ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645
Indian Journal of Science and Technology, *Vol 12(45), DOI: 10.17485/ijst/2019/v12i45/145722, December 2019*

A Comparison of Topic Modelling Approaches for Urdu Text

Siraj Munir, Shaukat Wasi and Syed Imran Jami*

Department of Computer Science, Mohammad Ali Jinnah University, Karachi, Pakistan; imran.jami@jinnah.edu, shaukat.wasi@jinnah.edu, sirajmunir93@gmail.com

Abstract

Objectives: Machine learning based approaches for topic modeling are successful in extracting logical and semantic topics from a given collection of text. We experimented topic modelling approaches for Urdu poetry text to show that these approaches perform equally well in any genre of text. **Methods:** 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. **Findings:** Before forwarding the raw text to aforementioned models, we did feature 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 our test samples, we used coherence and dominance model. **Applications:** Our experiment shows that LDA, and LSI performed well on CORPUS dataset but none of the mentioned approaches performed well on poetry text. This brings us to a conclusion that we need to devise sequence based models that allow users to define weights for poetry specific text. This work opens a new direction for the domain of text generation and processing.

Keywords: LDA, LSI, HDP, Urdu Poetry Processing, Urdu Poetry Collection, Topic Modelling.

1. Introduction

Machine learning is an approach to train machines on doing specific task efficiently. Machine learning has produced great results for different genre of real world problems. Natural Language Processing (NLP) is an immense branch of computer science which deals with understanding and processing human languages. Topic modelling is an area of Natural Language Understanding (NLU) based on statistical modelling to discover keywords which can represent complete/partial document using a dimension reduction technique which is applied on text data.[1] In this study, we have compared three different models for topic modelling that is Latent Dirichlet Allocation (LDA), Latent Semantic Indexing (LSI), and Hierarchical Dirichlet Process (HDP) for topic modelling on Urdu News and Urdu poetry corpuses. LSI is a topic modelling approach which uses low rank approximation over Single Value Decomposition (SVD). LSI uses term-document matrix integrated with SVD and occurrence matrix for its complete processing cycle. Occurrence matrix is same as term-frequency matrix, but here it is sparse in nature.

Finally, by using aforementioned techniques LSI reduces dimensions of given text data. LDA is an effective statistical model which allows us to extract latent representations from the document.[2,3]

LDA is a parametric Bayesian version of Probabilistic LSI (PLSI). LDA enforce two models that is topic-document and word-topic models. Dirichlet in LDA is a multinomial probability distribution with probability simplex. Probability simplex is a collection of numbers summed up to 1.[4] HDP is a nonparametric Bayesian and generalized Hidden Markov Model (HMM) approach for topic modelling. HDP clusters grouped data by using Dirichlet process. However, it is more like LDA

*Author for correspondence

at group level. For evaluation of proposed models, we used coherence model and dominant topic. Coherence model is used to measure the quality of generated topic words. Furthermore, after applying coherence model we computed top ten generated topic words with their respective probabilities. Topic model consist of multiple topics by using a dominant topic technique we can find out the dominant topic words. Most often topics contain a single dominant topic. This study is divided into the following sections. Introduction briefly discusses the topic modelling techniques. Methodology section discusses the adopted models in detail. In data acquisition section, we discuss the problems regarding collection of datasets. It also discusses a novel Urdu poetry dataset. The results and comparison section discuss our contribution and achieved results on each dataset. Finally, the conclusion section discusses the concluding remarks and future work.

