Impact of Distance Measures on Urdu Document Clustering

Zarmeen Nasim
Artificial Intelligence Lab
School of Mathematics and Computer Science
Institute of Business Administration (IBA), Karachi
Znasim@iba.edu.pk

Sajjad Haider
Artificial Intelligence Lab
School of Mathematics and Computer Science
Institute of Business Administration (IBA), Karachi
sahaider@iba.edu.pk

Abstract

Document Clustering aims to group similar documents based on the distance among them. A wide range of distance measures are available in the literature, and selecting an appropriate distance function is a non-trivial task. This paper empirically evaluates four distance measures: Euclidean, Manhattan, Cosine, and Pearson Correlation, on Urdu news headlines. In addition to distance measures, the effect of stemming and lemmatization techniques on clustering is also studied. Unigram-based features and word embedding-based features were used to build a feature matrix. The evaluation results indicate that the frequent unigram features yielded the highest Adjusted Rand Index (ARI) scores on average. Among the four distance measure, the Cosine distance metric was found to be more valuable. Furthermore, the stemming technique was identified to be more useful in contrast to lemmatization for clustering news.

1 Introduction

Document clustering is an unsupervised learning task that aims to group similar documents together and separate dissimilar documents from each other. Most of the clustering algorithms utilize distance functions to identify documents that are syntactically close to each other. The choice of distance measure is critical as it influences the result of clustering.

In literature, clustering and the impact of distance measures have been widely studied on the English language corpus. This paper studies the impact of various distance measures on the clustering of Urdu documents. Urdu is the national language of Pakistan, and despite being spoken by around 170 million people around the globe (Hamdani et al., 2020), it is considered a low-resource language. With the Urdu language support provided on various social media platforms, a huge corpus of user-generated content is now available in digital format. The availability of such enormous data provides opportunities to researchers working in Urdu language processing.

Urdu is a morphologically rich language in contrast to the English language. Several words have various inflections, which adds computational complexity and requires sophisticated models. As a result, the algorithms and techniques developed for the English language cannot be directly applied to the Urdu language corpus due to morphological, syntactical, and lexical differences between both languages. To address the issue of morphological richness of the Urdu language, stemming and lemmatization techniques can be applied to raw text. Therefore, this study also presents the empirical evaluation of stemming and lemmatization on document clustering.

The rest of the paper is organized as follows. Section 2 presents the literature review of document clustering. The methodology is described in Section 3, while Section 4 reports the results of empirical evaluation. Finally, Section 5 concludes the paper.

2 Related Work

This section presents a brief description of the previous research document clustering and the impact of distance functions on clustering.

(Huang, 2008) studied the impact of five different similarity measures, including Euclidean, Cosine, Jaccard, Pearson correlation, and averaged

Kullback-Leibler divergence on partitional clustering. The evaluations were performed on seven different datasets. It was found that the Euclidean distance measure results in the worst performance, whereas the performance of the remaining four measures was similar.

(Aggarwal et al., 2001) studied the impact of distance measures in high dimensional feature spaces. It was theoretically and empirically found that the performance of Lk norm decreases with the increasing value of k in high dimensional spaces. The authors suggested that the Manhattan distance is more appropriate in high dimensional feature space than Euclidean distance.

(Aggarwal et al., 2019) proposed the improvement in the K-means clustering algorithm to deal with the uncertainties in real-world datasets. Further, the authors studied the performance of the proposed clustering algorithm using four distance measures: Euclidean, CityBlock, Cosine, and correlation distance measures. The evaluations were conducted using Davies–Bouldin index and purity metrics. The experiments showed that the correlation distance performed best among other closeness measures as it results in the minimum value of Davies–Bouldin index and maximum purity value.

(A et al., 2013) compared the performance of four different distance measures, including Euclidean, Jaccard, Cosine, and Correlation distance on a clustering task. Purity was used to evaluate the performance of distance measures. It was revealed that Jaccard and correlation distance measures were performing better than Euclidean distance in most cases. (Subhashini and Kumar, 2010) discussed the impact of distance measures on information retrieval and document clustering. They experimented with three distance functions, including Euclidean, Cosine, and Jaccard. Purity metric was used to evaluate the performance of distance measures on clustering tasks. The results showed that the cosine similarity measure and Jaccard index achieved similar performance, whereas the Euclidean distance measure performed worst.

