EAI Endorsed Transactions
on Scalable Information Systems

Research Article EAI.EU

Unsupervised Machine Learning based Documents Clustering in Urdu

Atta Ur Rahman[1,*], Khairullah Khan[1], Wahab Khan[2], Aurangzeb Khan[1] and Bibi Saqia[1]

[1]Department of Computer Science, University of Science & Technology Bannu, Pakistan
[2]Department of Computer Science & Software Engineering, IIU, Islamabad 44000, Pakistan

Abstract

The volume of data on the web is growing rapidly, due to the proliferation of news sources, contents, blogs and journals etc. Like other languages, the Urdu language has also observed tremendous growth on the internet. As the volume of data is expanding, information retrieval (IR) is becoming complicated. Document clustering is an unsupervised ML approach, employed to group a huge number of dispersed documents into a small number of significant and consistent clusters, thus providing a base for indexing, IR and browsing mechanisms. Documents clustering has a long tradition in English as well as English like western languages, but Urdu lags behind in terms sophisticated natural language processing (NLP) tools and resources for documents clustering. Documents clustering becomes a challenging task in Urdu language having a rich morphology, particular structure, syntax peculiarities and cursive nature. In this study, we have developed a framework of document clustering and analysed various similarity measures for Urdu documents. We have also checked the effect of stop words removal in the process of Urdu document clustering.

Keywords: Urdu; Documents clustering; Similarity Measures; K-Means Algorithm

Received on 30 September 2018, accepted on 04 December 2018, published on 19 December 2018

doi: 10.4108/eai.19-12-2018.156081

1. Introduction

The extent of data on the cyberspace is expanding quickly due to the large-scale and rapid expansion of web technologies [1-4]. These databases are continuously upgraded for growth of documents and possess a high query stack. This unstructured information has asked a few new examinations to investigate this gigantic information, sort related data and to subsequently enhance the association of the content existing on the web [5]. Nearly every information one requests are currently accessible on the internet [6]. English and European languages have mainly dominated the web since its beginning [7]. However, in the past few years, a widespread range of information in the Indian local languages such as Urdu, Hindi, Bengali, Oriya, Tamil, and Telugu have been observed on the internet [6, 8]. The richness of data along with the vibrant and diverse nature of the Web makes Information retrieval (IR) a challenging task [9, 10]. Document clustering presents a structure for categorizing a large collection of documents [11, 12]. Document clustering is exploited to consequently find the intrinsic characteristics and native grouping amongst documents, to sort out them into various clusters [13, 14]. Documents clustering is an exciting approach, since it groups the documents exclusive of human intercession and exempts organizations from the prerequisite of manual categorization of documents, which might be an arduous and tedious process [15]. Various studies regarding document clustering, exploiting English language documents as input have been presented [16]. However, each language can generate distinct levels of exactness, depending on each natural language shapes and characteristics, like morphological and syntax peculiarities, use of antonyms and synonyms, and utilization of native expressions etc [17, 18].
Structure of this paper is organized as: section 2 highlights the importance and challenges of Urdu, section 3 describes

[*]Corresponding author. attacs9@gmail.com

un-supervised learning approach. In section 4 related work is presented, Section 5 describes proposed work, section 6 tells about the adopted unsupervised clustering algorithm while section is about the experimentation work and section 8 provides conclusion.

2. Urdu Language

Urdu is a lingua franca and national language of Pakistan [25, 26]. As per Wikipedia statistics there exist 100 million native speakers of Urdu in Pakistan and India and an additional 300 million speakers around the globe [25]. The development of computational sources is the elementary step in any Natural Language Processing (NLP) task. Urdu is broadly communicated languages of Asian sub-continent, though due to sources scarceness, not sufficient effort has been accomplished aimed at Urdu language processing [7]. The "daily Jang" stayed the leading newspaper which generates the Urdu scripts digitally in the Nastaliq script design. Currently, many Urdu journals and magazines are issued in Pakistan on daily bases. There exists a bulk quantity of tweets in the shape of Geo News, Jang News, Roznama Dunya, Dawn News, BBC, ARY, AJJ, and Abb Takk News etc. Moreover, India also distributed more than 3,000 Urdu publications on regular basis.

2.1 Challenges in Urdu Document Clustering

The great number of issues identified in Urdu language causes document clustering a difficult task. This section describes a few major constraints which diminish the execution of proposed structure.

- Resource Scarceness

A large number of complexities related with Urdu content makes it a rare dialect to be studied for NLP. A benchmark and an extensive corpus is the essential prerequisite for any NLP associated task. However there is no standard corpus accessible for Urdu language processing. The accessible Urdu NE labeled corpora are: Backer-Riaz (2002) and Emile (2003) corpus. Currently, there is not any dataset accessible for Urdu documents clustering.

- Context sensitive and Cursive Nature

In Urdu, the state of a character isn't just influenced by its position but additionally by its adjacent characters. Urdu characters have distinctive shapes at beginning, middle and end of the word. For example in the word (Love, محبت), the state of (te,ت) is changed in the word (gift,تحفہ). Thus the character "ل" has a diverse shape in the words (slave,غلا م), (Electricity, بجلی), (long, طویل), and (but, لیکن).

- Words segmentation problem

Segmentation is far problematic in Urdu dialect in light of the fact that here space is utilized for word limit. Space enclosure and exception are caused by utilization of space in Urdu content. For example space inclusion happens, such as, "حوب صورت" (hobsorat, beautiful) is a single word however because of space insertion the framework will assume it as two words like "حوب" and "صورت". Space omission issues happen, for example, "اس لیے" (aslye, therefore) is two words but because of space exclusion, the framework will consider it"اسلیے" like a single word.

- Compound Named issues

A compound named are made out of various words like (ولادی میر پوتنVladimir putin ,عمران احمد نیازی, Imran Ahmad Niazi), here both words refer to a single word but the system will consider each one as a three separate words. Such as "پوتن", "میر", "ولادی", and "نیازی", "احمد", "عمران".

- Large number of Synonyms

Urdu language possess a large number of synonyms like (جنت, heaven) has synonyms such as (فردوس، باغ ، بہشت) which create a great problem in documents clustering.

- Conjunction issues

Some entities are formed by utilizing conjunction word such as (پاکستان اور چین, Pakistan and China) and (علم و دانش, Knowledge and wisdom) etc.

- Acronym ambiguities

In English dialect acronym can be easily distinguished because of the upper casing principle, however in Urdu, it is very hard to perceive acronym, for example, (بی بی سی , BBC, سی پیک , CPEC ,یو این او, UNO) and so on.

The rest of the paper is structured as follows: Section 2 describe an extensive detailed of related work, Section 3 clarifies the proposed architecture employed for Urdu documents clustering, and Section 4 shows experimental analysis, result and evaluation metric while Section 5 concludes the paper.