2. Literature Survey

In this section, we will put some light on recent works proposed in the area of machine learning for topic modelling. For review, we just considered recent study from renowned databases including IEEE, Science Direct, and Nature journal. In Ref.,[5] literature proposed an event ranking algorithm based on daily news. Authors also proposed a novel event mining and feature generation approach. For evaluation authors tested the proposed model on real world large scaled data. In Ref.,[6] literature proposed a supervised temporal topic modelling approach. Proposed methodology was acquired for the topic modelling of internet news about different diseases. Authors evaluated their methodology in an outbreak disease report of USA, China, and India. In Ref.,[7] Chen et al. proposed two novel approaches for topic modelling including temporal distance and lexical similarity approach. Authors implemented a variation of LDA named (LapPLDA) Laplacian probabilistic latent semantic analysis. Trial results shown an excellent F1-score of 0.8 (80% accuracy). In Ref.,[8] Gui and Wang proposed an Apache Spark implementation of LDA model on MLIB. Proposed model was piloted with Scala a next generation functional programming language. In Ref.,[9] Zhang et al. proposed a novel story discovery model for news and Twitter feeds. Proposed model used three-step incremental model that includes discovering of essential information from data sources and then modelling topic words from raw data 3 for a scalable solution. In Ref.,[10] Larsen and Thorsrud discussed importance of news for economic development. Authors used LDA on Norwegian business newspaper and tried to find out latent topic words which may represent economic fluctuation. In Ref.,[11] the authors proposed Deep Learning model named Doc2Vec which uses Natural Language Processing and embedding approach. For visualization, authors used Markov Stability model. In Ref.,[12] the authors proposed LDA, LSI, and Doc2vecmodels using Bangla news corpus. Evaluations showed Doc2vec feasibly outperformed LDA and LSI models.

In Ref.,[13] Asadi Kakhki et al. proposed LDA model implementation for financial news. Authors considered one-year financial news as their dataset. Experiments showed that LDA outperformed standard classification approaches for financial topic modelling. In Ref.,[14] Hidayatullah et al. proposed LDA model for football news tweets in Bahasa language. Dataset was extracted from official Indonesian Twitter accounts. By applying LDA authors successfully inferred pre-match analysis, live match update, football club achievements, etc.

Recently in Ref.,[15] Shakeel proposed a Urdu LDA (ULDA) model. Authors claimed that the proposed model is the first attempt of topic modelling in Urdu language. Proposed model used pre-processing, standard LDA and Gibbs sampling for evaluation. In this work, we will compare the top notch topic modelling algorithms like LDA, LSI, and HDP.

For comparison, we used three datasets (i) Urdu News CORPUS dataset, (ii) Urdu poetry dataset 1, and (iii) Urdu poetry dataset 2. As per the available literature, this work is the first attempt towards topic modelling for Urdu poetry text. The next section of literature will discuss our methodology.

3. Methodology

This section will discuss our methodology. This section is divided into three sub-sections. Each sub-section will briefly discuss implementation of individual model. For each model, we have done some preprocessing including tokenization, lemmatization, and stop word removal. Figure 1 depicts the general model that we have adopted for each topic modelling technique. Whereas Figure 2 depicts dataflow pipeline for topic modelling. Our proposed models are fully unsupervised due to which we have trained our model on preprocessed raw text.

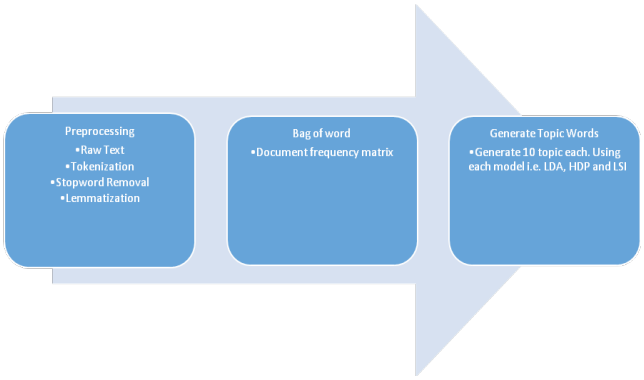

Figure 1. General roadmap model.

