Queensland University of Technology
Brisbane Australia

This may be the author's version of a work that was submitted/accepted for publication in the following source:

Kapugama Geeganage, Dakshi, Xu, Yue, & Li, Yuefeng
(2018)
Investigation of the quality of topic models for noisy data sources.
In Tao, X, Pasi, G, & Weber, R (Eds.) *Proceedings of the 2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI).*
Institute of Electrical and Electronics Engineers Inc., United States of America, pp. 488-493.

This file was downloaded from: https://eprints.qut.edu.au/128777/

© Consult author(s) regarding copyright matters

This work is covered by copyright. Unless the document is being made available under a Creative Commons Licence, you must assume that re-use is limited to personal use and that permission from the copyright owner must be obtained for all other uses. If the document is available under a Creative Commons License (or other specified license) then refer to the Licence for details of permitted re-use. It is a condition of access that users recognise and abide by the legal requirements associated with these rights. If you believe that this work infringes copyright please provide details by email to qut.copyright@qut.edu.au

Notice: *Please note that this document may not be the Version of Record (i.e. published version) of the work. Author manuscript versions (as Submitted for peer review or as Accepted for publication after peer review) can be identified by an absence of publisher branding and/or typeset appearance. If there is any doubt, please refer to the published source.*

https://doi.org/10.1109/WI.2018.00-48

Investigation of the Quality of Topic Models for Noisy Data Sources

Dakshi T. Kapugama Geeganage
School of Electrical Engineering and Computer Science
Queensland University of Technology
Brisbane , Australia
dakshi.geeganage@hdr.qut.edu.au

Yue Xu
School of Electrical Engineering and Computer Science
Queensland University of Technology
Brisbane , Australia
yue.xu@qut.edu.au

Yuefeng Li
School of Electrical Engineering and Computer Science
Queensland University of Technology
Brisbane , Australia
y2.li@qut.edu.au

***Abstract*— Latent Dirichlet Allocation (LDA) has become the most stable and widely used topic model to derive topics from collections of documents where it depicts different levels of success based on diversified domains of inputs. Nevertheless, it is a vital requirement to evaluate the LDA against the quality of the input. The noise and uncertainty of the content create a negative influence on the topic model. The major contribution of this investigation is to critically evaluate the LDA based on the quality of input sources and human perception. The empirical study shows the relationship between the quality of the input and the accuracy of the output generated by LDA. Perplexity and coherence have been evaluated with three data-sets (RCV1, conference data set, tweets) which contain different level of complexities and uncertainty in their contents. Human perception in generating topics has been compared with the LDA in terms of human defined topics. Results of the analysis demonstrate a strong relationship between the quality of the input and generated topics. Thus, highly relevant topics were generated from formally written contents while noisy and messy contents lead to generate meaningless topics. A considerable gap is noticed between human defined topics and LDA generated topics. Finally, a concept-based topic modeling technique is proposed to improve the quality of topics by capturing the meaning of the content and eliminating the irrelevant and meaningless topics.**

Keywords—Topic modeling, LDA, Content quality

I. INTRODUCTION

Information is overloaded with the rapid growth of online data and the volume of text data has been drastically increased in different media. People are too much relying on electronic text than ever before with the interactions with "web, social media, instant messaging to online transactions, government intelligence, and digitized libraries" [1]. Large volume of poor quality text data is transmitting in social media and it is important to understand and categorize them into relevant topics. Hence, capturing the semantics is essential to trust the contents with uncertain information. Information is generated in large volume and it is indeed difficult to guarantee the quality of the text contents. It is a critical problem to employ human involvement to read, understand and summarize trillions of noisy and uncertain text data available in different media. Topic modeling is the state of the art and the widespread technique to understand, organize and extract the topics from collections of documents, hence different topic modeling techniques have been applied to identify the topics from collections of documents. Traditional and existing topic modeling approaches use the probabilistic methods [2], [3], [4] to discover the topics and the frequency and co-occurrences of the words are taken into consideration. The probabilistic models run according to a generative process that consists of hidden variables and accordingly, the words of the documents and the topic structure will be the main building blocks. Latent Semantic Analysis (LSA) [5], [6], Probabilistic Latent Semantic Analysis (PLSA) [2] and Latent Dirichlet Allocation (LDA) [3] are the popular topic modeling approaches and LDA is the most enhanced version of probabilistic topic modeling technique among them.

