information

MDPI

Article

Urdu Documents Clustering with Unsupervised and Semi-Supervised Probabilistic Topic Modeling

Mubashar Mustafa [1], Feng Zeng [1,*], Hussain Ghulam [1] and Hafiz Muhammad Arslan [2]

[1] School of Computer Science and Engineering, Central South University, 410083 Changsha, China; mubashar.ch331@csu.edu.cn (M.M.); ghnagri@csu.edu.cn (H.G.)
[2] School of Software Engineering, Northeastern University, 110819 Shenyang, China; arslanwaheed82@yahoo.com
* Correspondence: fengzeng@csu.edu.cn

Received: 26 September 2020; Accepted: 23 October 2020 ; Published: 5 November 2020

Abstract: Document clustering is to group documents according to certain semantic features. Topic model has a richer semantic structure and considerable potential for helping users to know document corpora. Unfortunately, this potential is stymied on text documents which have overlapping nature, due to their purely unsupervised nature. To solve this problem, some semi-supervised models have been proposed for English language. However, no such work is available for poor resource language Urdu. Therefore, document clustering has become a challenging task in Urdu language, which has its own morphology, syntax and semantics. In this study, we proposed a semi-supervised framework for Urdu documents clustering to deal with the Urdu morphology challenges. The proposed model is a combination of pre-processing techniques, seeded-LDA model and Gibbs sampling, we named it seeded-Urdu Latent Dirichlet Allocation (seeded-ULDA). We apply the proposed model and other methods to Urdu news datasets for categorizing. For the datasets, two conditions are considered for document clustering, one is "Dataset without overlapping" in which all classes have distinct nature. The other is "Dataset with overlapping" in which the categories are overlapping and the classes are connected to each other. The aim of this study is threefold: it first shows that unsupervised models (Latent Dirichlet Allocation (LDA), Non-negative matrix factorization (NMF) and K-means) are giving satisfying results on the dataset without overlapping. Second, it shows that these unsupervised models are not performing well on the dataset with overlapping, because, on this dataset, these algorithms find some topics that are neither entirely meaningful nor effective in extrinsic tasks. Third, our proposed semi-supervised model Seeded-ULDA performs well on both datasets because this model is straightforward and effective to instruct topic models to find topics of specific interest. It is shown in this paper that the semi-supervised model, Seeded-ULDA, provides significant results as compared to unsupervised algorithms.

Keywords: topic modeling; document clustering; non-negative matrix factorization (NMF); Urdu language; latent dirichlet allocation (LDA)

1. Introduction

Document clustering has become important in the period of data explosion. The data on the computer network are expanding more than expected due to the large-scope and quick expansion of web technologies [1–4]. The majority of this information is available on the internet in text format and presented without labels. However, interpreting these data by hand is ordinarily a dreary task. Although automatic interpreting methods have been introduced, the related technology still needs to advancement. Therefore, clustering has become an important data mining technique, which presents a structure for categorizing a large collection of documents [5,6].

As the volume of data is expanding, information retrieval (IR) is becoming complex. IR from data sources is a significant task that is applied in many applications, particularly in data mining and natural language processing (NLP). NLP is a very difficult task, particularly for poor languages such as Urdu. There are several reasons behind it, such as resource scarceness, context sensitive, cursive nature, words segmentation problem, compound name issues, a large number of synonyms, conjunction issues and acronym ambiguities [7]. Therefore, to work in Urdu language has become a challenging task, and we proposed a semi-supervised model to deal with the challenges.

In this era, topic modeling has gained great attention in the field of study. It is being used in many fields, especially in IR and NLP. The purpose of topic modeling is to find a pre-defined number of topics, and this is an unsupervised method based on statistical concepts where no earlier knowledge about the data is needed [8]. It has great potential to help users understand text corpora. However, this potential is hindered due to its purely unsupervised nature, which regularly extracts topics that are neither fully meaningful nor effective in external tasks [9]. This study is inspired by the work in [10], which changed the nature of the topic model from the purely unsupervised to semi-supervised in guiding topic models to extract topics of specific interest. The purpose of that work was to obtain better results by providing seed word sets that represent the inherent topics in the corpus. They proposed the seed-LDA model, which uses these seed words to improve the topic–word distribution and document–topic distribution.

In our study, two datasets are used to evaluate models and find the deficiency of the existing works. One dataset has four categories of documents (e.g., Business, Entertainment, Sports, Health), and all documents have distinct natures, called "Dataset without overlapping". The other dataset has documents of the weird category, which comprises overlapping categories connected to heath and entertainment, called "Dataset with overlapping". Then, we apply unsupervised algorithms, such as LDA [11], NMF and K-means [12] for extracting topics from the dataset without overlapping. The results show that all these algorithms find meaningful topics. After that, we deploy these algorithms on the dataset with overlapping and this demonstrates that LDA, NMF and K-means find some meaningless topics. To make topic extraction effective on a dataset with overlapping, we proposed an effective semi-supervised topic model for Urdu language, called Seeded-ULDA, to extract topics of specific interest.

Clustering and topic modeling are much alike, and they need to define the number of categories ahead and there is no need to provide labels for operating. Topic models extract a group of topics and every topic has a set of words applying probabilistic techniques. They extract a set of topics from a text corpus, where every topic is described as a statistical distribution of a group of words. This is accomplished by employing probabilistic models. Several topic modeling [13] methods exist in the literature. Recently, the probabilistic topic model, LDA, has been used in clustering and obtained good results [8,14]. Due to the good performance of probabilistic topic models, we take the advantages of LDA and Seeded-LDA to cluster Urdu documents. Our work is based on the LDA and Seeded-LDA, and the motivation is to effectively identify topics and simultaneously classify the documents among different topics, and especially to improve the performance on the documents with overlapping topics.

The rest of the paper is organized as follows. Section 2 describes Urdu language and related work. The proposed clustering method is introduced in Section 3. Section 4 describes evaluation techniques. Section 5 introduces the data processing and dataset, and analyzes the experimental results. Finally, Section 6 concludes this work.

