Discovering Coherent Topics from Urdu Text

Mubashar Mustafa[1], Feng Zeng[1]*, Hussain Ghulam[1], and Wenjia Li[2]†

[1] Central South University,Changsha, China
fengzeng@csu.edu.cn
[2] New York Institute of Technology, New York, USA
wli20@nyit.edu

Abstract

Topic modeling (TM), detection of theme or aspect from documents is an important text processing method in natural language processing (NLP) for helping users to get insights from a large number of documents. In recent years, many unsupervised models have been used in TM, and these models often produce aspects that are not interpretable. To figure out this issue, few semi-supervised methods have been developed that allow users to input some prior domain knowledge to produce coherent aspects. Most of them are well adapted to the English corpus, but there is very little work in Urdu. TM becomes a challenge for Urdu language having their own morphological structure, semantics, and syntax. In this paper, we first propose an effective semi-supervised topic model "Seeded-Urdu Latent Dirichlet Allocation (seeded-ULDA)" for Urdu language. The model is proposed to produce coherent topics dealing with the morphological structure of Urdu language. The proposed Urdu topic model Seeded-ULDA combines preprocessing, seeded-LDA, and Gibbs sampling. Second, we introduce word2vec word embedding in Urdu and discover topics through clustering of semantic space. This work aims to evaluate and compare various topic modeling frameworks in the Urdu news dataset. After comprehensive experiments and evaluation, the results show that word embedding is unable to extract coherent topics in Urdu language. The proposed seeded-ULDA model is more than 39% efficient as compared to existing ULDA model based on coherence measure.

keywords: Topic Modeling, Coherent topics, Word embedding ,Seeded-LDA, Natural Language Processing

1 Introduction

In this era, the explosive growth of electronic file archives has attracted a lot of attention. The report predicts that data storage capacity will increase to 40 trillion gigabytes by 2022, 50 times more than in early 2010 [2]. The most important concern now is to determine effective tools or methods that automatically organize, index, search, and browse this unstructured electronic text data. TM is one of the most widely used technology to organize these types of data. TM is a well-known advanced Machine learning technology. Using this technology, we can discover patterns that usually reflect basic themes. Given D is a set of documents composed of a set of terms W and T is a set of latent topics, TM will find T based on statistical inference on the term W. Thus, the document is a mixture of topics where topics represent a statistical distribution of words. A graphical representation of topic modeling is shown in Figure 1.

*Corresponding: Feng Zeng, School of Computer Science and Engineering, Central South University,Changsha, China (Email:fengzeng@csu.edu.cn)
†Corresponding: Wenjia Li, Department of Computer Science, New York Institute of Technology, New York, USA (Email:wli20@nyit.edu)

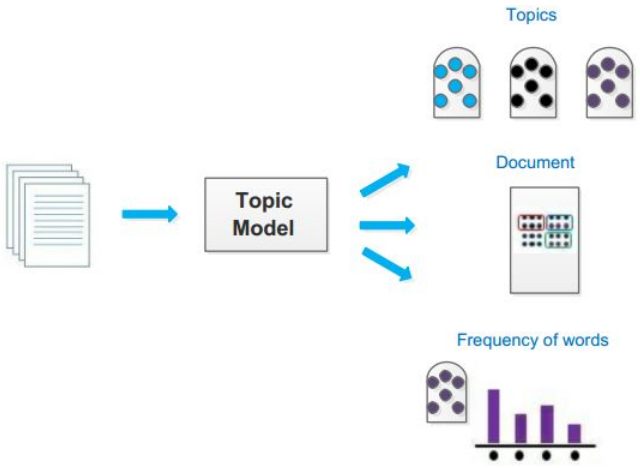

Figure 1: The graphical representation of topic modeling

In machine learning and NLP, the topic model is a statistical model used to discover the theme "topic" that occurs in the documents collection. TM is a widely used text mining tool for finding hidden semantic structures in text documents. A popular algorithm for modeling the text data is LDA, and different extensions have been proposed So far: Online variational inference for LDA in [5], Correlated topics Model (CorrLDA) [4], Hierarchical Topic Model (hLDA) [3], etc. Recently, Word2Vec [10] words Embedding has been used for theme extraction and achieved hopeful results [8][9]. However, the most commonly used topic model is LDA, which provides a powerful framework to extract hidden topics from text documents. But, the researchers found that extracted topics of unsupervised models are unexplainable or meaningless [12]. This is not a problem with LDA only: it is potentially a problem with any extension. Several knowledge-based models have been proposed to address this problem, such as seed-LDA, in which seed words are used as input to guide the model [6]. This model can produce coherent topics of particular interest to users.

