ADCAIJ: Advances in Distributed Computing and Artificial Intelligence Journal
Regular Issue, Vol. 10 N. 4 (2021), 381-399
eISSN: 2255-2863
DOI: https://doi.org/10.14201/ADCAIJ2021104381399

Urdu News Clustering Using K-Mean Algorithm On The Basis Of Jaccard Coefficient And Dice Coefficient Similarity

Zahid Rahman[a], Altaf Hussain[a, b*], Hussain Shah[c, d], and Muhammad Arshad[e]

[a] Institute of Computer Sciences & IT (ICS/IT), The University of Agriculture Peshawar, Pakistan
[b] Department of Accounting and Information Systems, College of Business and Economics, Qatar University, Doha, Qatar
[c] Shaykh Zayed Islamic Centre, University of Peshawar, Pakistan
[d] Department of Computer Science, Islamia College University Peshawar, Pakistan
[e] City University of Science and Information Technology Peshawar, Pakistan
zahidrahman182@gmail.com, altafkfm74@gmail.com, safihussain@uop.edu.pk, arshad12@aup.edu.pk
[*]Correspondence Email: altafkfm74@gmail.com

KEYWORDS

Urdu News; Clustering Mechanism; Jaccard Coefficient; Dice coefficient; Python; WEKA; K-mean; MSE.

ABSTRACT

Clustering is the unsupervised machine learning process that group data objects into clusters such that objects within the same cluster are highly similar to one another. Every day the quantity of Urdu text is increasing at a high speed on the internet. Grouping Urdu news manually is almost impossible, and there is an utmost need to device a mechanism which cluster Urdu news documents based on their similarity. Clustering Urdu news documents with accuracy is a research issue and it can be solved by using similarity techniques i.e., Jaccard and Dice coefficient, and clustering k-mean algorithm. In this research, the Jaccard and Dice coefficient has been used to find the similarity score of Urdu News documents in python programming language. For the purpose of clustering, the similarity results have been loaded to Waikato Environment for Knowledge Analysis (WEKA), by using k-mean algorithm the Urdu news documents have been clustered into five clusters. The obtained cluster's results were evaluated in terms of Accuracy and Mean Square Error (MSE). The Accuracy and MSE of Jaccard was 85% and 44.4%, while the Accuracy and MSE of Dice coefficient was 87% and 35.76%. The experimental result shows that Dice coefficient is better as compared to Jaccard similarity on the basis of Accuracy and MSE.

1. Introduction

Data mining evaluated and summarizes the data from multiple sights into valuable form. Generally the data mining also state that as Knowledge Discovery in Database. The automatic and useful mining of design demonstrating knowledge indirectly stored in huge databases or data warehouse and the web or other huge data sources. The topic of data mining is involves clustering which is the most famous data mining techniques and its background text has studied widely. This research i.e. Urdu news clustering is tackled as a text clustering problem. In experience it shows that the various applications of data mining or machine learning is the clear structures of knowledge which has the best performance ability on new examples which obtained structure of the reports. The data mining is used for the purpose to get knowledge and clusters the documents into groups on the basis of similarity, not just predictions (Kalmegh, 2015).

Urdu is the national language of Pakistan and a famous language of Ind-o-Pak countries. One out of twenty four authorized languages of India and 22nd utmost verbal language in the whole world. As to the vastly internet's application more data are created. Today urdu news documents are generated on the internet rapidly the users are interested to read exact kind of news according to their personal choice. To achieve this target machine learning algorithms is used on the basis of similarity techniques. Increase clustering accuracy by approximating similarity between Urdu news documents. The objective of this research is to get news clusters where each cluster comprises information about the same area (Usman et al., 2016).

The process through which the data is aggregated and then divided into separate groups is called clustering mechanism. Hence there are two methods for clustering, in one the similar types of data sets are clustered and in another different data objects are divided into groups. This technique is to be considered as to easily identify and similarity of the objects which belongs to their relate classes. Clustering is used as widespread way, to deal with data which are different kinds normally. The dataset are showed as n-dimensioned vector of attributes defined by nominal or numerical categorical values symbolic. Another concept for data where the objects are complex such as multi-categorical model and interval, usually new application can be used to describing more complex data. For example customer or user constraint, performance of desires. This type of data can showed more precise and concise using logic-based representations (Boudane et al., 2017).

Clustering examination is a significant tool for expanding and investigating to easily analyze achieve the aims by briefly exploring the major features of the data. Lots of procedures have been done in this field in the previous era. Algorithms of clustering i.e. k-mean algorithm has a past time history of fifty years. Methods of clustering have also been established to categorize data. These methods are used in image processing, pattern identification and retrieval of information. Clustering has a cost background in additional disciplines like psychology, psychiatry, biology, mines investigation, earth exploration, geology, geography and advertising. Clusters works according to the desired data objects that is needed and also use dissimilar methods which further from need based of the use. Clustering algorithms and methodologies are dependable on the amount of instances to be taken into account, size of a distinct instance and the best accuracy result is desired. The overall influences give growth to many algorithms and methods (Fahiman et al., 2017).

Every day the quantity of text is increasing at a high speed. When there is unstructured data in large amount of text can't be easily gets and processed by computer. The huge amounts of data spread and produce day by day due to technological enhancement and usage of social networks such as Facebook, twitter etc. These text data can be presented as three big V's i.e. Velocity, Volume and

Variety. For example, every day twenty petabyte of data needs to handle by Google; it means that when the data is in large shape i.e. breadth and depth, diversity, size in growth at the same time which occurred in unexpected events to handle. The data is arising fabulously more and more by usage of now a day's internet and electronic gadgets. Urdu text on the Internet also increases day by day, therefore today the users are attracted to discover the information from large datasets rapidly and efficiently, for the discovery of actual and efficient algorithms and techniques valuable patterns are mandatory (Liu & Li, 2015).

Urdu is a recognized language in Indo-Pak states. Urdu language is spoken by more than 100 million people all over the world. However, a limited research is conducted to develop Natural Language Processing (NLP) tools and APIs to process Urdu text easily. Most of the media houses publish news online to their websites and social networking sites to rapidly attract readers. Majority of the news readers are interested to read specific type of news according to their own interests. For example, politicians and businessmen are interested to remain updated with latest news especially related to politics and business. A news aggregation tool is required to automatically aggregate news from various news sources and cluster them for the readers to read same news from different sources (Khaliq et al., 1989).