2. Urdu Language

Urdu is a national language and one of the two official languages of Pakistan, along with English. It is used in education, literature, government and in lower positions of administration. Urdu language is closely related to Hindi language (Official language of India), developed from Indo-Iranian, Indo-European and Indo-Aryan [15]. It is extensively used in Nepal, Bangladesh, India, Afghanistan and various Asian countries [16]. Textual data of Urdu language are increasing day by day due to the extending of Urdu language, and there are many unlabeled Urdu text documents. However, manual interpreting of documents is usually a boring task for humans. Even automatic interpreting methods have been published, but the latest development for Urdu language is necessary.

2.1. Particularities

Urdu language owns 38 alphabets, 12 vowels and 25 consonants [17]. It is an Indo-European language but is dissimilar from other languages of Indo-European, and has extremely complicated grammar [18]. The main difference between Urdu language and European languages is that Urdu language has no idea of capitalization. The order of a sentence of Urdu is different from the order of a sentence of English, and sentence structure for Urdu follows an SOV (subject, object, verb) template. Sometimes, the verb appears with subject rather than object because Urdu represents the relativity free word order. There are many variants to define a single word in Urdu and several methods to express a sentence that delivers the same meaning. Urdu language is written in Arabic script which is different from English language, and the Urdu writing system goes from right to left, which is opposite to English scripting rules. Urdu language is new in the research communities of IR and NLP, hence little research has been done on it. Urdu language structure is totally contrary with other languages. Therefore, tools and models built for other languages cannot work properly with Urdu.

2.2. Urdu Language and Clustering

The nature of the Urdu language, the morphological structure, writing orientation and writing system has caused us to proceed more slowly in research works, particularly in automatic clustering or categorization. The previously published works in Urdu show that most of the researchers focused on the Urdu ligature [19–21] and developing preprocessing tools, such as stemmer. Khan et al. [22] developed a light weight stemmer for Urdu text by using a rule based approach. Agglomerative hierarchical clustering was proposed in [20] for Urdu ligature recognition, and the authors utilized Decision Tree, K-nearest Neighbor (KNN), Naive Bayes and Linear Discernment Analysis for classification. Urdu ligature organization is achieved using a deep neural network in [21]. The authors used a corpus of 2430 ligatures and gained an accuracy of 73.13%. A comprehensive study on Urdu document images was conducted by employing various clustering algorithms, such as a self organizing map (SOM), K-means and hierarchical clustering [19]. The authors segmented images into ligatures or partial words and then these ligatures were grouped together using different clustering methods. Asghar et al. [23] developed and examined a dataset for detecting Urdu text from images. Zarmeen et al. [24] conducted experiments to evaluate clustering models for Urdu tweets. In their experiments, they used three different clustering techniques and compared results with topic modeling. Their results showed that K-means with TF-IDF features performed well. In [25], the authors compared local weight and global weight approaches for Urdu language. Their experimental results showed that the local weight approach was better for text summarization. The authors used classification methods to categorize the websites and examined the records that are not properly classified [26]. As mentioned above, most of the research focused on Urdu ligature, and only a small part of the research focused on Urdu document clustering. We identified two works on Urdu document clustering. Recently, Rahman et al. [7] developed a framework of document clustering using the K-mean algorithm and analyzed various similarity measures for Urdu documents. Ehsan et al. [27] employed a fully probabilistic Bayesian method and Latent Dirichlet Allocation (LDA) for Urdu document clustering. Their experimental results indicated that LDA gains 90% accuracy on the Urdu corpus.

In our work, firstly, we used LDA, NMF and K-Means on datasets without overlapping and the results show that LDA has 95% accuracy. Compared with NMF and K-Means method, LDA performs better on the Dataset without overlapping. Secondly, LDA, NMF and K-Means were applied on the dataset with overlapping and the results show that these algorithms do not perform well on this dataset. Thirdly, we ran the proposed framework on the datasets, and the results show that it can get better performance on both datasets.

3. Clustering Methods

Document clustering is used for grouping a large collection of documents. In the clustering process, it first extracts the intrinsic characteristics of documents, and groups the documents as some topics, then sorts them into various clusters [28,29]. In the following section, we will describe the three unsupervised algorithms (LDA, NMF, K-means) and the proposed semi-supervised model (seeded-ULDA) which is used in this study for Urdu document clustering.

3.1. K-Means

This algorithm was first proposed in 1957 while the term ”K-means” was first used in 1967 [12]. K-means clustering is a method that is commonly used for cluster analysis in data mining. Its objective is to divide the n observations into k clusters, where each observation belongs to the cluster with the highest mean or weighted average, called the centroid. Calculate the centroid by using Equation (1), where C_k is the centroid of the cluster k, N_i are the elements of the cluster, and x_k is the number of these elements.

$$C_k = \frac{1}{x_k} \sum N_i \in jN_i \tag{1}$$

The basic steps of K-means are described below.

- Input: A set of N numbers and K.
- Output: A group of numbers into k clusters.
- Step 1: K random elements are randomly selected at the beginning and are regarded a centroid, such as C_k.
- Step 2: Assign each number to the cluster, which has minimum distance $d(N_i, C_k)$.
- Step 3: Assign all other elements to the nearest centroid and recalculate the new centroid.
- Step 4: Repeat steps 2 and 3 until no other elements move from one cluster to another.

3.2. Latent Dirichlet Allocation

LDA is a machine learning algorithm used to discover latent topics. It was originally developed in the context of population genetics [11]. In the context of machine learning, it was developed again in 2003 [30] and has now become the most widely used algorithm in data mining tasks. LDA is a generative model whose purpose is to find hidden topics. For this reason, they think that the document is a mixture of topics, and the words of the document are generated by the topics. LDA uses two types of distributions. One of them is “distributed over topic”, which means that each document in the corpus is distributed by topic, and each topic is assigned a probability. The sum of the probabilities of all topics should be 1. The other is “distributed over word”, each word of topic has a probability assigned to it, and the sum of the probabilities of all words should be equal to 1. The LDA model is shown in Figure 1, which has three layers: dataset $\sim$ (α, β), documents $\sim$ θ and terms $\sim$ (z, w).

