A study to enhance candidate screening process using similarity analysis

Anshul Ujlayan and Manisha Sharma*

School of Management,
Gautam Buddha University,
Greater Noida, India
Email: ujlayan@gmail.com
Email: manisha@gbu.ac.in
*Corresponding author

Abstract: Recruitment process is always continuous and open positions mostly have short time to get filled thereby for any such open position, organisations generally have to deal with higher incremental costs. Many advanced recruitment processes/methodologies are directly associated with the risk of failure, loss of effort, loss of time and money. Every organisation in any industry always aims to hire the best suitable candidate to the open position with right skillsets in the shortest possible time and with the lowest costs. Therefore, there is a need to experiment and explore the existing approaches, which can help in reducing the time and the screening effort in initial stage of recruitment practices. Consequently this paper attempts to implement the similarity analysis approach for a random sample of resumes from IT sector in National Capital Region (NCR), India to provide the best match of candidates' profile as per the required job description. The latent Dirichlet allocation (LDA) is used to find the similarity between job description and candidate's profile. The study will help the IT industry in identifying and selecting the best matched candidates' profile based on the key features in job description.

Keywords: recruitment process; similarity analysis; latent Dirichlet allocation; LDA; candidates' profile; IT sector.

Reference to this paper should be made as follows: Ujlayan, A. and Sharma, M. (2019) 'A study to enhance candidate screening process using similarity analysis', *Int. J. Business and Data Analytics*, Vol. 1, No. 1, pp.41–53.

Biographical notes: Anshul Ujlayan is a Research Scholar in the School of Management, Gautam Buddha University, India. He is doing a research in the area of business analytics and machine learning. He is a postgraduate in statistics and has more than ten years of work experience in analytics, data science and machine learning with various multinational companies.

Manisha Sharma is an Assistant Professor at the School of Management, Gautam Buddha University, India. She has organised many conferences, workshops in the area of decision sciences and research methodology and has many international publications to add to her credentials, and has also presented research papers at various conferences. Her disciplinary research interests lie in the fields of business analytics, supply chain management, decision sciences and application of statistical methods in management.

1 Introduction

Quality and productivity of the workforce are very important for people-centric and knowledge-intensive industries such as IT, BPO and services in general. It is the responsibility of the talent acquisition (TA) function within HR to recruit the workforce of highest possible quality. The TA function needs to attract the best possible talent from a complex supply chain of educational institutes (if experience is not required), job portals, employment agencies, recruitment consultants and direct sourcing through buddy, emails, advertisements, walk-ins and web.

The channels differ in terms of the number and quality of resumes sourced, time and cost for sourcing, selection ratio and joining ratio for sourced candidates, etc. The recruitments themselves need to be done under stringent goals such as shortest possible times-frames, lowest possible recruitment costs/efforts and working at many locations and dealing with diverse domains and technical skills. Moreover, a variety of human and economic factors affect recruitments (Srivastava et al., 2015). Methods of recruiting employees with proper competence can be divided into two basic types: external recruitment and recruitment within the organisation. Depending on which type of recruitment an organisation decides to engage in, there are various tools to choose from. Every organisation may carry out a simultaneous process of external and internal recruitment or may decide to carry out only one of those. Some companies in principle choose to recruit from within the company in the initial phase of recruitment. This course of action boosts the employees' engagement in their work, has a positive effect on the motivation of the whole team and accelerates the recruitment process. A lot of tools and recruitment techniques can be used in both external and internal recruitment. The basic recruitment tools used in the internal recruitment include interviews with potential candidates, obtaining references from the candidate's direct superior, and using the results of competence measurement, provided that the organisation has implemented human resources management system based on competence. The advanced recruitment tools used in the internal recruitment include assessment centre (a multidimensional process of competency assessment conducted by a team of judges) which may be carried out either with the support of a third party or by a dedicated unit from inside the organisation (Grabara et al., 2016).