Due to basic task, the similarity of document calculation has wide effect to document based grouping or classification and clustering. Bag of words of the document is representing by the old methods and its similarities are calculating using different measures like cosine or Jaccard and dice coefficient. For related words in documents it can be difficult for calculating entity expressions rather than single words of the documents. Furthermore, the relation between entities or words and types of entities are also informative. The measure to find how much two data objects are similar to each other is called similarity measurement. Basically the similarity, in the framework of data mining is defined as a distance with structures and dimensions representation of the objects. The similarity will be high when the distance is small, and vice versa. It is individual and vastly dependent on the presentation or area of the documents. e.g. the similarity between two fruits due to taste or size or color. Carefully distance should be calculated through the unrelated structures or dimensions. The comparative values of each document are necessity to normalize. Similarity is calculated in 0 to 1 range. Similarity=1 if x=y and similarity=0 when x not equals to y. (x and y are 2 objects) Following are some Similarity Measure techniques (Wang et al., 2015).

- Euclidean Distance

The utmost normally used distance for measurement is the Euclidean Distance. For the purpose, Euclidean distance is frequently just indicates to as «distance». When the data is continuous or thick, then this is the better nearness measurement. Euclidean distance shows the length between two points of the direction linked them. The given distance between 2 points is hereby described by the theorem of Pythagorean.

$$\mathbf{E.D} = \sqrt{S_{\mathbf{i=1}}^{\mathbf{k}} (\mathbf{xi} - \mathbf{yi})^2} \tag{1}$$

- Manhattan Distance

It is a math work in which distance between 2 points is the addition of exact of their Cartesian coordination differences. In other words it's the exact addition of alternation between two coordinates. Assume there are points A and B: to find out the Manhattan Distance (MD) between A and B, then only have to add up the exact and absolute 2-dimensional axis differences. To find out the MD between

these 2 points by calculating and determining as well as axis will be at the correct positions. In a plane with p1(x1, y1) and p2(x2, y2).

$$MD = |x1 - x2| + |y1 - y2| \tag{2}$$

- Cosine Similarity

Cosine similarity (CS) of metric calculates the standardized dot product of two attributes. To determine the CS it will efficiently try to find the cosine angle between two objects. Cosine of 0 degree is 1 and for any other angle it is less than 1. Hence it's an angle decision not a magnitude: a CS of two paths with the same direction is 1, while at 90 degree the similarity of two paths is 0, and two paths completely opposite to each other with the similarity of -1 and their magnitude is independent. CS is mostly used in positive space, the result is everywhere precisely constrained in [0,1]. The popularity of CS is due to especially the efficient calculation of sparse vectors.

$$\mathrm{Sim}(x, y) = \cos(\theta) = \frac{x \cdot y}{\|x\| \|y\|} \tag{3}$$

- Jaccard Similarity

The jaccard measure of similarity among sets or group is the division between the intersection of two sets and union of two sets and its cardinality. Suppose to calculate jaccard similarity between two sets X and Y is a ratio of cardinality (X∩Y) and (X∪Y).

- **Set:** A set is unordered group of items{x,y,z}. To represent the set notation is used as items which are separated by commas inside curly brackets{}.They are unordered so{x,y} = {y,x}.
- **Cardinality:** The symbol (|X|) is used for cardinality of X which is used to show or counts the total number of elements in the set X.
- **Intersection:** The intersection is used to find the common elements or words between two sets e.g. Two sets X and Y are represented by X∩Y.
- **Union:** Union represents the total number of items or elements which are in both sets. E.g. two sets X and Y are represented by X∪Y.

$$JS(X, Y) = \frac{|X \cap Y|}{|X \quad Y|} \tag{4}$$

- Dice Similarity Coefficient

The Dice Similarity Coefficient (DSC), also called by additional terms like (Sorensen index, Dice coefficient and similarity coefficient or index), is a statistical term which is used for the comparison of similarity of two samples. It was established in 1945 and 1948 by Lee Rymond Dice and Thurvald Sorensen independently. DSC is the proportion of comparison and ranges between (0, 1). It has to be observed as a similarity measure over sets. The Dice similarity coefficient of two sets A and B is expressed as:

$$DSC = \frac{2|X \cap Y|}{|X| + |Y|} \tag{5}$$

The rest of the paper is organized as follows. Section 2 consist of related work in which the previous work have been shown from different authors point of views. Section 3 defines and discusses the methodology, tools and techniques used for the proposed work. Section 4 shows the results along with their proper justification in graphical representation. And finally, section 5 shows the concluding part of the paper along with some future research directions.

2. Review of the Related Literature

Majority of the related work are stated in this section which shows the previous work in contrast with the proposed work.

- **Related Work**

(Bouras & Tsogkas, 2012) examined the usage of a high range of clustering algorithms as well as the similarity measure to news documents which create from the internet. In this paper the improvements of standard k-mean algorithm also proposed by the researcher from Word Net using outer knowledge. Before clustering procedure the bag of words is used to help in label generation process. Also this research examined the result that pre-processing of text has on clustering. Operated a corpus or documents of news articles derived from main news gateways. The evaluation of current clustering procedures exposed that K-means gives better total results when it comes to efficiency. The suggested Word Net enable W-k mean clustering algorithm meaningfully advances standard k-mean producing also valuable and high class cluster tags by using the obtainable labels clusters process. They have similarly obtained a new algorithm approach to improve the K-mean algorithm by information of external database. In this study the researcher have calculated 10 times enhancement on standard K-mean algorithm of low inter cluster and high intra cluster similarity. Also the subsequent tags are with high accuracy and the right one is related with their group labeling matching parts.

(Khaliq et al., 1989) Proposed and evaluated a new algorithm to automatically cluster Urdu news from different news agencies. This task was challenging as the researcher did not has language processing libraries for Urdu language. This experimental dataset consist of 500 news such as International, National, Health, Business and Entertainment from famous Pakistani media houses including Jang, Express, and voice of America Urdu (VOA). The proposed algorithm only used headlines to cluster the news. The researcher has developed a crawler that scraps Urdu news from various Urdu news broadcasting agencies. An algorithm get Related NewsList() has designed with specific threshold variable used by the algorithm to identify similar news based on matching number tokens between two given news headlines. This experimental evaluation showed micro and macro averages for precision 0.45 and 0.48 respectively to Identified similar news using headlines. The result looks far better than a random probability for precision. Therefore, the proposed algorithm is effective to cluster similar news.