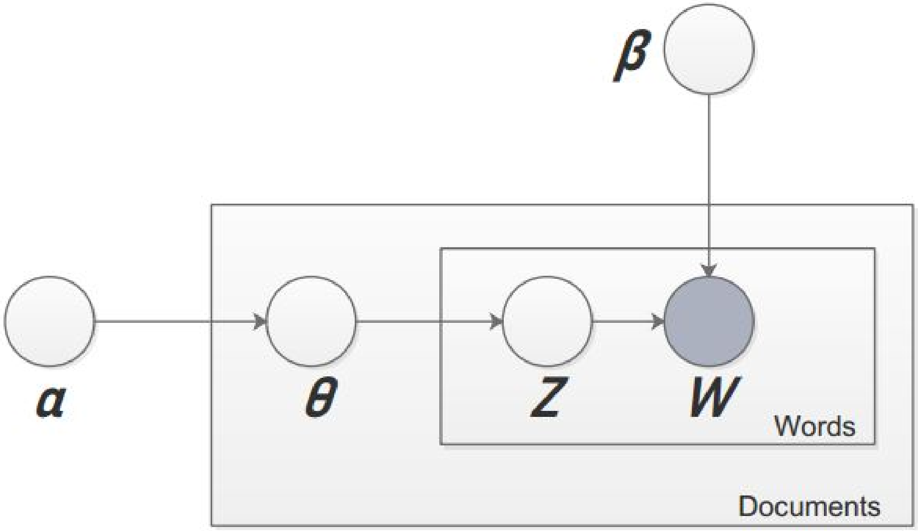


Figure 1. Generative Model LDA [30].

3.3. Non-Negative Matrix Factorization

NMF was first proposed in 1994 [31], and aroused people's attention in an article by Lee and Seung in 1999 [32]. Since then, the number of publications involving NMF has grown rapidly. NMF is an effective machine learning technique in which the matrix V is decomposed into (usually) two matrices W and H (such that $X \approx WH$). To the matrix X, each row corresponds to a word, and each column corresponds to a document. As shown in Equation (2), NMF will generate two matrices W and H, according to X, $W \geq 0$ and $H \geq 0$. The columns of the W matrix are regarded as basic documents (bags of words), and they represent topics (sets of found words in various documents at the same time). Matrix H shows how to aggregate the contributions of various topics to reconstruct the word combination of a given original document. Given a set of documents as input, NMF determines the topics and, at the same time, classifies the documents between these different topics.

$$\underbrace{X(:,j)}_{jth\text{ Document}} \approx \sum_{k=1}^{r} \underbrace{W(:,j)}_{kth\text{ Topic}} \qquad \overbrace{H(:,j)}^{\text{importance of } kth \text{ topic in } jth \text{ Document}} \tag{2}$$

3.4. Seeded-ULDA Framework

The creation of an impressive semi-supervised clustering model for Urdu documents, named Seeded-ULDA, is considered to be a challenging task in the field of data mining, because Urdu language has complex nature and due to the overlapping nature of the dataset. Figure 2 shows the comprehensive overview of the semi-supervised model, Seeded-ULDA.

The following techniques are involved in Seeded-ULDA.

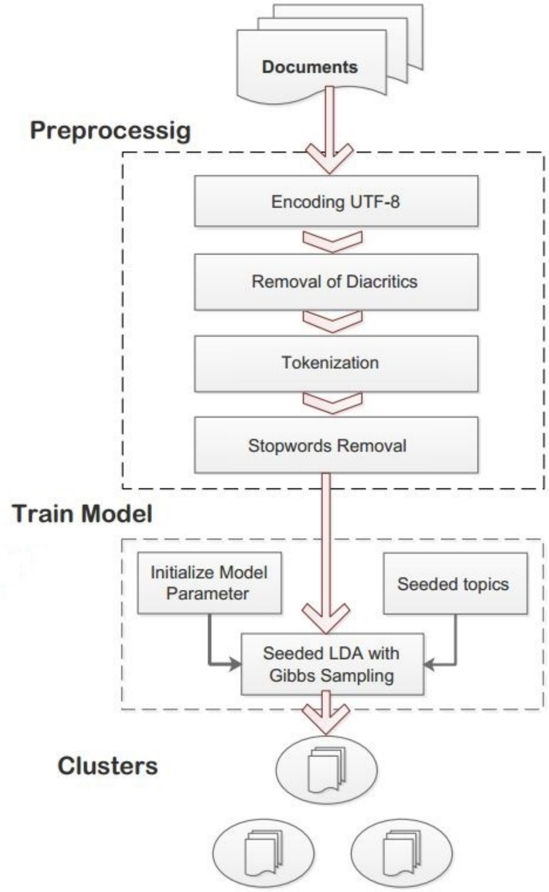


Figure 2. Proposed Seeded-ULDA Framework.

3.4.1. Preprocessing

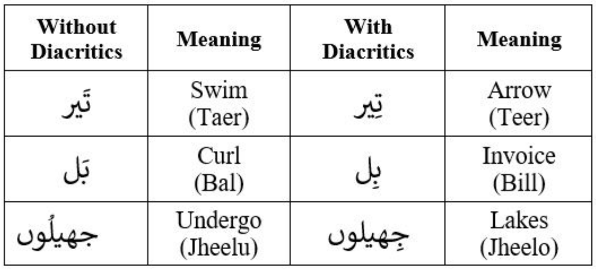
Data preprocessing is a basic phase combined before executing any IR, NLP and data mining task [18]. Preprocessing intends to standardize the representation of data to be classified. In this study, we preprocessed data by the following four techniques.

Enconding UTF-8

Computer programs face the problem to recognize the characters in the text of Urdu language. We employ Unicode Transformation Format 8 (UTF-8) encoding to recognize the Urdu characters. Unicode is a widely-used computing industry model that defines a comprehensive mapping of unique numeric code values to the characters. UTF-8 is a compromise character encoding that can contain any unicode characters.

Removal of Diacritics

A diacritic is a symbol placed above, through or below a letter to change the pronunciation. The sound it represents is different from that of letters without diacritics [33]. Urdu diacritics are Zer, Zabar and pesh, and they are called Aerab [18]. Some diacritics examples are given in Figure 3, which shows that the meaning of the word will change as diacritics change. Usually, only letters are used to mark Urdu, and diacritics are optional. Every word has a set of correct diacritics, but it can be written with or without diacritics. We discard diacritics to form a standardized corpus.