An organisation, which decides to recruit a new employee through, the process of external recruitment has at its disposal two basic methods: passive recruitment and active recruitment. Passive recruitment methods include re-examination of the curriculum vitae of candidates who participated in earlier recruitment processes. The main advantage of using the passive recruitment method over the active recruitment method is time and cost saving. Where the advantage of the active recruitment method is the possibility to assess the competence of a candidate in different situations and with the use of a variety of techniques and tools. The basic active recruitment methods include advertisements in the press, on the websites of recruitment agencies or in social media. The advertisement includes the employer's expectations that the candidate should meet, which is referred to as the candidate's profile.

In the recruitment of the middle and upper management staff, organisations often use the services of technology recruiter. It is an example of another active recruitment method. This method involves a precise specification of the employer's expectations regarding the candidate (they are usually discussed with the technology recruiter). In the next step, technology recruiter interview potential candidates. Within the active recruitment methods, organisations can also participate in special events such as local job fairs or university job fairs. Another active recruitment method is close cooperation with universities with such cooperation every organisation may offer students traineeship or paid internship. During such an internship the future employer has a full image of the potential worker's competence (Szałkowski, 2000). An organisation, which decides to recruit a new employee through, the process of external recruitment has at its disposal the methods described above. Applying these or other methods allows an organisation to recruit potential candidates for a given position. The literature on this subject describes at least a few ways of planning recruitment as a process (Breaugh, 2008). Every organisation may carry out the recruitment process in a way which is considered most suitable for it. It is essential, however, that the process be carried out fairly and accurately. The care taken during the recruitment process gives higher efficiency in the form of the acquisition of candidates with a higher level of competence.

2 Information technology recruitment industry: the scenario for technology skills hiring

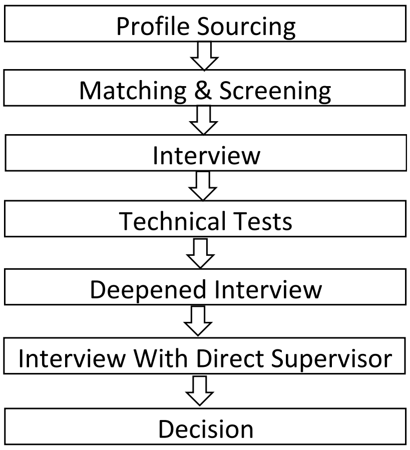
The global sourcing market in India continues to grow at a higher pace compared to the IT-BPM industry. The global IT and ITeS market (excluding hardware) reached US$1.2 trillion in 2016–2017, while the global sourcing market increased by 1.7 times to reach US$173–178 billion. India remained the world's top sourcing destination in 2016–2017 with a share of 55%. Indian IT and ITeS companies have setup over 1,000 global delivery centres in over 200 cities around the world. More importantly, the industry has led the economic transformation of the country and altered the perception of India in the global economy. India's cost competitiveness in providing IT services, which is approximately 3–4 times cheaper than the USA, continues to be the mainstay of its unique selling proposition (USP) in the global sourcing market. However, India is also gaining prominence in terms of intellectual capital with several global IT firms setting up their innovation centres in India. The IT industry has also created significant demand in the Indian education sector, especially for engineering and computer science. The Indian IT and ITeS industry is divided into four major segments – IT services, business process management (BPM), software products and engineering services, and hardware. India has come out on top with the highest proportion of digital talent in the country at 76% compared to the global average of 56% (IBEF Report, 2018).

Recruiters receive large numbers of applications through e-mails, online job portals, or through services provided by partner staffing companies. Online job portals like monster.com, indeed.com, dice.com, and careerbuilder.com and staffing firms like Manpower Inc., Adecco and Kelly Services draw in most of the applications. Résumés obtained from such diverse sources are thus difficult to process and store in a unified database format. It becomes very tedious to select the most appropriate ones. Since résumés are structured documents containing information based on the author's thinking and writing skills, they can be created in a multitude of formats (e.g., plain text or structured table), languages and file types (e.g., txt, pdf, and doc.). This makes the information extraction (IE) process highly complex. High precision and recall becomes complicated for this domain (Saxena, 2011).

Dynamic filtering techniques are used by the industry to extract relevant resumes. These filtering techniques match hundreds of resumes from the database to a single job posting. Resumes extracted by these filters are generally similar to each other as they satisfy the same search criteria, based mainly on keyword matching. The resume filtering becomes more challenging when the job requirement demands a specialised skill set. Thus, it becomes cumbersome to further analyse the short-listed resumes in order to select the most relevant resumes.

