig. 7).

	country	alpha-2	alpha-3	continent	anthem	label
80	Australia	AU	AUS	Oceania	Australians all let us rejoice For we are youn...	3
81	Papua New Guinea	PG	PNG	Oceania	O arise all you sons of this land, Let us sing...	1
82	New Zealand	NZ	NZL	Oceania	God of Nations at Thy feet, In the bonds of lo...	4
83	Fiji	FJ	FJI	Oceania	Blessing grant oh God of nations on the isles ...	0
84	Solomon Islands	SB	SLB	Oceania	God Save our Solomon Islands from shore to sho...	4

Figure. 7 continents with label.

VISUALIZATION OF CLUSTERS IN A MAP WITH FOLIUM

The plan now is to use a choropleth map to colour each country in its corresponding group. The Folium library can be used for this.

First, we shall create a palette for cluster show in (fig. 8)

Figure. 8palettes for cluster

We need to plot the location of every country on the map. Jason file with the multifaceted nature of each country identified with the ISO code.

Integrating each country's label with its polygon with the ISO code is now straightforward show in (fig. 9)!

	id	name	label	geometry
0	AFG	Afghanistan	3	POLYGON ((61.21082 35.65007, 62.23065 35.27066...
1	AGO	Angola	3	MULTIPOLYGON (((16.32653 -5.87747, 16.57318 -6...
2	ARE	United Arab Emirates	2	POLYGON ((51.57952 24.24550, 51.75744 24.29407...

Figure. 9country's label with its polygon

We can construct a map adduce using Folium, draw the polygons and paint the countries with colour of their corresponding labels, thanks to the polygons and labels now associated in the data frame above show in (fig. 10).

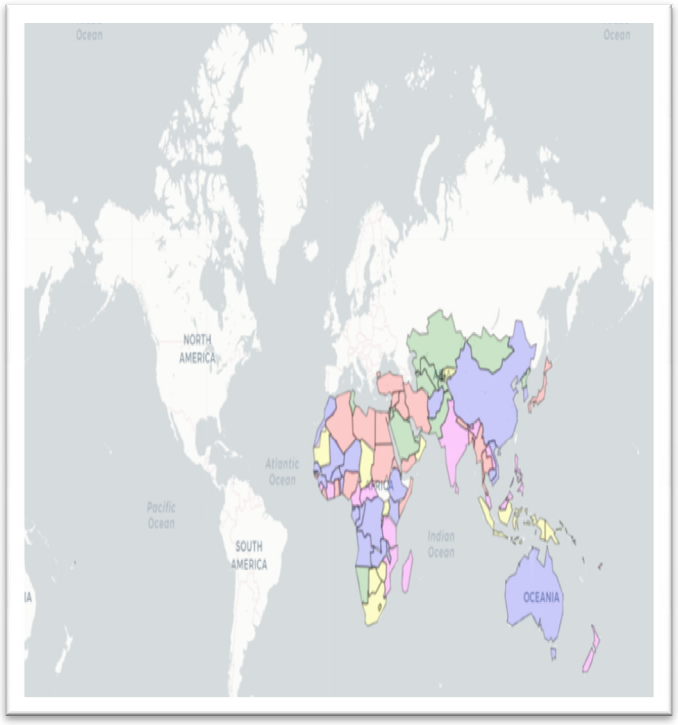


Figure. 10countries with groups

6. CONCLUSION

Text clustering is a process that concern the use of (NLP) and clustering algorithms. This technique of identifying clusters in unstructured texts can be used in a variety of applications, including feedback analysis, study segmentation, and so forth. Many factors affect K-mean results, including distance measurement, centrodes initial positions, and grouping analysis.

REFERENCES

[1]. Nasim, Zarmeen, and Sayeed Ghani. "Sentiment Analysis on Urdu Tweets Using Markov Chains." SN Computer Science 1.5 (2020): 1-13.

[2]. Laxmi Lydia, P.Govindaswamy, SK.Lakshmanaprabu, D.Ramya , "Document Clustering Based On Text Mining K-Means Algorithm Using Euclidean Distance Similarity ", Jour of Adv Research in Dynamical & Control Systems, Volume 10, 02-Special Issue, 2018

[3]. Ammar Ismael Kadhim, Yu-N Cheah, Nurul Hashimah Ahamed, "Text Document Preprocessing and Dimension Reduction Techniques for Text Document Clustering", 4th International Conference on Artificial Intelligence with Applications in Engineering and Technology, 2014

[4]. Kumbhar, Rutuja, et al. "Text Document Clustering Using K-means Algorithm with Dimension Reduction Techniques." 2020 5th International Conference on Communication and Electronics Systems (ICCES). IEEE, 2020.

[5]. Nasim, Zarmeen, and Sajjad Haider. "Cluster analysis of urdu tweets." Journal of King Saud University-Computer and Information Sciences (2020).

[6].Tkatek, Said, et al. "Artificial intelligence for improving the optimization of NP-hard problems: a review." International Journal of Advanced Trends Computer Science and Applications 9.5 (2020).