(Bsoul et al., 2014) studied the impact of stemming and lemmatization on Arabic document clustering. Furthermore, the authors also conducted the evaluation of five distance functions for the document clustering task. The distance functions include Cosine, Jaccard, Pearson Correlation, Euclidean, and averaged Kullback-Leibler divergence. The results indicated that the proposed stemming algorithm for the Arabic language yielded good results. Moreover, the experiments also showed that the cosine similarity and Euclidean distance functions achieved the best results compared to other distance measures. (Rahman et al., 2018) studied the effect of various distance measures on Urdu document clustering. The distance measures evaluated in their work included Levenshtein distance, Jaccard index, and Cosine function. The experiments demonstrated that the Jaccard index yielded good results in terms of purity.

In this paper, the impact of stemming and lemmatization is studied for document clustering. The focus of this research is on short-length documents such as News headlines written in the Urdu language. The short-length documents pose the challenge of sparsity in feature space. Furthermore, this paper also identified the best-performing distance measure through empirical evaluation. To the best of our knowledge, such thorough analysis of distance measures, stemming, and lemmatization on short-length document clustering is not performed for the Urdu language.

3 Methodology

This section describes the workflow of experiments conducted to evaluate the impact of distance measures, stemming, and lemmatization.

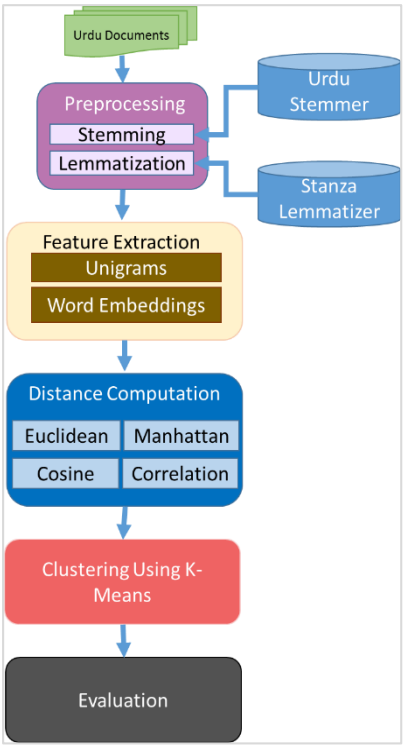

Figure 1: Workflow of Methodology

A pictorial representation of the workflow is presented in Figure 1.

The process involved the following five major steps:

1. Preprocessing
2. Feature extraction
3. Distance computation
4. Clustering
5. Evaluation

Each of these steps is described in the following subsections.

3.1 Preprocessing

The given input corpus is first preprocessed before feature extraction is performed. During preprocessing, the URLs, non-Urdu alphabets, and characters are removed. Punctuation marks and diacritics are also filtered from the input text. Moreover, stopwords are also removed on account of being not crucial for the clustering task. The stop words list, available in the Urdu Hack library[1], is used for the stopwords removal task. After cleaning raw text, stemming and lemmatization are also applied to the cleaned text.

a) Stemming: In natural language processing, stemming is the process of transforming a word into its root form. We implemented the stemming approach proposed by (Akram et al., 2009) for Urdu language using the Python language.

b) Lemmatization: Lemmatization is the process of grouping all inflections of a word to the base form called *Lemma*. For Urdu lemmatization, the Stanza[2] library is used as it supports the Urdu language along with other numerous languages.