The current standard process followed by IT industry is given by Grabara et al. (2016). The benchmark process which most of the recruiters are using across the globe is shown in Figure 1.

Figure 1 Standard recruitment process



3 Review of the literature

A detailed review was done to understand the different methods adopted by the researchers in order to enhance the recruitment process and they are summarised as below:

Kirimi and Moturi (2016) used predictive analytics and data mining to classify candidate based on their skills and performance profile for best screening process. Nedelcu (2017) performed a study that focused on identifying patterns which relates to human skills. Recently, with the new demand and increasing visibility, human resources are seeking a more strategic role by harnessing data mining methods. To investigate how applicant characteristics influence the use of impression management (IM) tactics in interviews, and how these behaviours affect interviewer perceptions of person-job fit (P–J fit) and applicant-interviewer similarity. Results from 72 applicants demonstrated that extraverted applicants made greater use of self-promotion during their interviews, while agreeableness was associated with non-verbal cues (Kristoff et al., 2002).

The explorative paper by Flecke (2015) aimed to link the underlying concept of the person-job fit to the opportunities which are provided to recruiters by utilising SNS as assistance tools to determine the degree of fit between a person and a job. Spaeth and

Desmarais (2013) did an exploration of text-mining techniques to improve classical collaborative filtering methods for a site aimed at matching people who are looking for expert advice on a specific topic. Results were compared from a LSA-based text similarity analysis, a simple user-user collaborative filter, and a combination of both methods used to recommend people to meet for a knowledge-sharing website

Hauff and Gousios (2015) proposed a pipeline that automates this process and automatically suggests matching job advertisements to developers, based on signals extracting from their activities on GitHub.

Rawashdeh and Ralescu (2013) stated that an important aspect of these applications relied on similarity measures between nodes in the network. Several similarity measures, described in the literature were surveyed with the goal of providing a guide to their selection in various applications. The further research conducted by Eras (2015) showed a semantic architecture for the extraction, comparison and feedback of professional and educational competencies. The main product was a scheme that facilitates the detection of competencies based on the skills and knowledge. The scheme carried out tasks of natural language processing and of similarity calculation, among others, in order to determine the differences among the professional and educational competencies.

Huang et al. (2007) proposed a new method to build the query model with latent state machine (LSM) which captures the inherent term dependencies within the query and the term dependencies between query and documents. The method firstly splits the query into subsets of query terms (i.e., not only single terms, but different combinations of multiple query terms). Secondly, these query term combinations are then considered as weighted latent states of a hidden Markov model to derive a new query model from the pseudo relevant documents. Thirdly, our method integrates the aspect model (AM) with the EM algorithm to estimate the parameters involved in the model.

Riedl and Biemann (2012) used general method to use information retrieved from the latent Dirichlet allocation (LDA) topic model for text segmentation: using topic assignments instead of words in two well-known text segmentation algorithms, namely TextTiling and C99, leads to significant improvements. Fang (2015) did an experiment to show that this supervised training procedure is able to produce a re-ranking model that improves significantly over the search ranking on common information retrieval performance metrics.

Shao and Qin (2014) observed that LDA topic model has been widely applied to text clustering owing to its efficient dimension reduction. The prevalent method is to model text set through LDA topic model, to make inference by Gibbs sampling, and to calculate text similarity with Jensen-Shannon (JS) distance. However, JS distance cannot distinguish semantic associations among text topics. For this defect, a new text similarity computing algorithm based on hidden topics model and word co-occurrence analysis is introduced. Tests are carried out to verify the clustering effect of this improved computing algorithm 'there is no explicit use of predictive analytics in the above process. Therefore, we are proposing a model for this research which will use predictive modelling approach to optimise the process followed by industry'.

Based on review done from the available literature, there is certainly a need of exploring and implementing the predictive analytics approach at each step of the recruitment process. The approach will help the recruiters to process the candidate application information and utilise the same information to reduce the effort in reviewing and evaluating the candidate profile for recruitment process.