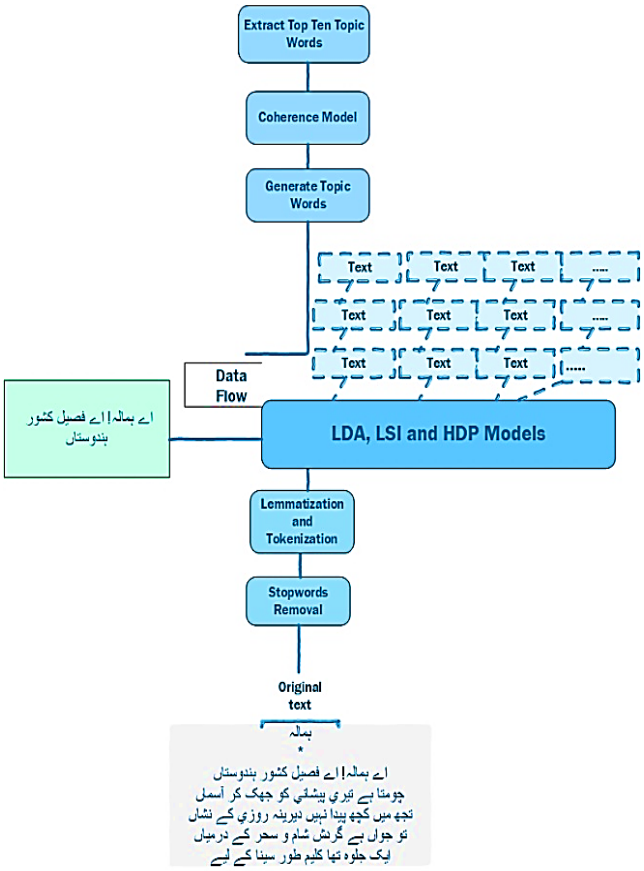

Figure 2. Data flow representation for topic modelling.

3.1. LSI Model

LSI is the first implemented model. LSI was briefly discussed in the previous section. For LSI model, we took preprocessed data. We used python's library for topic modelling named Gensim. Gensim has optimized implementation for each topic modelling model that is LDA, HDP, and LSI. LSI uses the following mathematical model for topic word extraction:

$$\omega_{i,j^1} = tf_{i,j} \times \log \frac{N}{df_j}$$

where ω in equation represents term-document score while $tf_{i,j}$ represents occurrence matrix and $\frac{N}{df_j}$ is total documents over documents containing topic words. From each model, we generated ten topic words. Figure 3 shows generated topic words along their respective probabilities using LSI model.

```
[0),
 '0.496*"عمر"*0.131 + "جائنہ"*0.131 + "انسٹیٹیوٹ"*0.131 + "افغان"*0.173 + "عمران"*0.181 + "پاکستان"*0.204 + "رپورٹ"*0.
('"ناع,
 1),
 '0.424*"نادر"*0.163 + ".ہوئے"*0.179- + "انٹیز"*0.179- + "میریز"*0.182- + "ملاقات"*0.186 + "ٹربیونل"*0.196 + "میٹیا"*0.
('"ا,
 2),
 '0.306*"ولیمس"*0.170 + "ساؤتھی"*0.170 + "کیویز"*0.204 + "افغانستان"*0.210- + ".ہوئے"*0.238 + "انٹیز"*0.238 + "عمران"
```

Figure 3. LSI generated topic words.

The next sub-section will briefly discuss LDA model implementation.

3.2. LDA Model

In this sub-section, we will discuss implementation of LDA model. The basic idea was briefly discussed in the previous section. Figure 4 explains the process how LDA works and generate topic words using Dirichlet distribution.

1. Choose $\theta_i \sim \text{Dir}(\alpha)$ (where $i = 1, \ldots, M$; $\theta_i \in \Delta_K$)
   - $\theta_{i,k}$ = probability that document $i \in \{1, \ldots, M\}$ has topic $k \in \{1, \ldots, K\}$.
2. Choose $\phi_k \sim \text{Dir}(\beta)$ (where $k = 1, \ldots, K$; $\phi_k \in \Delta_V$)
   - $\phi_{k,v}$ = probability of word $v \in \{1, \ldots, V\}$ in topic $k \in \{1, \ldots, K\}$.
3. Choose $c_{i,j} \sim \text{Polynomial}(\theta_i)$ (where $c_{i,j} \in \{1, \ldots, K\}$)
4. Choose $w_{i,j} \sim \text{Polynomial}(\phi_{c_{i,j}})$ (where $w_{i,j} \in \{1, \ldots, V\}$)

Figure 4. LDA model algorithm.[19]

Figure 5 depicts the generated topic words along with their probabilities. The next sub-section will discuss HDP model.

```
[0),
 '0.496*"عمر"*0.131 + "جائنہ"*0.131 + "انسٹیٹیوٹ"*0.131 + "افغان"*0.173 + "عمران"*0.181 + "پاکستان"*0.204 + "رپورٹ"*0.
('"ناع,
 1),
 '0.424*"نادر"*0.163 + ".ہوئے"*0.179- + "انٹیز"*0.179- + "میریز"*0.182- + "ملاقات"*0.186 + "ٹربیونل"*0.196 + "میٹیا"*0.
('"ا,
 2),
 '0.306*"ولیمس"*0.170 + "ساؤتھی"*0.170 + "کیویز"*0.204 + "افغانستان"*0.210- + ".ہوئے"*0.238 + "انٹیز"*0.238 + "عمران"
```

Figure 5. LDA generated topic words.

