IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received March 19, 2019, accepted March 28, 2019, date of publication April 18, 2019, date of current version June 4, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2912012

Knowledge Based Recommender System for Academia Using Machine Learning: A Case Study on Higher Education Landscape of Pakistan

HUMA SAMIN AND TAYYABA AZIM
Center of Excellence in IT, Institute of Management Sciences, Peshawar 25000, Pakistan
Corresponding author: Tayyaba Azim (tayyaba.azim@imsciences.edu.pk)

This work was supported by the Institute of Management Sciences

ABSTRACT Allocation of courses and research students based on faculty's subject specialization and area of interest has always remained a challenging task for university administration due to the presence of academics' cross-domain interests, stale faculty resumes at university portals and changing the skill set demands from the industry. Collaborative filtering and content-based recommender systems have already been in use by the industry for recommending things, such as movies, news, restaurants, and shopping items to the users, and however, no one has utilized these off-the-shelf models for enhancing the student experience and improving the quality of higher education in academia. This paper presents a case study showcasing the use of probabilistic topic models for generating recommendations to users in academia through appropriate course allocation and supervisor assignment. The proposed system coined as *ScholarLite* harnesses the power of machine learning to extract research themes from faculty members' past publications, mines research interests from their resumes, and combines it with their educational background to generate recommendations for course teaching, research supervision, and industry–academia collaboration. We have shown the recommendation results on real-world data gathered from the higher education commission of the country and demonstrated that the proposed techniques are scalable across various programs offered by the universities and could be deployed in a small budget by universities for automating course and supervisor allocation procedures. The experiments confirm our performance expectation by showing good relevance and objectivity in results, thus making this decision management system more appealing for large-scale deployment and use by academia.

INDEX TERMS Author topic model, higher education, knowledge management application, latent Dirichlet allocation, machine learning, perplexity, recommender systems, topic mining.

I. INTRODUCTION

With the rapid growth and spread of information and communication technology (ICT), recommender systems have evolved that have completely reshaped the web experience of users by providing meaningful, effective, and personalized recommendation of products and services to users. Through proper data modeling and analysis, recommender systems tend to support users in decision making processes by enhancing their ability and quality of thinking. This has been witnessed extensively in the area of e-commerce where recommender systems are used to enhance revenues by selling more products, whereas in scientific libraries, recommender systems provide support to users by moving beyond catalog searches. The domain has received an increased amount of attention in the recent years and has become an integral part of many frequently visiting web sites such as Amazon, Youtube, E-bay, CDNow, MovieFinder, Netflix, Last.fm, IMDb, etc.

While the popularity and usage of these systems remains integral in various domains [1]–[6], its utility for academia has been limited for suggesting research articles majorly [7]–[11]. Some researchers have also dedicated efforts in developing systems for recommending relevant academic jobs, conferences and scholarships on social academic networks [12]. Existing websites like CiteULike, Mendeley, ResearchGate and Academia allow researchers to create their own reference libraries for the articles and share them with other researchers. The suggestions given are either relevant

The associate editor coordinating the review of this manuscript and approving it for publication was Mehul S. Raval.

to ones area of interest or may also include cross domain topics to facilitate interdisciplinary collaborations [13]. Some of the other recommender systems in academia provide internet based portals for recruitment based facilities [14]. In these systems, candidates and employers can upload their profiles and as a result matching jobs are recommended to the candidates on the basis of their qualification and experience.

Despite the enormous interest of industry and researchers on the topic, recommender systems have suffered from the problem of cold-start, sparsity and scalability. Cross domain topic models handle the issue of sparsity [15] and skewness in topics just like author topic model based collaborative filters [16]. With the onset of success for deep learning paradigm, a number of deep learning based approaches like convolutional neural networks and collaborative deep learning model, etc. have also been introduced in combination with collaborative filtering techniques to design recommender systems. The usage of deep learning techniques with collaborative filtering deals with the problem of cold start in recommender systems [17]–[19].

The system proposed in this research lies at the cross roads of *topic models* and *recommender systems*. Topic models offer a statistical approach to discover underlying themes of text occurring in a collection of documents and are used to support the recommendation process [20]–[23]. This work introduces the use of two popular topic models: Latent Dirichlet allocation (LDA) and author topic model (ATM) to automate *supervisor recommendations* for submitted research proposals and *faculty recommendations* for courses offered at national universities. The conventional procedure used to allocate courses in an institute is manual and focuses more on equal distribution of teaching workload given the availability of relevant faculty. This relevance is determined either by faculty's previous experience of teaching specific courses or their teaching preferences. Such teaching assignment practices for the faculty are not standardized across the university and often do not comply with the higher education industry standards due to the lack of faculty's updated skill set information locally and the biased judgment of humans in charge. In addition to course assignment, the university management also faces a great challenge when assigning supervisors to research students. It is either a student's responsibility to reach out for a suitable supervisor or the management takes this decision on his part by looking into faculty's workload policy. Either way, the selection procedure takes the toll as the top management as well as the students are unaware of the expertise of local faculty members due to the presence of stale resumes at university portal. The task of connecting to the right research team is equally cumbersome for the faculty and industry as it is for the students. It requires a lot of drill and experience to locate an appropriate research collaborator from the academia or industry, thus causing a gap between the two fields. The recommendation task becomes even more challenging and less transparent when the size of the faculty in a department or school increases. Thus, a system that can automate these tasks may bring transparency and fairness in academic procedures besides improving the quality of research output and teaching in the country.

We believe topic models in the current scenario can provide a great deal of information about the faculty's interests from the content of their resumes submitted regularly to the higher education commission (HEC) of Pakistan and research papers indexed automatically by Google scholar and DBLP. The faculty resumes present at the HEC portal are up to date due to the administrative nature of the national organization, and are further complemented by authors' research papers indexed by Google Scholar and DBLP. In the absence of ScholarLite, the recommendation tasks were manually performed by designated program coordinators or students, thus leading to inefficient and sometimes biased course and research supervisor allocations. The proposed system *ScholarLite* can easily be scaled up and applied to the data set of any academic institution or industrial organization as discussed in detail ahead (See Section IV-F).

The main contributions of this work are summarized as below:

- Development of first supervisor & course recommender system for academia in the region. Recommender systems have been studied exhaustively worldwide, however to the best of our knowledge no one has explored the use of topic models for supervisor and course recommendation tasks in academia or industry before.
- Construction of a national database of computer scientists' resumes demonstrating their professional expertise in different areas. The real world data can help national organizations derive meaningful insights about the higher education landscape of the country and take data literate decisions focused on improving real challenges faced by academia. The data set labeled as *PakScholarScan* and features extracted from it are available online at https://github.com/tabzim/Data-Set.git.
- Identification and comparison of the topic models most suitable for developing recommender systems for academia.
- Explores the scalability, objectiveness and computational efficiency of the proposed recommender system *ScholarLite* for academia.

We have organized this paper as follows: Section II discusses the preliminary concepts required to understand the proposed recommender system as well as related recommender systems prevailing in academia. Section III describes the details of the experiments conducted for generating recommendations. This is followed by Section IV that discusses the empirical findings and Section V that concludes this case study with some directions to improve the quality of the proposed recommender system in future.

II. PRELIMINARIES AND RELATED WORK

This section discusses the preliminary concepts used for developing the proposed recommender system as well as the related algorithms used in developing such systems.

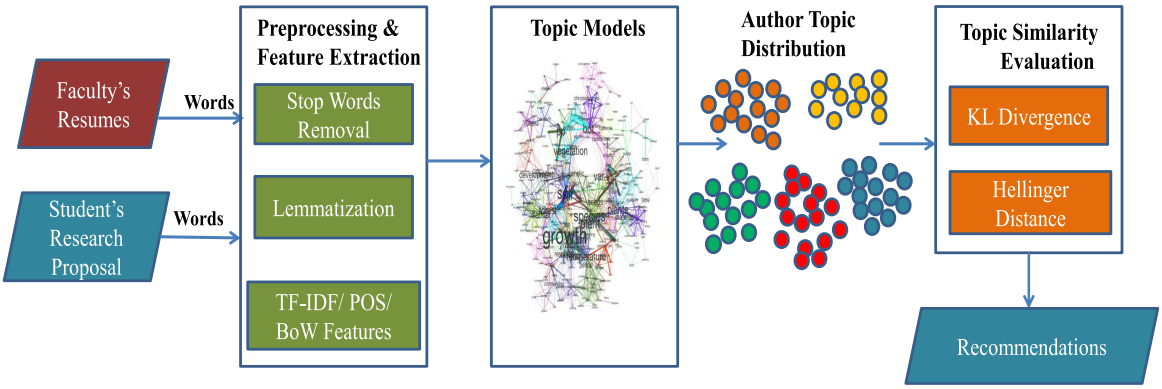


FIGURE 1. Graphical abstract of the proposed recommender system (ScholarLite) for academia. The content of faculty resumes and research proposals is pre-processed to learn various themes and author topic relationship through topic models. Once the distributions are modeled, distance metrics are used to find out author's similarity to any given topic in the distribution.