Thereby, the objective of this study is to implement the similarity analysis to find the best match of the candidate's profile from a set of random sample based on job description. For this, first the relevant resumes will be identified based on similarity analysis and then will be clustered based on the taxonomy of similar items.

4 Research methodology

Three different IT companies were approached to collect the candidates' profiles and job recruitment or description for software development job and thereby for the defined job description, the 100 profiles were considered for the similarity analysis with the help of LDA.

4.1 Job description

'Programming experience on Java J2EE two years programming experience with semantic technology ontology data base development experience with building end to end application passionate Java J2EE Architect with five years of software architect development experience in developing web-based applications hands on with J2EE application servers such as Tomcat/Apache/WebSphere hands on experience in Java based framework. Spring Hands on experience with REST web services API hands on experience with enterprise search indexing platform SolrLucene Siren Aspose Experience with NoSQL RDBMS Mongo DB application level programming using well versed in using source code control systems such as GIT SVN perforce hands on in designing developing unit test cases for various application personal attributes and behaviours should take initiatives self motivated and self directed performs in ambiguous environment team worker good interpersonal skills ability to interact with senior management'.

4.2 Method: LDA

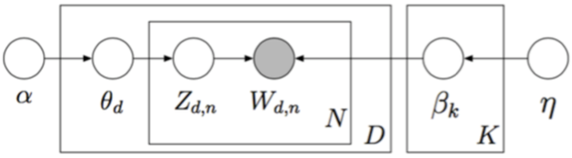
In machine learning and natural language processing, topic models are generative models, which provide a probabilistic framework (Zhou and Haiyi, 2016). Topic modelling methods are generally used for automatically organising, understanding, searching and summarising large electronic archives. The 'topics' signifies the hidden, to be estimated, variable relations that link words in a vocabulary and their occurrence in documents. A document is seen as a mixture of topics. Topic models discover the hidden themes throughout the collection and annotate the documents according to those themes. Each word is seen as drawn from one of those topics. Finally, a document coverage distribution of topics is generated and it provides a new way to explore the data on the perspective of topics.

Latent semantic analysis (LSA) given by Niraula and Banjade (2013) and LDA proposed by Vidhya and Aghila (2010), are two popular mathematical approaches to modelling textual data. Questions posed by algorithm developers and data analysts working with LSA and LDA models motivated to How closely do LSAs concepts correspond to LDAs topics? How comparable are the most significant terms in LSA ideas to the most imperative terms of relating LDA subjects? Are the same documents affiliated with matching concepts and topics? Do the report closeness diagrams delivered by the two calculations contain comparative record? LSA and LDA models, numerous other factor models of literary information, much in like manner. Both are use bag-of-words modelling, begin by transforming text corpora into term-document frequency matrices, reduce the high dimensional term spaces of textual data to a user-defined number of dimensions, produce weighted term lists for each concept or topic, produce concept or topic content weights for each document, and produce outputs used to compute document relationship measures. Yet despite these similarities, the two algorithms generate very different models. LSA uses vector index document (VID) to define a basis for a shared semantic vector space, in which the maximum variance across the data is captured for a fixed number of dimensions. In contrast, LDA utilises regards each record as a blend of latent fundamental subjects, every theme is displayed as a blend of word probabilities from a vocabulary. Although LSA and LDA outputs can be used in similar ways, the output values represent entirely different quantities, with different ranges and meanings. LSA produces term idea and record idea connection matrix. LDA produces term-topic and document-topic matrices. Direct comparison and interpretation of similarities and differences between LSA and LDA models is an important challenge in understanding which model may be most appropriate for a given analysis task.

LDA is a generative model that allows sets of observations to be explained by unobserved groups that explain why some parts of the data are similar. LDA has made a big impact in the fields of natural language processing and statistical machine learning and has quickly become one of the most popular probabilistic text modelling techniques in machine learning.

Intuitively in LDA, documents exhibit multiple topics. In text pre-processing, we exclude punctuation and stop words (such as, ‘if’, ‘the’, or ‘on’, which contain little topical content). Therefore, each document is regarded as a mixture of corpus-wide topics. A topic is a distribution over a fixed vocabulary. These topics are generated from the collection of documents. For example, the recruitment topic has word ‘experience’, ‘skillsets’ with high probability and the computer topic has word ‘data’, ‘network’ with high probability. Then, a collection of documents has probability distribution over topics, where each word is regarded as drawn from one of those topics. With this document probability distribution over each topic, we will know how much each topic is involved in a document, meaning which topics a document is mainly talking about.