(Usman et al., 2016) presented a technique by machine learning algorithms with max voting system to succeed maximum accuracy. To categorize news into various pre-defined classes such as business, science, health, culture, sports, weird and entertainment five various machine learning algorithm is used. This methodology contained a step by step process to collect Urdu language documents and then some preprocessing methods are applied such as stopwords removal, tokenization and stemming for feature collection to apply existing classification algorithms. These algorithms are Linear SVM / SDG, Multinomial / Bernoulli Naïve Bayes, Naïve Bayes and random forest classifier. In this research the MAX VOTING method provides better precision 0.942 and recall is 0.944 and

f1-score is 0.943. The stemming accuracy is 95% which is more than the other algorithms applied in this research.

(Arif et al., 2016) solved word sense disambiguation problem for Urdu language text. Work on language Urdu is till now in-progress or not yet properly accomplished. As Urdu language is very popular and has many shapes on the basis of speaking within the country. It's difficult to make decision on the language of Urdu which method has been applied on English Language can be effective for Urdu as well. Using the library of SVM standard many works have been done on Urdu Language documents so that to remove ambiguity from the words which shows the ambiguous meaning i.e. not a proper meaning. Two types of procedures are proposed in this study, one is for identification and other is for decreasing purposes. By lexicon ambiguity method is task of detection is performed. The specific collection of words belongs to Urdu text documents are processed by the identification methods to enhance and overcome the difficulties in the desired language of the study. Two types of algorithms for classification is been studies for this methodology, SVM and Naïve Bayes. Outcomes of this study has shown that the given SVM has many unique properties by which it can be best for the clustering and classification of the documents to show the accurate and meaningful similarities for that to handle large amount of data for further study and investigation and the datasets should be highly sensitive.

(Pratama et al., 2017) Studied about the graph which is the most proper data model to show, represent digital news it can be defined in simple method. This research worked on 200 Indonesian digital news which taken from website and write in json format. The researcher use Chines Whispers Algorithm (CWA) to solve the problem. The usage CWA that it is easy and has well for clustering based research. The CWA algorithm work like kid game where one kids says some words and the other kids repeatedly use this words. This algorithm big advantage is that it works on big data in a very fast time. This research focuses on the comparison of graphs quality intra and inter-cluster at each node. As results shows that 95% of nodes have intra-cluster weight higher than its inter-cluster weight. The researcher also worked on graph accuracy with the help of cluster based research and the results gives that of 80% chines whisper algorithm is used.

(Khademi et al., 2017) proposed an un-supervised technique to concise Persian documents. In the given procedure a new mixture mechanism is used to simplify the concept of clusters of the text using deep learning and statistical methods. The proposed method is language independent. In this work the researcher focused on Persian text summarization. First produced a word embedding based on Hamshari2 corpus and a dictionary of word frequencies. The Hamshahri2 corpus has 3206 text files in unlabeled texts sections. Each of these files is the Concatenation of hundreds news and articles. The subject of these news and articles is in different fields such as cultural, political, social and so on. In the proposed method the relationship between words and concepts is discussed in the input documents. In this research, the importance of sentences is determined Using semantic and syntactic similarities between words and Instead of using single words to express concepts, different related words are used. For this purpose, ROUGE 2.0 (Java implementation) tool is used. The proposed algorithm extracts the keywords of the document, clusters its concepts, and finally ranks the sentences to produce the summary. This proposed method achieves competitive performance. Rough-3 recall score for system summaries generated with 25% compression ratio on Pasokh corpus is 0.27.

(Patra & Saha, 2019) presented a novel word clustering technique to capture contextual similarity among the words. Related word clustering techniques in the literature rely on the statistics of the words collected from a fixed and small word window. For example, the Brown clustering algorithm is based on bigram statistics of the words. However, in the sequential labeling tasks such as named entity recognition (NER), longer context words also carry valuable information. To capture this longer

context information, they proposed a new word clustering algorithm, which uses parse information of the sentences and a non-fixed word window. This proposed clustering algorithm, named as variable window clustering, performs better than Brown clustering in our experiments. Additionally, to use two different clustering techniques simultaneously in a classifier, they proposed a cluster merging technique that performs an output level merging of two sets of clusters. To test the effectiveness of the approaches, they used two different NER data sets, namely, Hindi and BioCreative II Gene Mention Recognition. A baseline NER system is developed using conditional random fields' classifier, and then the clusters using individual techniques as well as the merged technique are incorporated to improve the classifier. Experimental results demonstrated that the cluster merging technique is quite promising.

(Khan et al., 2020) presented a feature-based clustering framework for the lexical normalization of Roman Urdu corpora, which includes a phonetic algorithm UrduPhone, a string matching component, a feature-based similarity function, and a clustering algorithm Lex-Var. UrduPhone encodes Roman Urdu strings to their pronunciation-based representations. The string matching component handles character-level variations that occur when writing Urdu using Roman script. The similarity function incorporates various phonetic-based, string-based, and contextual features of words. The Lex-Var algorithm is a variant of the k-medoids clustering algorithm that groups lexical variations of words. It contains a similarity threshold to balance the number of clusters and their maximum similarity. The framework allows feature learning and optimization in addition to the use of pre-defined features and weights. They evaluated their framework extensively on four real-world datasets and show an F-measure gain of up to 15 percent from baseline methods. They also demonstrated the superiority of UrduPhone and Lex-Var in comparison to respective alternate algorithms in their clustering framework for the lexical normalization of Roman Urdu.

(Anwar et al., 2019) presented for authorship identification in English and Urdu text using the LDA model with n-grams texts of authors and cosine similarity. -e proposed approach used similarity metrics to identify various learned representations of stylometric features and uses them to identify the writing style of a particular author. -e proposed LDA-based approach emphasizes instance-based and profile-based classifications of an author's text. Here, LDA suitably handles high-dimensional and sparse data by allowing more expressive representation of text. -e presented approach is an unsupervised computational methodology that can handle the heterogeneity of the dataset, diversity in writing, and the inherent ambiguity of the Urdu language. A large corpus has been used for performance testing of the presented approach. -e results of experiments show superiority of the proposed approach over the state-of-the-art representations and other algorithms used for authorship identification. -e contributions of the presented work are the use of cosine similarity with n-gram-based LDA topics to measure similarity in vectors of text documents. Achievement of overall 84.52% accuracy on PAN12 datasets and 93.17% accuracy on Urdu news articles without using any labels for authorship identification task is done.