A. PRELIMINARIES

The graphical abstract of the proposed recommender system can be seen in Figure 1. The proposed system makes use of two popular probabilistic topic models: Latent Dirichlet allocation (LDA) and author topic model (ATM) to facilitate course and supervisor recommendations. The probabilistic models and the evaluation metrics used to assess their learning power are discussed briefly in sections below.

1) LATENT DIRICHLET ALLOCATION (LDA)

LDA is a popular probabilistic generative model used for discovering the mixture of topics discussed in a document and in the entire corpus [24]. The model takes each document, d as a mixture of topics that are hidden and each topic, $\mathbf{z}$ is formulated by a vocabulary of fixed words, $\mathbf{w}$ which could be observed in a document. There are two groups of unknown parameters governing the model: The K topic distributions ϕ, the D document distributions θ. In addition, the model has latent variables that handle the assignment of words $\mathbf{w}$ to topics $\mathbf{z}$.

The document generation process can be broadly divided into two steps: 1) Choose a combination of topics for a document by using a Dirichlet distribution with parameters α over a fixed set of topics K, 2) Generate each word $\mathbf{w}_i$ in the document, d by first picking each topic (from the Dirichlet distribution calculated in the previous step) and using it to generate the word itself. The hidden and observed variables are given as a joint probability distribution as follows:

$$p(\beta_{1:K}, \theta_{1:D}, \mathbf{z}_{1:D}, \mathbf{w}_{1:D}) = \prod_{i=1}^{K} p(\beta_i) \prod_{d=1}^{D} p(\theta_d) \prod_{n=1}^{N} p(\mathbf{z}_{d,n}|\theta_d) \cdot p(\mathbf{w}_{d,n}|\beta_{1:K}, \mathbf{z}_{d,n}), \quad (1)$$

where $\mathbf{w}_d$ are the words observed for document d, $\beta_{1:K}$ are the word probabilities where each β_k represents a distribution of the vocabulary of words given a topic, θ_d stores the topic proportions for document d, $\mathbf{z}_{d,n}$ is the topic assignment for word n in document d. See Figure 2a for illustration of the generative model. The figure highlights three distinct levels of representation for model parameters and variables: 1) corpus level, 2) document level and 3) word level. The parameters α and β are corpus level parameters, the variable θ_d is a document-level variable and the variables $\mathbf{z}_{dn}$ and $\mathbf{w}_{dn}$ are word-level variables. Corpus level parameters are sampled only once while generating a corpus, document level variables are sampled only a single time for every document and word level variables are sampled only once for every word in each document.

According to the Bayesian formulation, the posterior distribution of the hidden variables after observing a document is given as below:

$$p(\theta, \mathbf{z}|\mathbf{w}, \alpha, \beta) = \frac{p(\theta, \mathbf{z}, \mathbf{w}|\alpha, \beta)}{p(\mathbf{w}|\alpha, \beta)}, \quad \text{where}$$
$$p(\theta, \mathbf{z}, \mathbf{w}|\alpha, \beta) = p(\theta|\alpha) \prod_{n=1}^{N} p(\mathbf{z}_n|\theta) p(\mathbf{w}_n|\mathbf{z}_n, \beta) \quad (2)$$

Unfortunately, this formulation is intractable for exact inference and is solved by using a wide variety of approximate inference algorithms such as variational approximation, Markov Chain Monte Carlo methods and Laplace approximation. In this work, we have used Variational Inference algorithm as shown by [24] to calculate this posterior distribution.

2) AUTHOR TOPIC MODEL (ATM)

Author topic model (ATM) [25] extends latent Dirichlet allocation (LDA) by including authorship information in addition to the topics as illustrated in Figure 2b. The model associates each word $\mathbf{w}$ in a document with two latent variables: $\mathbf{z}$ and $\mathbf{x}$, representing topic and an author respectively. These latent variables augment the N-dimensional vector w indicating topic and author assignments for the N words. The joint probability distribution of hidden and observed variables in author topic model is given as:

$$p(\theta, \phi, \mathbf{z}, \mathbf{x}, \mathbf{w}|\alpha, \beta, A) = p(\theta|\alpha) p(\phi|\beta) p(\mathbf{x}|A) p(\mathbf{z}|\mathbf{x}, \theta) p(\mathbf{w}|\mathbf{z}, \phi)$$

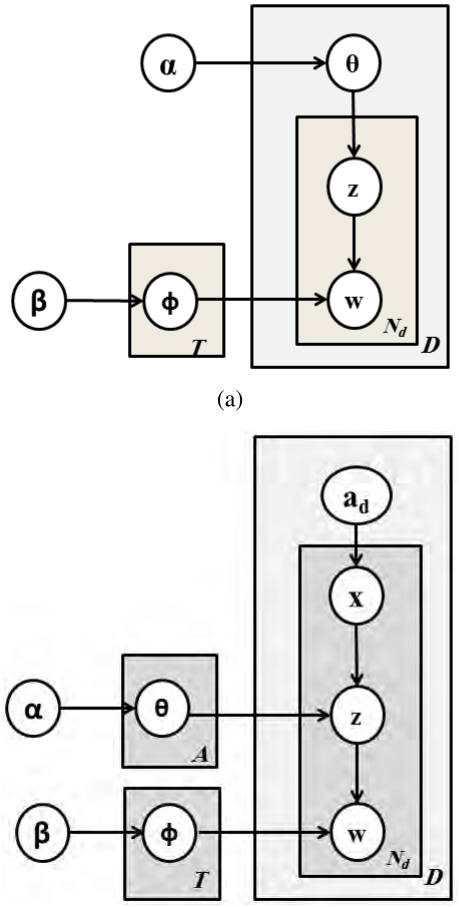


FIGURE 2. Probabilistic models shown in a plate diagram where the plates indicate repetition in the generative process and α, β are the parameters of the respective Dirichlet distributions. θ represents the document topic distribution matrix drawn on the basis of Dirichlet prior denoted by α. ϕ represents the topic distribution matrix for each of the K topics having a multinomial distribution over V vocabulary items. (a) Plate diagram of LDA graphical model depicting words, $\mathbf{w}$ as visible variables. (b) Plate diagram of ATM graphical model in which authors, $\mathbf{a}_d$ and words, $\mathbf{w}$ are visible variables.

$$= \prod_{a=1}^{A} \mathrm{Dir}(\theta_a|\alpha) \prod_{k=1}^{K} \mathrm{Dir}(\phi_k|\beta) \prod_{d=1}^{D} \prod_{n=1}^{Nd} \mathrm{Unif}(x_{dn}|A_d) \times \mathrm{Mult}(z_{dn}|\theta_a, x_{dn} = a)\mathrm{Mult}(w_{dn}|\beta_k, z_{dn} = k).$$

For calculation purpose, it is assumed that the group of authors of each document is observed. The model associates each author with a multinomial distribution over topics, where each topic is represented with a multinomial distribution over words. Thus, a multiple authored document is represented by a probability distribution over topics that is a mixture of the probability distributions associated with the authors. The posterior probability distribution is given as:

$$p(\theta, \phi, \mathbf{z}, \mathbf{x}|\mathbf{w}, \alpha, \beta, A) = \frac{p(\theta, \phi, \mathbf{z}, \mathbf{x}, \mathbf{w}|\alpha, \beta, A).}{p(\mathbf{w}|\alpha, \beta, A)} \quad (3)$$

The marginal likelihood $p(\mathbf{w}|\alpha, \beta, A)$ cannot be estimated analytically, thus making the posterior distribution mathematically intractable. We have used variational Bayes algorithm [26] that uses variational distribution to approximate this posterior distribution.

Given a group of document authors A and their distributions over topics θ, the process of generating a document can be summarized as follows: 1) Choose an author uniformly at random for each word present in the document, 2) Sample a topic for each word from the distribution over topics associated with the author of that word, 3) From the words' probability distribution associated with each topic $\mathbf{z}$, sample the words.

3) MODEL EVALUATION METRICS

a: PERPLEXITY AND LOG LIKELIHOOD

Perplexity is a standard evaluation metric for estimating the generalization performance of probabilistic models like LDA and ATM. It is computed by estimating the multiplicative inverse of geometric mean of the likelihoods of word tokens in the test corpus given the model. Smaller magnitude of perplexity indicates less misrepresentation of words of the test documents by the trained topics. The perplexity of LDA model [27] can be evaluated as:

$$\mathrm{Perplexity} = \exp\Big(-\frac{\sum_{m=1}^{M} \log p(\mathbf{w}|M)}{\sum_{m=1}^{M} N_m}\Big), \quad (4)$$

where M represents the trained model. In document m, $\mathbf{w}_m$ represents the word vector. The log-likelihood of LDA model can be evaluated as:

$$\log p(\mathbf{w}|M) = \sum_{t=1}^{V} n_m^{(t)} \log\Big(\sum_{k=1}^{K} \phi_{k,l}.\theta_{m,k}\Big), \quad (5)$$

where $n_m^{(t)}$ stores the occurrences of word t in document m, θ represents the document topic distribution matrix drawn on the basis of Dirichlet prior denoted by α and ϕ represents the topic distribution matrix for each of the K topics having a multinomial distribution over V vocabulary items [24].