Without Diacritics	Meaning	With Diacritics	Meaning
تَیر	Swim (Taer)	تِیر	Arrow (Teer)
بَل	Curl (Bal)	بِل	Invoice (Bill)
جھیلُوں	Undergo (Jheelu)	جِھیلوں	Lakes (Jheelo)

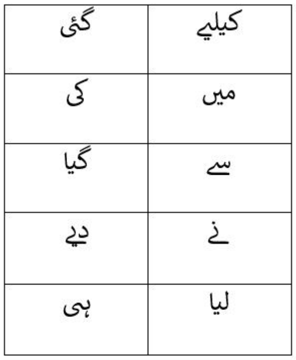
Figure 3. Urdu Diacritics Example.

Tokenization

As we know, machines—as advanced as they may be—are unable to understand words and sentences in the same manner as humans do. In order to make documents corpora more capable for computers, they must first be converted into some numerical form. There are a few methods used to achieve that, but in this study for Seeded-ULDA, we use count-vectorizer which is a very intuitive approach to figure out this problem, and the methods comprise of splitting the documents into tokens, assigning a weight to each token and creating a document–term matrix. By applying this method, we get document–term matrix with weight in integer format.

Stopwords Removal

In NLP, one of the main forms of preprocessing is to delete worthless data, and worthless words (data) are called stopwords. They are found frequently and do not add information or meaning to sentence. They can be safely avoided without losing the meaning of the sentence. We built our own stop words list and remove these stop words from our corpus to gain only meaningful words. A few of the most used stopwords in Urdu are shown in Figure 4.



گئی	کیلیے
کی	میں
گیا	سے
دیے	نے
ہی	لیا

Figure 4. List of some stop words of Urdu language.

3.4.2. Seeded-LDA

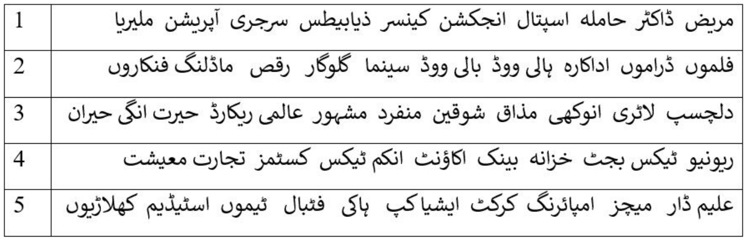
Seeded-LDA was proposed in [10], an extension of topic models. Topic models are typically unsupervised, which often results in topics that are either entirely meaningless or ineffective in extrinsic tasks (Chang et al., 2009 [9]). The purpose of seeded-LDA was to obtain better results by providing a set of seed words that represent the inherent theme of the corpus. The model used these seed word sets to improve the topic word distribution and improve the document topic distribution. The document clustering task reveals a major improvement when using seeded information.

3.4.3. Gibbs Sampling

Gibbs Sampling (GS) is a Markov-chain Monte Carlo (MCMC) method, which was proposed by Griffiths and Steyvers (2004) [34]. GS obtains the sequence of observations, which are approximated from multivariate probability distribution. Since GS is a straight-forward approach and quickly converges to a known ground-truth, it has been widely used in many probabilistic models and we combine it with seeded-LDA in our proposed model, called Seeded-ULDA.

3.4.4. Seed Topics

Seeded-ULDA allow a user to guide the topic discovery process. The user can give sets of seeded words that are representative of the given dataset. In our experiment, we use dataset, which contains five classes. We provide manually seeded topics. Figure 5 shows the first ten words of the seeded topic of each class.



1	مریض ڈاکٹر حامله اسپتال انجکشن کینسر ذیابیطس سرجری آپریشن ملیریا
2	فلموں ڈراموں اداکاره ہالی ووڈ بالی ووڈ سینما گلوگار رقص ماڈلنگ فنکاروں
3	دلچسپ لاٹری انوکھی مذاق شوقین منفرد مشہور عالمی ریکارڈ حیرت انگی حیران
4	ریونیو ٹیکس بجٹ خزانہ بینک اکاؤنٹ انکم ٹیکس کسٹمز تجارت معیشت
5	علیم ڈار میچز امپائرنگ کرکٹ ایشیا کپ ہاکی فٹبال ٹیموں اسٹیڈیم کھلاڑیوں

Figure 5. First ten words of seeded topic of each class.

3.5. Topic Modeling and Clustering

Generally, topic models are used in two ways for document clustering [35]. First, it is used to decrease the dimension of representation of documents, and then apply the standard clustering algorithm (such as NMF) to cluster the newly obtained representation. Second, it is used directly, the pre-defined number of topics in topic model is similar to the number of clusters, and the obtained topic, z, is considered to be consistent with the cluster. According to Equation (3), the topics with highest probability are assigned to the document.

$$arg\ max_z =_{1...k} \theta_z^d \tag{3}$$

In this study, we concentrated on the second method, because it makes it possible to check the performance of LDA and Seeded-ULDA compared to existing clustering methods (such as K-means).

4. Evaluation Techniques

In the experiment, we used two well-known evaluation measures—Rand index and F-measure. The values of these evaluation measures are between 0 and 1—higher is better.

4.1. F-Measure

The F-measure has a long history of cluster evaluation in the IR field [36]. It is defined as the harmonic mean of two measures—recall and precision [37]. Precision is the ratio of related instances to the retrieved instances (the attributes of the document from the cluster to the correct class). Recall represents a fraction of the total number of related instances to actually be retrieved. From the definition of Precision and

Recall, the values of them will not provide a right understanding to determine the quality of clustering. Many reasons are discovered in previous works, therefore a combination of precision and recall can give better understanding when seen in one value, named F-measure. In order to calculate precision, recall and F-measure, the confusion matrix is typically used. As shown in Table 1, the confusion matrix is composed of four values.

Table 1. Confusion Matrix.