(Munir et al., 2019) experimented topic modeling approaches for Urdu poetry text to show that these approaches perform equally well in any genre of text. Latent Dirichlet Allocation (LDA), Hierarchical Dirichlet Process (HDP), and Latent Semantic Indexing (LSI) were applied on three different datasets (i) CORPUS dataset for news, (ii) Poetry Collection of Dr. Allama Iqbal, and (iii) Poetry collection of miscellaneous poets. Furthermore, each poetry corpus includes more than five hundred poems approximately equivalent to 1200 documents. Before forwarding the raw text to aforementioned models, they did featured engineering comprising of (i) Tokenization and removal of special characters (if any), (ii) Removal of stop words, (iii) Lemmatization, and (iv) Stemming. For comparison of

mentioned approaches on their test samples, they used coherence and dominance model. Their experiment showed that LDA, and LSI performed well on CORPUS dataset but none of the mentioned approaches performed well on poetry text. This brought them to a conclusion that they needed to devise sequence based models that allow users to.

3. Methodology, Tools & Techniques

In this section, Urdu news document were clustered by using python programming language and Waikato Environment for Knowledge Analysis (WEKA) tool. This section actually shows that how the proposed work have been carried out by using the mentioned techniques and procedures to obtain the desired outcomes.

A. Data Collection

Data were collected from BBC urdu.com which contains different Urdu news such as Political, Business, Entertainment, Sports, weather forecasting and defense diplomacy etc. The dataset was available in text files i.e. (1.txt, 2.txt, and 3.txt…500.txt) and consist of 500 news of the current year 2018. Some sample of BBC Urdu news given in Table 1.

Table 1: Sample of BBC Urdu News

ایشیا کپ کرکٹ ٹورنامنٹ کے لیے پاکستان کی 16 رکنی ٹیم کا اعلان کردیا گیا ہے
یہ تو پاکستانی فلم انڈسٹری کے لیے خوش خبری ہے اب بالی وڈ کی فلمیں پاکستان میں دیکھنے کے ساتھ ساتھ بنائی بھی جا سکیں گی
موسم میں تیزی سے تبدیلی کا اثر، آسٹریلیا میں تارکول گاڑیوں کے ساتھ چپکنے لگی
جسمانی ورزش نہ کرنے کے سبب انھیں امراض قلب، ذیباطیس اور مختلف نوعیت کے سرطان لاحق ہونے کا خطرہ ہے

B. Data Preprocessing

Data preprocessing are used in machine learning to make the data meaningful. Initially, we needed to do some data cleaning before starting documents similarity measurement. In this phase, the several techniques like tokenization, stopwords removal and stemming are applied to the dataset for noise reduction and helping feature extraction. The given preprocessing step consists of three sub parts which are described below.

- **Tokenization**

The process of tokenization dividing the given texts into smaller parts called tokens. Tokens can be individual words, phrases or even whole sentences. Therefore, tokenization is used to splits whole documents or sentence into token separated by whitespace, line breaks or punctuation marks. Divided the data into tokens easily identified the words. e.g., میں پاکستانی ہوں the given sentence after tokenization was, ‘ہوں’ ,’پاکستانی’ ,’میں’.

- **Stopwords Removal**

Stop-words are basically a set of commonly used words in any language. Stop-words are those words that have no meanings in the sentence and no definitive list. The words that have no meaning should be removed during the preprocessing step. We focused on the important words instead. Stop-words are auxiliary words, adverbs, pronouns, and some conjunction words. Stop-words to be eliminated from the news documents are taken from the stop-words list. Stop-words refer to the most common or short function words in a language. e.g., گیا،گئ،کیسے،کو،میں،نے،پر،کی،کے،کا etc. Stop-words can cause problems when searching for expressions that include in it. So for easily searching these stop-words were removed from news documents.

- **Stemming**

The process to reduce modulated words to their stem or base or root form is called stemming. It is important in natural language processing (NLP) and natural language understanding (NLU) like the Urdu word 'منتظم' is the stem or root word of 'منتظیمین'. The stemming rules were applied to stem the Urdu news document to easily measure the accurate similarity score of the words. The architecture of stemming process is diagrammatically represented in Figure 1 which explains the detail of the stemming process for the Urdu language.

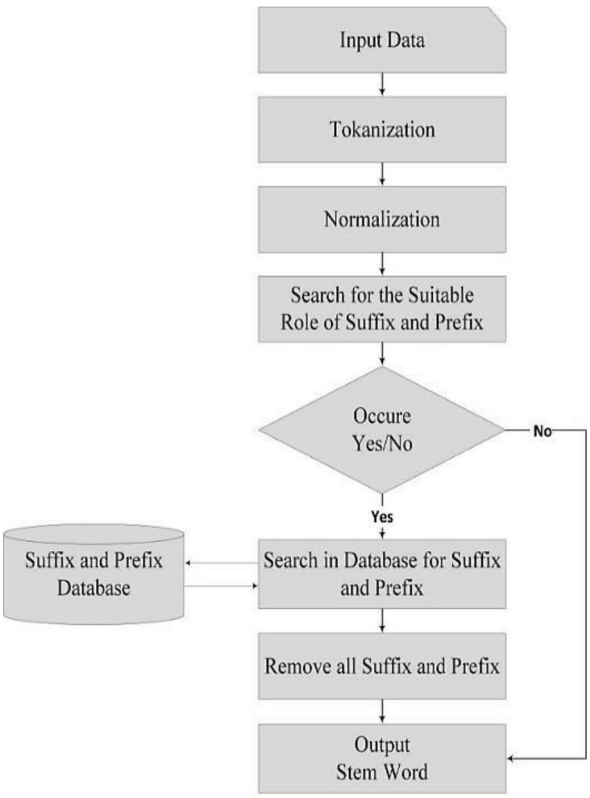

Figure 1: Architecture of stemming process

C. Similarity

Similarity is a technique that finds how two documents (news) are similar to each other. It means that how the news are closed to each other. In this research two similarity techniques are applied to find out the similarity between news documents, i.e. Dice Coefficient and Jaccard using python programming language.