3. Unsupervised Machine Learning

Supervised learning algorithm is typically used in subjective classification. This algorithm depends on manually labelled datasets and domain dependent. For that reason supervised algorithm is time consuming, required manual expertise and relatively difficult to understand a words of the human discourse. The few familiar examples of supervised learning algorithm are support vector machine (SVM), K-nearest neighbour (KNN) and Naive Bayesian classifier etc. While unsupervised algorithm working regardless of training data sets and its domain independent.

Thus, the purpose of unsupervised learning algorithm is to identify the actual classification of data and refine their structure. Some common examples of unsupervised learning algorithm are Association Rule Mining (AM), documents clustering, Likelihood Ratio Test (LRT) and K-mean clustering [19] etc.

Consequently, unsupervised algorithm is domain independent and do not required any labelled data sets over supervisor algorithm due to these two basic points we have used unsupervised algorithm for Urdu documents clustering. Since Urdu concerned to resource scarce languages and it's comparatively difficult to employed supervised technique for Urdu documents clustering.

3.1 Document Clustering

These days' data produced abundantly in the shape of news article, social networks analysis such as twitter etc, e-books, and financial analysis, etc. As predicted there exist 80% of the whole data on the Web in an unstructured fashion [19] . Conventional database query methods are not suitable to obtain a useful information from this large collection of data. Documents clustering is an unsupervised categorizing of a set of documents into self-relevant clusters such that each document is more identical to one another in the same cluster than with a document of other clusters [20] as shown in fig 1 [21]. These clusters are run-time constructed through the clustering procedure, rather than being labelled as in the instance of document classification, which is usually referred as supervised or pre-labeled categorization of documents [22, 23]. Document clustering has been utilized for various applications such as IR, indexing, surfing large document corpora, and extracting data on the cyberspace [24].

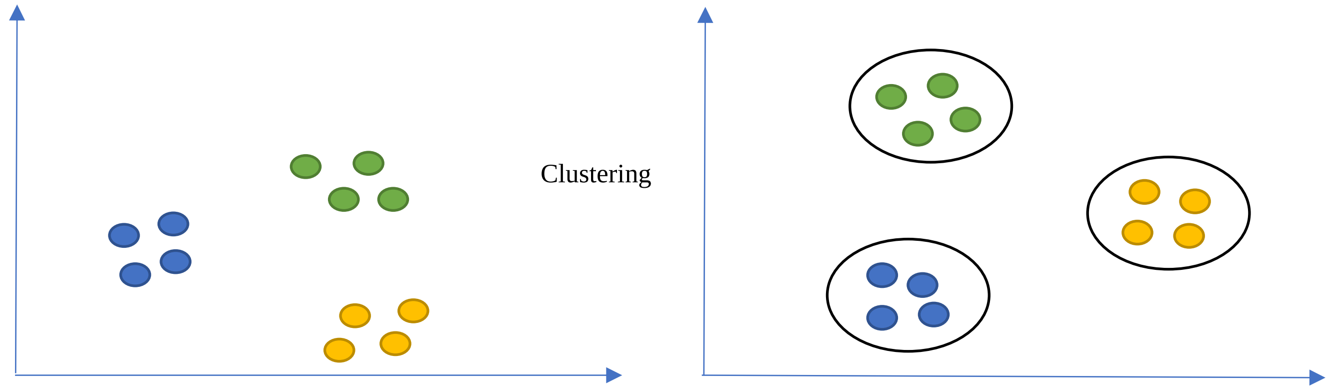

Figure 1 Documents clustering

4. Related Work

Document clustering has been broadly reviewed in data mining literature [1]. Enough research work has been explored in approving a well-organized document clustering techniques [2]. Hierarchical K-Means based clustering (HKM) is utilized for 242 Arabic documents and discover that the clustering-based IR has tremendous result over the traditional IR framework [3]. An experimental investigation has been carried out about Automated Text Clustering connected to Brazilian Portuguese text, the goal was to locate the best computational technique ready to cluster the documents [4]. Multilingual document clustering framework has been presented and tested on FIRE dataset by employing a bisect k-means algorithm [5]. Hierarchical clustering is broaden extended into divisive (top-down) and agglomerative (bottom-up) clustering [2, 6]. The divisive approach begins by taking all data objects in a unit cluster and divides them into different sub-clusters based on some splitting criterion until each data object makes a cluster of its own or some termination condition reaches [7]. The agglomerative approach begins by taking each data object as a separate cluster and combined them accordingly based on some proximity metric. The process remains continue until all data points are merged in a unit cluster or some closing condition reaches [8]. In partitional clustering, a dataset of n objects is directly decomposed into a set of K disjoint clusters, based on some optimization criterion [9]. K-means define by [10] and K-medoids describe by [11] are the two eminent algorithms of this type of clustering. A comparison has been made between k-means and k-mediods algorithms, by utilizing Arabic dataset of 242 documents. They observed that k-mediods has better performance than k-means algorithm [12]. The key idea of K-means is to revise the center of the cluster which is computed as the average point of the data objects in an iterative fashion until some closing criteria are reached [10]. K-mediods is an enhancement of K-means to deal with distinct data, which returns the adjacent points as the cluster centroids [11]. The commonly used clustering algorithms depend on partition also include PAM define by [13], CLARA describe by [14], and CLARANS introduce by [15]. In density-based clustering, the core objective of the clustering algorithm is, the document which is in the section with a prominent density of the document space is counted to fit in the similar cluster [16]. A density-based k-means algorithm is suggested to improve the performance of DBSCAN and K-means algorithms. They utilized a dataset of 250 documents and observed that DBK-means has outperforms the k-means and DBSCAN algorithms [17].

Clustering algorithm founded on density and distance is also utilized, which calculates the distance and the density of every data points and combined those data objects which have minimum distance and highest density, using a decision graph [18]. The COBWEB expresses by [19] and GMM outline by [20] depend on statistical learning and neural network. In Kernel-based clustering algorithms, the input space is converted into a feature of high dimension. The classic algorithms of this type of clustering are kernel K-means explains by [21], kernel FCM mark by [22], kernel SOM specify by [23], and SVC characterize by [24]. A clustering algorithm known as affinity propagation (AP) is offered in 2007, which relies upon "message passing" amongst information objects. In this kind of algorithm, the client can’t assign the quantity of groups as an input, such as a k-means algorithm. However, like a k-medoids, it can locate "exemplars", fellows of the input set that are illustrative of clusters [25]. Various strategies have been acquired to achieve semantic correlations amongst documents [26]. A famous tool such as WordNet has been utilized to improve the semantic association amongst words, such as synonyms etc [27]. Additional ontology made research are also incorporated [28, 29], which focuses on words semantic relationship. Chinese news-based clustering approach is proposed by utilizing a Neural network language model [30]. K-nearest neighbour, k-means and support vector machine are employed for Marathi news clustering [31]. Agglomerative hierarchical clustering is proposed for Urdu ligature recognition and they also utilized Naïve Bayes, decision tree, K-nearest neighbour and linear discernment analysis for classification [32]. A detailed study on Urdu document images has been conducted by utilizing various clustering algorithms such as Self organizing map, K-means and hierarchical clustering [33]. Urdu ligatures organization is accomplished using a deep neural network. They exploit a corpus of 2430 ligatures and achieved an accuracy of 73.13 % [34]. Table 1 shows the most related work about Urdu document clustering.

