Targeted Readings Recommender using Global Auto-Learning

Muhammad Irfan Malik[1], Muhammad Junaid Majeed[2], Muhammad Taimoor Khan[3]
FAST-NUCES, Peshawar
Pakistan
irfan.rala@gmail.com
junaidmajeed14@gmail.com
taimoor.muhammad@gmail.com

Shehzad Khalid
Bahria University, Islamabad
Pakistan
shehzad_khalid@hotmail.com

ABSTRACT: *Huge volume of content is produced on multiple online sources every day. It is not possible for a user to go through these articles and read about topics of interest. Secondly professional articles, blog and forum have many topics discussed in a single discussion. Therefore, a targeted readings recommender system is proposed that analyze all the documents and discussions to highlight key issues discussed as topics. The topics are extracted with an Automatic knowledge-based topic modeling that allows multiple users to help grow the knowledgebase of the model which benefit new readers as well. On selecting the issues of interest the user is taken to those sections of the articles and discussions that are specifically of interest to the reader in relevance to the issue selected. The application has an ever growing knowledge-base to which every task from every user help the model grow in experience and improve its quality of learning.*

Key words: Recommender Systems, Knowledge Models, Online Content, Learning, Reading

Received:7 March 2016, Revised 14 April 2016, Accepted 23 April 2016

1. Introduction

With the advent of social media platforms like Facebook, Twitter, LinkedIn, Amazon etc. users share their feelings, suggestions and trends with other like-minded users. They support discussions on topics of interest among multiple users arguing on topics through comments. It has made social media platforms a great resource for online content analysis to the benefits of an individual. By mining this information, applications can be developed to benefit society and individuals. Along with content analysis, the meta-information supported by the popular platforms can help to support the analysis. For example, *facebook* allow its users to *share*, *comment* and *like* while *twitter* users are more familiarized with terms like

tweet and *retweet*. *hashtags* and *trends* are also recently introduced to let individual know what are the hot issues. The users of twitter have limitation of 140 characters for their message, due to which it usually miss the context. Therefore, it is very important to have prior knowledge of the events in the background to which the *tweet* is referring. Its compactness or casualness of its users require its content to undergo Natural Language Processing (NLP) at pre-processing [15]. But the briefness of messages has greatly added to the success of *twitter*. It supports more than 40 languages and has 310 Million active users per month. As an individual user, the content relevant to his/her interest are to be focused to recommend targeted readings. Similarly a user may have different interests to read about, depending upon the mood and time.

Topic modeling is the best source for extraction of topics from larger scale text data [22]. Topic models are extended with a variety of approaches including supervised, semi-supervised, hybrid and transfer learning. However, these approaches require some