The perplexity of ATM for a set of test words $(\mathbf{w}_d, \mathbf{a}_d)$ for D belonging to D_{test} can be evaluated as follows:

$$\mathrm{Perplexity}(\mathbf{w}_d, \mathbf{a}_d) = \sum d\theta \sum d\phi p(\theta|D^{\text{train}})p(\phi|D^{\text{train}}) \times \prod_{m=1}^{N_d} \Big[\frac{1}{A_d} \sum_{i \in \mathbf{a}_d, j} \theta_{ij}\phi_{\mathbf{w}_{mj}}\Big]. \quad (6)$$

4) RESUME SIMILARITY MATCHING METRICS

a: HELLINGER DISTANCE

Hellinger distance is a similarity evaluation metric used to assess the resemblance of two probability distributions p and q [28] by the following computation:

$$H(p, q) = \frac{1}{2}\sqrt{\sum_{i=1}^{K} (\sqrt{p_i} - \sqrt{q_i})^2}. \quad (7)$$

The value of Hellinger distance metric varies from 0 to 1. The lower the value of Hellinger distance, the more similar are the probability distributions.

b: KULLBACK-LEIBLER DIVERGENCE

Kullback-Leibler (KL) divergence is a distance based metric used to evaluate the similarity of two probability distributions. The value of KL varies from 0 to 1, where a value close to 0 implies higher similarity than the value closer to 1. The KL divergence between two probability distributions $p(x)$ and $q(x)$ can be evaluated as described in [25]:

$$D_{KL(p(x)||q(x))} = \sum_{x \in X} p(x) ln \frac{p(x)}{q(x)}. \tag{8}$$

5) RECOMMENDATION EVALUATION METRIC

In order to measure the quality of recommendations returned by the topic models, we approach the retrieved recommendations as a ranking problem where they are rated according to their order of relevance. Among the standard recommendation evaluation metrics, we have estimated the normalized discounted cumulative gain (NDCG) metric to analyses how good the recommendations have been ranked by the deployed topic models.

a: NORMALISED DISCOUNTED CUMULATIVE GAIN

Normalised discounted cumulative gain (NDCG) takes into account two measures: (1) *Discounted cumulative gain* (DCG) of the search results showing how relevant each retrieved result is. This is a non-negative number that is generally derived from the user's implicit/explicit feedback. We have used the retrieved topic model probabilities as a measure of author/document gain here, (2) *Ideal discounted cumulative gain* (IDCGP) which ranks the top most retrieved results according to their position.

$$\text{NDCG}_p = \frac{\text{DCG}_p}{\text{IDCG}_p}. \tag{9}$$

The DCG accumulated at a particular rank position p is calculated as:

$$\text{DCG}_p = \sum_{i=1}^{p} \frac{\text{rel}_i}{\log_2 (i+1)} = \text{rel}_1 + \sum_{i=1}^{p} \frac{\text{rel}_i}{\log_2 (i+1)},$$

The ideal discounted cumulative gain (IDCGP) is given as follows:

$$\text{IDCGP}_p = \sum_{i=1}^{|\text{REL}|} \frac{2^{\text{rel}_i} - 1}{\log_2 (i+1)},$$

where |REL| represents the list of relevant documents (ordered by their relevance) in the corpus up to position p.

B. RELATED WORK

This section will take into account the development of techniques and their applications in recommender systems for academia.

A number of applications focus on proposing e-learning systems for the purpose of recommending articles and documents to the users using machine learning approaches. In [29], the system takes into account user preferences and uses information retrieval and collaborative filtering techniques along with utilizing key extraction algorithm and automatic query extraction from web in order to recommend documents on a given topic. In a similar way, some of the applications generate recommendations to the users by taking into account not only their own preferences but also the interests of their friends and faculty in academic social networks [12]. In [30] myPTutor, a system for planning and recommending learning routes is presented. The system uses artificial intelligence (AI) and case based planning techniques to suggest learning routes according to the students' requirements and is implemented on top of MOODLE but the system does not involve ontology based algorithms in order to increase the reusability of contents. In [31], a course recommendation system is presented that helps the learners in the selection of the courses according to their specified requirements. The system uses a hybrid methodology using ontology in order to generate recommendations.

The applications of recommender systems have also been seen in the area of English language learning. Hsu et al. [32] proposed an English reading material recommendation system based on the rule based knowledge engineering approach. The expert recommender system uses opinions and domain knowledge of English teaching experts in order to generate recommendations on the basis of preferences and knowledge levels of individual students.Furthermore, annotation module was also included in a personalized mobile language learning system in order to enable the students to take notes of English language vocabulary translation for reading the content [33]. The system does not support the updation of user profile and does not keep a record of the updates in the reading interests over time. In [34], a recommendation forum for English learning was presented. The study examined the impact of collaborative filtering on college students use of an online forum.

A number of systems focus on recommending scientific and scholarly articles to the users [7]–[10]. In [35], a system for recommending scholarly papers to the researchers has been proposed. The recommendations are generated by using latent information extracted from term frequency-inverse document frequency (TF-IDF) features and cosine similarity is measured between users research interests gathered from their past publication information and the papers that cite the users work, however the system does not provide recommendations for inter-disciplinary research work. Sun et al. [36], [37] proposed a recommendation approach that analyzes the semantic content of the articles by keyword similarity calculation and extracts online users' connections in order to support article voting and generating recommendations. The approach is also implemented in an online social network platform called ScholarMate. In order to resolve the mismatch problem and match irrelevance problem in recommending articles, they further integrated three similarity measures such as keyword similarity, journal similarity and author similarity. First of all the keyword similarity is used to generate a list of candidate articles and then journal and

author similarity is used to select the relevant articles from the list of candidate articles. Xia *et al.* [38] presented a scientific article recommendation system by using the information of relation between articles on the basis of common authors. The proposed system presents two forms of recommendations: One on the basis of common author based search pattern and other on the basis of frequently appeared author for the articles but it does not involve citation relationships between authors in order to perform recommendation on the basis of author based search patterns. In [39], a scientific article recommendation system is presented that uses a bi-relational graph representing article content similarity, researcher interest correlation and research article readership. In order to generate recommendations, an iterative random walk on the Bi-Relational Graph is conducted. The problem with the system is that it focuses on researcher-researcher relevance but the article-researcher relevance is not used for generating recommendations. Furthermore, in [40], a publication venue recommendation system is proposed that compares the title and abstract of the research article with the prospective venues for the publication.

Although, all of the recommender systems discussed above use a number of machine learning and data mining techniques, none of them have utilized topic models for the purpose of generating course and supervisor recommendations in academia. In this case study, we have done empirical analysis to identify the most suitable topic models for this task on real world data set [1] and explored their scalability, objectiveness and computational efficiency for implementation in national organizations.

III. EXPERIMENTS AND RESULTS

A. DATA COLLECTION

In order to explore the content of faculty resumes for various research themes in Computer Science, relevant data is required to train the topic models. Due to the non-availability of any relevant benchmark data set, we developed our own data set *PakScholarScan* comprising of relevant material for training and testing. To develop the train set, paper abstracts along with the author information from different areas of computer science were accumulated. The abstracts were collected from NIPS conference proceedings[2] of years 1999 and 2000, SIGCOMM conference proceedings[3] of years 2015 and 2016, KDD conference proceedings[4] of year 2016, and ICIP[5] proceedings of year 2016. The four are considered as the largest conferences in the areas of Artificial Intelligence, Computer Networks, Data Mining and are regarded as a melting pot of researchers from various areas of science such as Machine learning, Computer Vision, Statistics, Physics, Mathematics, Neuroscience, and Data Science. Our train data thus entails a wide range of themes and perspectives shared in these publications. The test data is built from the resumes of faculty members from the Institute of Management Sciences (IMSciences) Pakistan [6] and other national institutes whose faculty has been approved by the Higher Education Commission of Pakistan (HEC) for supervising post graduate level research. These resumes were extracted from their workplace websites and latest publication information revealed by the search engines: Google Scholar and DBLP. From the resumes, only the publication information is extracted that contains author information and title of the published research piece. The focus is on the publications' section of resume as this information signifies academic's research interests and potential of supervising students in a specific area. A collection of 698 resumes from HEC and IMsciences portal are accumulated out of which 612 are assigned to the train set and 86 documents are assigned to the test set. Please note that this test data set is prone to scale up with more faculty recruitments by HEC in future and as we increase the scope of the project to programs other than Computer Science. The data set used in this case study is available at https://github.com/tabzim/Data-Set.git.

B. DATA PRE-PROCESSING

Once the data has been collected, a number of pre-processing steps are performed to make it amenable for model training and testing. In the first step, *stop words* are removed from the collected resumes and paper abstracts. Stop words identify some of the most common, short functional words, such as 'the','is','at','on', etc. that cannot help in distinguishing one theme from another and hence are discarded. After stop word removal, the next steps involve *lemmatization* and *stemming* which facilitate in removing the inflectional forms of the word and extract the common root or dictionary form of the word. After performing these data cleaning steps, *unigram* features are extracted from the word tokens to make them amenable for statistical modeling. Some of the most popular feature extraction techniques used in natural language processing are as follows: n-gram (unigram, bigram, trigram) [41], bag of words (BoW) [42], term frequency-inverse document frequency (TF-IDF) [43] and parts of speech (POS) [44]. The unigram features are further weighted by the count of term occurrences in a document and thus entire corpus respectively.