3.3. HDP Model

HDP is a non-parametric approach that is auto-optimized. Auto-optimized means HDP adjust its parameters while training automatically. HDP uses the following mathematical model for extraction of topic words.

$$x_{j,i} \mid \mathcal{G}_j \sim \sum_{h}^{\infty} \pi_{j,h} \mathcal{F}\left(\theta_h^*\right)$$

where θ_h^* is mixture component parameter and $\pi_{j,h}$ is the mass of mixing proportion. By using this equation, we can interpret each component which is modelling clusters of data items.[16] The next section will discuss hurdles we faced during collection of data.

4. Data Acquisition and a Novel Dataset

In this section of the study, we will put some lights on data collection and the development of our dataset. One of the core tasks in the development of any machine learning based models is the collection of data, as whole learning/algorithm during training and testing will be dependent on data. From data machine learning model learn hidden features and patterns. By using learned patterns machine learning model produce effective results. However, the collection of data is not trivial due to annotation, labeling, and feature engineering. The corpus of Urdu dataset lacks volume and variation for training. This problem is resolved by using the poetry of Allama Iqbal. In[1] this dataset was then cleaned for experiments.

For Urdu news dataset, we simply used a collection of news corpus.[17] For pilots we also contributed a novel dataset. Proposed dataset consist of four different styles of poetry that is romantic poetry, religious poetry, serious poetry and humorous poetry. We collected poetry text from different blogs and webpages. The next section will discuss our results and compare proposed models.

5. Results and Comparison

In this section of the study, we will put some lights on our achieved results. Following sub-sections discuss obtained results.

5.1. Urdu News Corpus Results (Unsupervised)

In the first model of our work, we implemented three different variants of topic modelling algorithms namely LDA, LSI, and HDP. All the mentioned approaches where applied to preprocessed raw data as mentioned in Figure 1.

We used CORPUS dataset which includes news headlines and description. The dataset contains 600 different Urdu News like showbiz, sports, international, etc.

We implemented each algorithm in unsupervised manner that is after preprocessing we passed extracted text to LDA, LSI and HDP and generated topic words from it. After generating topic words from LDA, we observed that each topic represents same class like sports, international/national, etc. Figure 6 depicts achieved results from LDA which shows that topic number 4 contains topic words from sports category. Furthermore, the generated topic words also maintain good semantic relation between topic words.

Figure 7 depicts LSI model results achieved on Urdu news dataset. Results show that LDA performed better than LSI. LDA generated topic words have better semantic relation than topic words generated by LSI. Generated topic also maintains semantic relation between topic words. LSI topic 1 show keywords extracted from national and international news category.

Figure 8 depicts HDP model results achieved on an Urdu news dataset. After overseeing results, we concluded that HDP did not perform well. HDP generated topics and topic words were over-fitted to national and international category of dataset. After implementation of each topic

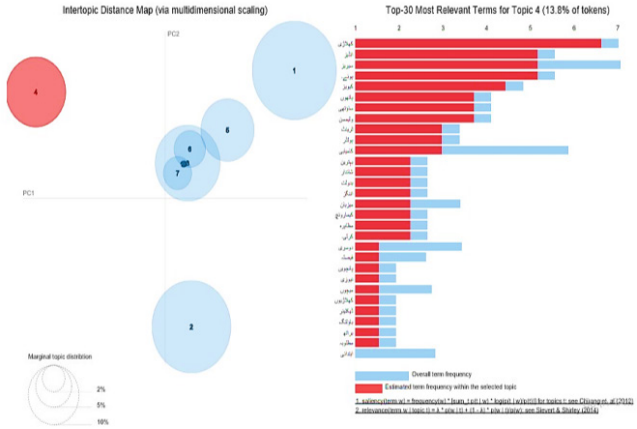

Figure 6. LDA results on Urdu news corpus (unsupervised).