Most topic models are designed, developed, and implemented for English text corpus. Therefore, these techniques are very effective for the English corpus. But Urdu has distinct nature from famous languages (such as Chinese, English, Arabic, etc.), and has distinct grammatical forms, Synonyms, antonyms of various words, morphological structure, Semantics, and syntax. Therefore, TM becomes a challenging task in Urdu and the limited contribution is committed to Urdu in NLP. There are many research communities for English and most software, application programming interface (API), and tools are specially developed for English which do not work effectively for Urdu language. To use these tools and software in Urdu language, further work will be required for better performance.

According to the literature review, there is little work on topic modeling in Urdu Language [15][16][7]. However, there is no work to extract coherent topics in Urdu language and it is first work for the extraction of coherent topics from Urdu documents. In this paper, first, we apply our proposed semi-supervised topic model Seeded-ULDA. Second, we introduce word embedding in Urdu language to discover topics by clustering of semantic space; generated through word2vec word embedding. After intensive examination, the results show that word embedding is unable to extract coherent topics in Urdu language and the semi-supervised model Seeded-ULDA outperforms ULDA based on coherence measure.

2 Methodology

In this section, we present the techniques used in this study for topic modeling. We will first focus on using the Word2Vec word embedding model and then discuss the Seeded-ULDA approach. All these techniques are implemented and made the performance comparison in our experiment.

2.1 Seeded-Urdu Latent Dirichlet Allocation (Seeded-ULDA)

We start with a short explanation of the effectiveness of the Seeded-ULDA. It is considered a challenging task to develop an efficient semi-supervised Urdu topic model "Seeded-ULDA" that combines preprocessing, seeded-LDA and Gibbs sampling [21]. Figure 2 gives a complete overview of the proposed model Seeded-ULDA. We introduce the technologies involved in Seeded-ULDA in the following subsections.

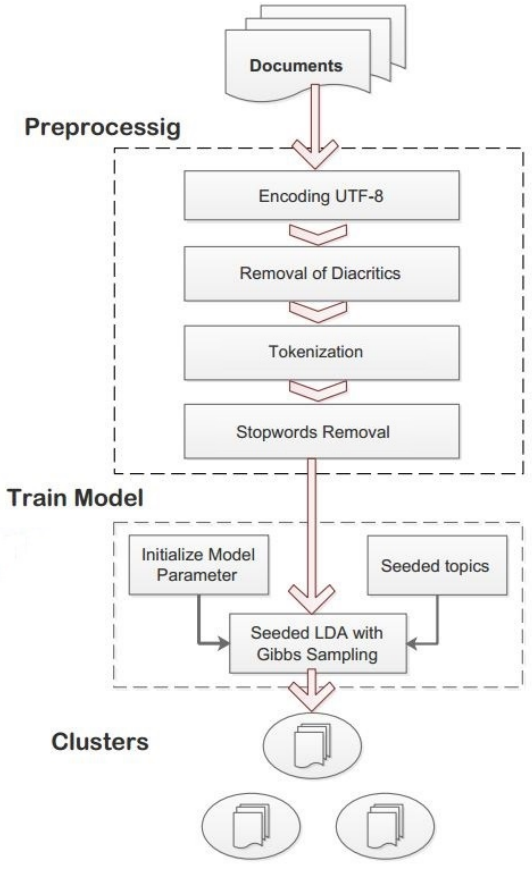

Figure 2: Proposed methodology Seeded-ULDA

2.1.1 Text PreProcessing

Text preprocessing aims to standardize the representation of texts to improve the accuracy of the topic detecting models. In this step, encoding UTF-8, diacritics removal, tokenization, and stop words removal will be performed to standardize the input dataset.

Enconding UTF-8 Computer programs face the problem of character recognition in Urdu text. We are using Unicode Transformation Format 8 (UTF-8) encoding for Urdu character recognition. Unicode is one of the most widely used encoding in the computer industry. UTF-8 means that every character is mapped by unique variable-width numeric code.