Figure 2 Graphical model for LDA



As the figure illustrated, we can describe LDA more formally with the following notation. First, α and η are proportion parameter and topic parameter, respectively. The topics are $\beta_{1:k}$, where each β_k is a distribution over the vocabulary. The topic proportion for the d^{th} document are θ_d, where $\theta_{d,k}$, is the topic proportion for topic k in document d. The topic assignments for the d^{th} document are Z_d, where $Z_{d,n}$ is the topic assignment for the nth word in document d. Finally, the observed words for document d are W_d, where $W_{d,n}$ is the n^{th} word in document d, which is an element from the fixed vocabulary. With this notation, the generative process for LDA corresponds to the following joint distribution of the hidden and observed variables:

$$p(\beta_{1:K}, \theta_{1:D}, \omega_{1:D}) = \prod_{i=1}^{K} p(\beta_i) \prod_{d=1}^{D} p(\theta_d) \left(\prod_{n=1}^{N} p(Z_{d,n} | \theta_d) p(W_{d,n} | \beta_{1:K}, Z_{d,n}) \right)$$

This distribution specifies a number of dependencies. The topic assignment $Z_{d,n}$ depends on the per-document topic distribution θ_d and the observed word $W_{d,n}$ depends on all of the topic assignment $Z_{d,n}$ and all the topics $\beta_{1:k}$ and the topic assignment $Z_{d,n}$ (Liu, 2016).

Operationally, it will be defined by looking up which topic $Z_{d,n}$ refers to and looking up the probability of the word $W_{d,n}$ within that topic. These dependencies define LDA in the particular mathematical form of the joint distribution, and it can defined in the *probabilistic graphical model* for LDA. Probabilistic graphical models provide a graphical language for describing families of probability distributions. The graphical model for LDA is in Figure 1. These two representations are equivalent ways of describing the probabilistic assumptions behind LDA (Blei, 2012). Obtaining a minimum perplexity value with optimal number of topics is the base criterion for selecting the different values of the parameters.

4.3 Design implications

The importance of candidate profile screening process in recruitment is really high. The above results suggested a way to optimise the time, cost and manual effort by implementing topic modelling (i.e., LDA). In above study we need to have well defined job description and set of candidates profile to for screening process. The LDA algorithm consists of tokenisation, stop word removal. After the preprocessing step is complete LDA algorithm can be match similarity of input text file with the collection of document and giving list of matching document link. Our results have implications for the design of initial screening process to select relevant profile of the candidates. The effort required to find out the most suitable profile of the candidate during the hiring process. Our results suggest it is important to provide all required technical skills, project experience and academic qualification in job description.

5 Data analysis

5.1 Data pre-processing

The next step is text-cleaning process. The purpose of text cleaning is to simplify the text data, eliminating as much as possible language dependent factors. Articles are written in natural language for human to understand. But in text mining, those data are not always easy for computers to process. In this experiment, there are three steps in text cleaning namely

1 Tokenisation: a document is treated as a string, removing all the punctuations and then partitioned into a list of tokens.

2 Removing stop words: stop words such as 'the', 'if', 'and' are frequently occurring but no significant meanings which need to be removed.

3 Stemming word: stemming word that converts different word form into similar canonical form.

For example, computing to compute, happiness to happy. This process reduces the data redundancy and simplifies the later computation (Khan et al., 2010).

On filtering content archives they are changed over into a feature vector. A progression utilises TF-IDF calculation. Every token is relegated a weight, as far as recurrence (TF), mulling over a solitary research dad for each. IDF considers every one of the papers, scattered in the database and figures the opposite recurrence of the token showed up in all examination papers. So, TF is a local weighting function, while IDF is global weighting function.