[1]https://en.wikipedia.org/wiki/Muhammad_Iqbal

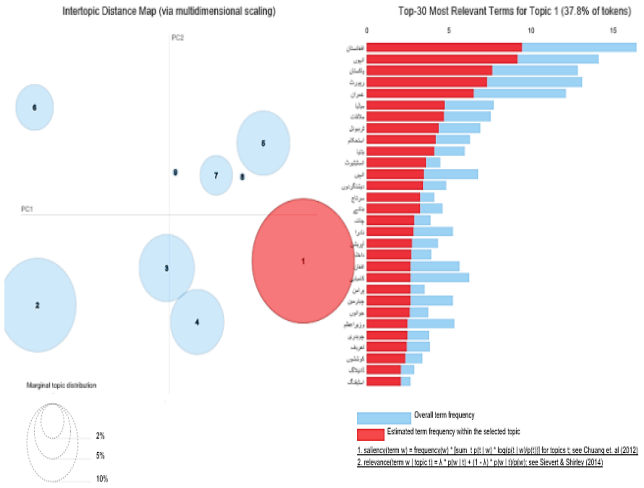

Figure 7. LSI results on Urdu news corpus (unsupervised).

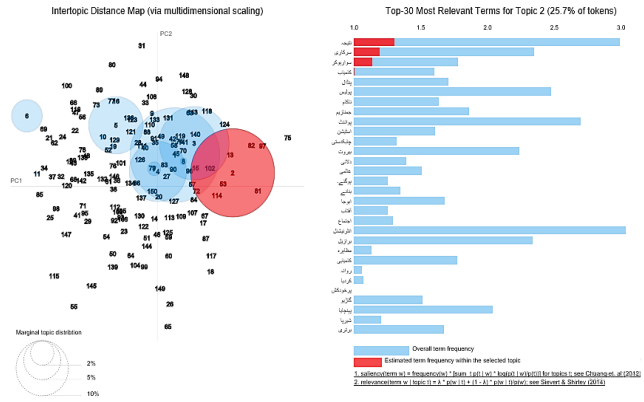

Figure 8. HDP results on Urdu news corpus (unsupervised).

modelling technique we implemented coherence model. Through coherence model we carried out top ten topic generated by each model.

These topics were sorted by highest to lowest probability. Furthermore, after applying models we also evaluated them by measuring the contribution and dominance of generated topics respectively as shown in Figure 8. Comparison shows that LDA outperformed LSI and HDP in extracting useful topics. In next section we will discuss implementation of same topic modelling approaches and there results when applied over poetry dataset 1.

5.2. Allama Iqbal Urdu Poetry Corpus Results (Unsupervised)

In this section, we will discuss the implementation of LDA, LSI, and HDP model over the Allama Iqbal Urdu Poetry Corpus. Corpus contains poetry collection of Dr. Muhammad Allama Iqbal who has the vast collection of Urdu poetry. For each model, we implemented methodology mentioned in Figure 1. After retrieving tokenized text, we implemented each of the mentioned models. The general observation is topic modelling approaches did not perform well on Urdu poetry text. The major reason for this failure is the difference in level of complexity. In normal text we do not need to maintain any sort of regime between sentences. While in poetry text we need to maintain regime that is the thing which makes semantic difference between poetry and normal text. In normal text we have sentences which tells us story about anything but in poetry text we need equal lengthen stanzas, semantical connection between stanzas etc. Moreover, in the case of Urdu poetry text the aforementioned problems became complex to achieve. Figure 9 depicts the performance of LDA on Allama Iqbal Urdu Poetry Corpus which clearly shows that the generated topic words do not relate to each other. If we compare the generated topic words in contrast to overall frequency it shows that the overall frequency of generated words in topic 1 was 12.9% of total tokens but the algorithm was failed to give any suitable estimated frequency to the topics. There can be several reasons for the following results (i) Urdu poetry text is enough complex that well-known topic modelling approaches fail to find latent patterns from it. (ii) We need to test it in any other version of Machine Learning algorithms like semi-supervised or supervised. (iii) Test these approaches on different poetry corpus. Other approaches like LSI and HDP shown similar patterns to LDA which make us to believe that these topic modelling approaches did not