Removal of Diacritics A diacritic is a sign which is added with a letter to change the pronunciation. Urdu diacritics are a subset of Arabic diacritics. The most widely used diacritics in Urdu are Zabar, Pesh, and Zer which are called Aerab [20], and other diacritics are seldom used. When it is attached to a word, the sound and meaning of the word change [19]. Urdu is usually written with words only and diacritics are left to personal choice. But we discard diacritics to form the dataset standardise. Finger 3 shows some Urdu diacritics examples.

Without diacritics	Meaning	With diacritics	Meaning
تیر	Swim (Taer)	تیر	Arrow (Teer)
سونا	Gold (Soona)	سونا	Sleep (Sona)
میں	I (main)	میں	In (mein)

Figure 3: Urdu diacritics Example

Tokenization Tokenization plays an important role in text analysis tasks. Tokenization is the process to split text documents into tokens, and the token is an individual instance of a sequence of characters in natural language. Tokenization is a component of the methodology, and then we use it to teach machines to understand words. Many tokenization techniques have been proposed, but in this work, we use count-vectorizer to split the text into tokens.

Stopwords Removal In NLP, stopwords removal is the part of pre-processing to remove unworthy data and unworthy words (data) are regarded as stopwords. They do not make the addition of meaning or information and are found frequently in a sentence. We can safely ignore them without losing the information of the sentence. In order to get meaningful data, we exclude stopwords from our corpus. Few mostly used stopwords of Urdu are shown in figure 4.

کیلیے	گئی
میں	کی
سے	گیا
نے	دیے
لیا	ہی

Figure 4: Stopwords of Urdu

2.1.2 Seeded-LDA

This approach allows a user to guide the topic discovery process by letting him provide seed words that are representative of the corpus [6]. This model can use the seed words in two ways: to improve both topic-word and document-topic probability distributions. To improve topic-word distributions, the model is set up in which each topic prefers to generate terms that are same to the terms in a seed set. To improve document-topic distributions, the model is encouraged to select document-level topics based on the existence of input seed words in that document. Our work aim is to produce coherent topic. So, we use the first way to improve topic-word probability distribution. In traditional topic models, Multinomial distribution ϕ_k expresses each topic k over words. This notion is extended and the topic is defined as an inter-mixture of two different distributions: a regular topic distribution and a seed topic distribution. In seed topic distribution, the words are generated from the given seed set. In regular topic distribution, any words can be generated including seed words. It is emphasized that, like ordinary topics, all words of seed topics have non-uniform probability distribution. The model takes a set of seed words as input, and then outputs the probability distribution of these words.

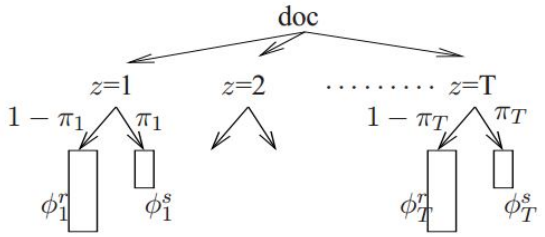

Figure 5: Tree representation of a document in seeded-LDA model

For simplicity, the model is explained by considering one-to-one compatibility between reg-

ular and seed topics. when regular topics are more, then this consideration can be simply relaxed by making copies corresponding to the seed topics. The figure 5 shows that documents are a mixture of topics T where these topics are a mixture of seed topics ϕ^s and regular topics ϕ^r. The probability of picking a term from the regular topic distribution versus the seed topic distribution is controlled by parameter π_k. The graphical notation is shown in Figure 6 and the generative process of seeded-LDA is as follow:

- For each topic $k = 1...T$
  - Draw regular topic $\phi^r k$ ~Dir(β_r).
  - Draw seed topic $\phi^s k$ ~Dir(β_s).
  - Draw $\pi_k \sim \beta(1, 1)$.
- For each document d, Choose θ_d ~Dir(α)
  - Select a topic z_i ~Mult(θ_d).
  - Select a indicator x_i ~Burn(π_{zi}).
  - if $x_i = 0$, Select a word w_i ~Mult(ϕ^r_{zi}). // choose from regular topic.
  - if $x_i = 1$, Select a word w_i ~Mult(ϕ^s_{zi}). // choose from seed topic.