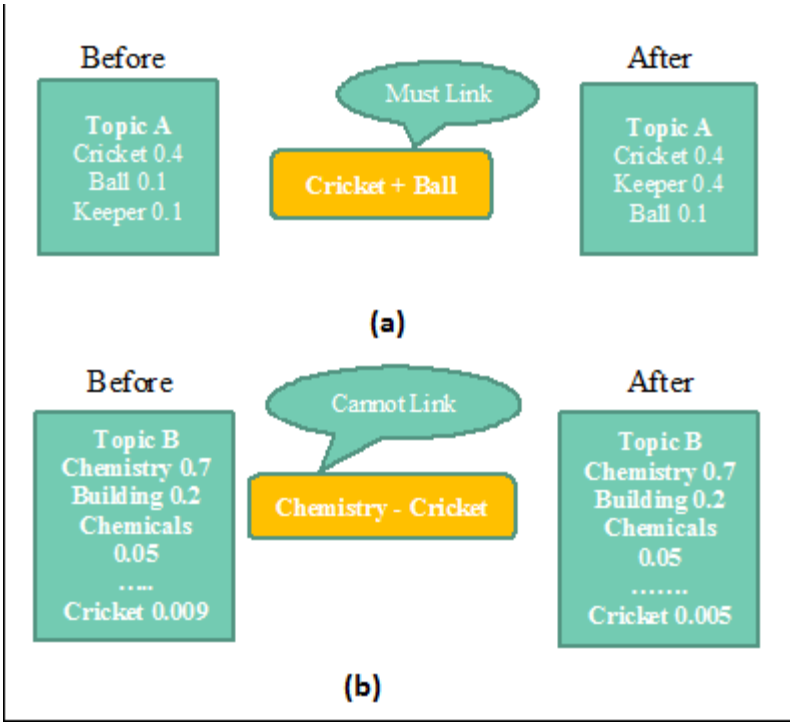

Figure 1. Use of mustlinks and cannotlinks to improve quality of topics

level of domain specific manual intervention from expert and therefore, they cannot be applied to large-scale data consisting of many domains that are not known prior to analysis [19]. Similarly due to the variety of content produced online, its expensive and time-consuming to identify experts for each domain and help them improve the quality of concepts explored from the data as popular topics. For this purpose Automatic knowledge-based topic models are used which are scalable to large-scale data [7]. The model processes one domain after another to provide relevant topics to its users and learns trending rules and patterns for its. This information is retained as knowledge rules. The model produce improved results, represented by the quality of its topics as it matures with the number of tasks. These models benefit from the tasks performed by different users, as they contribute towards growing the knowledge-base of the model. The model learn more number of rules from consistent domains having thousands of documents and incorporate the rules learnt to help improve the quality of topics for domains having fewer documents or noise in the dataset.

Knowledge-based topic models were initially introduced with semi-supervised approach. The knowledge rules were provided to the model by domain expert in the form of mustlinks and cannotlinks. However, when Automatic knowledge-based topic models took over, these models are equipped with an automatic learning mechanism. It empowered these models to eliminate the need of domain experts and learn for itself. The learning of the model improves its performance when higher number of tasks are performed. A must-link state that two words belongs to the same topic or must co-exist e.g. In topic A, *cricket* and *keeper* has

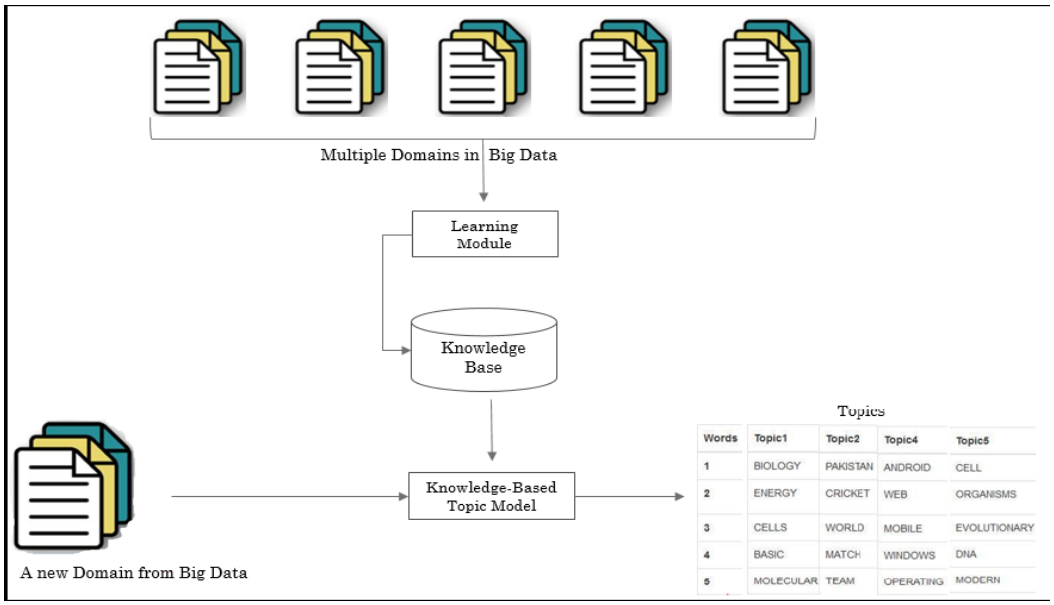

Figure 2. Lifelong learning topic models

a mustlink used by knowledge-based topic model as shown in Fig. 1(a). Where due to the use of mustlink the probability of *keeper* is increase in topic A. Similarly, in topic B, *chemistry* and *cricket* has a cannotlink, then after using the cannotlink, their probability is decreased for the same topic, as shown in Fig. 1(b). The existing knowledge-based topic models [1, 2, 9] make good use of prior domain knowledge to produce coherent topics. The words in a coherent topic hold well together to represent the underlying concept. Automatic knowledge-based topic models exploit the huge volume of content and the variety in its domains, to learn knowledge rules and use these rules to improve the quality of key concepts extracted from within.