3.2 Feature Extraction

After preprocessing the text, the next step is to convert the text into a feature matrix. For transforming text into a feature matrix, the following two methods are employed.

a) Unigram Features: In this method, the text is first tokenized into words. After tokenization, a vocabulary of unique words is built. The length of vocabulary represents the size of the feature matrix. The input text is then transformed into a feature vector. The term frequency-inverse document frequency (TF-IDF) metric is used to weight the feature vector. The matrix built using unigram features represents a sparse matrix where most of the entries are zero.

b) Word Embeddings: In recent years, word embeddings have shown tremendous improvements over the bag of words model in various NLP tasks. Word embedding refers to the distributed vector representation of a word in a dense feature space. In this work, pre-trained word embeddings (Kanwal et al., 2019) trained using Word2Vec (Mikolov et al., 2013) algorithm on Urdu news corpus are used. The word embedding model produces a vector representation of a single word. To generate the sentence embeddings, the average of word embeddings is computed.

3.3 Distance Computation

Once the feature matrix is built, a distance function computes the distance among documents. From several distance measures, four functions are used in this paper on account of their popularity. This includes Manhattan, Euclidean, Cosine, and Pearson Correlation distance functions. The details of each of the distance metric are given below:

a) Manhattan Distance:

The Manhattan distance, also known as CityBlock, between two data points A (x_1, y_1) and B (x_2, y_2) is the sum of absolute difference. It is computed as:

$$Manhattan\ Distance\ (A,B) = \sum_{i=1}^{N} |A_i - B_i| \quad (1)$$

Where N is the number of dimensions.

b) Euclidean Distance:

The Euclidean distance represents the shortest distance between two data points, A (x_1, y_1) and B (x_2, y_2). It is calculated as follows:

$$Euclidean\ Distance\ (A,B) = \sqrt{\sum_{i=1}^{N} (A_i - B_i)^2} \quad (2)$$

c) Cosine Distance:

Cosine Similarity measures the cosine of the angle between the two data points A (x_1, y_1) and B (x_2,

[1] https://github.com/urduhack/urduhack

[2] https://stanfordnlp.github.io/stanza/

y_2). The maximum value of similarity represents highly similar documents. This value is subtracted from one to get the distance between two data points. It measures the orientation of the document instead of the magnitude as in Euclidean distance.

The formula given below calculates the cosine distance between A (x_1, y_1) and B (x_2, y_2).

$$Distance\ (A,B) = 1 - \frac{\sum_{i=1}^{N} A_i\, B_i}{\sqrt{\sum_{i=1}^{N}(A_i)^2}\sqrt{\sum_{i=1}^{N}(B_i)^2}} \quad (3)$$

d) Pearson Correlation Distance:

Pearson Correlation distance measure is based on the linear correlation between two data points, A (x1, y1) and B (x2, y2). It is limited to only linear associations between the variables. The following formula computes correlation distance.

$$Distance\ (A,B) = 1 - \frac{(A-\bar{A}).(B-\bar{B})}{\|(A-\bar{A})\|_2 \|(B-\bar{B})\|_2} \quad (4)$$

3.4 Clustering

In the previous step, distance measures are used to compute the distance between the documents. In this step, clustering is performed using the K-Means clustering algorithm (Lloyd, 1982). K-Means is a partitional clustering algorithm that assigns documents to different clusters such that the resultant clusters are non-overlapping. The algorithm works as follows:

a) Initialize k centroids randomly

b) Calculate the distance of centroids from each document using the distance function

c) Assign the document to the closest centroid

d) Take the average of the documents to update centroid

e) Reiterate steps (b) – (d) for n number of iterations.

The clustering result produced by the K-Means algorithm depends upon the initial centroids and varies with different seeds. Therefore, all the experiments conducted in this chapter report the average result of five independent runs of the K-Means algorithm. In each run, a different seed value was chosen.

3.5 Evaluation

In this work, the adjusted rand index is used to evaluate and compare the effect of distance measures, stemming, and lemmatization on clustering. Adjusted Rand Index (Hubert and Arabie, 1985) is the measure of similarity between the true cluster labels and the predicted cluster labels. The Rand Index is computed as follows:

$$Rand\ Index = \frac{a+b}{\binom{n}{2}} \quad (5)$$

Where,
n is the number of documents in the clustering,
a refers to the number of documents that are in the same clusters in actual and predicted clustering,
b refers to the number of documents that are in different clusters in actual and predicted clustering.