This paper presents an extensive analysis to evaluate the quality of the topics generated by LDA with the variation of input sources and then a new approach will be proposed to grab the semantic meanings from the collections of documents. It is indeed difficult to rely on the quality of the topics due to the noise, uncertainty, complexity of the contents, large volume and different written patterns. Most of the available text contents are poor quality, noisy and not up to the standard where content processing has been a difficult task. This analysis focuses to analyze three levels of inputs which are formally written contents belong to diversified areas of domains, formally written contents for a specific set of areas and messy and noisy social media contents which cover wide range of domains. Reuters Corpus version 1 (RCV1)[7] and abstracts of data mining conferences are professionally written documents while tweets extracted from twitter represent the noisy, uncertain social media data. It is beneficial to conduct an empirical analysis to evaluate the LDA with respect to the quality of the inputs and generated outputs hence there could be a negative impact for the topic generation when it processes poor quality contents. Most probably, tweets like social media contents are not trustful data which do not have a properly written content instead of set of characters or terms to share the messages among colleagues or in public forums.

An experimental analysis was conducted to evaluate the relevancy of the generated topics with the content of the collections of documents in three different datasets (RCV1, conference dataset and tweets). Qualitative and quantitative analysis were conducted to evaluate the LDA. Quantitative analysis calculates the perplexity and the coherence of the topic model. Qualitative analysis was done by comparing the LDA generated topics with the human defined topics. The contribution of this paper is to investigate the relationship among the quality of the input contents with the topics generated by LDA model and evaluate the relevancy and meanings of the topics in terms of the collections of documents. Furthermore, a new approach [8] is introduced to overcome two limitations of LDA model by applying a concept layer to filter more relevant and semantically meaningful topics.

Evaluating the LDA topic model to filter meaningful and relevant topics, exploring the characteristics of collections of documents which lead to success of the topic modelling

process and discovering a methodology to overcome the problems of existing LDA by generating semantically meaningful topics are the main objectives of this research. Next sections of the paper are organized as follows; Section 2 demonstrates an overview of LDA while section 3 reviews the related researches in the topic modeling and evaluation. Section 4 illustrates the experimental setup with the results and proposed concept embedded topic modeling technique [8] will be described in the section 5. Finally, section 6 explains the conclusion of the investigation.

II. OVERVIEW OF LDA

LDA [3] is designed for text corpora which considers as bag of words and the content of the collection is captured as "random mixtures over latent topics" [3]. LDA is based on a three-level hierarchical Bayesian model and topics will be described as a distribution over words. LDA arbitrarily selects the topic distribution and grasps topics which represent the topic distribution. Since the topic modeling focused to automatically generate the set of topics, LDA examines the documents where the topic organization, topics, document topic distribution and document word distribution are hidden in the collection. [4] Fig. 1 represents the topic generation process and (1) illustrates the probability distribution of LDA model.

$$\mathrm{p}\ (\beta_{1:K}, \theta_{1:D}, Z_{1:D}, W_{1:D}) = \prod_{i=1}^{K} p(\beta_1) * \prod_{d=1}^{D} p(\theta_d)$$

$$* \ (\prod_{n=1}^{N} p(Z_{d,n}|\theta_d) p(W_{d,n}|\beta_{1:K}, Z_{d,n})) \qquad (1)$$

LDA topic modelling process is based on 3 assumptions; Consider the corpus as the "bag of words" is the first assumption and accordingly the sequence of the words will not be taken into consideration. The second assumption specifies that the entire collections of documents are considered without a proper order or the sequence of documents will not be a problem to generate the topics. The final assumption states that the number of topics should be mentioned before starting the topic modeling process [4]. Hence, LDA grasps the word frequency and co-occurrences of words from a corpus with the assumption of "bag of words", the accuracy level of the output can be changed according to the quality of the input source.