C. MODEL TRAINING

After preprocessing the entire corpus, we use the train data to train the LDA and ATM models respectively. Training the model requires the use of optimal hyper-parameters which are not known beforehand for every data set. In order to find their best values, we divide the train set further into train and cross validation sets with 420 documents left out for training and 192 documents separated for model validation task. The new training set is used to train the model for

[1] https://github.com/tabzim/Data-Set.git
[2] https://archive.ics.uci.edu/ml/datasets/NIPS+Conference+Papers+1987-2015
[3] http://www.sigcomm.org/ccr/papers
[4] http://www.kdd.org/kdd2016/
[5] http://2016.ieeeicip.org/
[6] https://www.imsciences.edu.pk

TABLE 1. Hyper-parameters tweaking via grid search method to find out optimal values for LDA model training. (a) LDA's perplexity and likelihood on cross validation set with fixed number of topics, $K = 15$ and varying α and β. The best value for α and β are 1 and 0.01 respectively. (b) LDA's perplexity and likelihood on cross validation set with optimal hyper-parameters, $\alpha = 1.0$ and $\beta = 0.01$. By changing the values of K, we observe the perplexity and choose the optimal number of topics $K = 5$ for LDA.

Alpha(α)	Beta(β)	Perplexity	Likelihood
0.01	0.001	1.05375428571e+29	-96.4114532505
0.01	0.01	272.562820666	-8.09044497239
0.01	0.1	325.30246979	-8.34563796555
0.01	1.0	240.385312395	-7.90920493948
0.01	2.0	257.813521171	-8.01018411839
0.1	0.001	9.57550665944e+28	-96.2733354818
0.1	0.01	253.999847428	-7.98868382018
0.1	0.1	288.166726272	-8.17075995133
0.1	1.0	249.263143729	-7.9615257702
0.1	2.0	268.148924414	-8.06689065628
1.0	0.001	8.01564446043e+28	-96.016805175
1.0	**0.01**	**237.782059**48	**-7.89349605845**
1.0	0.1	319.427863181	-8.31934635199
1.0	1.0	308.477308	-8.26902055649
1.0	2.0	350.795699669	-8.45448725197

(a)

K	Perplexity	Likelihood
5	**228.669312528**	**-7.83711895852**
10	231.574367166	-7.85533176085
15	237.78205948	-7.89349605845
20	242.900587689	-7.9242221703

(b)

different possible values of α and β by keeping the number of topics K fixed. See Table 1a and Table 1b for illustration of hyper-parameter setting in LDA. The learning of model is assessed by observing the values of perplexity and log likelihood. The smaller the perplexity, the higher is the ability of the model to generate the documents/resumes. The bold values in the table show that the optimal values of α and β are 1 and 0.01. Once suitable values of α and β are found, we keep these values fixed and change the value of number of possible topics K. Table 1b reveals that the optimal number of topics discovered by LDA are 5.

For ATM, optimal hyper-parameters are searched in a similar way by first fixing the value of K and iterating α and β. Once optimal values of these hyper-parameters are found, we treat them as constants and vary the value of K to observe its perplexity on the held out set. As shown in Table 2, the optimal hyper-parameter values for ATM are $\alpha = 1$ and $\beta = 2$. ATM also shows lowest perplexity and highest likelihood on 5 topics. Please note that the data given to both the models for assessment of top topics and top authors with topics is the same and hence could be compared for models' performance evaluation.

D. MODEL EVALUATION

After determining optimal values of hyper-parameters and number of topics, the cross validation and train sets are merged again for training the two topic models on the entire train set. The performance of these trained models is evaluated on unseen test data set using perplexity and log likelihood scores shown in Table 3. In order to explore hidden

TABLE 2. Hyper-parameters tweaking via grid search method to find out optimal values for ATM training. (a) Perplexity and likelihood of ATM on cross validation set with $K = 15$ and varying α, β. The best value for α and β are 1 and 2 respectively. (b) ATM's perplexity and likelihood on cross validation set with $\alpha = 1.0$, $\beta = 2.0$ and varying K. By changing the values of K, we observe the perplexity and choose the optimal number of topics K for ATM.

Alpha(α)	Beta(β)	Perplexity	Likelihood
0.01	0.001	1.89648035047e+16	-54.0741739421
0.01	0.01	1.77747712554e+12	-40.6929681325
0.01	0.1	9914.26521893	-13.2752901379
0.01	1.0	1257.40320891	-10.2962316347
0.01	2.0	910.494104668	-9.83050586574
0.1	0.001	19410509768.7	-34.1761189562
0.1	0.01	4616704.43774	-22.1384319431
0.1	0.1	9488.21273614	-13.2119206417
0.1	1.0	1402.32918834	-10.4536093377
0.1	2.0	1079.72838768	-10.0764527239
1.0	0.001	1300574386.83	-30.2765017711
1.0	0.01	497639.858321	-18.9247425168
1.0	0.1	2347.0186379	-11.1966135831
1.0	1.0	1272.37733493	-10.3133108629
1.0	**2.0**	**888.877434482**	**-9.79584069223**

(a)

K	Perplexity	Likelihood
5	**614.393879526**	**-9.26302003405**
10	768.458201371	-9.58582297953
15	888.877434482	-9.79584069223
20	986.022895376	-9.94547733604

(b)

TABLE 3. Performance of the topic models on test resumes. (a) LDA perplexity and likelihood on test set with $\alpha = 1.0$ and $\beta = 0.01$. The best results are obtained at $K = 5$ as indicated on the cross validated set. (b) ATM perplexity and likelihood on test set with $\alpha = 1.0$ and $\beta = 2.0$. The best results are obtained at $K = 5$ as indicated on the cross validated set.

(a)

K	Perplexity	Likelihood
5	**196.618639253**	**-7.61925628401**
10	261.813799454	-8.03239732938
15	277.237718605	-8.11497974141
20	292.098907775	-8.19031315338
50	402.689812815	-8.65352516544
100	612.106419392	-9.25763868809

(b)

K	Perplexity	Likelihood
5	**4692.26620223**	**-12.1960691475**
10	4988.6827926	-12.2844432226
15	6449.89192073	-12.6550592705
20	7665.72843586	-12.9042071752
50	7889.41799946	-12.9457031615
100	6385.02435319	-12.6404764072

themes in resumes, LDA produces top words with high probabilities for each theme. The results can be seen in Table 4. In comparison to LDA, the author topic model produces author ranking with respect to each topic discovered from the resumes data set. For each of the discovered research themes, we can observe the top words for each topic as well as the faculty member's relevance to that area of research.

IV. DISCUSSION & ANALYSIS

On comparing the perplexity/likelihood of LDA and ATM on test data set, we observe that the generative power of LDA is much better than ATM and it is much likely to reproduce documents (resumes) with the same theme seen before. In contrast, the ATM focuses on joint distribution of author/s with topics and hence their margin of error for joint

TABLE 4. Top five topics returned by LDA topic model.

Topics	Words
0	0.032×"network" + 0.016× "comput" + 0.015× "applic" + 0.015× "mobil" + 0.013×"energi" + 0.013× "user" + 0.010× "secur" + 0.010 × "cloud" + 0.010× "scheme" + 0.010× "system"
1	0.013× system + 0.013×"softwar" + 0.012× "present" + 0.011 × "languag" + 0.010 × "develop" + 0.010 × "model" + 0.010 × "demand" + 0.009 × "constraint" + 0.009 × "busi" + 0.008 × "transmiss")
2	0.041 × "imag" + 0.021 × "propos" + 0.019 × "featur" + 0.019 × "method" + 0.014 × "base" + 0.013 × "system" + 0.013 × "detect" + 0.011 × "extract" + 0.010 × "recognit" + 0.009 × "techniqu";
3	0.036 × "model" + 0.020 × "algorithm" + 0.015 × "data" + 0.014 × "perform" + 0.012 × "predict" + 0.011 × "channel" + 0.010 × "process" + 0.010 × "sequenc" + 0.010 × "method" + 0.010 × "structur"
4	0.021 × "base" + 0.018 × "approach" + 0.015 × "data" + 0.014 times"propos" + 0.013 × "process" + 0.013 × "graph"+ 0.012 × "techniqu" + 0.012 × "inform" + 0.010 × "protein" + 0.010 × "comput"

TABLE 5. Top five topics and top ten authors for each topic retrieved by author topic model (ATM).