Table 1 Summaries of Related Work

Study	**Application**	**Clustering Algorithm**	**Dataset**	**Result**	**Language**
(Ghwanmeh, 2007)	Information Retrieval System	Hierarchical K-means (HKM), Traditional IR	242 Arabic documents	Precision of HKM for 2 cluster 0.62 and for 5 clusters 0.59, Traditional IR 0.49	Arabic
(Mumtaz & Duraiswamy, 2010)	Document Clustering	DBSCAN, K-means, and DBK-means	250 documents	Rand Index Such as DBSCAN 0.37, K-means 0.60, and DBK-means 0.73	English
(Kumar, Santosh, & Varma, 2011)	Multilingual Document Clustering	Bisecting K-means	FIRE Dataset	F measure 0.57 and Purity 0.68	English and Hindi
(Alkoffash, 2012)	Arabic text clustering	k-means and k-mediods	242 documents	Precision of K-means 0.56 and K-mediods 0.69	Arabic
(Afonso & Duque, 2014)	Clustering of Newspaper and Scientific Text	K-means, sIB and EM	Scientific and Newspaper Corpus of 36 documents	sIB correctness, Scientific 77.8 % and Newspaper 68.9 % , EM 53 %	Brazilian Portuguese
(Fan, Chen, Zha, & Yang, 2016)	A Chinese news based Clustering Approach	Neural network language model	Size of data 600 MB	F measure 0.93	Chinese
(Dangre, Bodke, Date, Rungta, & Pathak, 2016)	System for Marathi News Clustering	K-means, KNN, and SVM	Marathi Text	Not Available	Marathi
(Khan, Adnan, & Basar, 2017)	Multi-level Agglomerative Clustering for	Decision Tree, LDA, Naïve	A corpus of 2430 ligature	Accuracy of Decision Tree 62, LDA 61,	Urdu

	Ligature Recognition	Bayes, and KNN		Naïve Bayes 73, and KNN 90% respectively	
(Shabbir, Javed, Siddiqi, & Khurshid, 2017)	Clustering Technique for Urdu Ligature	K-means, SOM and Hierarchical Clustering	Scanning Images of Urdu book ‘Zawiya’	Calinski-Harabasz Index ,Davies-Bouldin Index, and Dunn Index are calculated such as K-means 13503, 1.0841, 0.0085, Hierarchical clustering 741, 0.75, 0.07 and SOM 10343, 1.10, and 0.002 respectively	Urdu
(Rafeeq, ur Rehman, Khan, Khan, & Jadoon, 2018)	Ligature Categorization	Deep Neural Network,	A corpus of 2430 ligature	Accuracy, Neural Network 73.13%	Urdu

5. Proposed Architecture

The demand for Urdu document clustering turns out to be necessary because this data is growing to be very popular on the cyberspace. The frequently expanding use of documents clustering and the extensive span of its appliances directed us to accomplish an investigational analysis of Urdu documents clustering on the basis of various similarity measures. The proposed methodology cconsists of the following steps; Step 1 Data collection and Pre-processing, Step 2 Document representations as a Bag of Words Model Step 3 Similarity measurements and Step 4 Documents clustering using the k-means algorithm as shown in figure 2.

5.1 Dataset

The presence of benchmark dataset is required for every natural language processing task [7]. It is essential to provide a significant extent of the pre-labeled dataset to train the model effectively. However Urdu language droughts in owning such a linguistic resources for natural language processing task [59, 60]. The datasets exploited in this study is collected from BBC Urdu news portal http://www.bbc.com/urdu and saved it in Notepad in UTF-8 design. This dataset comprised of 1000 documents of five distinct classes, such as Arts, International, National,

Sports, and scientific news while each document contains different number of sentences and tokens as shown in table 2.

Table 2 Consolidated Statistics of Dataset

Document Type	**Number of document**	**Number of sentence**	**Number of tokens**
Arts News	200	3800	57726
International News	200	3300	55321
National News	200	3900	60781
Scientific News	200	3100	58934
Sports News	200	3700	63104

5.2 Pre-processing

Data pre-processing is an essential phase incorporated before executing any NLP, IR, and data mining task [7]. The setting of every Natural Language contents comprises of two kinds of words such as functional words and contented words. The contented words present lexical meaning while the functional words provide a syntactic role [25]. The stop words belong to the category of functional words. Stop words exclusion offer a total cut of 30% in the file size and indexing of documents [61]. This is typically realized that stop words did not provide any lexical implication to the contents, however their greater existence triggers an obstruction in the documents clustering. Mostly, every sentence comprises of contented words and stop words. Here we represent contented words as a key words as shown in Table 3.

Table 3 Example of Urdu Stop words and Key words

Urdu Documents	Key words	Stop words
جدید تحقیق کے مطابق وٹامن ڈی کی گولیاں ہڈیوں کی صحت بہتر نہیں کرتیں اور نہ ہی ان سے ہڈی ٹوٹنے سے بچاؤ میں مدد ملتی ہے-	جدید، تحقیق، وٹامن ڈی، گولیاں، ہڈیوں، صحت، ، ہڈی، ٹوٹنے، بچاؤ، مدد	کے، مطابق، کی، بہتر ، نہیں، کرتیں، اور، نہ، ہی، اان، سے، میں، ملتی، ہے
ایک نئی تحقیق میں کہا گیا ہے کہ دوسری جنگ عظیم میں اتحادی افواج کی جانب سے جو بم برسائے گئے وہ اتنے طاقتور تھے کہ ان سے کرۂ ہوائی کو نقصان پہنچا ہے-	تحقیق، دوسری جنگ عظیم، اتحادی افواج، بم، برسائے، طاقتور، کرۂ ہوائی، نقصان، پہنچا	ایک ، نئی، میں، کہا، گیا، ہے، کہ، کی، جانب، سے، جو، گئے، وہ، اتنے، تھے، ان، کو،
برطانیہ میں قائم لیورپول یونیورسٹی کے پروفیسر فرانسس مکلیون نے بتایا کہ 'مچھر یا دیگر کیڑے مکوڑے اور پودے انسانی جلد پر ٹاکسن یعنی حیوانی یا نباتاتی زہر چھوڑتے ہیں- جسکی وجہ سے اعصاب کی جانب سے دماغ کو خارش کا سگنل ملتا ہے اور ہم خارش کرنے لگتے ہیں-	برطانیہ، قائم، لیورپول یونیورسٹی، پروفیسر، فرانسس مکلیون، مچھر، کیڑے مکوڑے، پودے، انسانی، جلد، ٹاکسن، حیوانی، نباتاتی، زہر، چھوڑتے، اعصاب، دماغ، خارش، سگنل	میں، کے، نے، بتایا، کہ، یا، دیگر، اور، پر، یعنی، ہیں، جسکی، وجہ، سے، کی، جانب، کو، کا، ملتا، ہے، ہم، کرنے، لگتے