Even though, LDA is defined as the most stable, widely used, trusted and reputed probabilistic approach to generate list of topics from collections of documents, it depicts different levels of success based on diversified contents of inputs. Same time meaningless and irrelevant topics will be generated due to not capturing the meanings of the words to conclude the set of topics. Since, number of topics should be given before the topic generation process, irrelevant and meaningless topics will be generated to fulfill the number of topics. This hinders the process of finding most suitable and semantically meaningful topics from the collection of documents.

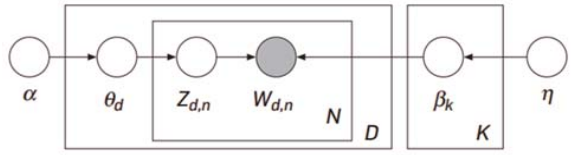

Fig. 1. Generative Process of LDA [4]

In analysing the previous researches conducted to evaluate the LDA, a less attention is paid to study the behavior of LDA model when generating topics from various types of input sources.

III. RELATED WORK

With the enhancement of topic modeling researches, there was a requirement to evaluate the developed topic models in different perceptions [9-11]. Several researches have been conducted to evaluate the set of topics against the relevancy of the topics in terms of the collections of documents and human perception. Topic models were evaluated against the human perception [12] and further highlighted the significance of human understanding of the contents with the semantic meaning. They considered LDA [3], PLSI [2] and Correlated Topic Model (CTM) [13] in their evaluation and went through two assessments based on the human perception named as "word intrusion" and "topic intrusion" [12]. Words and topics were separately considered to measure the semantic coherence and the relevancy to the collections of documents. An analysis [14] was conducted to compare the topics generated by "NMF (Non-negative Matrix Factorization)" in various collections of documents in terms of their coherence and associated generality. "Weighted" and "unweighted" topic descriptor techniques were used to evaluate the topics and according to their analysis, NMF has generated topics with high coherence. Coherence based evaluation [15] has been conducted to evaluate the topics extracted from topic models. They have conducted their analysis based on variety of subject domains and elaborated a scoring model based on "point wise mutual information (PMI)". Then human evaluators assessed the coherence of the produced topics and thereafter the coherence of the topics were forecasted based on WordNet [16], Wikipedia and Google.

It has been a proven fact that the result of human perception contains a significant difference when compared to the results of traditional topic modeling techniques. Therefore, there was a requirement[17] to find an approach which gives similar type of results as the human understanding process. WordNet [16] has been used in several researches [17] to grasp the semantic meanings of the words. WordNet concept hierarchy [17] was used to verify the relationship in between the human perception and ontology based topic interpretation. Gibbs sampling and annealed importance sampling (AIS) have been used as the evaluation techniques [18] to evaluate the LDA. Yi and Allan [19] focused on the information retrieval capability of topic models and assessed the performance of them. An evaluation metric-based topic model [20] was introduced to enhance the quality of topics in huge collections of documents from the "National Institutes of Health (NIH)". In their analysis [20], they have defined a coherence metrics to filter high/low quality topics and identified improving the semantic quality of topics as the most important challenge in future topic modeling. There are various approaches which combined the LDA with external knowledgebases to introduce semantic based topic models. [21], [22].

IV. ANALYSIS AND RESULTS

LDA can be evaluated through different aspects like performance, accuracy, topics quality and behavioural changes of parameters with different characteristics. Since

the poor quality contents lead for wrong decisions due to the uncertainty of the data, we have conducted the analysis to evaluate the LDA to measure the quality of topics with the quality of the input contents and human perception. Three types of datasets which belong to different quality levels were used to assess the quality of the topics generated by LDA model. They represent different level of complexities and written patterns.

Reuters Corpus version 1 (RCV1) [7], conference dataset and tweets were applied to generate topics from LDA to observe the reflection of quality of the contents against the generated topic models. RCV1 is a dataset which contains newswire stories about different domains like corporate, industrial, economics, government and social etc. Conference data set contained the abstracts of Conference on Information and Knowledge Management (CIKM), Conference on Hypertext and Social Media (HT), Knowledge Discovery and Data Mining (KDD) and Conference on Research and Development in Information Retrieval (SIGIR) from year 2002 to 2011. RCV1 and conference datasets are formally written collection which contain less noise when compared to the tweets, while tweets contain lots of crutches, pillar words, meaningless terms and symbols.