Using the formula (5) in python programing language the similarity of the Urdu news was found. In python program we will take documents (i, j) i.e., Di = i1, i2, i3…in and Dj = j1, j2, j3… jn

Each document (i) represent news which are compared with another document (j). The given documents (news) are in the form of rows and columns.

Similarly the same procedure is applied for Jaccard similarity to measure the news similarity by using the formula (4) in python programing language.

e.g., the similarity score results will be look like in the Table 2.

Table 2: Similarity results

Di,j	j1	j2	j3	j4	j5	jn
i1	1	0.2	0.3	0.13	0.26	0.09
i2	0.21	1	0.51	0.07	0.37	0.07
i3	0.03	0.21	1	0.2	0.45	0.25
i4	0.05	0.15	0.35	1	0.33	0.05
i5	0.2	0.3	0.34	0.21	1	0.36
In	0.17	0.29	0.16	0.21	0.28	1

The following is the Pseudo Code of the proposed method.

Algorithm-1

INPUT:
T: Text File (Urdu News Documents)
OUTPUT:
Similarity score between news
INITIALIZATION:

BEGIN

1. Var1 = String 1
2. Var2 = String 2
3. Intersection _ Cardinality = len(Var1) ∩ (Var2)
4. Union _ Cardinality = len(Var1) U (Var2)
5. Similarity Value = Intersection _ Cardinality / Union _ Cardinality

Print = Similarity Value

END

D. Input Data to WEKA

WEKA selects to input data in the Attribute Relation File Format (ARFF). It is an extension of the Comma Separated Values (CSV) file format in which a header is used that offers metadata about the data-types in the columns, e.g. Dataset in CSV format shown in Table 3.

Table 3: Comma Separated Values format

0.01, 0.05, 0.30, 0.40, bbcurdu
0.09, 0.10, 0.40, 0.70, bbcurdu
0.07, 0.24, 0.35, 0.61, bbcurdu
0.01, 0.05, 0.30, 0.40, bbcurdu
0.30, 0.46, 0.54, 0.72, bbcurdu

WEKA has a precise computer science centric terminology when labeling data.

- **Instance:** It is a data which belongs to the row in table.
- **Attribute:** The characteristic of an entity is known as attribute. Every attribute may have a different type, for example:
- **Real** for actual mathematical values like 1.3.
- **Integer** for mathematical values which have no fractional part like 6.
- **Nominal** for categorical data like 'cat' and 'mouse'
- **String:** for lists of arguments, like the sentence.

Using the format in Table 3 the calculated similarity results (data) were loaded to WEKA and apply the clustering algorithm to cluster the data.

E. K-Mean Algorithm

The most famous separating approach of clustering is the k-mean. The McQueen was the first one who suggested it in 1967. It is an un-supervised, statistical, non-deterministic or iterative method of clustering. In it every cluster is denoted by the mean value of items in the cluster. Now it has to divide a set of n item into k-cluster for that inter-cluster similarity is less and intra-cluster similarity is maximum. The Similarity can be calculated in term of mean value of items in a cluster. The algorithm contains of distinct two phases.

1st Phase: Select k centroid arbitrarily, where the value of k is fixed already.

2nd Phase: Every item in the given dataset is linked to the adjacent centroid. To calculate or measure the distance between every data item in cluster centroid the Euclidean distance is used.

Input

K: number of chosen cluster

D: {d1, d2,...dn} a dataset consisting n items.

Output

The set of k-cluster as indicated in input

i. Randomly choose k data item from D dataset as first cluster centroid.
ii. Repeat
iii. Allocate every data item di to the cluster to which item is utmost similar on the basis of mean value of the item in cluster.
iv. Analyze the novel mean value of the data objects for every cluster and update the mean value.
v. When there is no variation.

F. Results Collection

From the k-mean algorithm the output data (results) were created for further analysis and comparison. The collection procedure is used for the purpose of results evaluation.

G. Evaluation Criteria

Based on performance parameters i.e. Accuracy and Mean Square Error (MSE) the collected results of dice coefficient and Jaccard similarity techniques using k-mean algorithm is carried out to show which technique gives accurate and best results.

- **Accuracy**

Determine the percentage of the total number of correctly clustered documents to the whole number of documents. Some important relations or terms for accuracy measurements are.

Truly Clustered (TC) total number of documents belongs to clusters that were correctly clustered. These documents are shown using their class's i.e.

TC = TC1+TC2+TC3 … TCn. Here n is the number of clusters in the dataset and TC is the number of documents of cluster Ci which belongs to correctly/truly clustered.

N is total number of documents which are clustered. By using the formula (6) the accuracy of clusters were calculated.

$$\textbf{Accuracy} = \frac{\textbf{TC}}{\textbf{N}} \tag{6}$$

- **Mean Square Error (MSE)**

The Mean Squared Error (MSE) or mean squared deviation (MSD) of an estimator measures the mean of the squares of the errors that is the average squared differences between the actual values and predicted values. The MSE is the second instant (about the origin) of the error, and thus includes both the variance of the actual and its predicted values. To find the Mean Square Error the formula (7) were used.

$$\textbf{MSE} = \frac{\sum(\textbf{A} - \textbf{TC})^2}{\textbf{n}} \tag{7}$$

A = Actual value
TC = Predicted value
n = Number of time period
Firewall VPN

H. Tools

To cluster Urdu news document the PYTHON and WEKA tools were used. The PYTHON used for similarity measurement while WEKA is used for clustering purposes.

4. Results Analysis & Discussion

This section explains the results and output of the proposed mechanism similarity based clustering of Urdu news using K-mean algorithm. In this research study two similarity techniques i.e., Jaccard and Dice coefficient performance are checked. The mechanism for similarity is constructed and presented using Python language while for clustering Urdu news documents k-mean algorithm is used in the environment of WEKA tool.

The experiment was conducted over news documents collected from BBC urdu.com. News documents similarity of Jaccard and Dice coefficient were found. Total 500 news in which each and every single news or document (D) compare with all news i.e. (500x500) gives 250000 values output. The similarity results are carried out using python programming language for both similarity techniques i.e. Jaccard and Dice coefficient. In our proposed mechanism we build a strongest method for making to find the accurate similarity between Urdu news documents. The proposed method first pre-processed the dataset i.e. tokenized the news then stopping words are removed then stemming rules are applied for data cleaning and noise reduction. After pre-processing the next step is to find the similarity score of the news, for that a model compares every single news or document (D) with all the rest of the news. The results of the news documents are calculated according to the Jaccard and Dice coefficient specific formulas and displayed the output in the form of rows and columns. The result shows that the Dice coefficient technique is a little better as compared to Jaccard similarity. Figure 2 show the cluster results for Jaccard while Figure 3 show the cluster results for Dice Coefficent.