5.2 The analysis framework

The below analysis used python and web Gensim packages to analyse the similarity between job description and candidate profile. The output of the similarity analysis shown in Table 1 describes the similarity of the select profile where % similarity is greater then 10% with the job description. Profile 22 has the highest similarity with job description 88%, this will help the recruiter to refer the profile for further action. In this study we selected the optimal number of topics (k), by fixing the value for hyper parameters.

Table 1 Similarity between candidates' profile and job description

Profile number	*Similarity with job description*
Profile 45	11%
Profile 9	11%
Profile 20	11%
Profile 42	17%
Profile 68	20%
Profile 94	23%
Profile 14	27%
Profile 97	27%
Profile 5	31%
Profile 22	36%
Profile 25	*88%*

Figure 3 is the graph between highest similarities of candidates' profile with job description from Table 1. In Figure 4, profile 25 has highest similarity with job description and the profile 9, 20 and 45 has the same percentage similarity with job description

Percentage similarity between job description and 72 profiles is very low. This shows that processing these profiles for screening will not help much to the recruiter. This also implies that the candidate's profile is not matching with job description or there is very less match between them.

Figure 3 Analysis framework

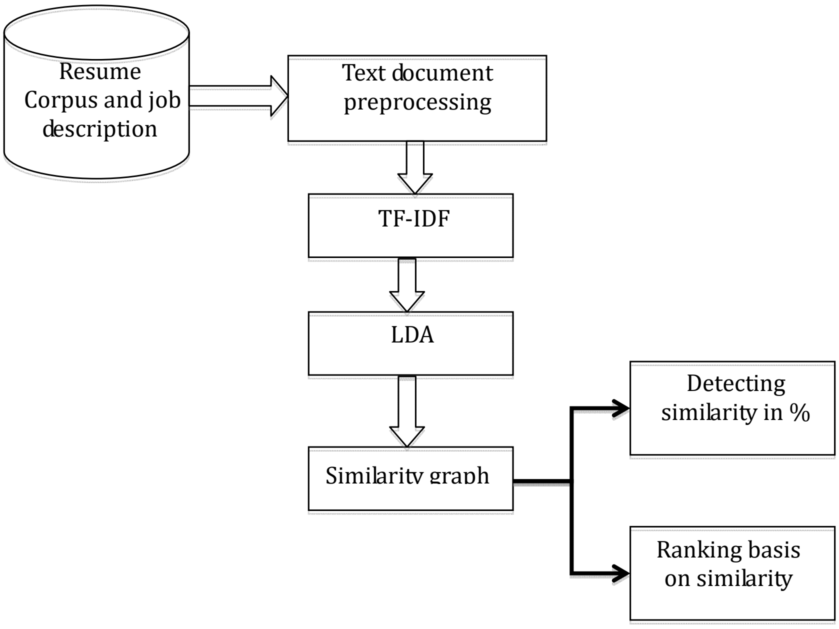
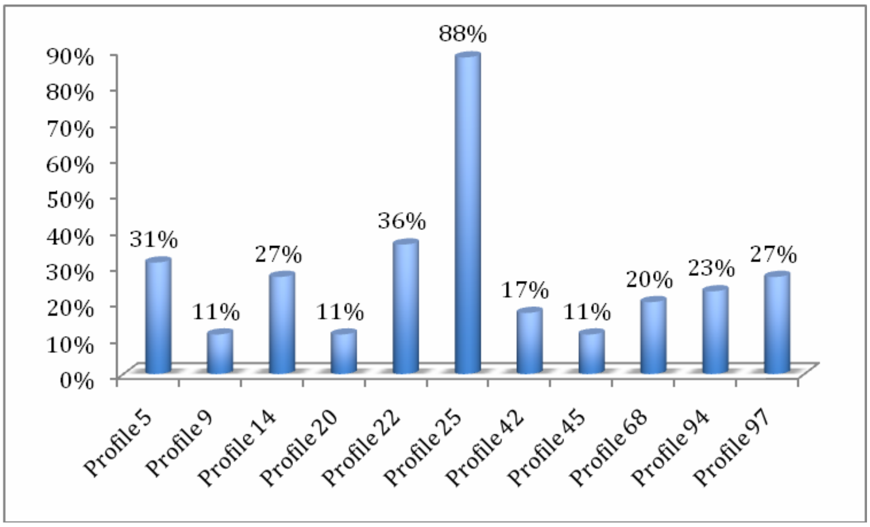

Resume Corpus and job description
Text document preprocessing
TF-IDF
LDA
Similarity graph
Detecting similarity in %
Ranking basis on similarity

Figure 4 Similarity between job description and candidates profiles (see online version for colours)



The output of the similarity analysis shows that the similarity between profile 38 and profile 87 is more than 32%. This result will help to identify the similar profile for further process and knowledge repository. Profile 60 and profile 37 have 100% similarity, which shows that these profiles are almost similar to each other.