The analysis focused to assess the topic distribution with the perplexity and coherence. Further, the relevancy and meanings of the topics were compared to the human perception. Quantitative analysis was performed to evaluate the quality of the topics produced by LDA and values of α and β were set to 0.5 and 0.01 respectively. Number of iterations (passes) was 50 and number of topics was changed during the evaluation to observe the maximum topic quality of the LDA.

Perplexity and coherence were measured to assess the quality of topics in terms of likelihood and the relevance to the topic word distribution. Table 1 depicts the important attributes related to the RCV1, conference dataset and tweets. Further, a qualitative analysis was conducted to evaluate the quality of topics in terms of human perception by comparing the human defined topics with the LDA generated topics. Conference dataset and tweets were evaluated to compare the human defined topics.

TABLE I. IMPORTANT ATTRIBUTES OF THE DATASETS

Dataset	Number of Documents	Total number of tokens after pre-processing	Average number of tokens per document
RCV1	27463	3351025	123
Conference dataset	5494	406637	74
Tweets	1308468 (No of tweets)	48413316	37

A. Calculating the Perplexity

Perplexity calculates the log likelihood of a held out test in a dataset. Each dataset was divided into two random parts by considering 90% of training dataset and 10% of testing dataset and every time the test data set and training dataset were exchanged to cross validate the results. Finally, the average perplexity is calculated by considering the perplexity value of each round. Perplexity distribution of each dataset is shown in the Fig. 3.

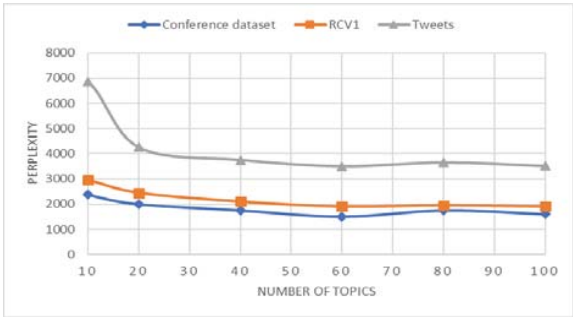

Fig. 2. Perplexity of datasets

Let the collections of documents $D= w_1, \ldots w_m$. In here, the testing dataset has not been seen by the LDA model and test data set considers as the collection of unseen documents (w_d). Accordingly log-likelihood will be evaluated according to the "collection of unseen or test documents w_d". The perplexity of held out documents can be defined as follows in (2) [3].

$$\text{Perplexity } (\mathrm{D}_{test}) = \exp -\frac{\sum_{d=1}^{M} logp(w_d)}{\sum_{d=1}^{M} N_d} \qquad (2)$$

The perplexity of RCV1, conference dataset and tweets has been measured for variations of number of topics. Accordingly, conference data set contains the lowest perplexity and Tweets contains the highest perplexity which contains the less quality of topic models.

B. Calculating the Coherence

Topic coherence evaluates the relevancy among the generated topics with the topic interpretation. Pairwise similarity of words will be considered among the topic-word distribution [21] and sum of the score is equivalent to the coherence. Pairwise scores of the words $v_1,..,v_n$ interprets the topics and (3) [21] specifies the coherence calculation.

$$\text{Coherence} = \sum_{(v_t, v_j) \in V} score(v_i, v_j) \qquad (3)$$

Many approaches have discussed different techniques to measure the topic coherence with benefits and limitations of each coherence method. Accuracy and performance criteria were evaluated to discover a bench mark for coherence measure. In this analysis, UMass coherence [20], was used to measure the coherence as it considers the cooccurrence of documents and checks the relevancy of the main word of the topic with each preceding topic words.

Further, UMass coherence evaluates the conditional probability of rare word with frequent/common words and checks the possibility to occur a rare word with a common word. "$D(v)$ is the document frequency (number of documents) of word type v where at least a single token of v included. $D(v, \tilde{v})$ is the co-document frequency of words v and . $V^{(t)} = v_1^{(t)}, \ldots, v_M^{(t)}$ is a list of M most probable words in topic t". [17] The equation of UMass coherence is stated in (4) [17], and coherence of each dataset depicts in the graph in Fig. 3. UMass coherence of topics generated by the datasets was measured for 50 iterations by changing the number of topics time to time.