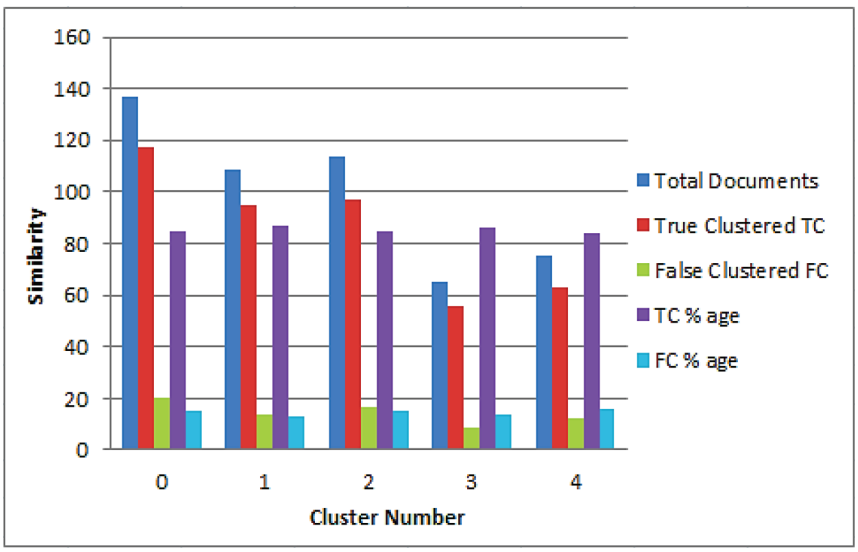

Figure 2: Cluster results of Dice Coefficient similarity

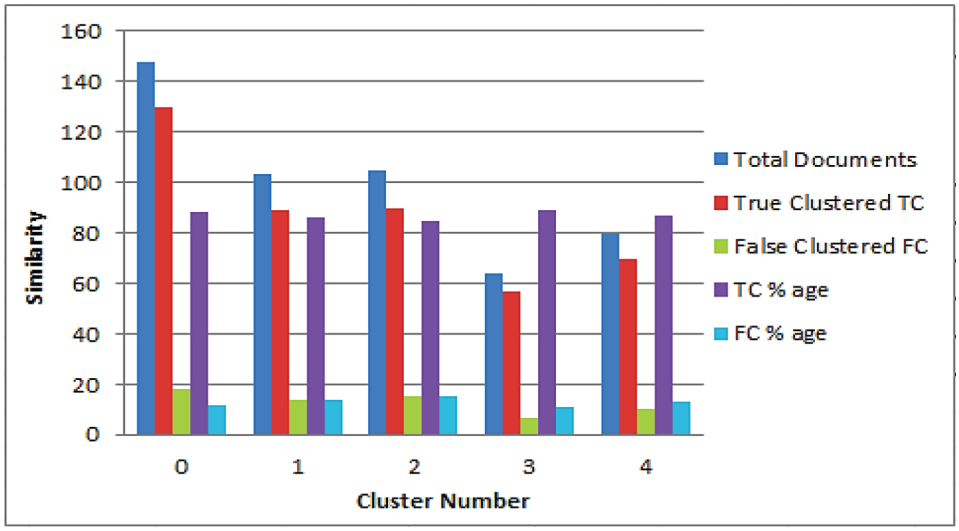

Figure 3: Contingency results for JS Accuracy

In Figure 4a the value of Total Clustered is 428, Falsely Clustered is 72 and Total Documents are 500.

In Figure 4b the value of Total Clustered is 436, Falsely Clustered is 64 and Total Documents are 500.

Accuracy has been calculated by the contingency Figure 4a and Figure 4b using the formula (6) which gives the accuracy 85% and 87%.

The overall Squared Error $(A\text{-}TC)^2$ is $\Sigma=222$ in Figure 5.