Table 2 Profiles depicting low similarity with job description

Similarity with job description	*Candidates' profile number*	*Total no. of profiles*
1%	10, 44, 56, 61, 62, 87, 99	7
2%	4, 6, 15, 21, 30, 32, 37, 38, 39, 48, 49, 60, 71, 77, 80, 81, 82, 91	18
3%	16, 17, 19, 26, 27, 28, 31, 35, 36, 47, 52, 67, 69, 74, 79, 84, 86, 98	18
4%	7, 13, 18, 33, 46, 53, 54, 65, 73, 75, 76, 83, 85, 92, 93	15
5%	2, 3, 24, 34, 41, 43, 50, 55, 59, 64, 66, 72, 88, 90	14

Table 3 Similarity between profiles

	Similarity between profiles	
Profile 25	Profile 5	34%
Profile 22	Profile 25	40%
Profile 24	Profile 30	44%
Profile 24	Profile 31	42%
Profile 30	Profile 31	35%
Profile 24	Profile 36	34%
Profile 17	Profile 52	33%
Profile 70	Profile 72	32%
Profile 71	Profile 72	48%
Profile 75	Profile 55	54%
Profile 75	Profile 58	33%
Profile 76	Profile 55	32%
Profile 87	Profile 38	33%
Profile 37	Profile 60	100%
Profile 52	Profile 15	33%
Profile 62	Profile 61	61%

Based on the output of similarity analysis all the profiles are grouped in four different categories. The criterion is the similarity value between job description and the candidate profile. The 'category 1 consists of all the profiles with greater than 27% similarity', 'category 2 consists the profiles with greater than 11% and less than 27% similarity', 'category 3 consists all the profiles with greater than 5% and less than 11% similarity', 'category 4 consists all the profiles with less than 5% similarity'.

Table 4 Category distribution of profiles

	Similarity	*No. of profile*
Category 1	High	5
Category 2	Moderate	6
Category 3	Low	17
Category 4	Very low	72

6 Conclusions and scope for further research

In this paper, we implemented similarity analysis on a random sample of candidate's profile to demonstrate the usefulness of latent Dirichlet algorithms in recruitment screening process. We proposed two experiments, which built up on finding similarity and use of the similarity value for further categorisation based on business threshold value. Due to the limitation of data availability, this research is based on a relatively small random sample. However, the result is quite convincing even with the small size. Applying to a larger dataset will more likely achieve better results.

In this work, we examined how a proper defined job description can help to save time money and manual effort. The research experiment provides similarity value for all profiles with job description. Based on high similarity values one can directly pick up the suitable candidate for further processing. Grouping of candidates' profile based on similarity between profiles will also help to quickly pick similar profile for future recruitment process. Based on the findings one can extend application of topic modelling to manage, search and explore offline candidate's profile. In future, it can be generalised and applied to any number of candidates' profile or huge dataset.

References

Blei, D. (2012) 'Probabilistic topic model', *Communication of the ACM*, Vol. 55, No. 4, p.77.

Breaugh, A. (2008) 'Employee recruitment: current knowledge and important areas for future research', *Human Resource Management Review*, Vol. 18, pp.103–118.

Eras, A. (2015) 'Semantic architecture for the analysis of the academic and occupational profiles based on competencies', *Contemporary Engineering Sciences*, Vol. 8, No. 33, pp.1551–1563.

Fang, M. (2015) *Learning to Rank Candidates for Job Offers Using Field Relevance Models, University of Groningen* [online] https://lct-master.org/getfile.php?id=1342&n=1&dt=TH&ft=pdf (accessed 7 February 2018).