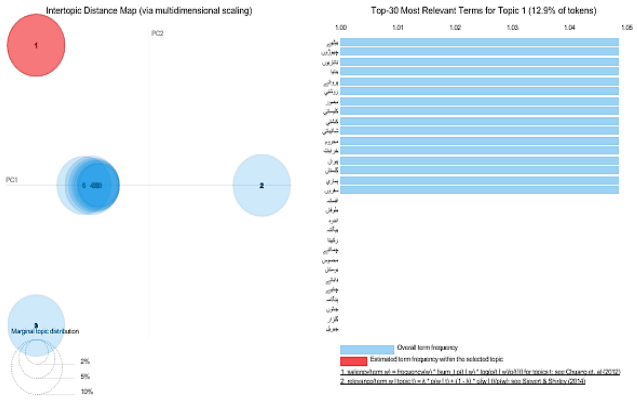

Figure 9. LDA results on Allama Iqbal Urdu poetry corpus.

perform well on Urdu poetry text. Figure 10 depicts LSI model results achieved on the mentioned corpus.

In next section, we will discuss the results of topic modelling approaches on mixed Urdu poetry corpus.

5.3. Mixed Urdu Poetry Corpus (Unsupervised)

In this section, we will discuss implementation and results achieved on mixed Urdu poetry corpus. Mixed Urdu poetry corpus is a collection of poetry written by different authors. We collected this poetry from different blogs, websites, and web pages.

The corpus contains poetry collection having four different genres (i) Romance poetry. (ii) Serious poetry. (iii) Religious poetry, and (iv) Hilarious poetry. We followed the same model as proposed in Figure 1. Furthermore, we implemented same topic modelling approaches on the mentioned corpus. Surprisingly, when we implemented topic modelling approaches to mixed Urdu poetry corpus we observed the same deficiency. This makes our intuition more robust that topic modelling approaches did not perform well on Urdu poetry text. Figure 11 depicts the results of LDA model applied on mixed Urdu poetry corpus as discussed in previous sub-section. Whereas Figure 12 shows dominance and contribution score for respective topics.

Figure 13 depicts results of LSI model on same corpus. Now as per results we can stay with our claim that topic modelling approaches are unable/not enough good to find latent topic words at least for unsupervised version.

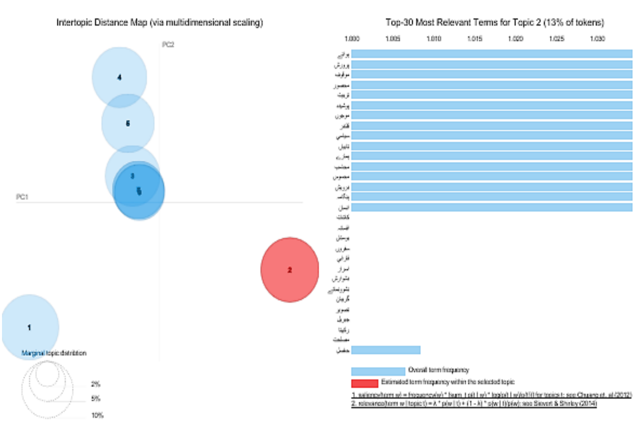

Figure 10. LSI results on Allama Iqbal Urdu poetry corpus.

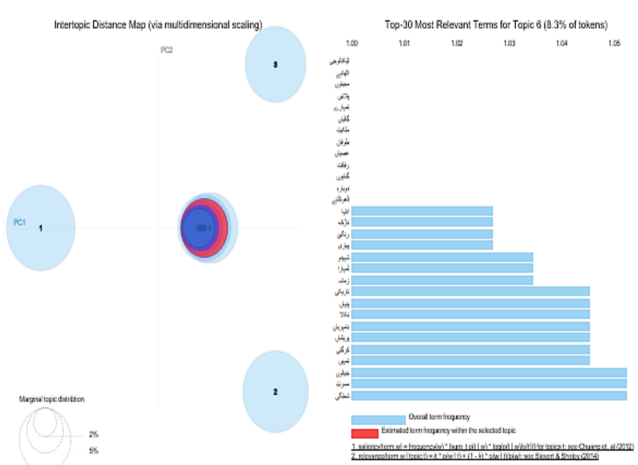

Figure 11. LDA results on mixed Urdu poetry corpus.