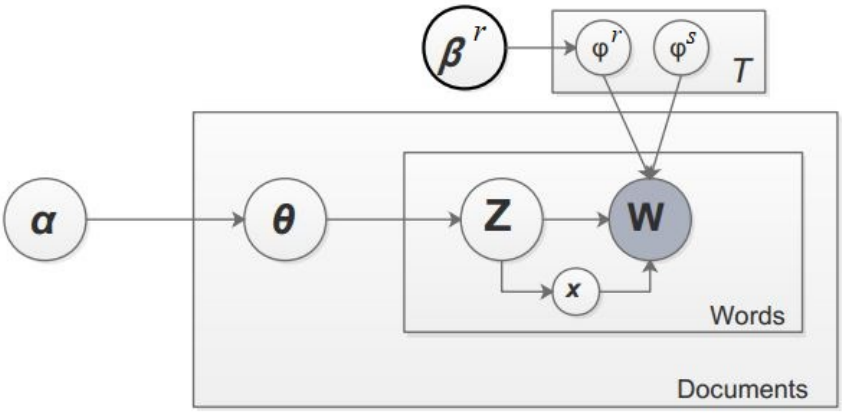

Figure 6: The graphical model of seeded-LDA [6]

2.1.3 Gibbs sampling

When direct sampling is hard, then Gibbs sampling is used to get a sequence of observations by Markov Chain Monte Carlo (MCMC) which is a methodology to sample from statistical distribution [17][18]. It is an algorithm that uses the logic of randomness means it sample randomly, and is widely employed as a statistical inference. In this paper, we apply Gibbs sampling with the probabilistic generative model to discover coherent topics.

2.2 Word2vec

The Word2vec model is a two-layer neural network that can be trained to recreate the language context of words [10]. It captures the context, semantic and syntactic similarity of words in a document. Word2Vec is one of the most widely used method of learning word embedding using shallow neural networks. It takes a large number of text corpora as input, and generates a vector space that can have hundreds of dimensions, and assigns a corresponding vector in the space to each unique word in the lexicon. This technique is different from other topic models, which use documents as context. Word2Vec learns the distributed representation of each target word by identifying the context as surrounding terms.

3 Experimental studies

We presented some experiments to demonstrate the effectiveness of the above defined topic modeling techniques. These experiments were performed using two corpuses discussed in the following section.

3.1 Dataset

Urdu language is not owning any benchmark dataset for NLP tasks. Therefore, we create our own corpus, which contains text articles in Urdu of a widely known news website https://www.express.pk. It is now publicly available at https://github.com/Mubashar331/Urdu-corpus for research purposes. The collected dataset has five categories of documents named Health, Sports, Science, Entertainment, and Business. We also collected a dataset of English having four categories. After the completion of preprocessing steps that are discussed in the above subsections, we applied above defined topic modeling techniques on these corpora and evaluated performance. Table 1 is briefly describing the corpus.

Table 1: Detail of Dataset

Sr.	**Corpora**	**Description**	**No. of Classes**	**Total words**
1	**Corpus 1**	Urdu news articles from Urdu Express Newspaper	5	20289
2	**Corpus 2**	English news articles from different Newspaper such as Dawn news, Express	4	11771

3.2 Experiments

We presented three experiments to evaluate the performance of topic modeling techniques. These experiments are performed on dataset or corpus discussed in above subsection.

3.2.1 Experiment 1: Topic Modeling by word2vec

The purpose of this experiment is to evaluate the accuracy of word2vec on Urdu text documents. Word2vec has two types of architecture, Skip Gram Model and Continuous Bag of Words (CBOW), to gain vectors of features. In this study, we use the CBOW model to build vectors of a given pre-processed dataset. Then, we cluster the gained vectors of features using K-means method. The process of topic modeling by word2vec is shown in figure 7.