The automatic knowledge-based topic models are also known as Lifelong learning models. The model continuously learns and are therefore, also known as human-like learning models. Unlike traditional machine learning techniques, these models carry something from each task known as knowledge rules. It helps the model to behave maturely in new situations by apply previously learnt rules [7, 9]. The rules are applied in the new situation based on their relevance of context. This is to resolve word sense disambiguation, as a word in Natural Language can be used in different senses [10]. Existing Lifelong learning models extract key concepts as topics. By looking into these topics the user can learn about the issues of concern in a domain. Similarly these models are supported with single user, therefore, their learning is a slow process. This is because in order to mature in results the model require to perform a fair number of tasks to grow in experience and build a consistent knowledge-base. But for a single user it is expected to take time. Similarly, in case of a single user the model will only learn in relevance to the tasks performed by that user. The contribution of this research work is to develop a targeted readings recommender application, that uses lifelong learning models which not only provide key concepts but recommend targeted readings as well. The contributions are:

• The model associate each topic from a domain to top document sections for further reading. It helps the user to explore the targeted topics by reading multiple relevant sections from different articles as professional articles and user discussions have many topics discussed.

• The model globally maintains an auto-learning knowledge-base. In order to quickly mature with experience, the tasks performed by many users contribute to learning that helps provide targeted readings to new users as well.

2. Literature Review

Topic models are initially introduced for text analysis and are used in many application areas to high their significance [22]. Latent Dirichlet Allocation (LDA) is basically used to extract topics which are more relevant to users' needs. The topics extracted represent domain features which are usually referred to as aspects in commercial domains [1, 3]. Aspects are the sentiment

targets in Aspect-based Sentiment Analysis. It can help as an efficient data mining technique for unknown domain exploration. The problems related to Natural Language are not addressed by the machine learning techniques and are resolved at pre-processing [10]. In order to improve the quality of topics for specific domains, they are extended with semisupervised, hybrid and transfer learning approaches [11]. The domain specific user intuition helps the model to group words into more contextually co-related topics. The applications of topic models have outgrown from commercial to social, political and management fields [15].

Semi-supervised topic modeling allows topics to incorporate expert guidance through manual annotation. The extracted concepts as topics are restricted to predefined words as seeds while the model populate other contextually related words around them [23]. The number of topics may vary depending upon the requirements of the users. Fewer topics focus on global topics only while more number of topics has the model focus on specialized topics [7]. In semi-supervised topic models, few of the topics are left unlabeled which are targeted to identify the concepts that are missed by the expert. In this way the model shows high accuracy for topics with guidance and can also explore unknown and less frequent concepts. These models are used to improve results for sensitive domains, where high accuracy is require [24]. However, they cannot be extended to large-scale data, where the model is applied to extract topics from many unknown domains. Knowledge-based topic models are initially introduced with a semi-supervised approach, where the quality of topics is related to the quantity and quality of rules provided by the expert [11]. Platform specific meta-information help improve results but they limit the application to only one type of data source.

Knowledge-based topic models (KBTM) are introduced as an extension of topic models that make use of relevant domain knowledge [3]. These models are different from transfer learning models as there is no single target domain and they are not known prior to analysis. There are different knowledge-based topic models that use mustlink type of knowledge, cannotlink type of knowledge or both [1, 3, 4]. Automatic learning models face the problem of learning wrong rules as well, discussed in [3, 7]. Wrong rules lead the distribution of topic models in the wrong direction which results in reducing the quality of topics in terms of topic coherence. It is efficient as it uses less resource utilization and high performance. The accuracy and performance of the model is strongly dependent on the learning mechanism of the model [11]. Automatic knowledgebased topic models exploit the huge volume of data and the variety of its domains and benefit from it to learn more consistent rules. Different types of approaches are used for learning mustlinks and cannotlinks in these models [11]. For word sense disambiguation the context of knowledge rule is stored to ensure that the rule is used in the same context in which it was learnt.