	Same Cluster	**Different Cluster**
Similar Documents	True Positive (TP)	False Negative (FN)
Different Documents	False Positive (FP)	True Negative (TN)

In Table 1, *TP* is the number of pairs of documents that are in the same class and the same cluster, *FN* is the number of pairs of documents that are in different classes and the same cluster, *FP* is the number of pairs of documents that are in different clusters and the same class and *TN* is the number of pairs of documents that are in different classes and different clusters. Then, we can measure *recall*, *precision* and *F-measure* based on these values, according to the equations, as given below:

$$Percision = \frac{TP}{TP + FP} \tag{4}$$

$$Recall = \frac{TP}{TP + FN} \tag{5}$$

$$F - measure = \frac{2 \times Percision \times Recall}{Percision + Recall} \tag{6}$$

4.2. Rand Index

The *Rand* statistic or *Rand index* [38] is a direct criterion for comparing the obtained cluster labels with predefined labels. The former is calculated by considering all couple of documents in the collection after clustering. If two documents are located in different positions or the same position in both the predefined classes and clustering results, it is regarded as an agreement. Otherwise, there is a disagreement. The Rand Index is calculated according to the following formula.

$$Rand - Index = \frac{TP + TN}{TP + FP + TN + FN} \tag{7}$$

5. Experiment and Evaluation

In this section, we present some experiments on two types of datasets using unsupervised and semi-supervised techniques and demonstrate the effectiveness of the proposed Seeded-ULDA framework. Then, we evaluate the obtained results from these experiments by using above defined measures.

5.1. Dataset

For every NLP task, the presence of a benchmark dataset is required. However, Urdu language does not provide such linguistic resources for the NLP task. For the experiments, we built our corpus that contains Urdu text articles from news websites, which is now publicly available at https://github.com/Mubashar331/Urdu-corpus for research purpose. The datasets exploited in this study were collected from Express Urdu news portal https://www.express.pk and saved in Notepad in UTF-8 encoding. For the dataset, two conditions were considered in this study. First, the dataset had four categories of documents, such as Business, Entertainment, Sports, Health and all had distinct nature, called "Dataset without overlapping". The document categories—Business, Entertainment, Sports and Health—related to finance, art, games and disease news, respectively. Table 2 shows detail of the dataset without overlapping, and these four categories have a thematically distinct nature and are not connected to each other.

Table 2. Detail of Dataset without overlapping.

Topic	Documents	Tokens
Business	200	37,933
Entertainment	210	29,621
Sports	200	32,224
Health	200	55,934
Total	810	155,712

Second, when the documents of the weird category are added into the dataset, then the new dataset is called "Dataset with overlapping". Table 3 shows the detail of the dataset with overlapping—these five categories have a thematically distinct nature, but the documents of the weird category are connected to health and entertainment; therefore, it is named the dataset with overlapping.

Table 3. Detail of Dataset with overlapping.

Topic	Documents	Tokens
Business	200	37,933
Entertainment	210	29,621
Sports	200	32,224
Health	200	55,934
Weird	200	32,365
Total	1010	188,077

5.2. Experiment

We verified our proposed Seeded-ULDA framework by performing some experiments on given datasets. We performed three experiments, and in the first experiment, we employed unsupervised methods, such as LDA, NMF and K-means on the preprocessed dataset without overlapping. In the second experiment, we employ unsupervised methods on the preprocessed dataset with overlapping. In the third experiment, we used our proposed Seeded-ULDA framework to illustrate the effectiveness of the model.

5.3. Results and Discussion

Class labels exist for data used in this study. We use these labels to assess the results of document clustering. The labels are retrieved by using the various techniques compared with these existing labels. First, we assess the results of clustering techniques which are traditional techniques for grouping documents. These techniques perform well on the dataset without overlapping, but fail on the dataset with overlapping. The reason behind this is that some retrieved clusters from the dataset with overlapping consist of irrelevant terms. Second, we assess the results of unsupervised and semi-supervised probabilistic topic models. The unsupervised probabilistic model performed well on the dataset without overlapping and poorly perform on dataset with overlapping. However, our proposed semi-supervised probabilistic topic model outperformed on both datasets.

5.4. Evaluation

First, we evaluate the performance of unsupervised techniques, such as LDA, NMF and K-means on the dataset without overlapping. The experimental results show that all these give satisfying results. Figure 6 shows the accuracy of LDA, NMF and K-means for each class on the dataset without overlapping, which indicates that, in two classes, LDA is better and, in the other two classes, NMF is better. Overall, other evaluation methods (such as rand-index, F-measure, precision and recall) show that LDA is a little better than others, as shown in Figure 7.

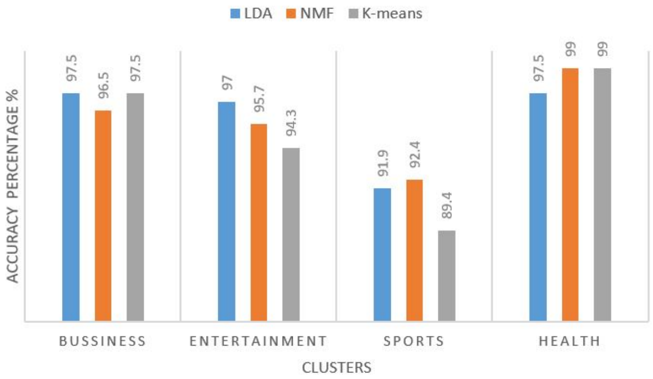


Figure 6. Accuracy of LDA, NMF and K-means for each class on dataset without overlapping.

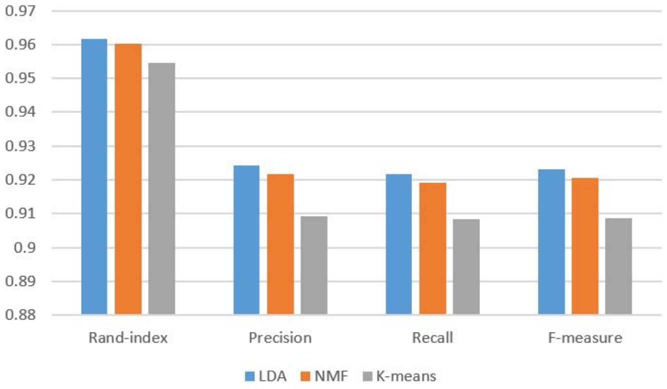


Figure 7. Performance measure of LDA, NMF and K-means by Rand-index, precision, recall and F-measure on dataset without overlapping.