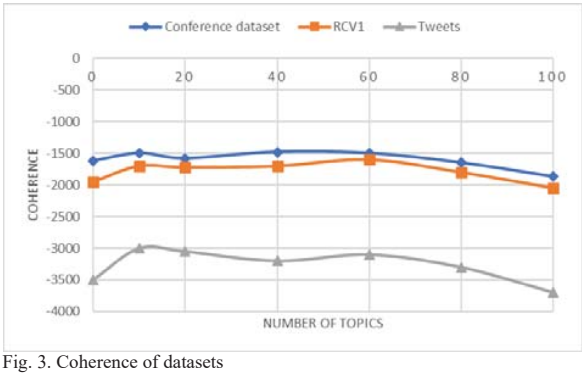

Fig. 3. Coherence of datasets

$$\text{Coherance}_{u\_mass} = \sum_{m=2}^{M} \sum_{l=1}^{m-1} log \frac{D(v_m^{(t)}, v_l^{(t)}) + 1}{D(v_l^{(t)})} \quad (4)$$

According to the graph and coherence measurements, the conference data set contains the highest coherence while RCV1 is in the middle and tweets dataset contains the lowest coherence values.

C. *Evaluate according to the Human Perception*

It is important to evaluate the topics generated by LDA with the human perception, since the basic principles of two processes are not closely aligned with each other. Human understands, summarizes the content and discovers the topics from document collections while LDA calculates the word frequency and co-occurrence to generate the set of topics. Further, human does not consider a pre-defined assumption of number of topics to be generated and most related topics will be found by understanding the content. Conference dataset and tweets were used for this experiment and human defined research paper titles, research tracks and hash tags (#) were taken to compare with the human perception. Conference dataset was divided according to the conference name (CIKM, HT, KDD, SIGIR) and then topics were generated from the abstracts. Human defined titles and conference tracks were combined as the human defined topics. Further, WordNet was embedded with the human defined topics and LDA generated topics to discover the synonyms and closely related words. Finally, human defined topics and LDA generated topics were compared to evaluate the human perception with the LDA. Number of topics generated by LDA were changed during the experiment and compared with the human defined topic. Table 2 demonstrates the organization of the conference dataset. Number of topics were changed during the experiment and applied the LDA model to each conference type. Finally matching proportion is calculated as follows ;

$$\text{Matching Proportion} = \frac{\text{Human Defined Topics}}{\text{LDA Generated Topics}}$$

Matching Proportion = Matching proportion of human defined topics with LDA generated topics

Human Defined Topics = Number of matched words with human defined topics

LDA Generated Topics = Number of unique Words in LDA generated topics

The matching proportion was calculated by changing the number of topics Fig. 4 and 5 depicts the matching proportions of Conference dataset and tweets respectively.

TABLE II. STRUCTURE OF THE CONFERENCE DATASET

Conference Name	Number of Documents	Number of unique words in the human defined category (after pre-processing)
CIKM	2048	4326
HT	483	1343
KDD	1227	2919
SIGIR	1736	3509

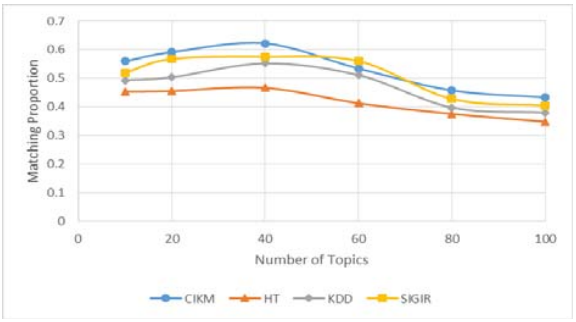

Fig. 4. Matching proportion according to the conference type

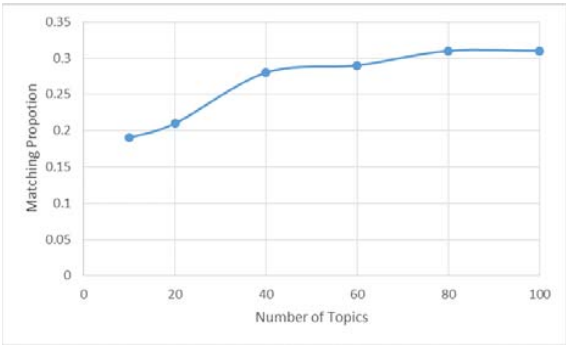

Fig. 5. Matching proportion of hashtags with generated topics

It is clearly noticed that the matching proportion is varying according to the conference and a significance change can be noticed in between the Conference on Information and Knowledge Management (CIKM) and Conference on Hypertext and Social Media (HT). This might happened due to the quantity of the documents and the number of unique words available in the abstracts of two conferences.