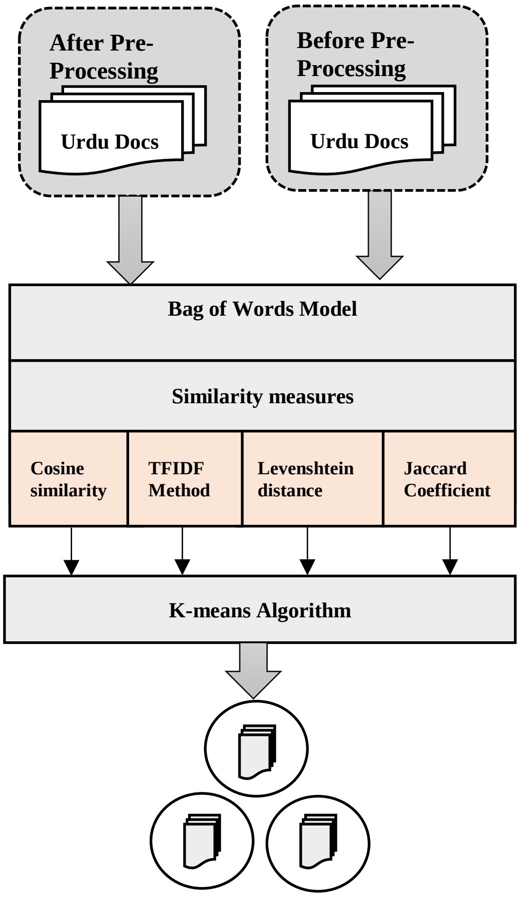

Figure 2 Proposed Architecture

5.3 Bag-of-Words Model

We have represented each document as a Bag-of-words model in this study. This technique is employed for name entity extraction, opinion targets and documents clustering. The Bag-of-words model make good use of Term Frequency Inverse Document Frequency for document clustering and describes their frequencies without any contextual as well syntactic association of words in documents. The Bag-of words model is well known and frequently used method for object classification, information retrieval (IR), similarity measuring and natural language processing (NLP).This model extract a document as pack of its words ignoring word sequence. [62].

5.4 Similarity Measures

Document clustering needs a correct description of the proximity amongst a set of documents, concerning of both, the pairwise likeness or space [63]. The similarity measure is utilized to implicitly capture the alikeness amongst documents or records and allocates it a specific value in the range of 0 to 1 [21]. The documents preserve to be similar in two different ways; either lexically or semantically. They are lexically comparable, if contain the uniform dictionary words, although semantically related if they illustrate the identical concept [64, 65]. Still up-to date there is no agreed similarity metric that is suitable for long range of clustering practices [66]. This section describes some frequently used similarity metrics, utilized in the development of documents clustering.

5.4.1 Cosine Similarity

Cosine similarity treat each document as a terms or features vector and the likeness of the documents is calculated as the cosine of the angle amongst them [65-67]. The terms or words of the document are known as the features or dimension of the documents [30]. Cosine similarity is a widely used similarity metric in a different area of IR, such as clustering, classification and pattern recognition etc. Mathematically it can be described as

$$\mathrm{Cosine_{sim}}(\mathrm{doc_i}, \mathrm{doc_j}) = \frac{\mathrm{doc_i} \ . \ \mathrm{doc_j}}{\|\mathrm{doc_i}\| \ \times \ \|\mathrm{doc_j}\|} \tag{1}$$

Where doc i and doc j represent random documents.

Let we have two documents such as doc 1 and doc 2 as shown in table 4, then its Cosine similarity would be calculated as shown in table 5.

5.4.2 TF-IDF (Term Frequency Inverse Document Frequency)

Numerous term weighting methods have been suggested to compute the weight of a terms in a specific document and in the entire corpus. However, TF-IDF is the highly utilized term weighting scheme. It determines the weight of a term by its occurrence inside a document and the inverse of its document occurrences within the corpus [68]. In practice, the terms which arise repeatedly in a few documents but infrequently in the remaining documents incline to exist more significant for that individual set of documents [30]. Mathematically TF-IDF can be represented as

$$(\mathrm{Tf} - \mathrm{Idf})_{\mathrm{td}} = \mathrm{Tf_{td}} * \log(\frac{\mathrm{n}}{\mathrm{df_{td}}}) \tag{2}$$

Here $\mathrm{Tf_{td}}$ indicates term frequency, $\mathrm{Idf_{td}}$ represents inverse document frequency and $(\mathrm{Tf} - \mathrm{Idf})_{\mathrm{td}}$ express the total aggregate of term t inside document d [65, 69]. We have found the Tf-Idf of doc 1 and doc 2 as shown in table 6.

5.4.3 Levenshtein distance

The Levenshtein distance is utilized to find the character based likeness between strings or documents [64]. The closeness between two strings is computed by finding the number of activity or operations executed such as insertion, deletion, or substitution required to change one strings into another strings [70]. The Levenshtein distance between two strings s and t where s demonstrates a source string and t represents a target strings can be computed as

$$lev_{s,t}(i,j) = \begin{cases} \max(i,j) & if \min(i,j) = 0 \\ \min(i,j) \begin{cases} lev_{s,t}(i-1,j) + 1 \\ lev_{s,t}(i,j-1) + 1 \\ lev_{s,t}(i-1,j-1) + 1_{s \neq t} \end{cases} & otherwise \end{cases} \tag{3}$$

Where i represents the first character of s and j represents t The Levenshtein distance between two strings (درحواست, application) and (درست, right) can be calculated as shown in table 7, 8 and 9 respectively.

5.4.4 Jaccard Coefficient

The Jaccard coefficient also stated as the Tanimoto coefficient, calculates the similarity of the two documents such as, the sum of the weight of common terms is divided to the sum of the weight of those terms that are existing in any of the two document but are not the common terms [71-73]. It can be mathematically described as

$$\mathrm{Jaccard}(\mathrm{d_i}, \mathrm{d_j}) \ = \frac{\mathrm{T_i} \cap \mathrm{T_j}}{(\mathrm{T_i} \cup \mathrm{T_j}) - (\mathrm{T_i} \cap \mathrm{T_j})} \tag{4}$$

Here $\mathrm{T_i}$ and $\mathrm{T_j}$ represent the terms of documents $\mathrm{d_i}$ and $\mathrm{d_j}$ correspondingly. The Jaccard Coefficient of doc 1 and doc 2 would be calculated as shown in table 10.