Topic 0	Topic 1
Words	**Words**
0.019×"network" + 0.006×"control" + 0.005×"system" + 0.005×"data" + 0.004×"internet" + 0.004×"design" + 0.004×"present" + 0.004×"applic" + 0.004×"servic" + 0.004×"implement"	0.009×"neuron" + 0.004×"inform" + 0.004×"cell" + 0.004×"nois" + 0.003×"visual" + 0.003×"signal" + 0.003×"respons" + 0.003×"neural" + 0.003×"synapt" + 0.003×"activ"
Authors	**Authors**
(Salman Ahmad, 0.96624458)	(Saqib Ali, 0.96400599)
(Naima Iltaf, 0.64941235]	(AkhtarNawazKhan, 0.95594332)
(Muhammad Wasif Tanveer, 0.40642871)	(Attaur Rehman, 0.93468727)
(Tayyaba Azim, 0.22840075)	(Abid Sohail, 0.17558429)
(Furqan Muhammad Khan, 0.12415263)	(Ijaz Haider Naqvi, 0.06335061)
(MushtaqAli, 0.10196346)	(Fayyazul Amir Afsar Minhas, 0.05956543)
(Muhammad Shiraz, 0.08690942)	(Mahmood Ashraf, 0.04799168)
(Tariq Umar, 0.08087225)	(Amina Jameel, 0.03960921)
(Tauseef Jamal, 0.068107556)	(Muhammad Inamul Haq', 0.030140541)
(Sadaf Abdul Rauf, 0.049766426)	(Shariq Hussain, 0.028265943)

Topic 2	Topic 3	Topic 4
Words	**Words**	**Words**
0.015× "model" + 0.010×"learn" +0.010×"algorithm"+ 0.009×"method" + 0.009×"data" + 0.008× "base"+ 0.008×"propos" + 0.007×"imag" + 0.006×" problem" + 0.006×"network"	0.003×"price" + 0.002×"attractor" + 0.002×"bid" + 0.001×"log" + 0.001×"vc" + 0.001×"smo"+0.001×"polynomi" + 0.001×"lexi" + 0.001×"spot" + 0.001×"lower"	0.001×"margin" + 0.001×"light" + 0.001×"adaboost" + 0.001×"ep" + 0.001×"doom" + 0.001×"vr" + 0.001×"environ" + 0.001×"stop" + 0.001×"rl" + 0.001×"ure"
Authors	**Authors**	**Authors**
(Akhlaque Ahmad, 0.960342848)	(Khurram Jawad, 0.0284556484)	(Khurram Jawad, 0.029993866)
(Adeel Yousaf, 0.95593347)	(Muhammad Zubair', 0.027861176906828839)	(Muhammad Zubair, 0.028564983298252272)
(Saima Farhan, 0.954375192)	(Najeeb Ullah, 0.027058475)	(Muhammad Sajjad, 0.028160421)
(Ibrar Ali Shah, 0.953780173)	(Abid Sohail, 0.026388260863771759)	(Abid Sohail, 0.027869010708312363)
(Mohammad Nauman, 0.953594430)	(Muhammad Ashraf, 0.02484433)	(Najeeb Ullah, 0.027441727)
(Syed Sajid Hussain, 0.950753093)	(Muhammad Sajjad, 0.0238056104)	(Muhammad Ashraf, 0.0252259519)
(Syed Ali Abbas, 0.950582136)	(Muhammad Wasif Tanveer, 0.022293101)	(Muhammad Wasif Tanveer, 0.023476160)
(Irfana Memon, 0.950370783)	(Shariq Hussain, 0.021436022)	(Shariq Hussain, 0.021724904)
(Osman Khalid, 0.950209344)	(Ijaz Haider Naqvi, 0.0201463222)	(Ijaz Haider Naqvi, 0.020165151)
(Saleem Aslam, 0.950090571)	(Aftab Khan, 0.019389792)	(Aftab Khan, 0.019822892)

author-topic generation is higher. We have observed both the models for overfitting by using cross validation on held out resumes and selecting the number of topics for which the model is least perplexed. It was found that as we increased the number of topics K, the documents in the cross validation set get partitioned into very small topic collections producing more words with smaller probabilities. Such small probabilities lead the perplexity to explode for large K. With large number of topics K, the probability that the test document covers the same proportion of topics as train set decreases causing perplexity to grow and model to overfit as shown in Tables 1b and 2b earlier.

Next, we monitor the quality of topics and authors' entitlement to each topic with LDA and ATM respectively. We can compare the quality of topics retrieved by both the models from the semantics conveyed by most probable words for a theme and their probability strength for a topic. We have formally not given a specific label to the retrieved topics 0 to 4, however it seems like the areas of research identified by topic models are Computer Networks/Mobile Communication, Software Engineering and Modeling, Image Processing/Computer Vision, Machine learning/Artificial Intelligence and DataBases/Data Mining respectively. Note, that we have not optimized the semantic coherence of words for a topic, however research has been carried out in this area already [45], [46] and this avenue could be explored for further analysis in future. It is also important to note that better labels do not make a bad topic good. We need to check the co-occurrence information of words for all topics too.

A. SUPERVISOR RECOMMENDATION TO STUDENTS

In order to generate supervisor recommendations for the students, we preprocess the abstracts of project proposals

TABLE 6. Author topic model shows best supervisor match for the student interested in working on content based image retrieval and computer networks. We have chosen supervisors from Higher education commission's national database and checked their research relevance with the proposal of the student. (a) Supervisor recommendations at national level are made on the basis of Hellinger distance and KL divergence. (b) Submitted project proposal's probability for each topic is mentioned along with the list of recommended supervisors whose author topic probabilities with respect to each topic are given.

Student Name	Potential Supervisors	Hellinger Distance	KL Distance
1.Wajahat Amin	Amina Jamil	0.011327345	0.00050046
	Amjad Ali	0.011784324	0.00055963
	Furqan Aziz	0.014573251	0.00083250
	Zahoor Jan	0.019377082	0.00144147
	Tayyaba Azim	0.423309059	0.522304
2.Aziz Rehman	Ayesha Hakim	0.010254106	0.00042489
	Maqsood Hayat	0.017095374	0.00121428
	Shabbar Naqvi	0.017885818	0.00130442
	Sadaf Abdul Rauf	0.024080198	0.00242492

(a)

Student Name	Topic 0	Topic 1	Topic 2	Topic 3	Topic 4
1.Wajahat Amin	0.016776	0.011288	0.949139	0.011517	0.011278
Potential Supervisors					
Amina Jamil	0.013865	0.019306	0.891644	0.060669	0.014513
Amjad Ali	0.012732	0.013453	0.945960	0.014567	0.013285
Furqan Aziz	0.013615	0.014045	0.943874	0.014063	0.014400
Zahoor Jan	0.012182	0.013484	0.949067	0.012757	0.012507
Tayyaba Azim	0.010350	0.213886	0.753752	0.010431	0.011578
2.Aziz Rehman	0.0167696	0.852585	0.102150	0.014545	0.013949
Potential Supervisors					
Ayesha Hakim	0.022360	0.016669	0.920660	0.022360	0.020942
Maqsood Hayat	0.013533	0.015395	0.942138	0.015130	0.013801
Shabbar Naqvi	0.016391	0.029342	0.920363	0.016420	0.017480
Sadaf Abdul Rauf	0.018005	0.055160	0.888298	0.021192	0.017342

(b)

submitted by the students where students appear as document authors in the test set. The faculty's resumes are also preprocessed likewise as discussed above and augmented with the students' submissions in the test set. This test data is given to the trained ATM for author topic modeling of information. The results of the experiment showing recommendation for two students interested in working on *content based image retrieval* and *networks* domain are illustrated in Table 6. Subject to the content of the submitted proposal, the most similar topic for Wajahat ul Amin is Topic 2 (0.949139), whereas Aziz Rehman's proposal is highly relevant to Topic 1 (0.852585).

Through ATM, one can observe the topics where students and supervisors are appearing jointly with a strong probability and make recommendations accordingly. Keeping in account the content of the proposal, the system first finds relevant faculty members in the student's area of interest and then assesses which faculty member has higher relevance to the student's proposal. We can observe one such successful recommendation for Wajahat. The second example of Aziz Rehman is shared on purpose to demonstrate what happens when there are fewer/no faculty members working in student's area of interest. There were altogether fewer network experts on campus with publications not matching to the student's proposal. The model therefore suggests alternate supervisors not necessarily belonging to the student's area of interest yet their research record promises to support the idea student is interested to work on. Such a recommendation might not be the best recommendation, as the model is offering an unpopular and non-expert supervisor to the student, however things could not be made much different for this

TABLE 7. Author topic model showing researchers with similar areas of interest in the national and international research community. The developed recommender system reveals opportunities of possible collaborations between the researchers based on their expertise. (a) Best research collaborator match at national level on the basis of Hellinger distance and KL divergence. (b) Best researcher match at international level on the basis of Hellinger distance and KL divergence distance.