Then, hashtags of the tweet database were filtered and categorized as the human defined topics. WordNet was embedded and LDA generated topics for Tweets were compared with the hash tags by changing the number of topics. Finally, matching proportion is calculated as Fig. 5.

D. *Discussion*

It is noticed that there is a strong relationship exists among the quality of the topics and the quality of the input sources. Noisy and uncertain data leads to derive meaningless topics while structured and trustworthy contents generate acceptable set of topics. Topic distribution of the conference dataset is acceptable due to applying a formally written set of abstracts focused on data mining research domain. Even though the contents of RCV1 are formally written newswire stories, the topic distribution is not much quality as the conference dataset and it might happened due

to covering a wide range of domains. Finally, the topic distribution of tweets has been spread among different topics due to a large collection of noisy and messy contents when compared to the other two datasets. Further, it is noticed that the word distribution of conference dataset contained more general words about the researches such as approach, result, paper, research, users, propose etc rather than data mining or information retrieval related words. With the evidence of the word- topic distribution, it is clearly noticed that the topic modeling requires more attention to capture the semantic meaning of the content without only grasping the frequency and co-occurrences of the words.

The perplexity of good quality topic model is comparatively low and according to this experiment highest quality set of topics have been generated from the conference dataset. Even though the RCV1 and conference abstracts are formally written documents perplexity of conference dataset is lowest. This may be happened due to two reasons; conference dataset contains less number of documents compared to RCV1 and conference dataset narrow down to the field of data and information mining. The coherence of good quality topics will be high and speciality here in this study is, the coherence values of good quality input sources are high when compared to the low-quality input contents. Conference data set contains the highest coherence due to generating set of highly related topics. Since, conference dataset contains formally written abstracts, and all are related to the data mining researches, the coherence is higher than RCV1 and tweets. Word-topic distribution of tweets covered a wide range of domains and most of the words were not related to each other. Therefore, tweets contain the lowest coherence due to the poor quality of the contents and the large volume.

In the evaluation with the human defined topics, it is evidenced that there is a considerable gap in between human understanding and topic generation with word frequency and co-occurrence. Even though the experiment attempted to relate the human defined topics and LDA generated topics using WordNet synonyms, still there is a difference due to not capturing the semantics of the words in LDA topic generation process. Further, matching proportion of conference dataset is higher than the tweets due to conference dataset contains meaningful and less ambiguous words which described in the data mining domain.

V. TOWARDS A CONCEPT EMBEDDED TOPIC MODELING APPROACH

According to the results of the analysis, there are strong evidences to say the some of the topics produced by the LDA are meaningless and irrelevant. Considering the word frequency and co-occurrence instead of the meaning of the contents and focusing pre-specified number of topics rather than the most related topics from the content are the major two problems existing in the LDA model. The proposed approach aims to address the existing issues and planes to introduce a concept embedded topic modeling technique [8] which generates the most suitable topics by capturing the semantics of the content from collections of documents.

The concept embedded topic modeling approach contains techniques to grab the semantic elements from the content, categorize the concept and terms considering the meanings and provide most suitable topics without getting the number of topics from the user prior to the topic modeling process.

The research is conducted in three phases and collections of documents will be the input of the research. After going through the three phases, the proposed approach will generate semantically meaningful set of topics as the output.

A. Phase 1: Semantic elements extraction from collections of documents

Main objective of this phase is to identify and extract the semantic elements from collections of documents. Semantically meaningful elements are identified from the collections of documents and WordNet is used to find the related words together. Pre-processing is the first step of this phase and tokenizing, stop word removing and stemming the words to ignore the variational forms can be defined as the most important tasks under the pre-processing. Term-frequency and Inverse document frequency (TF-IDF) have been used to filter the frequent terms and important terms are considered as the inputs for the next steps. Then WordNet is used as the lexical database to find the related terms and synonyms (synsets), entailments, and key words of the first definition are applied to make cliques of related words together. Probase [22] is used to identify and interpret the concepts in more meaningful manner and derive the matching concepts for related terms. Patterns are generated based on the matching concepts and cliques of related terms.