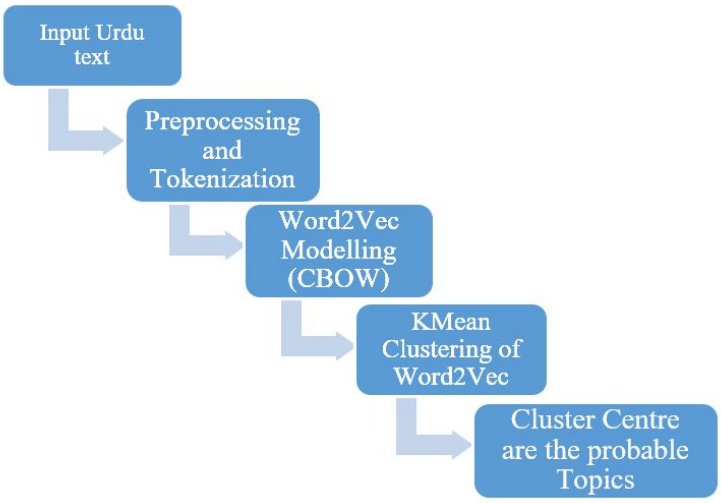

Figure 7: The process of topic modeling by word2vec

3.2.2 Experiment 2: Topic Modeling by Seeded-ULDA

Seeded-ULDA is a semi-supervised technique that integrates previous domain knowledge into topic models to contribute in producing consistent topics. In order to extract coherent topics, it allows a user to guide the model by giving a set of seed words as input that are representative of the given corpus. In this experiment, we use Urdu dataset which contains 5 classes. So, we incorporate five seeded topics manually into topic models to contribute in generating consistent topics. The first ten words of the seeded topic of each class are shown in figure 8.

1	فلموں ڈراموں اداکارہ ہالی ووڈ بالی ووڈ سینما گلوگار رقص ماڈلنگ فنکاروں
2	خزانہ بینک اکاؤنٹ انکم ٹیکس کسٹمز ریونیو ٹیکس بجٹ تجارت معیشت
3	مریض ڈاکٹر حاملہ اسپتال انجکشن کینسر ذیابیطس سرجری آپریشن ملیریا
4	علیم ڈار میچز امپائرنگ کرکٹ ایشیاکپ ہاکی فٹبال ٹیموں اسٹیڈیم کھلاڑیوں
5	فون موبائل ناسا فیس بک ستارے روبوٹس سٹلائٹ مریخ اسٹیفن ایپل

Figure 8: First ten words of seeded topic of each class

3.2.3 Experiment 3: Topic Modeling by ULDA

In this experiment, we compare the accuracy of our proposed model with the ULDA model proposed in an article [15]. This model was proposed to discover topics from a corpus of Urdu news articles. We employ the ULDA model on our own corpus. We ran this model several times to evaluate results on topics discovered from Urdu text documents.

4 Evaluation and Result Discussion

In the NLP research community, evaluation of discovered topics from the topic model is considered an open challenge [13]. Some researchers use internal evaluation methods such as perplexity or likelihood for evaluating topic models. But these evaluation methods cannot measure the consistency of discovered topics. Through large-scale user studies, an author [13] argued that the topic model which performed well on perplexity or likelihood failed to produce coherent topics. Therefore, we do not use these evaluation methods to evaluate above defined topics models.

We evaluate topic models by using a manual evaluation technique named Coherence Measure (CM), it is the ratio of relevant words to total candidate words of a topic [14]. For this manual evaluation technique, we take 20 words with the highest values of every topic and request 5 students to examine and label them. First, they need to examine the words to label a topic as interpretable or irrelevant. When the topic is interpretable, then they need to inquire about the words of a topic that are relevant. CM is calculated using equation 1 where x is relevant words and n is total candidate words.

$$CM = \frac{x}{n} \tag{1}$$

We presented the experiments to demonstrate the effectiveness of topic modeling techniques based on CM. First, we find topics by word2vec and mostly extracted topics are labeled irrelevant. Second, we extract five topics by LDA and ULDA from Urdu corpus and find out that all topics extracted by LDA are labeled irrelevant. We show seven words of one topic in figure 9 which is extracted by LDA and ULDA. Topic extracted by LDA is irrelevant and does not belong to any class of our given dataset. But Topic extracted by ULDA is relevant and belongs to the Health class of our given dataset.

Topic by LDA		Topic by ULDA	
سال	Year	ریڈیو	Radio
فیصد	Percentage	افراد	Persons
فلم	Film	صحت	Health
کام	work	کینسر	Cancer
حکومت	Government	ڈاکٹر	Doctor
پاکستان	Pakistan	علاج	Treatment
وقت	Time	جسم	Body

Figure 9: Topic extracted by LDA and ULDA from Urdu corpus

Third, we extract five topics by Seeded-LDA and our proposed model Seeded-ULDA from Urdu corpus and find out that all topics extracted by Seeded-LDA are labeled irrelevant. As

shown in figure 10, the topic extracted by Seeded-LDA is irrelevant and does not belong to any class of our given dataset. But the topic extracted by Seeded-ULDA is relevant and belongs to the Health class of our given dataset.