(a)

Main Author	Similar National Researchers	Hellinger Distance	KL Divergence
Abdul Nasir Khan	Saleemullah	0.001418	0.047355
	Amjad Ali	0.002359	0.037984
	Mian Muhammad Hamayun	0.002672	0.000874
Shariq Hussain	Aftab Khan	0.015942	0.561349
	Ayesha Hakim	0.017833	0.480081
	Abid Sohail	0.248017	0.301228
Zuhaib Ashfaq Khan	Babar Nazir	0.006668	0.010842
	Qaiser Abbas	0.010217	0.007923
	Muhammad Azhar Iqbal	0.012705	0.019388
Tayyaba Azim	Muhammad Wasif Tanveer	0.409334	0.645781
	Furqan Aziz	0.274419	0.427204
	Zahoor Jan	0.285873	0.471292
Furqan Aziz	Zahoor Jan	0.017280	0.001245
	Tayyaba Azim	0.274419	0.231696
	Fawad Hussain	0.3765187	0.432342

(b)

Main Author	Similar International Researchers	Hellinger Distance	KL Divergence
Abdul Nasir Khan	Yang Zhang	0.669570	1.85886
	Zoubin Ghahramani	0.598371	1.3337
	Sebastian Thrun	0.735915	0.781245
Shariq Hussain	Charles Sutton	0.191524	0.134918
	Jiawei Han	0.609260	1.4884
	Adrian Trapletti	0.449186	0.833707
Zuhaib Ashfaq Khan	Peng Cui	0.748778	2.75366
	Chang Lan	0.601732	1.45448
	Jay Chen	0.641850	1.68199
Tayyaba Azim	Rodney Douglas	0.121426	0.39010
	Zoubin Ghahramani	0.547535	1.78895
	Sebastian Thrun	0.603578	1.12389
Furqan Aziz	Charles Sutton	0.552091	1.09285
	Wenwu Zhu	0.737583	2.36591
	Zoubin Ghahramani	0.596839	1.32787

student due to the novelty of his research idea and lack of subject experts in the university. We have also calculated the NDCG scores of supervisor recommendations given to Wajahat and Aziz. The NDCG score for Wajahat is 0.9954 and the NDCG score for Aziz Rehman is 0.9886, both reflecting the relevance of recommendations returned to the students subject to the provided data set.

B. RESEARCH COLLABORATOR RECOMMENDATIONS

In order to generate recommendations for project collaborators in academia, we have also computed KL Divergence and Hellinger distance based similarity metrics for academics registered in our data base. A list of some selected collaborators at national level could be seen in Table 7a.

The table shows that for Abdul Nasir Khan, the most matching researcher is Saleemullah having the lowest Hellinger distance value of 0.0014186, the second closest researcher is Amjad Ali having a distance value of 0.0023596 and the third relevant researcher is Mian Muhammad Humayun with a distance value of 0.002672. The results of these distance based evaluation metrics have been verified by checking the area of specialization of queried and recommended collaborator. Such evaluation metrics powered by topic models can help us achieve multiple objectives: 1) Find relevant co-supervisors or collaborators for a

TABLE 8. Application of the proposed recommender system for identifying faculty eligible to teach courses offered in the home institutes. (a) Recommended national faculty for the courses taught in university. (b) Recommended international faculty for the courses taught in university. (c) Topic association of recommended national and international faculty.

Course Name	Recommended National Faculty	Hellinger Distance	KL Distance
Computer Vision	Shariq Hussain	0.011458	0.000536
	Zareena Kausar	0.0348404	0.005012
	Babar Nazir	0.038832	0.006238
Information Security	Muhammad Inam ul Haq	0.003709	5.4525e-05
	Imran Sarwar Bajwa	0.003723	5.5264e-05
	WaqasJadoon	0.004218	7.1063e-05
	Usama Ijaz	0.006051	0.000143
Machine Learning	Muhammad Sajjad	0.013365	0.000718
	NajeebUllah	0.014640	0.000892
	Ayesha Hakim	0.017583	0.001233

(a)

Course Name	Recommended International Faculty	Hellinger Distance	KL Distance
Computer Vision	Matthias Seeger	0.007718	0.000235
	David Picard	0.020471	0.001737
	Joshua B Tenenbaum	0.020690	0.001765
Information Security	Nobuo Suematsu	0.002878	3.3210e-05
	Akira Hayashi	0.002878	3.3210e-05
	Simone Croci	0.004339	7.4214e-05
Machine Learning	Michael C Mozer	0.010391	0.000440
	Wei Liu	0.013804	0.000779
	H Attias	0.014544	0.000849

(b)

Course Name	Topic 0	Topic 1	Topic 2	Topic 3	Topic 4
Computer Vision	0.048082	0.013695	0.912346	0.012938	0.012937
Recommended National Faculty					
Shariq Hussain	0.021760	0.020285	0.909167	0.028220	0.020565
Zareena Kausar	0.014751	0.016473	0.934467	0.018518	0.015789
Babar Nazir	0.017289	0.029339	0.916024	0.018827	0.018518
Recommended International Faculty					
Matthias Seeger	0.057223	0.012860	0.903086	0.013673	0.013155
David Picard	0.056343	0.011868	0.907996	0.011884	0.011907
Joshua B Tenenbaum	0.057653	0.010275	0.909682	0.011315	0.011072
Information Security	0.011673	0.011637	0.014318	0.950666	0.011703
Recommended National Faculty					
Muhammad Imam ul Haq	0.012125	0.013293	0.933946	0.028866	0.011768
Imran Sarwar Bajwa	0.0123125	0.013534	0.949570	0.012216	0.012366
WaqasJadoon	0.011233	0.015293	0.949108	0.012343	0.012020
Recommended International Faculty					
Nobuo Suematsu	0.013901	0.012162	0.951391	0.011284	0.011260
Akira Hayashi	0.242292	0.025238	0.680088	0.026435	0.025945
Simone Croci	0.013171	0.012278	0.950138	0.012206	0.012205
Machine Learning	0.031806	0.014997	0.923048	0.015019	0.015127
Recommended National Faculty					
Muhammad Sajjad	0.025756	0.019687	0.904927	0.025857	0.023770
NajeebUllah	0.040821	0.026661	0.873893	0.029449	0.029173
Ayesha Hakim	0.017606	0.017632	0.923844	0.020647	0.020268
Recommended International Faculty					
Michael C Mozer	0.027148	0.016167	0.923467	0.016723	0.016493
Wei Liu	0.016376	0.017128	0.931024	0.017858	0.017612
H Attias	0.017642	0.016774	0.937386	0.0140296	0.014167

(c)

project, 2) Gauge our research expertise by comparing the distance with key researchers in the field at national and international level. See Table 7b for the illustration of foreign research experts recommended by the proposed system. It is important to note that these experts only consists of researchers whose papers were published in the proceedings from where the data was gathered. If we increase our database size, the results shall give us a picture of global research landscape featuring various experts with similar research interests. The proposed application could also be extended for locating industrial experts to promote industry academia linkages.

C. COURSE TEACHER ALLOCATION

We also assessed the proposed system for allocating courses to faculty members present in national institutes. In order to provide this functionality, the standard course outlines approved by Higher Education Commission (HEC) were used as documents authored by their course titles. These author topic documents were used as a test set and passed to the author topic model to find out faculty who could teach those subjects nationally as well as internationally. Some of the results of this experiment are shown in Table 8.

In order to increase the reliability of the results, we have also demonstrated its performance with international researchers who may/may not be a part of academia, yet their subject specializations revealed through their publications make them a relevant candidate for teaching the course. It is important to note that several courses can fall together in a particular topic due to their overlapping course contents and fellowship to a broad area of science. For example, Machine Learning and Computer Vision both are specialized courses of Artificial Intelligence and hence show strong probability to Topic 2. To evaluate the quality of these recommendation, we have also calculated the NDCG scores of recommended national and international faculty. For Computer Vision, the NDCG scores for the national and international faculty are 0.9947 and 0.9994 respectively. For Information Security, we attained NDCG scores of 1.00 and 0.9979 respectively, whereas for Machine learning, the NDCG scores for the national and international faculty are 0.9973 and 0.9962 respectively. These results endorse the recommendations provided to the users based on their relevancy and order of search retrieval.

D. ASSESSMENT OF HIGHER EDUCATION LANDSCAPE OF THE COUNTRY

Using the retrieved topics and faculty's research associations returned by ATM, we also take a look at the country's higher education landscape featuring the strengths as well as weaknesses of our institutes in ICT sector. One can have a bird's eye view of the human resource deficit country is facing in various areas of research and can therefore necessitate the need of making policies for the eradication of expertise gap in higher education institutes. Figure 3 reports the mean probability of researchers belonging to each topic category in author topic model. We noticed that the country has a large pool of researchers in Topic 2, yet there is a scarce arrangement of faculty in other areas of Computer Science. Please note that these observations only hold true with the deployed data set that involves academic researchers approved by the Higher Education Commission (HEC) of Pakistan. There may exist a pool of additional Computer Science experts who have neither registered nor are approved by HEC for supervising research and are therefore part of the unexplored market segment assisting the students in country.

E. OBJECTIVENESS & COMPUTATIONAL NEEDS OF THE PROPOSED RECOMMENDER SYSTEM

The automation proposed by the current system adds a new dimension of transparency and fairness in managing research projects and allocating courses/supervisors in academia. The lack of human engagement in the entire process ensures that the system is not influenced due to the likes/dislikes of

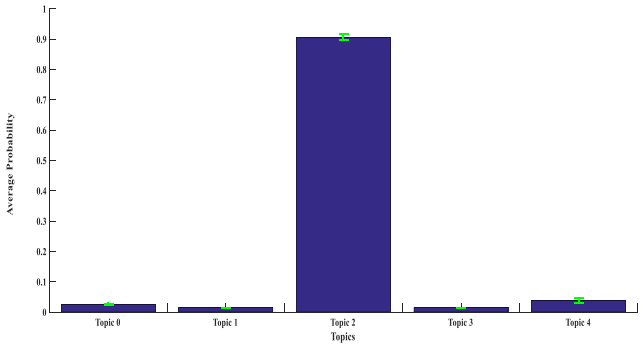


FIGURE 3. Mean probability of researchers' relevance to a specific subfield in Computer Science.

administrative staff, rather all the recommendations are based on merit determined by matching the scholarly content of CV/proposal with the content of papers' database showcasing existing and emerging research themes.