Second, we evaluate the performance of these unsupervised techniques on dataset with overlapping. The results of Experiment 1 show that all these give satisfying results, but this experiment shows that NMF and K-means do not perform well on the dataset with overlapping, because on this dataset, these algorithms find some topics that are either entirely meaningless or ineffective in extrinsic tasks. We deployed NMF on the dataset with overlapping and it did not label the documents well according to the topics, as shown in Table 4. According to our experiment, topic 0 belongs to Business class and NMF labels 85% accurate documents of Business category; topic 1 related to Entertainment class and NMF labels 93% accurate documents of Entertainment category; topic 2 belongs to Sports class and NMF labels 90% accurate documents of Sports category; topic 3 belongs to health class and NMF labels 99% accurate documents of Health category, but NMF failed to classify the documents of weird class and labeled them with a topic which belongs to health class. Because NMF cannot find the topic which relates to the weird class.

Table 4. Performance of NMF in labeling documents by using Topic.

Categories	Topic 0	Topic 1	Topic 2	Topic 3	Topic 4
Business	85%	0%	0%	2%	12.5%
Entertainment	3%	93%	1%	3%	0%
Sports	3.5%	5.5%	90%	1%	0%
Health	1%	0%	0%	99%	0%
Weird	6%	4.5%	0.5%	86.5%	2.5%

We deployed K-means on the dataset with overlapping and the results are shown in Table 5, which demonstrate that K-means also failed to classify the documents of weird class and labeled them with a topic which belongs to the health class, because it also cannot find the topic which relates to the weird class.

Table 5. Performance of K-means in labeling documents using Topic.

Categories	Topic 0	Topic 1	Topic 2	Topic 3	Topic 4
Business	22%	0%	7.5%	70.5%	0%
Entertainment	0%	6%	17%	6%	71%
Sports	0.5%	90%	7%	0.5%	2%
Health	0%	0%	99%	1%	0%
Weird	2%	0%	97.5%	0%	0.5%

We deployed LDA on dataset with overlapping which shows better performance as compared to NMF and K-means, LDA gives better results in each class, as shown in Table 6, but needs to improve the results in the weird category. In this study, we first extracted a set of topics from text documents where every topic was described as a statistical distribution over a group of words and then simultaneously classified the documents among these different topics. These NMF and k-means algorithms failed to extract the topic which had sets of words related to the weird category of documents. To solve this problem, we proposed Seeded-ULDA, which manually accepted topics of users interest and then classified the documents among these topics.

Table 6. Performance of LDA in labeling documents by using Topic.

Categories	Topic 0	Topic 1	Topic 2	Topic 3	Topic 4
Business	1%	2%	96.5%	0%	0.5%
Entertainment	96%	3%	0%	1%	0%
Sports	6%	0.5%	2%	91%	0.5%
Health	0%	5%	3%	1%	91%
Weird	15%	71%	2%	0.5%	11.5%

Third, we evaluate the performance of our proposed model Seeded-ULDA on the dataset with overlapping. The results show that Seeded-ULDA gives better results, as compared to LDA. As shown in Figure 8, Seeded-ULDA gives better results for each class and is 19.5% better in the weird category as compared to LDA. Overall, other evaluation methods (such as rand-index, F-measure, precision and recall) show that Seeded-ULDA is better than LDA, as shown in Figure 9.

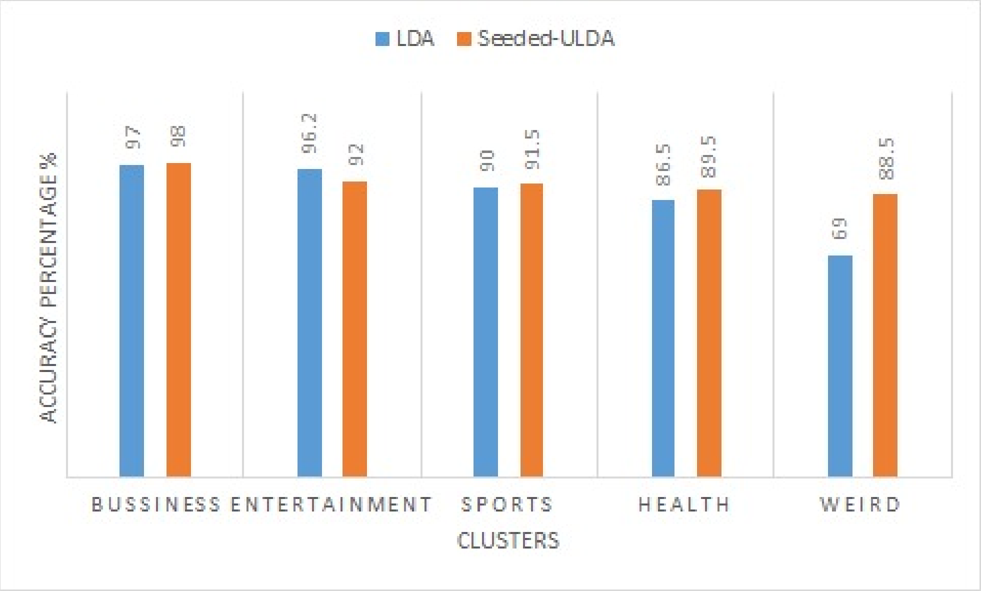


Figure 8. Accuracy of of Seeded-ULDA and LDA on dataset with overlapping.