The automatic knowledge-based topic models require good number of tasks, depending upon the complexity of domain, to start producing good quality topics. Working of automatic knowledge-based topic model or Lifelong learning model is shown in Fig. 2. For commercial domains after processing 15 to 20 domains the model has a consistent knowledge-base developed with which it improved the results of other domains about commercial products [7]. This is due to high overlap as common aspects, in commercial products. Similarly the topics extracted are presented as key concepts from the given data, they are not associated to their supporting documents. Therefore, a learning based application is proposed that uses global learning, to generate high quality topics while associate those topics to their respective content sections in documents that could be suggested readings.

Algorithm 1 proposedModel(D^i, V, N, Knowledge-base)

1: **procedure** PROPOSEDMODEL

2: *topics* ← TopicModeling(D^i, V, N, Knowledge-base)

3: *candidateRules* ← LearningModule(topics, D^i, V)

4: *rules* ← evaluateRules(candidateRules)

5: *keyIssues* ← results(topics)

6: *readings* ← suggestReadings($keyIssues_j$)

7: *targetSections* ← suggestReadings(readings)

8: **end procedure**

3. Proposed Model

The proposed model uses topic model at its core. However, in order to produce topics with higher accuracy as more coherently

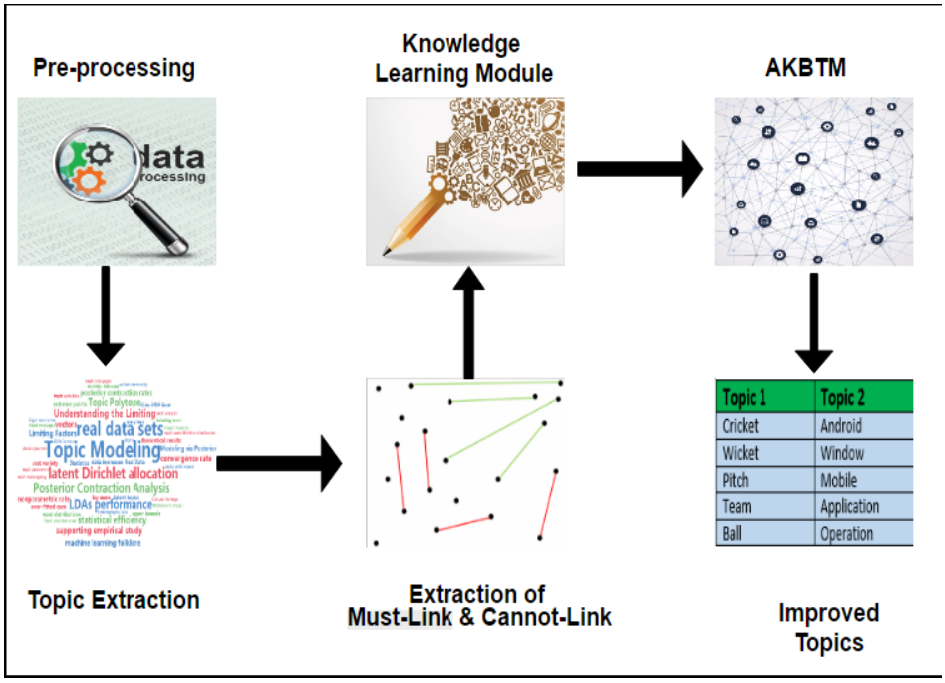

Figure 3. Workflow of the proposed readings recommender system

representing a key concept in the given domain, it uses Lifelong learning approach. With lifelong learning the model maintains it own knowledge-base and improve the quality of topics without any user intervention. Since lifelong learning models require experience to get mature in results, the global learning approach has enabled the model to facilitate multiple users where they benefit from the learning of each Fig. 3. Workflow of the proposed readings recommender system others' tasks. The model provide the user with more coherent concepts in the given domain area. However, if the user is interested to read further, the model suggests targeted readings from each concept. This approach allows the users to be very specific about reading the content of their interest only. Such target content may exist as part of multiple articles and discussions but are grouped together as reading suggestions. Working of the application is presented in Algorithm 4. The model consists of 6 major steps as shown in the Fig. 3. The steps are explained below.