The adjusted rand index accounts for adjustments due to the number of clusters. It is calculated as:

$$ARI = \frac{RI - Expected\ RI}{\max(RI) - Expected\ RI} \quad (6)$$

The value of ARI is between 0 and 1, where 0 refers to worst quality clusters, and 1 refers to the best quality clusters.

4 Experiments and Results

This section first describes the dataset on which evaluations were performed. Later in this section, we describe the series of experiments that were conducted in this research work.

Cluster Labels	**Count**
کرپشن (Corruption)	491
کرونا وائرس (Coronavirus)	200
سی پیک (CPEC Agreement)	163
نیشنل ٹی ٹونٹی (National T20 Cup)	42
الیکشن کمیشن (Election Commission)	157
جوبائیڈن (Joe Biden)	139
کراچی کے مسائل (Problems of Karachi)	37
موسم (Weather)	355
ڈینگی (Dengue)	166
Total: 1750 Headlines	

Table 1: Dataset Description

Experiment	Features	Euclidean	Cosine	Pearson	Manhattan
Raw Text	All Unigrams	0.708	0.674	0.24	0.544
	Frequent Unigrams	0.7	**0.778**	0.12	0.648
	Word Embeddings	0.41	0.5	0.504	0.424
Stemmed Text	All Unigrams	0.722	0.728	0.308	0.554
	Frequent Unigrams	0.797	**0.825**	0.11	0.617
	Word Embeddings	0.432	0.55	0.522	0.418
Lemmatized Text	All Unigrams	0.668	0.676	0.294	0.618
	Frequent Unigrams	**0.752**	0.742	0.092	0.716
	Word Embeddings	0.522	0.534	0.462	0.482

Table 2: Clustering Results on Urdu News Headlines dataset

4.1 Dataset

The dataset used for empirical evaluation comprised 1750 Urdu news headlines on various topics. The news headlines were fetched from the RSS feeds from the popular Urdu news agencies, including Express[3], UrduPoint[4], Nawae Waqt[5], Voice of America[6] , and BBC Urdu[7]. Table 1 shows basic statistics of the dataset containing news headlines on nine selected keywords. The keywords are used as true cluster labels for the extrinsic evaluation of clustering experiments.

4.2 Clustering Experiments

In this series of experiments, the K-Means clustering algorithm was applied using four different distance measures on raw text, stemmed text and lemmatized text. The raw text refers to the cleaned and preprocessed text without stemming and lemmatization. Three different feature extraction techniques were used to build a feature matrix. The clustering results reported are the average of five independent runs of the K-Means algorithm initialized with different random seeds. K-Means algorithm requires the number of clusters before applying clustering. The value of clusters (k) was set to nine (9) as there were nine topics present in the dataset.

In the first experiment, the raw text was passed to the feature extraction module for extracting various features. K-Means clustering algorithm was then applied on feature matrix for each distance measure. Table 2 presents the results of clustering evaluation on raw text.

It was found that, on average, when frequent unigram features were considered, the highest ARI value was obtained. Furthermore, the results also indicated that the average performance of the clustering algorithm was maximum when cosine distance was used to compute the distance between the news headlines. The Pearson Correlation distance function performed worst on unigram features. However, its performance is almost similar to the cosine distance function on word embedding-based features. This is due to the reason that the mean across word embedding dimensions is zero and the computation of Pearson Correlation distance becomes approximately equal to the cosine distance function.

The second experiment applied stemming to the preprocessed text before the feature extraction stage. Afterward, three different feature extraction techniques were applied, similar to the previous experiment. Clustering was performed on the resultant feature matrix with four distance metrics. As shown in Table 2, on average, frequent unigram

3 https://www.express.pk/

4 https://www.urdupoint.com/daily/

5 https://www.nawaiwaqt.com.pk/

6 https://www.urduvoa.com/

7 https://www.bbc.com/urdu

features performed better, as was the case with the first experiment. In addition, it was found that, on average, the clustering algorithm obtained the maximum ARI score when cosine distance was used as a distance function.