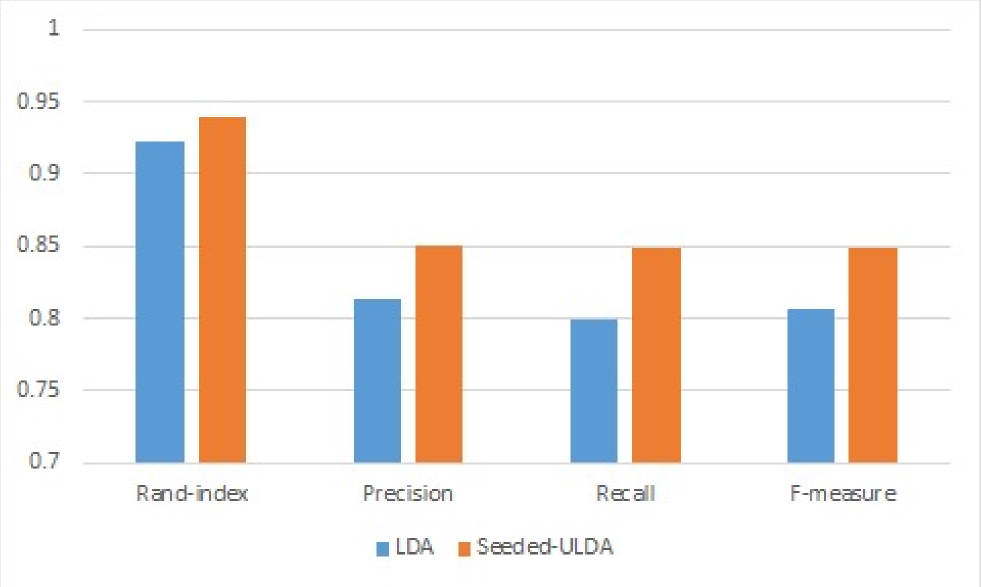


Figure 9. Performance of Seeded-ULDA and LDA by Rand-index, precision, recall and F-measure on dataset with overlapping.

6. Conclusions and Future Plan

Document clustering aims to assign documents to different topics when the topics are not known in advance. It is a cluster analysis task and is well-investigated in English language when compared to Urdu. The peculiarities of Urdu (such as lack of resources, their own morphological structure, semantics and syntax) make Urdu document clustering more complex. Some unsupervised models have been used to cluster Urdu documents. However, regular unsupervised clustering models might not give good results in some cases due to the purely unsupervised nature of these models. In this study, we proposed Seeded-ULDA for Urdu document clustering, which is a semi-supervised framework. The experiment was conducted in three phases. First, we used unsupervised algorithms, such as LDA, NMF and K-means to cluster the dataset without overlapping. The experimental results show that the probabilistic topic model LDA gives better results as compared to clustering techniques NMF and K-means. Second, these algorithms are used to cluster the dataset with overlapping. The results demonstrate that clustering techniques NMF and K-means fail on the overlapping category, and LDA gives 69% accuracy on the overlapping category. Third, we applied the Seeded-ULDA framework on dataset with overlapping and the results show that Seeded-ULDA gives better results, as compared to LDA, and it gives 88.5% accuracy on the overlapping category. This is confirmed by evaluating the results of experiments through other validation techniques.

Working with Urdu text is a challenging task in the data mining research community, yet there is a huge gape for development and improvement. In the future, academic research might cover developing word-embedding techniques for Urdu text, which require big data to gain considerable results. Existing word embedding techniques have been deployed with success to few languages, such as English, and have obtained significant results.

Author Contributions: M.M. and F.Z. conceived and designed the experiments; M.M. performed the experiments; F.Z. analyzed the data; H.G. and H.M.A. contributed reagents/materials/analysis tools; M.M. and F.Z. wrote the paper. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported in part by the Nature Science Foundation of Hunan Province (Grant No. 2020$JJ$4756)

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Kumar, K.; Santosh, G.S.K.; Varma, V. Multilingual Document Clustering Using Wikipedia as External Knowledge. In *Multidisciplinary Information Retrieval*; Hanbury, A., Rauber, A., de Vries, A.P., Eds.; Springer: Berlin/Heidelberg, Germany, 2011; pp. 108–117.
2. Jain, A.K. Data Clustering: 50 Years Beyond K-means. *Pattern Recognit. Lett.* **2010**, *31*, 651–666. [CrossRef]
3. Peng, M.; Zhu, J.; Wang, H.; Li, X.; Zhang, Y.; Zhang, X.; Tian, G. Mining Event-Oriented Topics in Microblog Stream with Unsupervised Multi-View Hierarchical Embedding. *ACM Trans. Knowl. Discov. Data* **2018**, *12*, 1–26. [CrossRef]
4. Peng, M.; Zhu, J.; Li, X.; Huang, J.; Wang, H.; Zhang, Y. 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, CIKM '15, Melbourne, Australia, 19–23 October 2015; ACM: New York, NY, USA, 2015; pp. 1611–1620. [CrossRef]
5. Ghosh, J.; Strehl, A. Similarity-Based Text Clustering: A Comparative Study. In *Grouping Multidimensional Data: Recent Advances in Clustering*; Springer: Berlin/Heidelberg, Germany, 2006. [CrossRef]
6. Liu, L.; Kang, J.; Yu, J.; Wang, Z. A comparative study on unsupervised feature selection methods for text clustering. In Proceedings of the International Conference on Natural Language Processing and Knowledge Engineering, Wuhan, China, 30 October–1 November 2005; pp. 597–601. [CrossRef]
7. Rahman, A.U.; Khan, K.; Khan, W.; Khan, A.; Saqia, B. Unsupervised Machine Learning based Documents Clustering in Urdu. *EAI Endorsed Trans. Scalable Inf. Syst.* **2018**, *5*. [CrossRef]
8. Alhawarat, M.; Hegazi, M. Revisiting K-Means and Topic Modeling, a Comparison Study to Cluster Arabic Documents. *IEEE Access* **2018**, *6*, 42740–42749. [CrossRef]
9. Chang, J.; Boyd-Graber, J.; Wang, C.; Gerrish, S.; Blei, D.M. Reading Tea Leaves: How Humans Interpret Topic Models. In *Neural Information Processing Systems*; Curran Associates, Inc.: Chatsworth, ON, Canada, 2009.
10. Jagarlamudi, J.; Daumé III, H.; Udupa, R. Incorporating Lexical Priors into Topic Models. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*; Association for Computational Linguistics: Avignon, France, 2012; pp. 204–213.
11. Pritchard, J.K.; Stephens, M.; Donnelly, P. Inference of Population Structure Using Multilocus Genotype Data. *Genetics* **2000**, *155*, 945–959.