3.1 preparing data

In order to recommend readings to the user, the data is to be prepared so that topic modeling can be efficiently used. Since topic modeling is a machine learning technique, it doesn't consider linguistic problems which are to be addressed at preprocessing. In order to prepare the data, since the readings recommender system only support English language, therefore, a language identification module helps in allowing the content in English language only. The unwanted words from documents i.e. is, am, are, the which do not contribute towards any concept are filtered at preprocessing. Similarly, the symbols not contributing like , #, $, %, , etc. are ignored as well. Lemmatization is applied to shorten words to their stems.

3.2 Topic Extraction

Since the domains are unknown and the expert guidance is not available, unsupervised topic modeling is applied to multiple domains. It helps the model to provide initial distribution and topics for each domain [22]. The unsupervised topics are used as initial intuition to learn from. Each domain has a vast vocabulary and making all possible combinations is an exhaustive process. Therefore the topics from unsupervised topic modeling helps to provide candidate rules.

3.3 Extraction of rules

In third step, the candidate rules are extracted by using the topic model distribution. They consists of a pair of words. They are considered candidate rules, consisting of both mustlinks and cannotlinks. Mustlinks word pairs are generated from the top words in the same topic while the cannotlink pairs are generated from top words in different topics. However, not all candidate rules can be learnt and therefore, they have to undergo an evaluation criteria. The evaluation criteria associate a score to each

candidate rule. It associates a confidence score to rules. The confidence in a rule is a measure for the quality of rule.

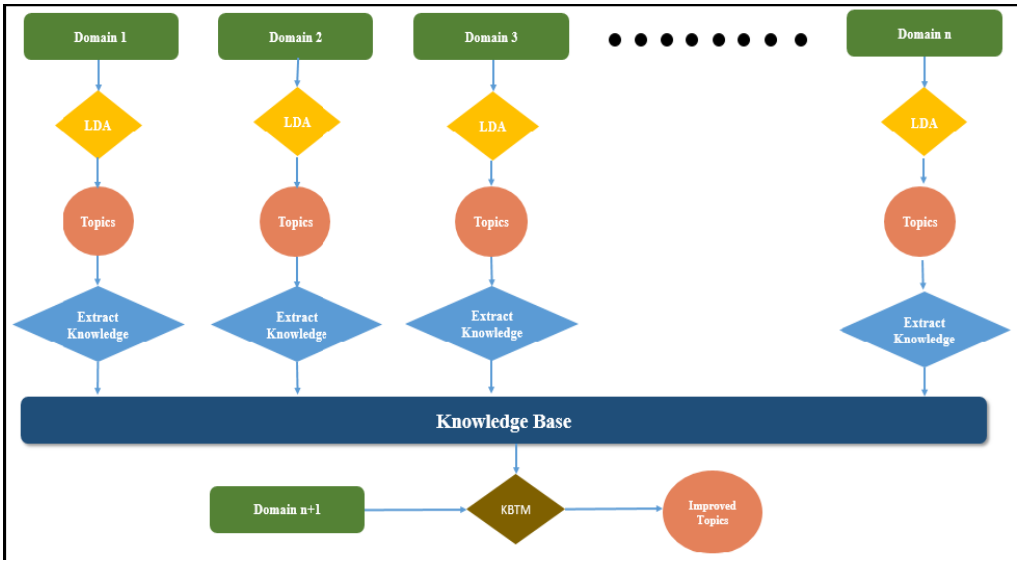

Figure 4. Working of the Automatic knowledge-based topic model used for readings recommender