Topic by Seeded-LDA		**Topic by Seeded-ULDA**	
ماہرین	Experts	مرض	Disease
پانی	Water	صحت	Health
زیادہ	Many	ڈاکٹر	Doctor
تحقیق	Research	علاج	Treatment
کراچی	Karachi	کینسر	Cancer
بات	Talk	دوائی	Medicine
روپے	Rupee	مریض	Patient

Figure 10: Topic extracted by Seeded-LDA and Seeded-ULDA from Urdu corpus

Finally, we apply LDA and sedded-LDA topic modeling techniques to the English corpus and find out that all extracted topics are labeled irrelevant. Then, we combine both models with Gibbs Sampling(GS) and extract topics from the English corpus. The results demonstrate that topics extracted by seeded-LDA(GS) are more coherent as compare to LDA(GS). As shown in table 2, the topic extracted by LDA and seeded-LDA does not belong to any class of given corpus. But the topic extracted by LDA(GS) and seeded-LDA(GS) belong to the health class of the English corpus.

Table 2: Topic words extracted from English corpus

LDA	**LDA(GS)**	**seeded-LDA**	**seeded-LDA (GS)**
Study	Brain	Pakistan	Cancer
Cancer	Health	Apple	Health
year	Risk	Tax	Blood
People	Says	Coronavirus	Found
Mice	People	Google	Disease
Week	Blood	Million	Studies
Blood	Research	People	Glucose

Now, we calculate the CM of Seeded-ULDA and ULDA. As can be seen in Table 3, the average CM calculated by Seeded-ULDA surpasses the ULDA model. We calculate CM of Seeded-LDA(GS) and LDA(GS) from English corpus and results are shown in Table 4. The results demonstrate that Seeded-LDA(GS) gives better results than LDA(GS) based on CM. Now, we examine the influence of minimum documents frequency ($min - df$) parameter on both model Seeded-ULDA and ULDA from Urdu corpus. We set the value of $min - df$ 1 and 2, then we examine the influence based on CM. As shown in figure 11, Both models produce more coherent topics with $min - df = 2$ and our proposed model Seeded-ULDA is better than ULDA.

Table 3: Results comparison of ULDA and seeded-ULDA with Urdu corpus

Class	**ULDA**	**seeded-ULDA**
annotator 1	0.34	**0.49**
annotator 2	0.28	**0.42**
annotator 3	0.42	**0.57**
annotator 4	0.39	**0.53**
annotator 5	0.40	**0.55**
average	0.366	**0.512**

Table 4: Results comparison of LDA(GS) and seeded-LDA(GS) with English corpus

Class	**LDA(GS)**	**seeded-LDA(GS)**
annotator 1	0.39	**0.55**
annotator 2	0.44	**0.47**
annotator 3	0.41	**0.52**
annotator 4	0.37	**0.58**
annotator 5	0.43	**0.50**
average	0.408	**0.524**

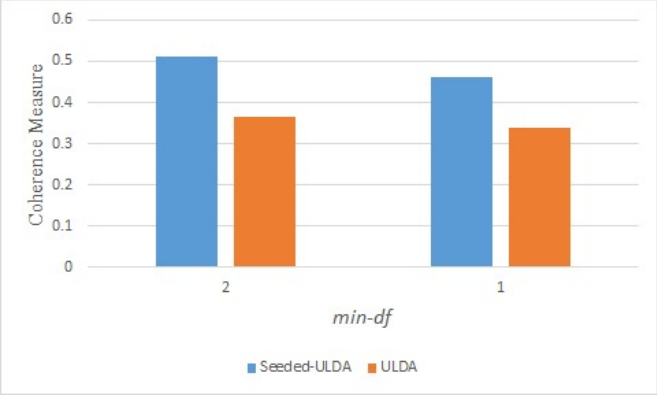

Figure 11: Influence of minimum documents frequency parameter

5 Conclusion

Regular unsupervised topic models might not produce coherent topics due to their pure unsupervised nature. Several knowledge-based topic models have been proposed to address this problem, but most of them are for English. Therefore, NLP research involving the Urdu language is comparatively hard, compared to other popular languages, due to the speciality of Urdu such as syntax, semantics, and morphological structure. The main motivation behind

this research is the lack of NLP resources and tools for the Urdu language. There are no important studies for extracting coherent topics from Urdu texts in literature. To meet the challenges of Urdu, we have proposed a topic model Seeded-ULDA for Urdu language to produce coherent topics. In order to evaluate the performance and effectiveness of our proposed Seeded-ULDA model, we conducted three experiments using the Urdu dataset generated by ourselves. The results demonstrate that unsupervised models produce less coherent or meaningless topics compared to semi-supervised framework. First, we apply word2vec word embedding and result shows that it is unable to extract coherent topics. In the second and third experiments, we apply Seeded-ULDA and ULDA respectively and results show that semi-supervised model Seeded-ULDA produces more than 39% coherent topics compared to unsupervised model ULDA.