	Document_No	Dominant_Topic	Topic_Perc_Contrib	Keywords	Text
0	0	0.0	0.100	سلسلے, زندگی, کہانیاں, شعلگی, مسرت, خیالوں, ط...	[]
1	1	0.0	0.100	سلسلے, زندگی, کہانیاں, شعلگی, مسرت, خیالوں, ط...	[]
2	2	1.0	0.700	ہمیں, دامہرباں, پریشاں, کرگئی, دکالا, پنہاں ...	[ہمیں, دکالا]
3	3	0.0	0.100	سلسلے, زندگی, کہانیاں, شعلگی, مسرت, خیالوں, ط...	[]
4	4	3.0	0.700	ڈھونڈئے, دوبارہ, رفاقت, گناہوں, عصیاں, طوفان ...	[دوبارہ, ڈھونڈئے]
5	5	0.0	0.100	سلسلے, زندگی, کہانیاں, شعلگی, مسرت, خیالوں, ط...	[]
6	6	0.0	0.100	سلسلے, زندگی, کہانیاں, شعلگی, مسرت, خیالوں, ط...	[]
7	7	0.0	0.100	سلسلے, زندگی, کہانیاں, شعلگی, مسرت, خیالوں, ط...	[]
8	8	3.0	0.550	ڈھونڈئے, دوبارہ, رفاقت, گناہوں, عصیاں, طوفان ...	[عصیاں]
9	9	9.0	0.775	رنگین, پیاری, دازک, انتہا, ہمارے, طوفان, عصی...	[رنگین, دازک, پیاری]

	Topic_Num	Topic_Perc_Contrib	Keywords	Representative Text
0	0.0	0.775	سلسلے, زندگی, کہانیاں, شعلگی, مسرت, خیالوں, طوفان, محبتوں, عصیاں, بلائیں	[سلسلے, زندگی, کہانیاں]
1	1.0	0.775	ہمیں, دامہرباں, پریشاں, کرگئی, دکالا, پنہاں, تاریکی, طوفان, گناہوں, محبتوں	[پریشاں, کرگئی, دامہرباں]
2	2.0	0.775	شیدو, زمانہ, تمُہارا, بلائیں, تاریکی, اٹھائے, ٹیکنالوجی, عصیاں, رفاقت, طوفان	[زمانہ, تمُہارا, شیدو]
3	3.0	0.700	ڈھونڈئے, دوبارہ, رفاقت, گناہوں, عصیاں, طوفان, بلائیں, ملکیت, گالیاں, محبتوں	[دوبارہ, ڈھونڈئے]
4	4.0	0.550	گالیاں, ملکیت, طوفان, اٹھائے, پنہاں, تاریکی, ٹیکنالوجی, بلائیں, خیالوں, ہمارے	[ملکیت]
5	5.0	0.550	بلائیں, انتہا, ٹیکنالوجی, گناہوں, طوفان, عصیاں, ملکیت, ہمارے, محبتوں, گالیاں	[بلائیں]
6	6.0	0.550	ہمارے, طوفان, تاریکی, انتہا, گالیاں, اٹھائے, گناہوں, بلائیں, پنہاں, عصیاں	[ہمارے]
7	7.0	0.550	ٹیکنالوجی, محبتوں, اٹھائے, عصیاں, خیالوں, طوفان, گالیاں, پنہاں, رفاقت, انتہا	[اٹھائے]
8	9.0	0.775	رنگین, پیاری, دازک, انتہا, ہمارے, طوفان, عصیاں, بلائیں, اٹھائے, خیالوں	[رنگین, دازک, پیاری]

Figure 12. Generated topics respective to their contribution.

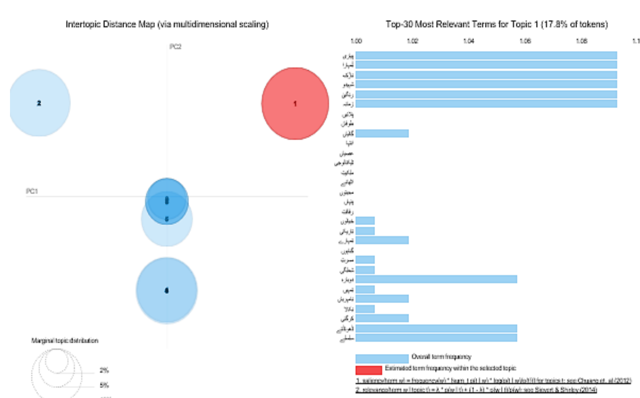

Figure 13. LSI results on mixed Urdu poetry corpus.