The proposed system, *ScholarLite* was developed and executed on Intel Core i-5 with 2.40 GHz processor and 4GB RAM. As one could observe, the computational needs of the model are not beyond the financial reach of government organizations in developing states and the proposed system could easily be deployed using a basic standalone machine. It is important to note that we haven't deployed any deep model for the task at hand due to their heavy requirements in terms of computational power and parameter optimization. We believe that deploying such models would be expensive in resource constrained environment present in developing countries. Therefore, the experimentation was confined to use of LDA and ATM models only.

The total running time taken for training LDA and ATM topic models is 835.8065577 seconds and 640. 1657035999999 seconds respectively. The time complexity incorporates CPU time only as we haven't implemented the code on graphics processing unit (GPU) to accelerate performance. It is possible to utilize the graphics processing units (GPUs) for training the topic models and utilizing them for generating recommendations [47], [48] when the size of the data set scales up. The current generation of GPUs provides higher computational capability and higher memory bandwidth than a commodity multi-core CPU thus allowing one to offload compute intensive portions of their program to run on GPU while running the remainder code on CPU.

F. SCALABILITY OF PROPOSED RECOMMENDER SYSTEM

ScholarLite demonstrates the proof of concept for faculty in IT/Computer Science only, however the proposed recommender system is scalable to other programs/degrees offered in the national/international institutes. In order to assure scalability of the proposed recommender system, we have deployed variational inference algorithm instead of Gibbs sampling for inference in LDA as well as ATM to assist topic discovery from a large corpus of examples. Topic models have shown their flexibility to scale up due to Variational Inference algorithm and parallel programming techniques on distributed computing architecture in the past [49], [50]. Thus, one can leverage such techniques to deliver a scalable recommender system encompassing all the disciplines of study in our national institutes. In order to ensure reproducibility and transparency of the research results, an open source code of the project is shared at https://github.com/tabzim/Recommender-System/.

V. CONCLUSION AND FUTURE WORK

This work aims at deploying machine learning learning tools for developing a recommender system useful for faculty and students in academia. Choosing relevant research supervisors, course teachers, and project investigators is considered quite a challenging task due to staff's cross domain interests, stale and varying templates of resumes and changing industrial expectations from the academia. The proposed solution demonstrates how to automate these practices efficiently while keeping transparency and relevancy intact. The proposed methodology is scalable to different disciplines and can easily adapt to new market trends by augmenting its 'train model' with the latest publications of researchers. The system ensures that students willing to work on new research ideas are able to find out relevant supervisors either in their enrolled institutes nationally or internationally. We have shown the results generated by two popular probabilistic models: LDA and ATM on real world data set and found out that the generative performance of LDA is much better than ATM, however ATM gives semantically more useful information than LDA and proves more suitable for the recommendation task at hand. ATM gives author information jointly with the discovered themes, thus making our recommendation procedures more pragmatic and logical.

The proposed system is the first recommender system of its kind whose proof of concept is shown for registered Computer Science staff in Pakistan only. In future, we aim to scale this system for faculty of other programs too. Scaling the system to other programs may help national organizations derive meaningful insights about the higher education landscape of the country and take data literate decisions focused on improving real challenges faced by academia. We would also like to explore the effect of using other types of features such as bigrams, term frequency-inverse document frequency (TF-IDF) to represent the word tokens and see their impact on supervisor recommendation and course allocation task. The system at the moment does not encompass time dynamics which may prove useful to represent authors changing research interests over a period of time. We would like to explore this avenue of research in future so that the recommendations remain relevant and up to date with respect to the dynamics of research and industry.

ACKNOWLEDGMENT

The authors would like to thank the Higher Education Commission (HEC) of Pakistan for maintaining and sharing the database of all faculty members eligible for supervising stu-

dents at graduate and post graduate levels. They also appreciate the Institute of Management Sciences for their continuous support in conducting this research and sharing up-to-date resumes of their Computer Science staff. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

REFERENCES

[1] I.-Y. Song, M. Song, T. Timakum, S.-R. Ryu, and H. Lee, "The landscape of smart aging: Topics, applications, and agenda," *Data Knowl. Eng.*, vol. 115, pp. 68–79, May 2018.

[2] G. Engin *et al.*, "Rule-based expert systems for supporting university students," *Procedia Comput. Sci.*, vol. 31, pp. 22–31, May 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1877050914004189

[3] M. Diaby, E. Viennet, and T. Launay, "Toward the next generation of recruitment tools: An online social network-based job recommender system," in *Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining (ASONAM)*, New York, NY, USA, Aug. 2013, pp. 821–828. doi: 10.1145/2492517.2500266.

[4] J. Xu and R. He, "Expert recommendation for trouble ticket routing," *Data Knowl. Eng.*, vol. 116, pp. 205–218, Jul. 2018.

[5] H. Mezni and T. Abdeljaoued, "A cloud services recommendation system based on fuzzy formal concept analysis," *Data Knowl. Eng.*, vol. 116, pp. 100–123, Jul. 2018.

[6] D. Mican and N. Tomai, "Association-rules-based recommender system for personalization in adaptive Web-based applications," in *Current Trends in Web Engineering*. Berlin, Germany: Springer, 2010, pp. 85–90. doi: 10.1007/978-3-642-16985-4_8.

[7] M. Dhanda and V. Verma, "Recommender system for academic literature with incremental dataset," *Procedia Comput. Sci.*, vol. 89, pp. 483–491, Aug. 2016.

[8] J. Beel and S. Dinesh. (2017). "Real-world recommender systems for academia: The pain and gain in building, operating, and researching them [long version]." [Online]. Available: https://arxiv.org/abs/1704.00156

[9] M. Mirzaeibonehkhater, "Developing a dynamic recommendation system for personalizing educational content within an e-learning network," Ph.D. dissertation, Dept. Elect. Comput. Eng., Purdue Univ., Lafayette, IN, USA, 2018.

[10] Y. Seroussi, I. Zukerman, and F. Bohnert, "Authorship attribution with topic models," *Comput. Linguistics*, vol. 40, no. 2, pp. 269–310, 2014. doi: 10.1162/COLI_a_00173.

[11] H. Liu, X. Kong, X. Bai, W. Wang, T. M. Bekele, and F. Xia, "Context-based collaborative filtering for citation recommendation," *IEEE Access*, vol. 3, pp. 1695–1703, Oct. 2015.

[12] V. A. Rohani, Z. M. Kasirun, and K. Ratnavelu, "An enhanced content-based recommender system for academic social networks," in *Proc. IEEE 4th Int. Conf. Big Data Cloud Comput. (BdCloud)*, Dec. 2014, pp. 424–431.

[13] J. Tang, S. Wu, J. Sun, and H. Su, "Cross-domain collaboration recommendation," in *Proc. 18th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, New York, NY, USA, 2012, pp. 1285–1293. doi: 10.1145/2339530.2339730.

[14] Y. Zhang, C. Yang, and Z. Niu, "A research of job recommendation system based on collaborative filtering," in *Proc. 7th Int. Symp. Comput. Intell. Design*, vol. 1, 2014, pp. 533–538.

[15] G. Adomavicius and A. Tuzhilin, "Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions," *IEEE Trans. Knowl. Data Eng.*, vol. 17, no. 6, pp. 734–749, Jun. 2005. doi: 10.1109/TKDE.2005.99.

[16] S. Jiang, X. Qian, J. Shen, Y. Fu, and T. Mei, "Author topic model-based collaborative filtering for personalized POI recommendations," *IEEE Trans. Multimedia*, vol. 17, no. 6, pp. 907–918, Jun. 2015.

[17] H. Wang, N. Wang, and D.-Y. Yeung, "Collaborative deep learning for recommender systems," in *Proc. 21th ACM Int. Conf. Knowl. Discovery Data Mining (KDD)*, New York, NY, USA, 2015, pp. 1235–1244. doi: 10.1145/2783258.2783273.