Table 4 Urdu Documents

وزیراعظم عمران خان آج خیبرپختونخوا کا دورہ کریں گے	doc 1
وزیراعظم عمران خان گورنراور وزیراعلیٰ خیبرپختونخوا سے بھی ملاقات کریں گے	doc 2

Table 5 Cosine Similarity

Terms	ملاقات	بھی	سے	وزیراعلیٰ	اور	گورنر	گے	کریں	دورہ	کا	خیبرپختونخوا	آج	عمران خان	وزیراعظم	Cosine similarity = 0.57
doc 1	0	0	0	0	0	0	1	1	1	1	1	1	1	1	
doc 2	1	1	1	1	1	1	1	1	0	0	1	0	1	1	
doc 1.doc 2 = 5				doc 1_SQ 7.0 = 2.64			doc 2_SQ 11 = 3.31				Cosine = doc 1.doc 2/ doc1_SQ x doc 2_SQ				

Table 6 TF-IDF

Terms	ملاقات	بھی	سے	وزیراعلیٰ	اور	گورنر	گے	کریں	دورہ	کا	خیبرپختونخوا	آج	عمران خان	وزیراعظم
doc 1	0	0	0	0	0	0	1	1	1	1	1	1	1	1
doc 2	1	1	1	1	1	1	1	1	0	0	1	0	1	1
Idf	0.47	0.47	0.47	0.47	0.47	0.47	0.30	0.30	0.47	0.47	0.30	0.47	0.30	0.30
Tf-Idf doc 1	0	0	0	0	0	0	0.30	0.30	0.47	0.47	0.30	0.47	0.30	0.30
Tf-Idf doc 2	0.47	0.47	0.47	0.47	0.47	0.47	0.30	0.30	0	0	0.30	0	0.30	0.30

Table 7 Representation of String in Levenshtein distance

Source String: درحواست	String representation						
	ت	س	ا	و	ح	ر	د
Target String: درست	ت	س	ر	د			

By applying Levenshtein distance we will performed three deletion operations such as (delt 1, delt 2 and delt 3 on highlighted characters) to convert source string into target string.

Table 8 Applying Deletion Operation of Strings

ت	س	ا	و	ح	ر	د
ت	س	ر	د			

After applying Levenshtein distance we will obtained the strings such as

Table 9 Levenshtein Distance

Source String = Target String: درست	ت	س	ر	د

Levenshtein distance = 3

Table 10 Jaccard Coefficient

Terms	ملاقات	بھی	سے	وزیراعلیٰ	اور	گورنر	گے	کریں	دورہ	کا	خیبرپختونخوا	آج	عمران خان	وزیراعظم	Jaccard Coefficient = 0.38
doc 1	0	0	0	0	0	0	1	1	1	1	1	1	1	1	
doc 2	1	1	1	1	1	1	1	1	0	0	1	0	1	1	
doc1 ∩ doc2 = 5				doc 1 ∪ doc 2 = 7+11= 18					Jaccard Coefficient = 5/18-5 = 0.38						

6. Clustering Algorithm

6.1 K-means Algorithm

This algorithm was first offered by Stuart Lloyd in 1957 while the term “K-means” was first utilized by James MacQueen in 1967 [74]. K-means is an unsupervised algorithm which groups the given information index into a particular set of clusters [75]. A centroid-based methodology is exploited in this type of algorithm which can be determined through the average of objects allocated to the clusters [2]. The aim of the K-Means algorithm is to reveal the finest separation of n data points in k number of clusters to such an extent that the aggregate separation amongst the data points and its relating centroids are minimized [76].

Basic Steps are listed below:

- Input: A set of numbers such as N= $\{ n_i , i = 1, \dots m\}$ and k.
- Output: A group of numbers into k clusters.
- Initiate by randomly selecting k number of center such as c_k.
- Assign each number to the cluster G_k, which has minimum distance d (n_i, c_k).
- Recalculate each c_k as the mean of all numbers of G_k.
- Repeat step 4 and step 5 until the centroids c_k and group members G_k no longer change.

7. Experiment

A dataset of 1000 documents comprising of five distinct classes, as shown in table 2 are crawled from BBC Urdu News Portal. This dataset is then pre-processed for stop words removal. After pre-processing, Bag of words model and similarity measures are employed to analyze the closeness of each document. We have achieved the similarity of each document in the form of numeric value ranging from 0 to 1. These values are then passed to the K-means algorithm as an input for clustering. The K-means algorithm cluster the dataset. This clustering result is then compared with the manually classified clusters.

7.1 Results and Discussion

The physically made classification is ordinarily utilized as a standard basis for assessing the result of documents clustering. Hence, the groups of documents formed utilizing various similarity measures and K-means algorithm are compared with the manually classified documents clusters. This kind of assessment accepts that the objective of grouping is to imitate human reasoning. A grouping arrangement is reasonable if the groups are persistent with the physically generated clusters.

7.2 Evaluation Metric

In the proposed framework, a purity metric is employed to evaluate the result. The purity metric calculates the consistency of a group or cluster such that the extent to which a cluster includes largest documents from a specific group [65]. Assume a certain cluster Ci of extent n_i , then the purity of Ci is mathematically expressed as

$$P(C_i) = 1/n_i \max(n_i^h) \qquad (5)$$

Where $\max(n_i^h)$ represent the set of documents that belongs to the dominant group. The purity values of each similarity measures in each cluster, utilizing a K-means algorithm are shown in table 11.

Table 11 Result of Similarity Measures before Pre-Processing

Number of Clusters	**Similarity Measures**			
	Cosine Similarity	TF-IDF	Levenshtein distance	Jaccard Coefficient
Cluster 1	0.45	1	1	0.05
Cluster 2	0.60	0.20	0.35	0.95
Cluster 3	0.40	0.65	1	0.75
Cluster 4	0.85	0.25	0.95	0.80
Cluster 5	1	0.1	0.05	1
Average	0.66	0.44	0.67	0.71

Table 11 described the result of five clusters by different techniques of similarity measures before pre-processing. The average accuracy of Cosine Similarity, TF-IDF, Levenshtein distance and Jaccard Co-efficient is 0.66 0.44, 0.67 and 0.71 respectively.

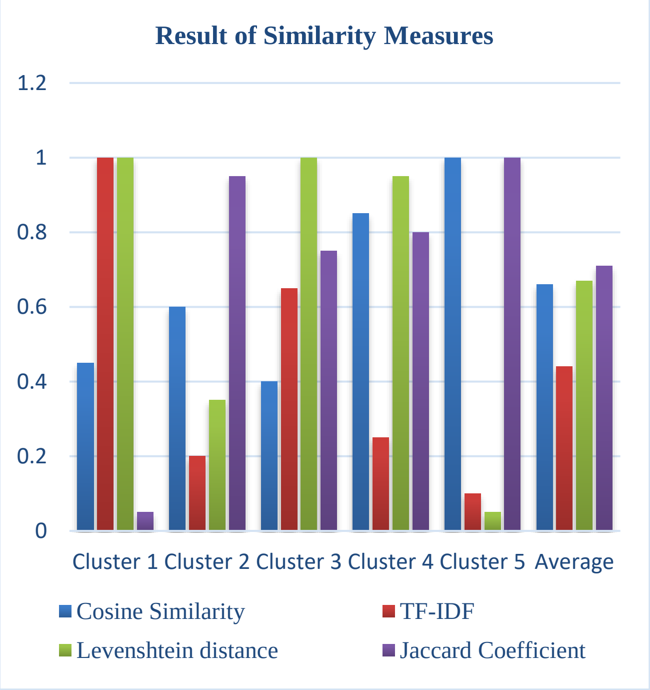

Figure 3 Result of Similarity Measures Before Pre-Processing

Figure 3 demonstrates the average result of five different clusters through several similarity measures techniques before pre-processing.