B. Phase 2: Semantic concept categorisation and domain clustering

The concepts and the cliques identified in the phase-1 will be further processed to determine the category of each clique. A new algorithm will be presented to identify the category of each concept and semantic clique. Then semantic representation will be introduced to represent the cliques in terms of the concepts. Finally, the collection of documents may contain set of identified semantic cliques and the concepts associated with each category. Then each semantic clique will be refined by merging similar cliques together or deleting the meaningless cliques to eliminate the redundancies. The result will contain only the high-quality cliques extracted from a certain document collection. In existing topic modeling techniques, pre-specified number of topics are required for the topic generation process and irrelevant topics will be generated to fulfill the number of topics to be generated. But in this phase, a fuzzy based clustering mechanism will be applied to cliques and cliques will be clustered to determine the number of representative topics in the collection.

In this research, a concept embedded topic modeling technique will be developed to identify the semantic elements or meaningful terms from collections of documents and ontology driven approach is used to understand and interpret the concepts. Further, fuzzy based automatic clustering mechanism will be introduced to generate the most related topics without considering a pre-specified number of topics.

C. Phase 3: Generate a concept embedded topic modelling

A novel algorithm will be introduced to weight the terms and concepts according to the phase-1 and phase-2 outcomes. A concept layer will be formulated to express the relationship and association among the terms, concepts and topics. The main research finding of this phase is the concept embedded topic modeling approach. The topic model will be generated based on the concepts and concepts associated with the collection of documents. Finally, semantically meaningful topics will be derived based on the concept embedded topic modeling technique and the most relevant topics will be generated. In most of the existing topic models, user need to provide the number of topics beforehand and due to that meaningless and less relevant topics will be generated. Nevertheless, specialty here in this research is, most related set of topics will be generated based on the relevance to the collection of documents instead of focusing the number of topics to be generated.

VI. CONCLUSIONS

LDA can be defined as the most stable probabilistic topic model to extract the topics from collections of documents by considering the frequency and co-occurrences of the words. This study was conducted to analyze the quality of topic models with respect to the quality of the input contents. Accordingly three different datasets (RCV1, conference dataset, and tweets) which contains different complexities in the content were used to evaluate the topic distribution, perplexity, coherence and human perception. The conference dataset performed better than the other two datasets due to the formally written contents which belong to the data mining research domain. The conference dataset was the highest quality and smallest dataset and due to that the perplexity values were lowest and coherence values were highest. Since tweets contain large collection of noisy, messy and poor quality contents, perplexity and coherence were highest and lowest respectively. Topic evaluation parameters of RCV1 dataset were in the middle but more close to the conference dataset due to the similarity of formally written contents. Accordingly, both topic evaluation parameters indicated that the quality of the input content has a strong impact towards the quality of the generated topics. With the evaluation and analysis, it is proved that the noisy and uncertain data leads to misinterpretation of contents. Further, it is noticed that there is a gap between the LDA generated topics and human defined topics. LDA model doesn't consider the semantic meanings of the contents and number of topics need to be specified beforehand the topic modeling process. Therefore, LDA has generated some meaningless and irrelevant topics to fulfil the number of topics. A concept embedded topic modeling approach is proposed to overcome the above two limitations of LDA. The concepts reflect the meanings of the contents and relationships will express the in depth explanations about the concepts. In this research, we develop a concept embedded topic modeling technique that can identify the semantic elements from collections of documents and understand the concepts and relationships using an ontology driven approach. Further, the concepts are categorised and list of topics will be determined by the related terms. Terms and concepts are weighted to interpret the semantic meaning of the contents. A concept layer is incorporated to generate semantically meaningful topics somewhat similar to human perception and understanding.

REFERENCES

[1] J. Evans and P. Aceves, "Machine translation: Mining text for social theory", Annual Review of Sociology, vol. 42, no. 1, pp. 21–50, 2016.