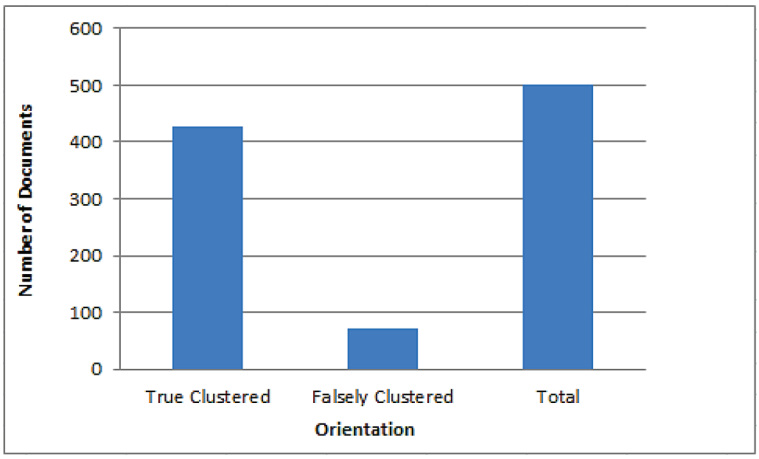

Figure 4a: Contingency results for DCS Accuracy

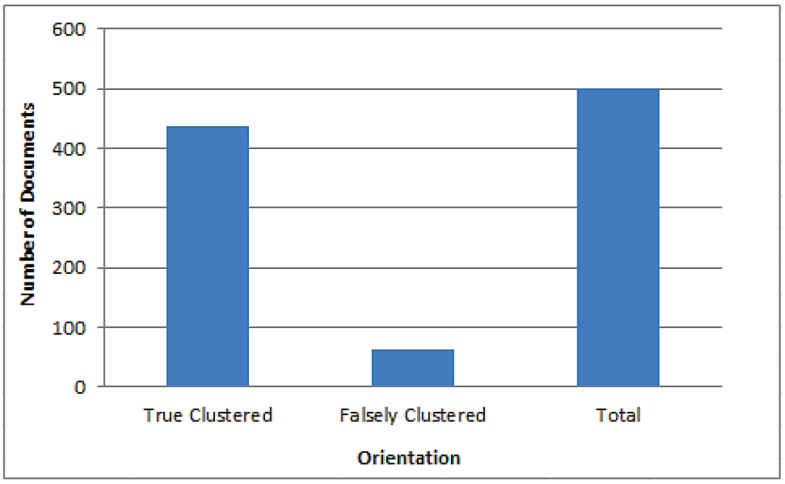

Figure 4b: Contingency results for JS Accuracy

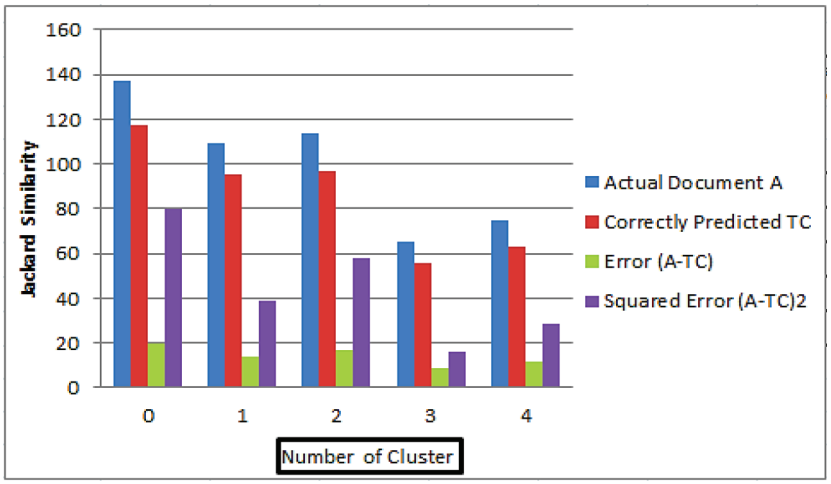

Figure 5: Contingency results for JS MSE

The overall Squared Error (A-TC)^2 is Σ=178.8 of Figure 6.

Mean Square Error (MSE) has been calculated by the contingency Figure 7a and Figure 7b using the formula (7) which give the MSE 44.4% and 35.76%.

In Figure 7a and 7b the comparison of both techniques using k-mean algorithm in terms of Accuracy and MSE shows that Dice coefficient is better as compared to Jaccard because the Accuracy for Dice coefficient is high than Jaccard while the MSE is less than Jaccard.

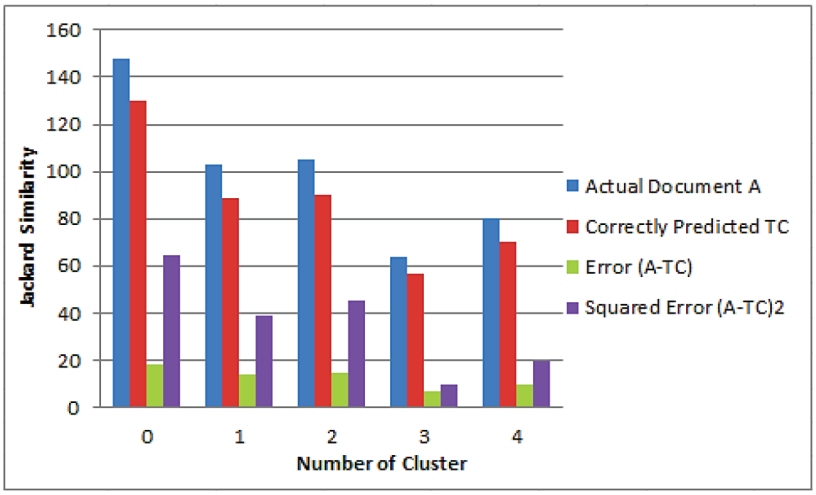

Figure 6: Contingency results for DCS MSE

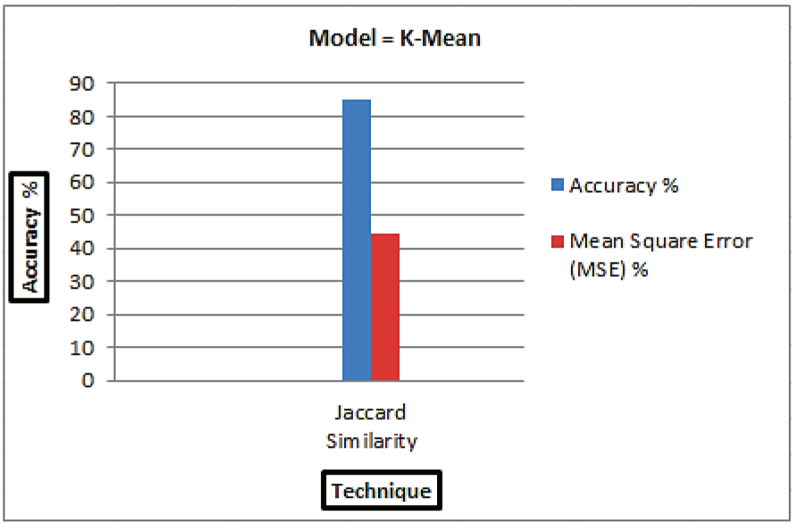

Figure 7a: Comparison of JS and DCS using K-Mean algorithm

5. Conclusion and Future Work

The objectives of this research were to clustered Urdu news for the readers to read specific type of news according to their own interests. Document clustering is being studied for many years. As Urdu documents are generated at great speed from variety of fields due to technological enhancement and usage of social networks such as Facebook, twitter etc. Grouping Urdu news manually is almost impossible, and there is an utmost need to device a mechanism which cluster Urdu news documents based on their similarity. The aim of this study is to find the similarity between Urdu news documents