In the third experiment, lemmatization was applied to the cleaned text. The results indicated that the frequent unigram features were more effective as they obtained maximum ARI score on average. Moreover, it was found that the clustering algorithm achieved maximum ARI score on average when Euclidean distance was used for distance computation.

To summarize the experimental results presented in Table 2, clustering results were optimal when cosine distance measure was used to compute distance matrix in most experiments. Furthermore, the stemming technique was helpful in contrast to lemmatization as it achieved the highest ARI scores of 0.825, respectively, with frequent unigram features on the dataset.

The aforementioned finding is also supported through the experiments performed by (Bsoul et al., 2014) on Arabic documents. (Bsoul et al., 2014) identified that cosine distance measure produced better clustering in contrast to Euclidean distance measure on clustering task. Furthermore, the authors also highlighted that the stemming obtained better results in comparison to lemmatization on the Arabic document clustering task. Similarly, in our work, stemming yielded optimal ARI score when evaluated on Urdu News headlines clustering.

5 Conclusion

This paper evaluates the impact of stemming and lemmatization on Urdu document clustering. In addition to stemming and lemmatization, the effect of distance measures on clustering short text was also studied. The experiments were performed on the Urdu news headlines corpus. Unigram-based features and word embedding-based features were used to build a feature matrix. The results showed that the frequent unigram features yielded the highest ARI score on average. Among the four distance measure, the Cosine distance metric was more valuable. Furthermore, the stemming technique was identified to be useful in contrast to lemmatization for clustering news headlines. Several extensions to this work are planned for the future. First, we plan to conduct a similar study on Urdu tweets corpus. Second, we would experiment with the recent state-of-the-art feature extraction techniques such as contextualized word embeddings. Lastly, we intend to experiment with various other distance measures to identify the optimal distance metric for the clustering task.

References

Kavitha Karun A, Mintu Philip, and Lubna K. 2013. Comparative Analysis of Similarity Measures in Document Clustering. In *2013 International Conference on Green Computing, Communication and Conservation of Energy (ICGCE)*.

Charu C. Aggarwal, Alexander Hinneburg, and Daniel A. Keim. 2001. On the surprising behavior of distance metrics in high dimensional space. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*.

Swati Aggarwal, Nitika Agarwal, and Monal Jain. 2019. *Performance analysis of uncertain k-means clustering algorithm using different distance metrics*.volume 798. Springer Singapore.

Qurat-ul-Ain Akram, Asma Naseer, and Sarmad Hussain. 2009. Assas-Band , an affix-exception-list based Urdu stemmer . (January):40–46.

Qusay Bsoul, Eiman Al-Shamari, Masnizah Mohd, and Jaffar Atwan. 2014. Distance measures and stemming impact on arabic document clustering. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 8870:327–329.

Saboor Hamdani, Rachel Kan, Angel Chan, and Natalia Gagarina. 2020. The Multilingual Assessment Instrument for Narratives (MAIN): Adding Urdu to MAIN. *ZAS Papers in Linguistics*.

Anna Huang. 2008. Similarity measures for text document clustering. In *New Zealand Computer Science Research Student Conference, NZCSRSC 2008 - Proceedings*.

Lawrence Hubert and Phipps Arabie. 1985. Comparing partitions. *Journal of Classification*.

Safia Kanwal, Kamran Malik, Khurram Shahzad, Faisal Aslam, and Zubair Nawaz. 2019. Urdu named entity recognition: Corpus generation and deep learning applications. *ACM Transactions on Asian and Low-Resource Language Information Processing*.

Stuart P. Lloyd. 1982. Least Squares Quantization in PCM. *IEEE Transactions on Information Theory*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their

compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Atta Rahman, Khairullah Khan, Wahab Khan, Aurangzeb Khan, and Bibi Saqia. 2018. Unsupervised Machine Learning based Documents Clustering in Urdu. *ICST Transactions on Scalable Information Systems*.

R. Subhashini and V. Jawahar Senthil Kumar. 2010. Evaluating the performance of similarity measures used in document clustering and information retrieval. *Proceedings - 1st International Conference on Integrated Intelligent Computing, ICIIC 2010*:27–31.