[2] T. Hofmann, "Probabilistic latent semantic analysis", Fifteenth Conference on Uncertainty in Artificial Intelligence, pp. 289–296, 1999.

[3] D. M. Blei, A.Y. Ng, and M. I. Jordan, "Latent dirichlet allocation", Journal of Machine Learning Research, vol. 3, pp. 993–1022, 2003.

[4] D. M. Blei , "Probabilistic topic models", Communications of the ACM,vol. 55, no. 4, pp. 77–84, 2012.

[5] T. Landauer, P. Foltz and D. Laham, "An introduction to latent semantic analysis", Discourse Processes, vol. 25, no. 2, pp. 259–284, 1998.

[6] M. Steyvers and T. Griffiths, "Latent semantic analysis: A road to meaning," T. Landauer, S. D. McNamara, and W. Kintsch, Eds. Laurence Erlbaum, 2007.

[7] D. D. Lewis, Y. Yang, T. G. Rose, and F. Li, "RCV1: A new benchmark collection for text categorization research", The Journal of Machine Learning Research, vol. 5, pp. 361–397, 2004.

[8] D. Kapugama Geeganage,"Concept embedded topic modeling technique", The Web Conference 2018, WWW '18, 2018.

[9] N. Aletras and M. Stevenson, "Evaluating topic coherence using distributional semantics", In Proceedings of the Tenth International Workshop on Computational Semantics (IWCS-10), pp. 13–22, 2013.

[10] S. Bhatia, J. H. Lau and T. Baldwin, " An automatic approach for document-level topic model evaluation", In Proceedings of the 21st Conference on Computational Natural Language Learning , 2017

[11] E. H. Ramirez, R. Brena, D. Magatti, and F. Stella, "Topic Model Validation". Neurocomputing, vol. 76, no. 1, pp. 125–133, 2012.

[12] J. Chang, J. B. Graber, C. Wang, S. Gerrish, and D. M. Blei, "Reading tea leaves: how humans interpret topic ,models", Neural Information Processing Systems, pp. 288–296, 2009.

[13] D. M. Blei and J. D. Lafferty , "Correlated topic models" , Neural Information Processing Systems, 2005.

[14] D. OCallaghan, D. Greene, J. Carthy and P. Cunningham, "An analysis of the coherence of descriptors in topic modeling", Expert Systems with Applications, vol. 42, no. 13, pp. 5645–5657, 2015.

[15] D. Newman, Y. Noh, E. Talley, S. Karimi and T. Baldwin, "Evaluating topic models for digital libraries", Proceedings of the 10th annual joint conference on Digital libraries - JCDL '10, 2010.G.

[16] A. Miller, "WordNet: a lexical database for English", In Communications of the ACM, vol. 38, no. 11, pp. 39–41, 1995.

[17] C. Musat, J. Velcin, S. Trausan-Matu and M. A. Rizoiu, "Improving topic evaluation using conceptual knowledge", 22nd International Joint Conference on Artificial Intelligence, vol.3, pp.1866–1871, 2011.

[18] H. Wallach, I. Murray, R. Salakhutdinov and D. Mimno, "Evaluation methods for topic models", Proceedings of the 26th Annual International Conference on Machine Learning - ICML '09, 2009.

[19] X. Yi and J. Allan, "Evaluating topic models for information retrieval", Proceeding of the 17th ACM conference on Information and knowledge mining - CIKM '08, 2008.

[20] D. Mimno, H. M. Wallach, E. Talley, M. Leenders, and A. McCallum,"Optimizing semantic coherence in topic models", Conference on Empirical Methods in Natural Language Processing , Association for Computational Linguistics, pp. 262–272, 2011.

[21] D. Newman, J. Han Lau, K. Grieser, and T. Baldwin, "Automatic evaluation of topic coherence", In Human Language Technologies: Annual Conference of the North American Chapter of the Association for Computational Linguistics (HLT '10).pp. 100–108 , 2010.

[22] W. Wu, H. Li, H. Wang and K. Q. Zhu, "Probase: A probabilistic taxonomy for text understanding", Special Interest Group on Management of Data Conference (SIGMOD), pp. 481-492. ACM, 2012.