As continuation to the topic modelling approaches in future we will pilot with semi-supervised and supervised version of the proposed algorithms. Till now semi-supervised version of LDA has performed very well on English language text.[18] Next section will discuss conclusion of this literature.

6. Conclusion

In this work, we piloted three different variants for topic modelling including LDA, LSI, and HDP. Topic modelling helps us to extract unexposed latent words which can represent complete documents. We also proposed a novel poetry dataset. Proposed model has shown that all three models were good at extraction of latent patterns for general text. But topic modelling techniques are not suitable for Urdu poetry text. This does not mean that topic modelling will fail in poetry text. Previous literature has also achieved some excellent results on English poetry text. However, Urdu poetry text is comparatively more complex in nature. In future work, we will propose a semi-supervised and novel deep Learning approach for modelling Urdu Language text.

References

1. Topic mode. [cited 2019]. https://en.wikipedia.org/wiki/Topic_mode.
2. Latent dirichlet allocation. [cited 2019 May 14]. https://en.wikipedia.org/wiki/Latent_Dirichlet_allocation.
3. Information retrieval. [cited 2008 May 27]. https://www.math.unipd.it/~aiolli/corsi/0910/IR/irbookprint.pdf.
4. What is an intuitive explanation of the dirichlet distribution? [cited 2014]. https://www.quora.com/What-is-an-intuitive-explanation-of-the-Dirichlet-distribution.
5. Modeling event importance for ranking daily news events. [cited 2017 Feb]. https://www.researchgate.net/publication/313263535_Modeling_Event_Importance_for_Ranking_Daily_News_Events.
6. Temporal topic modeling to assess associations between news trends and infectious disease outbreaks. [cited 2017 Jan 19]. https://www.nature.com/articles/srep40841.
7. Chen H, Xie L, Leung C, Lu X, Ma B, Li H. Modeling latent topics and temporal distance for story segmentation of broadcast news. IEEE/ACM Trans Audio Speech Lang Process. 2017;25(1):112–23.
8. Gui J, Wang Q. Topic modeling of news based on spark Mllib. In:International computer conference on wavelet active media technology and information processing (ICCWAMTIP); 2017. P. 224–8.
9. Zhang X, Zhao L, Chen Z, Boedihardjo AP, Dai J, Lu C. Trendi: tracking stories in news and microblogs via emerging, evolving and fading topics. In: IEEE international conference on big data (Big Data); 2017. P. 1590–99.
10. Larsen VH, Thorsrud LA. The value of news for economic developments. J Econ. 2019;210(1):203–18.
11. Content-driven, unsupervised clustering of news articles through multiscale graph partitioning. [cited 2018 Aug 03]. https://arxiv.org/abs/1808.01175.
12. Bangla news recommendation using doc2vec. [cited 2018 Sep]. https://www.researchgate.net/publication/328043719_Bangla_News_Recommendation_Using_doc2vec.
13. Asadi Kakhki SS, Kavaklioglu C, Bener A. Topic detection and document similarity on financial news. In: Book topic detection and document similarity on financial; 2018. P. 322–8.
14. Hidayatullah AF, Pembrani EC, Kurniawan W, Akbar G, Pranata R. Twitter topic modeling on football news. In: Book twitter topic modeling on football news; 2018. P. 467–71.
15. Shakeel K, Tahir GR, Tehseen I, Ali M. A framework of Urdu topic modeling using latent dirichlet allocation (LDA). In: IEEE 8th Annu. Comput. Commun. Work. Conf.;2018. P. 117–23.
16. Hierarchical dirichlet process. [cited 2019 Feb 16]. https://en.wikipedia.org/wiki/Hierarchical_Dirichlet_process.
17. Sharjeel M, Nawab RMA, Rayson P. COUNTER: corpus of Urdu news text reuse. Lang Resour Eval. 2017;51(3):777–803.
18. How we changed unsupervised LDA to semi-supervised guided LDA. [cited 2017 Oct 16]. https://www.freecodecamp.org/news/how-we-changed-unsupervised-lda-to-semi-supervised-guidedlda-e36a95f3a164/.
19. Topic modeling with LSA, PLSA, LDA & lda2Vec. [cited 2018 May 25]. https://medium.com/nanonets/topic-modeling-with-lsa-psla-lda-and-lda2vec-555ff65b0b05.