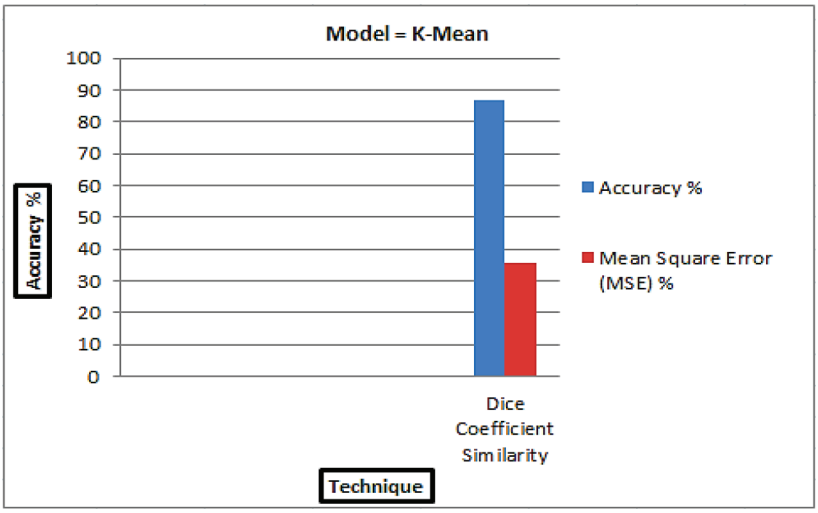

Figure 7b: Comparison of JS and DCS using K-Mean algorithm

based on Jaccard and Dice coefficient and cluster them using k-mean algorithm. The significance of this research is to group Urdu news documents from multiple fields for the users to read the same news easily from different sources which is related to their personal interest. In addition to that an own code are developed which find the similarity score for both techniques. This research proposed an architecture that uses Natural Language Processing technique to find the similarity between Urdu news i.e., Jaccard and Dice coefficient. After execution process the results are obtained which shows that the Dice Coefficient is better as compared to Jaccard in terms of accuracy and Mean Square Error. The results of both techniques clearly shows that the truly clustered documents in Dice coefficient is more than Jaccard while the falsely clustered documents is less than Jaccard, that's why the Dice coefficient is better. The experimental result shows that the system clustered the documents successfully. The Accuracy for Jaccard is 85% and Mean Square Error is 44.4% while the Accuracy for Dice Coefficient is 87% and Mean Square Error is 35.76%.

The obtained results in the experiment are very promising. However, there is still needs some work to be done in future. This can be stated that other similarity techniques and algorithm can also be used for best similarity and clustering. The same method can also be applied for Urdu news classification. This research can further be improved by classifying the clustered news, and designed automated software for easy use. Although the proposed model performs efficiently, clustered the news and obtains higher accuracies but needs still some improvements.

Conflict of Interest: All authors declare no conflict of interest.

References

Anwar, W., Bajwa, I. S., & Ramzan, S. (2019). Design and implementation of a machine learning-based authorship identification model. *Scientific Programming, 2019*.

Arif, S. Z., Yaqoob, M. M., Rehman, A., Jamil, F., & Jamil, F. (2016). Word sense disambiguation for Urdu text by machine learning. *International Journal of Computer Science and Information Security (IJCSIS), 14*(5).

Boudane, A., Jabbour, S., Sais, L., & Salhi, Y. (2017). *Clustering complex data represented as propositional formulas.* Paper presented at the Pacific-Asia Conference on Knowledge Discovery and Data Mining.

Bouras, C., & Tsogkas, V. (2012). A clustering technique for news articles using WordNet. *Knowledge-Based Systems, 36*, 115–128.

Fahiman, F., Erfani, S. M., Rajasegarar, S., Palaniswami, M., & Leckie, C. (2017). *Improving load forecasting based on deep learning and K-shape clustering.* Paper presented at the 2017 International Joint Conference on Neural Networks (IJCNN).

Kalmegh, S. (2015). Analysis of weka data mining algorithm reptree, simple cart and randomtree for classification of indian news. *International Journal of Innovative Science, Engineering & Technology, 2*(2), 438–446.

Khademi, M. E., Fakhredanesh, M., & Hoseini, S. M. (2017). Conceptual Text Summarizer: A new model in continuous vector space. *arXiv preprint arXiv:1710.10994*.

Khaliq, S., Iqbal, W., Bukhari, F., & Malik, K. (1989) Clustering Urdu News Using Headlines. *Language & Technology*, 89.

Khan, A. R., Karim, A., Sajjad, H., Kamiran, F., & Xu, J. (2020). A clustering framework for lexical normalization of Roman Urdu. *Natural Language Engineering*, 1-31.

Liu, Y., & Li, L. (2015). *Similarity Based Hot Spot News Clustering.* Paper presented at the 2015 IEEE 12th Intl Conf on Ubiquitous Intelligence and Computing and 2015 IEEE 12th Intl Conf on Autonomic and Trusted Computing and 2015 IEEE 15th Intl Conf on Scalable Computing and Communications and Its Associated Workshops (UIC-ATC-ScalCom).

Munir, S., Wasi, S., & Jami, S. I. (2019). A Comparison of Topic Modelling Approaches for Urdu Text. *Indian Journal of Science and Technology, 12*, 45.

Patra, R., & Saha, S. K. (2019). A novel word clustering and cluster merging technique for named entity recognition. *Journal of Intelligent Systems, 28*(1), 15–30.

Pratama, M., Kemas, R., & Anisa, H. (2017). *Digital news graph clustering using Chinese whispers algorithm.* Paper presented at the Journal of Physics: Conference Series.

Usman, M., Shafique, Z., Ayub, S., & Malik, K. (2016). Urdu text classification using majority voting. *International Journal of Advanced Computer Science and Applications, 7*(8), 265–273.

Wang, C., Song, Y., Li, H., Zhang, M., & Han, J. (2015). *Knowsim: A document similarity measure on structured heterogeneous information networks.* Paper presented at the 2015 IEEE International Conference on Data Mining.

Author's Biography

Altaf Hussain received his B.S and M.S degrees in Computer Science from University of Peshawar, Pakistan (2013) and The University of Agriculture Peshawar, Pakistan (2017) respectively. He worked at The University of Agriculture as a Student Research Scholar from 2017-2019. During his MS Degree he has completed his research in Computer Networks especially in Routing Protocols in Drone Networks. His recent approach is PhD in Computer Science & Technology. He has served as a Lecturer in Computer Science Department in Govt Degree College Lal Qilla Dir L, KPK Pakistan from 2020-2021. He has published many research papers including survey/review and conference papers. He was Research Scholar in (Career Dynamics Research Academy) Peshawar, Pakistan for one and a half year. Currently, he is working as a Research Assistant with the Department of Accounting & Information Systems, College of Business and Economics, Qatar University. His research interest includes Wireless Networks, Sensor Networks, Radio Propagation Models, and Unmanned Aerial Drone Networks.