[18] W. Jian, J. He, K. Chen, Y. Zhou, and Z. Tang, "Collaborative filtering and deep learning based recommendation system for cold start items," *Expert Syst. Appl.*, vol. 69, pp. 29–39, Mar. 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0957417416305309

[19] A. van den Oord, S. Dieleman, and B. Schrauwen, "Deep content-based music recommendation," in *Proc. 26th Int. Conf. Neural Inf. Process. Syst. (NIPS)*, vol. 2, 2013, pp. 2643–2651. [Online]. Available: http://dl.acm.org/citation.cfm?id=2999792.2999907

[20] X. Zhao, Z. Niu, and W. Chen, "Interest before liking: Two-step recommendation approaches," *Knowl.-Based Syst.*, vol. 48, pp. 46–56, Aug. 2013. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0950705113001214

[21] M. Pennacchiotti and S. Gurumurthy, "Investigating topic models for social media user recommendation," in *Proc. 20th Int. Conf. Companion World Wide Web (WWW)*, New York, NY, USA, 2011, pp. 101–102. doi: 10.1145/1963192.1963244

[22] M. Córdova, P. Raman, L. Si, and J. Fish, "Relevancy prediction of micro-blog questions in an educational setting," in *EDM*, 2014, pp. 1–2.

[23] C. Wang and D. M. Blei, "Collaborative topic modeling for recommending scientific articles," in *Proc. 17th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, New York, NY, USA, 2011, pp. 448–456. doi: 10.1145/2020408.2020480.

[24] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Mar. 2003.

[25] M. Rosen-Zvi, T. Griffiths, M. Steyvers, and P. Smyth, "The author-topic model for authors and documents," in *Proc. 20th Conf. Uncertainty Artif. Intell. (UAI)*, Arlington, VA, USA: AUAI Press, 2004, pp. 487–494. [Online]. Available: http://dl.acm.org/citation.cfm?id=1036843.1036902

[26] O. Mortensen, "The author-topic model," M.S. thesis, Dept. Appl. Math. Comput. Sci., Tech. Univ. Denmark, Lyngby, Denmark, 2017. [Online]. Available: http://www.compute.dtu.dk/English.aspx

[27] W. Chen, Z. Niu, X. Zhao, and Y. Li, "A hybrid recommendation algorithm adapted in e-learning environments," *World Wide Web*, vol. 17, no. 2, pp. 271–284, 2014. doi: 10.1007/s11280-012-0187-z.

[28] D. Hand, "Text mining: Classification, clustering, and applications edited by Ashok Srivastava, Mehran Sahami," *Int. Stat. Rev.*, vol. 78, no. 1, pp. 134–135, 2010. doi: 10.1111/j.1751-5823.2010.00109_1.x.

[29] E. Mangina and J. Kilbride, "Evaluation of keyphrase extraction algorithm and tiling process for a document/resource recommender within e-learning environments," *Comput. Educ.*, vol. 50, no. 3, pp. 807–820, 2008.

[30] A. Garrido, L. Morales, and I. Serina, "On the use of case-based planning for e-learning personalization," *Expert Syst. Appl.*, vol. 60, pp. 1–15, Oct. 2016.

[31] Z. Gulzar, A. A. Leema, and G. Deepak, "PCRS: Personalized course recommender system based on hybrid approach," *Procedia Comput. Sci.*, vol. 125, pp. 518–524, Jan. 2018.

[32] C.-K. Hsu, G.-J. Hwang, and C.-K. Chang, "Development of a reading material recommendation system based on a knowledge engineering approach," *Comput. Educ.*, vol. 55, no. 1, pp. 76–83, 2010.

[33] C.-K. Hsu, G.-J. Hwang, and C.-K. Chang, "A personalized recommendation-based mobile learning approach to improving the reading performance of EFL students," *Comput. Educ.*, vol. 63, pp. 327–336, Apr. 2013.

[34] P.-Y. Wang and H.-C. Yang, "Using collaborative filtering to support college students' use of online forum for English learning," *Comput. Educ.*, vol. 59, no. 2, pp. 628–637, 2012.

[35] K. Sugiyama and M.-Y. Kan, "Scholarly paper recommendation via user's recent research interests," in *Proc. ACM 10th Annu. Joint Conf. Digit. Libraries*, 2010, pp. 29–38.

[36] J. Sun, J. Ma, Z. Liu, and Y. Miao, "Leveraging content and connections for scientific article recommendation in social computing contexts," *Comput. J.*, vol. 57, no. 9, pp. 1331–1342, 2014.

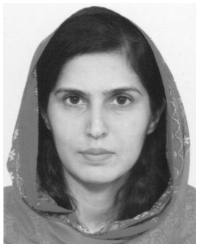
[37] J. Sun *et al.*, "A novel approach for personalized article recommendation in online scientific communities," in *Proc. IEEE 46th Hawaii Int. Conf. Syst. Sci. (HICSS)*, Jan. 2013, pp. 1543–1552.

[38] F. Xia, H. Liu, I. Lee, and L. Cao, "Scientific article recommendation: Exploiting common author relations and historical preferences," *IEEE Trans. Big Data*, vol. 2, no. 2, pp. 101–112, Jun. 2016.

[39] G. Tian and L. Jing, "Recommending scientific articles using bi-relational graph-based iterative RWR," in *Proc. 7th ACM Conf. Recommender Syst.*, 2013, pp. 399–402.

[40] E. Medvet, A. Bartoli, and G. Piccinin, "Publication venue recommendation based on paper abstract," in *Proc. IEEE 26th Int. Conf. Tools Artif. Intell. (ICTAI)*, Nov. 2014, pp. 1004–1010.

[41] A. Tripathy, A. Agrawal, and S. K. Rath, "Classification of sentiment reviews using n-gram machine learning approach," *Expert Syst. Appl.*, vol. 57, pp. 117–126, Sep. 2016.
[42] G. Salton and J. McGill, *Introduction to Modern Information Retrieval*. New York, NY, USA: McGraw-Hill, 1986.
[43] P. Barnaghi, P. Ghaffari, and J. G. Breslin, "Opinion mining and sentiment polarity on Twitter and correlation between events and sentiment," in *Proc. IEEE 2nd Int. Conf. Big Data Comput. Service Appl. (BigDataService)*, Mar./Apr. 2016, pp. 52–57.
[44] R. Ghosh, K. Ravi, and V. Ravi, "A novel deep learning architecture for sentiment classification," in *Proc. 3rd Int. Conf. Recent Adv. Inf. Technol.(RAIT)*, 2016, pp. 511–516.
[45] D. Mimno, H. Wallach, E. Talley, M. Leenders, and A. McCallum, "Optimizing semantic coherence in topic models," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2011, pp. 262–272. [Online]. Available: http://dl.acm.org/citation.cfm?id=2145432.2145462
[46] Q. Mei, X. Shen, and C. Zhai, "Automatic labeling of multinomial topic models," in *Proc. 13th ACM Int. Conf. Knowl. Discovery Data Mining (KDD)*, New York, NY, USA, 2007, pp. 490–499. doi: 10.1145/1281192.1281246.
[47] X. Xie, Y. Liang, X. Li, and W. Tan. (2018). "CuLDA_CGS: Solving large-scale LDA problems on GPUs." [Online]. Available: https://arxiv.org/abs/1803.04631
[48] M. Lu, G. Bai, Q. Luo, J. Tang, and J. Zhao, "Accelerating topic model training on a single machine," in *Proc. Asia–Pacific Web Conf.* Berlin, Germany: Springer, 2013, pp. 184–195.
[49] K. Zhai, J. Boyd-Graber, N. Asadi, and M. L. Alkhouja, "Mr. LDA: A flexible large scale topic modeling package using variational inference in Mapreduce," in *Proc. ACM 21st Int. Conf. World Wide Web*, 2012, pp. 879–888.
[50] A. Smola and S. Narayanamurthy, "An architecture for parallel topic models," *Proc. VLDB Endowment*, vol. 3, nos. 1–2, pp. 703–710, 2010.
[51] M. Salehi, I. N. Kamalabadi, and M. B. G. Ghoushchi, "Personalized recommendation of learning material using sequential pattern mining and attribute based collaborative filtering," *Edu. Inf. Technol.*, vol. 19, no. 4, pp. 713–735, Dec. 2014. doi: 10.1007/s10639-012-9245-5.



HUMA SAMIN received the bachelor's degree in information technology from the University of Peshawar, Peshawar, Pakistan, and the master's degree in computer science, majoring in software engineering, from the Lahore University of Management Sciences (LUMS), in 2009. She is currently pursuing the Ph.D. degree in software engineering with the School of Engineering and Applied Sciences, Aston University, Birmingham. She was a Lecturer with the Edwardes College, Peshawar, Pakistan. She joined the Institute of Management Sciences, in 2013. Her research interests include topic models, reinforcement learning, agile models, and mobile application development.

TAYYABA AZIM was born in Rawalpindi, Pakistan. She received the M.S. degree in computer science from the National University of Computer and Emerging Sciences, Islamabad, in 2009, and the Ph.D. degree in computer science from the University of Southampton, Southampton, U.K., in 2014. She was a Data Scientist with the Horizon Research Institute, University of Nottingham, and a Research Associate at Cortexica Vision Systems, Imperial College of London, U.K. Since 2015, she has been a tenured-track Assistant Professor with the Institute of Management Sciences. She has authored a book and has several conference and journal publications in the area of computer vision and machine learning. Her research interests include deep learning, topic models, kernel methods, and real-time systems. She was a recipient of a Startup Research Grant, a National Grassroots ICT Research Initiative Fund, a National ICT Research and Development Grant, and an Overseas Ph.D. Scholarship and received the Best Paper Award at ICPRAM, in 2017.