Table 12 Result of Similarity Measures after Pre-Processing

Number of Clusters	**Similarity Measures**			
	Cosine Similarity	TF-IDF	Levenshtein distance	Jaccard Coefficient
Cluster 1	0.55	1	0.90	0.95
Cluster 2	0.45	0.20	1	1
Cluster 3	0.95	0.65	0.15	1
Cluster 4	0.95	0.25	1	0
Cluster 5	1	0.1	0	0
Average	0.78	0.44	0.61	0.59

Table 12 shows the average result of five various clusters after pre-processing via different techniques of similarity measures. . The average accuracy of Cosine Similarity, TF-IDF, Levenshtein distance and Jaccard Co-efficient is 0.78 0.44, 0.61 and 0.59 correspondingly

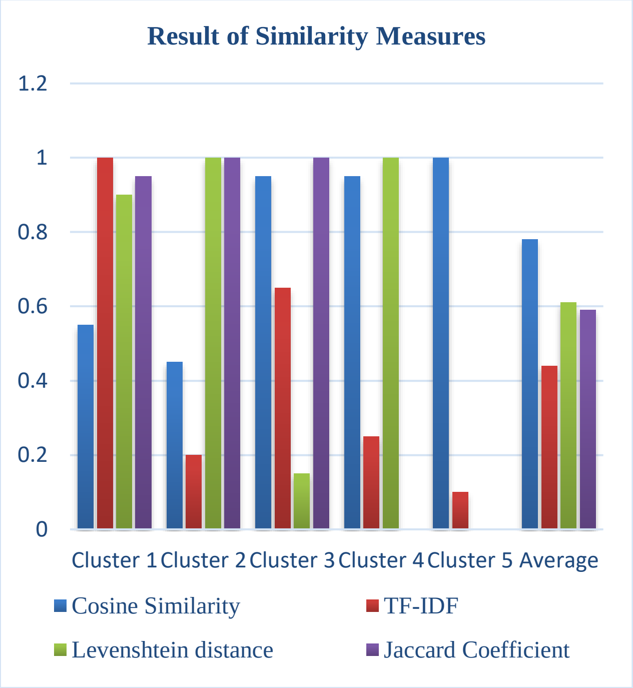

Figure 4 Result of Similarity Measure after Pre-processing

Figure 4 represents the average result of five clusters by utilized four various similarity measures after pre-processing.

8. Conclusion

Now a days, the progressive feelers that are widely adopted around the globe for the development of NLP framework, in almost all languages including Urdu, are machine learning approaches. The core reason behind its wide usage is based on four features: a) the capability of automatic learning b) the degree of accuracy c) the speed of processing and d) generic nature. Document clustering which aims to organize a huge number of documents distributed over different sites is well-investigated task from ML perspectives in Western language when compared to Urdu. The peculiarities of Urdu such as lack of resources, rich morphology, lack of capitalization and many other tasks makes Urdu document clustering more complex when compared with the language having script writing style from left to right. In this study, we attempted to propose the ever first adoption of ML approach e.g. the K-Mean clustering model with various similarities measure for Urdu documents. The Four similarity/distance measures which we experimentally analyzed in this study are: Cosine similarity, Jaccard coefficient, Levenshtein distance and TF-IDF. After conducting glare experiments, we observed that each similarity measures have a remarkable impact on Urdu documents clustering except for the TF-IDF measure. We obtained the purity values for each similarity measure, such as Cosine 0.66, Jaccard Coefficient 0.71, and TF-IDF 0.44, and Levenshtein distance 0.67 respectively.

Additionally, we also analyze the impact of stop words removal in the process of Urdu document clustering. We obtained the purity value of 0.78 for Cosine, 0.61 for Levenshtein, 0.59 for Jaccard Coefficient and 0.44 for TF-IDF respectively. The result obtained indicate that Jaccard Coefficient (0.71) before stop words removal and Cosine similarity (0.78) after stop words removal outperforms the remaining similarity measures in Urdu documents clustering. We also found that the outcomes of Levenshtein distance (0.67, 0.61) before and after pre-processing are notable than the outcomes of TF-IDF (0.44) respectively. In future work, we will likely to apply semantics similarity for document matching to demonstrate the relationship among documents more effectively.

References

[1] K. Kumar, G. Santosh, and V. Varma, "Multilingual document clustering using wikipedia as external knowledge," in *Information Retrieval Facility Conference*, 2011, pp. 108-117.

[2] A. K. Jain, "Data clustering: 50 years beyond K-means," *Pattern recognition letters,* vol. 31, pp. 651-666, 2010.

[3] M. Peng, J. Zhu, H. Wang, X. Li, Y. Zhang, X. Zhang, *et al.*, "Mining Event-Oriented Topics in Microblog Stream with Unsupervised Multi-View Hierarchical Embedding," *ACM Transactions on Knowledge Discovery from Data (TKDD),* vol. 12, p. 38, 2018.

[4] M. Peng, J. Zhu, X. Li, J. Huang, H. Wang, and Y. Zhang, "Central topic model for event-oriented topics mining in microblog stream," in *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, 2015, pp. 1611-1620.

[5] H. M. Alghamdi and A. Selamat, "Arabic Web page clustering: A review," *Journal of King Saud University-Computer and Information Sciences,* 2017.

[6] M. S. Husain and I. Siraj, "ALanguage Independent Approach To Develop Urdu IR System," *Computer Science and Information Technology (CS & IT),* vol. 10, pp. 397-406, 2013.

[7] A. Daud, W. Khan, and D. Che, "Urdu language processing: a survey," *Artificial Intelligence Review,* vol. 47, pp. 279-311, 2017.

[8] M. Madankar, M. Chandak, and N. Chavhan, "Information retrieval system and machine translation: a review," *Procedia Computer Science,* vol. 78, pp. 845-850, 2016.

[9] N. Oikonomakou and M. Vazirgiannis, "A review of web document clustering approaches," in *Data mining and knowledge discovery handbook*, ed: Springer, 2005, pp. 921-943.

[10] H. Anaya-Sánchez, A. Pons-Porrata, and R. Berlanga-Llavori, "A document clustering algorithm for discovering and describing topics," *Pattern Recognition Letters,* vol. 31, pp. 502-510, 2010.