12. Gad, W.K.; Kamel, M.S. Enhancing Text Clustering Performance Using Semantic Similarity. In *Enterprise Information Systems*; Filipe, J., Cordeiro, J., Eds.; Springer: Berlin/Heidelberg, Germany, 2009; pp. 325–335.
13. Blei, D.M. Probabilistic Topic Models. *Commun. ACM* **2012**, *55*, 77–84. [CrossRef]
14. Kelaiaia, A.; Merouani, H.F. Clustering with Probabilistic Topic Models on Arabic Texts. In *Modeling Approaches and Algorithms for Advanced Computer Applications*; Amine, A., Otmane, A.M., Bellatreche, L., Eds.; Springer International Publishing: Cham, Switzerland, 2013; pp. 65–74. [CrossRef]
15. Humayoun, M. *Urdu Morphology, Orthography and Lexicon Extraction*; CAASL-2, the Second Workshop on Computational Approaches to Arabic Script-based Languages; Linguistic Institute, Stanford University: Stanford, CA, USA, 2007.
16. Mukund, S.; Srihari, R.; Peterson, E. An Information-Extraction System for Urdu—A Resource-Poor Language. *ACM Trans. Asian Lang. Inf. Process. (TALIP)* **2010**, *9*, 1–43. [CrossRef]
17. Patil, A.; Pharande, K.; Nale, D.; Agrawal, R. Article: Automatic Text Summarization. *Int. J. Comput. Appl.* **2015**, *109*, 18–19.
18. Daud, A.; Khan, W.; Che, D. Urdu language processing: A survey. *Artif. Intell. Rev.* **2017**, *47*, 279–311. [CrossRef]
19. Shabbir, S.; Javed, N.; Siddiqi, I.; Khurshid, K. A comparative study on clustering techniques for Urdu ligatures in nastaliq font. In Proceedings of the 13th International Conference on Emerging Technologies (ICET), Islamabad, Pakistan, 27–28 December 2017; pp. 1–6.
20. Khan, N.H.; Adnan, A.; Basar, S. Urdu ligature recognition using multi-level agglomerative hierarchical clustering. *Clust. Comput.* **2018**, *21*, 503–514. [CrossRef]
21. Rafeeq, M.J.; Rehman, Z.; Khan, A.; Khan, I.A.; Jadoon, W. Ligature Categorization Based Nastaliq Urdu Recognition Using Deep Neural Networks. *Comput. Math. Organ. Theory* **2019**, *25*, 184–195. [CrossRef]
22. Khan, S.A.; Anwar, W.; Bajwa, U.I.; Wang, X. A Light Weight Stemmer for Urdu Language: A Scarce Resourced Language. In Proceedings of the 3rd Workshop on South and Southeast Asian Natural Language Processing, Mumbai, India, 8–15 December 2012; The COLING 2012 Organizing Committee: Mumbai, India, 2012; pp. 69–78.
23. Chandio, A.A.; Asikuzzaman, M.; Pickering, M.; Leghari, M. Cursive-Text: A Comprehensive Dataset for End-to-End Urdu Text Recognition in Natural Scene Images. *Data Brief* **2020**, *31*, 105749. [CrossRef]
24. Nasim, Z.; Haider, S. Cluster analysis of urdu tweets. *J. King Saud Univ. Comput. Inf. Sci.* **2020**, in press.
25. Nawaz, A.; Bakhtyar, M.; Baber, J.; Ullah, I.; Noor, W.; Basit, A. Extractive Text Summarization Models for Urdu Language. *Inf. Process. Manag.* **2020**, *57*, 102383. [CrossRef]
26. Bruni, R.; Bianchi, G. Website categorization: A formal approach and robustness analysis in the case of e-commerce detection. *Expert Syst. Appl.* **2020**, *142*, 113001. [CrossRef]
27. Ehsan, T.; Asif, H.M.S. Finding Topics in Urdu: A Study of Applicability of Document Clustering in Urdu Language. *Pak. J. Eng. Appl. Sci.* **2018**, *23*, 77–85.
28. Allahyari, M.; Pouriyeh, S.A.; Assefi, M.; Safaei, S.; Trippe, E.D.; Gutierrez, J.B.; Kochut, K. A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques. *arXiv* **2017**, arXiv:1707.02919,2017,
29. Aggarwal, C.C.; Zhai, C. A Survey of Text Clustering Algorithms. In *Mining Text Data*; Aggarwal, C.C., Zhai, C., Eds.; Springer US: Boston, MA, USA, 2012; pp. 77–128. [CrossRef]
30. Blei, D.M.; Ng, A.Y.; Jordan, M.I. Latent Dirichlet Allocation. *J. Mach. Learn. Res.* **2003**, *3*, 993–1022.
31. Paatero, P.; Tapper, U. Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. In Proceedings of the Fourth International Conference on Statistical Methods for the Environmental Sciences, Espoo, Finland, 17–21 August 1992.
32. Lee, D.D.; Seung, H.S. Learning the parts of objects by non-negative matrix factorization. *Nature* **1999**, *401*, 788–791. [CrossRef]
33. Wells, J.C. Orthographic Diacritics and Multilingual Computing. *Proc. Lang. Probl. Lang. Plan.* **2001**, *47*, 279–311. [CrossRef]
34. Griffiths, T.L.; Steyvers, M. Finding scientific topics. *Proc. Natl. Acad. Sci. USA* **2004**, *101*, 5228–5235, [CrossRef]
35. Lu, Y.; Mei, Q.; Zhai, C. Investigating task performance of probabilistic topic models: An empirical study of PLSA and LDA. *Inf. Retr.* **2011**, *14*, 178–203. [CrossRef]

36. Larsen, B.; Aone, C. Fast and Effective Text Mining Using Linear-time Document Clustering. In Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD ’99, San Diego, CA, USA, 15–18 August 1999; ACM: New York, NY, USA, 1999; pp. 16–22. [CrossRef]
37. Rijsbergen, C.J.V. *Information Retrieval*, 2nd ed.; Butterworth-Heinemann: Newton, MA, USA, 1979.
38. Rand, W.M. Objective Criteria for the Evaluation of Clustering Methods. *J. Am. Stat. Assoc.* **1971**, *66*, 846–850, doi:10.1080/01621459.1971.10482356. [CrossRef]

Publisher’s Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.