References

[1] David Blei, Andrew.NG and Michael Jordan,.(2003). Latent Dirichlet Allocation. Journal of Machine Learning Research. 3. 993-1022. 10.1162/jmlr.2003.3.4-5.993.

[2] J. Ganz and D. Reinsel, THE DIGITAL UNIVERSE IN 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the far East, Technical Report 1. IDC, Framingham, Dec. 2012, pp. 1-16

[3] T. Griffiths, M. Jordan, J. T.-A. in neural, and undefined 2004, Hierarchical topic models and the nested chinese restaurant process, papers.nips.cc.

[4] D. Blei, J. L. the 18th I. C. on N., and undefined 2005, Correlated topic models, papers.nips.cc.

[5] C. Wang, J. Paisley, D. B. the F. I. C. on, and undefined 2011, Online variational inference for the hierarchical Dirichlet process, jmlr.org.

[6] J. Jagarlamudi, H. Daum III, and R. Udupa, Incorporating lexical priors into topic models, in Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics. Avignon, France: Association for Computational Linguistics, Apr. 2012, pp. 204-213.

[7] A. Ur Rehman, A. H. Khan, M. Aftab, Z. Rehman and M. A. Shah, "Hierarchical Topic Modeling for Urdu Text Articles," 2019 25th International Conference on Automation and Computing (ICAC), Lancaster, United Kingdom, 2019, pp. 1-6. doi: 10.23919/IConAC.2019.8895047

[8] Sabitra Sankalp Panigrahi, Modelling of Topic from Hindi Corpus using Word2Vec 2018 Second International Conference on Advances in Computing, Control and Communication Technology (IAC3T)

[9] Esposito, Fabrizio et al. Topic Modelling with Word Embeddings. CLiC-it/EVALITA (2016).

[10] T Mikolov and J Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. Advances in neural information processing systems.

[11] Marco Baroni, Georgiana Dinu, and German Kruszewski. 2014. Dont count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In ACL (1), pages 238-247.

[12] Mimno, D., Wallach, H.M., Talley, E., Leenders, M. and McCallum, A. 2011. Optimizing semantic coherence in topic models. EMNLP, 262-272.

[13] Jonathan Boyd-Graber, Jordan Chang, Sean Gerrish, Chong Wang, and David Blei. Reading tea leaves: how humans interpret topic models. In Proceedings of the 23rd Annual Conference on Neural Information Processing Systems, 2009.

[14] Xie, Pengtao and Xing, Eric. (2013). Integrating Document Clustering and Topic Modeling.

[15] Shakeel, Khadija and Tahir, Ghulam Rasool and Tehseen, Irsha and Ali, Mubashir. (2018). A framework of Urdu topic modeling using latent dirichlet allocation (LDA). 117-123. 10.1109/C-CWC.2018.8301655.

[16] Rehman, Anwar Ur and Rehman, Zobia and Akram, Junaid and Ali, Waqar and Shah, Munam and Salman, Muhammad. (2018). Statistical Topic Modeling for Urdu Text Articles.

[17] Walter R Gilks. Markov chain monte carlo. Wiley Online Library, 2005.

[18] Robert P Dobrow. Markov chain monte carlo. Introduction to Stochastic Processes With R, pages 181-222, 2016.

[19] Wells, J. C. Orthographic Diacritics and Multilingual Computing. In Proceedings of Language Problems and Language Planning, 2001.

[20] A. Daud, W. Khan, and D. Che, Urdu language processing: a survey, Artif. Intell. Rev., vol. 47, no. 3,pp. 279-311, Mar. 2017.

[21] Mustafa, M.; Zeng, F.; Ghulam, H.; Muhammad Arslan, H. Urdu Documents Clustering with Unsupervised and Semi-Supervised Probabilistic Topic Modeling. Information 2020, 11, 518.