[11] J. Ghosh and A. Strehl, "Similarity-based text clustering: A comparative study," in *Grouping Multidimensional Data*, ed: Springer, 2006, pp. 73-97.

[12] L. Liu, J. Kang, J. Yu, and Z. Wang, "A comparative study on unsupervised feature selection methods for text clustering," in *Natural Language Processing and Knowledge Engineering, 2005. IEEE NLP-KE'05. Proceedings of 2005 IEEE International Conference on*, 2005, pp. 597-601.

[13] M. Allahyari, S. Pouriyeh, M. Assefi, S. Safaei, E. D. Trippe, J. B. Gutierrez, *et al.*, "A brief survey of text mining: Classification, clustering and extraction techniques," *arXiv preprint arXiv:1707.02919*, 2017.

[14] C. C. Aggarwal and C. X. Zhai, "a survey of text clustering algorithms, Mining text data (2012) 77–128," *Google Scholar,* 2012.

[15] K. Premalatha and A. Natarajan, "A literature review on document clustering," *Information Technology Journal,* vol. 9, pp. 993-1002, 2010.

[16] S. Montalvo, R. Martínez, A. Casillas, and V. Fresno, "Multilingual news document clustering: two algorithms based on cognate named entities," in *International Conference on Text, Speech and Dialogue*, 2006, pp. 165-172.

[17] A. R. Afonso and C. G. Duque, "Automated text clustering of newspaper and scientific texts in brazilian portuguese: analysis and comparison of methods," *JISTEM-Journal of Information Systems and Technology Management,* vol. 11, pp. 415-436, 2014.

[18] M. H. Ahmed and S. Tiun, "K-means based algorithm for islamic document clustering," in *Proceedings of International Conference on Islamic Applications in Computer Science and Technologies (IMAN 2013)*, pp. 2-9.

[19] B. Saqia, K. Khan, A. Khan, W. Khan, F. Subhan, and M. Abid, "Impact of Anaphora Resolution on Opinion Target Identification," *INTERNATIONAL JOURNAL OF ADVANCED COMPUTER SCIENCE AND APPLICATIONS,* vol. 9, pp. 230-236, 2018.

[20] H. Xia, S. Wang, and T. Yoshida, "A modified ant-based text clustering algorithm with semantic similarity

measure," *Journal of systems science and systems engineering*, vol. 15, pp. 474-492, 2006.

[21] S. M. Weiss, N. Indurkhya, T. Zhang, and F. Damerau, *Text mining: predictive methods for analyzing unstructured information*: Springer Science & Business Media, 2010.

[22] P. Bide and R. Shedge, "Improved Document Clustering using k-means algorithm," in *Electrical, Computer and Communication Technologies (ICECCT), 2015 IEEE International Conference on*, 2015, pp. 1-5.

[23] M. Peng, J. Huang, Z. Sun, S. Wang, H. Wang, G. Zhuo, *et al.*, "Improving distant supervision of relation extraction with unsupervised methods," in *International Conference on Web Information Systems Engineering*, 2016, pp. 561-568.

[24] R. Ali, M. A. Khan, M. Bilal, and I. Rabbi, "Reciprocal anaphora resolution in pashto discourse," in *Emerging Technologies, 2008. ICET 2008. 4th International Conference on*, 2008, pp. 1-5.

[25] K. Riaz, "Baseline for Urdu IR evaluation," in *Proceedings of the 2nd ACM workshop on Improving non english web searching*, 2008, pp. 97-100.

[26] M. Usman, Z. Shafique, S. Ayub, and K. Malik, "Urdu Text Classification using Majority Voting," *International Journal of Advanced Computer Science and Applications*, vol. 7, pp. 265-273, 2016.

[27] S. Karol and V. Mangat, "Evaluation of text document clustering approach based on particle swarm optimization," *Open Computer Science*, vol. 3, pp. 69-90, 2013.

[28] S. H. Ghwanmeh, "Applying Clustering of hierarchical K-means-like Algorithm on Arabic Language," *International Journal of Information Technology IJIT*, vol. 3, pp. 168-172, 2007.

[29] S. Al-Anazi, H. AlMahmoud, and I. Al-Turaiki, "Finding similar documents using different clustering techniques," *Procedia Computer Science*, vol. 82, pp. 28-34, 2016.

[30] H. Froud and A. Lachkar, "Agglomerative hierarchical clustering techniques for arabic documents," in *Advances in Computational Science, Engineering and Information Technology*, ed: Springer, 2013, pp. 255-267.

[31] J. Prakash and P. K. Singh, "Partitional algorithms for hard clustering using evolutionary and swarm intelligence methods: a survey," in *Proceedings of Seventh International Conference on Bio-Inspired Computing: Theories and Applications (BIC-TA 2012)*, 2013, pp. 515-528.

[32] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, 1967, pp. 281-297.

[33] H.-S. Park and C.-H. Jun, "A simple and fast algorithm for K-medoids clustering," *Expert systems with applications*, vol. 36, pp. 3336-3341, 2009.

[34] M. S. Alkoffash, "Automatic Arabic Text Clustering using K-means and K-mediods," *International Journal of Computer Applications*, vol. 51, 2012.

[35] L. Kaufman and P. J. Rousseeuw, "Partitioning around medoids (program pam)," *Finding groups in data: an introduction to cluster analysis*, pp. 68-125, 1990.

[36] L. Kaufman and P. J. Rousseeuw, *Finding groups in data: an introduction to cluster analysis* vol. 344: John Wiley & Sons, 2009.

[37] R. T. Ng and J. Han, "CLARANS: A method for clustering objects for spatial data mining," *IEEE transactions on knowledge and data engineering*, vol. 14, pp. 1003-1016, 2002.

[38] H. P. Kriegel, P. Kröger, J. Sander, and A. Zimek, "Density-based clustering," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 1, pp. 231-240, 2011.

[39] K. Mumtaz and K. Duraiswamy, "A novel density based improved k-means clustering algorithm–Dbkmeans," *International Journal on computer science and Engineering*, vol. 2, pp. 213-218, 2010.

[40] A. Rodriguez and A. Laio, "Clustering by fast search and find of density peaks," *Science*, vol. 344, pp. 1492-1496, 2014.

[41] D. H. Fisher, "Knowledge acquisition via incremental conceptual clustering," *Machine learning*, vol. 2, pp. 139-172, 1987.

[42] C. E. Rasmussen, "The infinite Gaussian mixture model," in *Advances in neural information processing systems*, 2000, pp. 554-560.

[43] B. Schölkopf, A. Smola, and K.-R. Müller, "Nonlinear component analysis as a kernel eigenvalue problem," *Neural computation*, vol. 10, pp. 1299-1319, 1998.

[44] Z.-d. Wu, W.-x. Xie, and J.-p. Yu, "Fuzzy c-means clustering algorithm based on kernel method," in *Computational Intelligence and Multimedia Applications, 2003. ICCIMA 2003. Proceedings. Fifth International Conference on*, 2003, pp. 49-54.