Flecke, L. (2015) 'Utilizing Facebook, LinkedIn and Xing as assistance tools for recruiters in the selection of job candidates based on the person-job fit', *5th IBA Bachelor Thesis Conference*, Enschede, 2 July.

Grabara, J. et al. (2016) 'Recruitment process optimization: chosen findings from practice in Poland', *Journal of International Studies*, Vol. 9, No. 3, pp.217–228.

Hauff, C. and Gousios, G. (2015) 'Matching GitHub developer profiles to job advertisements', *Proceedings of the ACM* [online] http://dl.acm.org/citation.cfm?id=2820563 (accessed 1 January 2018).

Huang, Q. et al. (2007) 'Learning and optimization of an aspect hidden Markov model for query language model generation', *Proceedings of the 1st International Conference on the Theory of Information Retrieval.*

IBEF Report (2018) *Report by India Brand Equity Foundation, IT & ITeS Industry in India* [online] https://www.ibef.org/archives/detail/b3ZlcnZpZXcmMzc2NTMmODk= (accessed 4 February 2018).

Khan, A. et al. (2010) 'A review of machine learning algorithms for text-document classification', *Journal Information Technology*, Vol. 1, No. 1, pp.4–20.

Kirimi, J. and Moturi, C. (2016) 'Application of data mining classification in employee performance prediction', *International Journal of Computer Applications*, Vol. 146, No. 7, pp.28–35.

Kristof-Brown, A.L., Barrick, M.R. and Franke, M. (2002) 'Applicant impression management: dispositional influences and consequences for recruiter perceptions of fit and similarity', *Journal of Management*, Vol. 28, pp.27–46.

Liu, Q. (2016) 'Employing latent Dirichlet allocation model for topic extraction of Chinese text', *International Journal of Database Theory and Application*, Vol. 9, No. 7, pp.51–66.

Nedelcu, B. (2017) 'Human talent forecasting', *Proceedings of the 11th International Conference on Business Excellence*, pp.437–447.

Niraula, N. and Banjade, R. (2013) 'Experiments with semantic similarity measures based on LDA and LSA', *International Conference on Statistical Language and Speech Processing Springer*, pp.188–199.

Rawashdeh, A. and Ralescu, A. (2013) 'Similarity measure for social networks – a brief survey', *Modern AI and Cognitive Science Conference (MAICS)* at Greensboro, 1353.

Riedl, M. and Biemann, C. (2012) 'Text segmentation with topic models', *Journal for Language Technology and Computational Linguistics*, Vol. 27, No. 1, pp.47–69.

Saxena, C. (2011) *Enhancing Productivity of Recruitment Process Using Data Mining & Text Mining Tools*, Master's Projects, p.324 [online] http://scholarworks.sjsu.edu/etd_projects/324 (accessed 7 January 2018).

Shao, M. and Qin, L. (2014) 'Text similarity computing based on LDA topic model and word co-occurrence', *2nd International Conference on Software Engineering, Knowledge Engineering and Information Engineering (SEKEIE 2014)*, Atlantis Press, pp.199–203.

Spaeth, A. and Desmarais, C. (2013) 'Combining collaborative filtering and text similarity for expert profile recommendations in social websites, from book user modeling, adaptation, and personalization', *21th International Conference, Proceedings*, Vol. 10, No. 14, pp.178–189.

Srivastava, R., Palshikar, G.K. and Pawar, S. (2015) 'Analytics for improving talent acquisition processes', *Proceedings of 4th international conference on advanced data analysis, business analytics and intelligence (ICADABAI 2015)*.

Szałkowski, A. (2000) *Wprowadzenie do zarządzania personelem*, Wydawnictwo Akademii Ekonomicznej w Krakowie, Cracow.

Vidhya, K. and Aghila, G. (2010) 'Text mining process, techniques and tools: an overview', *International Journal of Information Technology and Knowledge Management*, Vol. 2, No. 2, pp.613–622.

Zhou, T. and Haiyi, Z. (2016) 'A text mining research based on LDA topic modelling', *The Sixth International Conference on Computer Science, Engineering and Information Technology*, pp.201–210.