[45] D. MacDonald and C. Fyfe, "The kernel self-organising map," in *Knowledge-Based Intelligent Engineering Systems and Allied Technologies, 2000. Proceedings. Fourth International Conference on*, 2000, pp. 317-320.

[46] A. Ben-Hur, D. Horn, H. T. Siegelmann, and V. Vapnik, "Support vector clustering," *Journal of machine learning research*, vol. 2, pp. 125-137, 2001.

[47] B. J. Frey and D. Dueck, "Clustering by passing messages between data points," *science*, vol. 315, pp. 972-976, 2007.

[48] S. Arch-int, "Web document clustering using semantic link analysis," in *Computational Intelligence for Modelling, Control and Automation, 2005 and International Conference on Intelligent Agents, Web Technologies and Internet Commerce, International Conference on*, 2005, pp. 13-18.

[49] U. Sridevi and N. Nagaveni, "Semantically enhanced document clustering based on pso algorithm," *European Journal of Scientific Research*, vol. 57, pp. 485-493, 2011.

[50] L. Jing, L. Zhou, M. K. Ng, and J. Z. Huang, "Ontology-based distance measure for text clustering," in *Proc. of SIAM SDM workshop on text mining, Bethesda, Maryland, USA*, 2006.

[51] U. Sridevi and N. Nagaveni, "Ontology based semantic measures in document similarity ranking," in *Advances in Recent Technologies in Communication and Computing, 2009. ARTCom'09. International Conference on*, 2009, pp. 482-486.

[52] Z. Fan, S. Chen, L. Zha, and J. Yang, "A text clustering approach of Chinese news based on neural network language model," *International Journal of Parallel Programming*, vol. 44, pp. 198-206, 2016.

[53] M. Peng, Q. Xie, H. Wang, Y. Zhang, and G. Tian, "Bayesian sparse topical coding," *IEEE Transactions on Knowledge and Data Engineering*, 2018.

[54] M. Peng, Q. Xie, Y. Zhang, H. Wang, X. Zhang, J. Huang, *et al.*, "Neural sparse topical coding," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018, pp. 2332-2340.

[55] N. Dangre, A. Bodke, A. Date, S. Rungta, and S. Pathak, "System for Marathi News Clustering," *Procedia Computer Science*, vol. 92, pp. 18-22, 2016.

[56] N. H. Khan, A. Adnan, and S. Basar, "Urdu ligature recognition using multi-level agglomerative hierarchical clustering," *Cluster Computing*, pp. 1-12, 2017.

[57] S. Shabbir, N. Javed, I. Siddiqi, and K. Khurshid, "A comparative study on clustering techniques for Urdu ligatures in nastaliq font," in *Emerging Technologies (ICET), 2017 13th International Conference on*, 2017, pp. 1-6.

[58] M. J. Rafeeq, Z. ur Rehman, A. Khan, I. A. Khan, and W. Jadoon, "Ligature categorization based Nastaliq Urdu recognition using deep neural networks," *Computational and Mathematical Organization Theory*, pp. 1-12, 2018.

[59] W. Khana, A. Daudb, J. A. Nasira, and T. Amjada, "Named Entity Dataset for Urdu Named Entity Recognition Task," *LANGUAGE & TECHNOLOGY*, p. 51.

[60] K. Riaz, "Rule-based named entity recognition in Urdu," in *Proceedings of the 2010 named entities workshop*, 2010, pp. 126-135.

[61] B. S. Aqil Burney, N. Mahmood, Z. Abbas, and K. Rizwan, "Urdu Text Summarizer using Sentence Weight Algorithm for Word Processors," *International Journal of Computer Applications*, vol. 46, pp. 38-43, 2012.

[62] D. M. El-Din, "Enhancement bag-of-words model for solving the challenges of sentiment analysis," *International Journal of Advanced Computer Science and Applications*, vol. 7, pp. 244-247, 2016.

[63] N. Sandhya, Y. S. Lalitha, V. Sowmya, D. K. Anuradha, and D. A. Govardhan, "Analysis of stemming algorithm for text clustering," *IJCS*, vol. 8, pp. 1694-0814, 2011.

[64] W. H. Gomaa and A. A. Fahmy, "A survey of text similarity approaches," *International Journal of Computer Applications*, vol. 68, 2013.

[65] A. Huang, "Similarity measures for text document clustering," in *Proceedings of the sixth new zealand computer science research student conference (NZCSRSC2008), Christchurch, New Zealand*, 2008, pp. 49-56.

[66] M. R. Masuma and V. Losarwar, "A Similarity Measure for Text Processing," *International Journal for Research in Engineering Application & Management (IJREAM)*, vol. 02, pp. 1-6, 2016.

[67] J.-Y. Jiang, W.-H. Cheng, Y.-S. Chiou, and S.-J. Lee, "A similarity measure for text processing," in *Machine Learning and Cybernetics (ICMLC), 2011 International Conference on*, 2011, pp. 1460-1465.

[68] A. K. Murugesan and B. J. Zhang, "A new term weighting scheme for document clustering," in *7th Int. Conf. Data Min.(DMIN 2011-WORLDCOMP 2011), Las Vegas, Nevada, USA*, 2011.

[69] N. Sandhya, Y. S. Lalitha, A. Govardhan, and K. Anuradha, "Analysis of similarity measures for text clustering," *International Journal of Data Engineering*, vol. 2, 2008.

[70] T. A. Runkler and J. C. Bezdek, "Web mining with relational clustering," *International Journal of Approximate Reasoning*, vol. 32, pp. 217-236, 2003.

[71] R. Subhashini and V. J. S. Kumar, "Evaluating the performance of similarity measures used in document clustering and information retrieval," in *Integrated Intelligent Computing (ICIIC), 2010 First International Conference on*, 2010, pp. 27-31.

[72] N. Sandhya and A. Govardhan, "Analysis of similarity measures with wordnet based text document clustering," in *Proceedings of the International Conference on Information Systems Design and Intelligent Applications 2012 (INDIA 2012) held in Visakhapatnam, India, January 2012*, 2012, pp. 703-714.

[73] N. Sandhya, Y. S. Lalitha, A. Govardhan, and K. Anuradha, "Analysis of similarity measures for text clustering," *CSC Journals*, vol. 2, 2008.

[74] W. K. Gad and M. S. Kamel, "Enhancing text clustering performance using semantic similarity," in *International Conference on Enterprise Information Systems*, 2009, pp. 325-335.

[75] S. Na, L. Xumin, and G. Yong, "Research on k-means clustering algorithm: An improved k-means clustering algorithm," in *Intelligent Information Technology and Security Informatics (IITSI), 2010 Third International Symposium on*, 2010, pp. 63-67.

[76] M. S. T. Deokar, "Text documents clustering using k means algorithm," *International Journal of Technology and Engineering Science [IJTES] TM*, vol. 1, pp. 282-286, 2013.