3.4 Knowledge Learning Module

In order to select good quality rules from all candidate rules, a pair of thresholds are applied on their confidence scores. The candidate rule with confidence above the threshold are learnt while others are ignored. The threshold for mustlinks is usually kept lower than that cannotlinks because mustlinks are more effective as compared to cannotlinks. While cannotlinks may also be confused with no co-relation in specific domains and require more care to harvest.

topic0	topic1	topic2	topic3	topic4
Cricket	Wicket	building	runs	fitness
Team	out	area	target	fielding
Pitch	catch	class	score	athletic
Weather	LBW	boundary	century	running
Stadium	lost	land	runrate	timing

Table 1. Key Concepts As Topics, In a Domain

3.5 Apply the knowledge

The rules learnt are added to the knowledge-base. Whenever a contextually relevant scenario appears, the model incorporate the rules from the knowledge-base and use it to resolve the issue in hand. It has a mechanism of adding the bias of the rules in the topic modeling results. By applying the knowledge, the quality of topics is improved in terms of topic coherence. The model uses Gibbs sampling as an inference technique which repeatedly sample words and change their topics by incorporating rules until the words gets stabilized. The Automatic knowledge-based topic model is shown in Fig. 4. The model trained from n domains can process n + 1 domain by utilizing all the knowledge from its past experience.

3.6 Visualization

An important part of the application is to highlight key concepts to the users, where the user could pick a concept and the system suggest top sections from different documents as targeted readings. Similarly the rules used to produce those results can also be viewed by the user.

mustlinks	
cricket	team
fitness	fielding
runs	score
fielding	athletic
cannotlinks	
out	runs
fitness	pitch
weather	boundary
LBW	stadium

Table 2. Mustlinks And Cannotlinks

4. Experiment And Results

The experiment is carried out on 20 different domains with 200 reviews of each domain and apply Knowledge-based topic modeling to improve topic result. By maintaining a knowledge base it suggests much targeted readings effectively, as shown in Table I. The rules used to generate these topics are shown in Table 2. The rules helped improve the quality of topics for less common domains having as few as 200 review documents. They are not enough for topic modeling which usually need documents in thousands. The extracted topics are presented to the user as key concepts. In the application developed, the user can select a topic to explore further about it where the application suggest relevant readings. With a global learning mechanism, the model can incorporate knowledge from tasks of other users to improve the results for user in current task.

5. Conclusion

The readings about an issue has many other issues discussed in relation to them. Professional authors use different background scenarios to lay a plot while discussion of forums and blogs go astray of the main topic. Therefore, a targeted readings recommender system is proposed that takes the user only to the specific sections of multiple articles and blogs that have addressed the issue of interest to the author. However, in order to have the issues expressed more compactly, automatic knowledge-based topic models are used. As a remote application it serves multiple users and maintains a global knowledge base. Thus, users using the application are helping future users to have the model grow in experience and knowledge, in order to serve them better the users in future tasks.

References

[1] Andrzejewski, D., Zhu, X., Craven., M., (2009). Incorporating domain knowledge into topic modeling via Dirichlet Forest priors, *In*: ICML, p. 25-32.

[2] Andrzejewski, D., Zhu, X., Craven, M., Recht., B., (2011). A framework for incorporating general domain knowledge into latent Dirichlet allocation using first-order logic, *In:* IJCAI, p. 1171-1177.

[3] Chen, Z., Mukherjee, A., Liu, B., Hsu, M., Castellanos, M., Ghosh., R., (2013) Discovering Coherent Topics Using General Knowledge, *In*: CIKM, p. 209-218.

[4] Chen, Z., Mukherjee, A., Liu, B., Hsu, M., Castellanos, M., Ghosh., R., (2013) Exploiting Domain Knowledge in Aspect Extraction, *In*: EMNLP, p. 1655-1667.

[5] Blei D. M., McAuliffe., J. D., (2009) Supervised Topic Models, *In:* NIPS, p. 121-128.

[6] D. Ramage, D., Hall, R. ,Nallapati, C. D. Manning., Labeled LDA: a supervised topic model for credit attribution in multi-labeled corpora, *In:* EMNLP, p. 248-256.

[7] Khan, Muhammad Taimoor, et al. (2016). Online Knowledge- Based Model for Big Data Topic Extraction, Computational intelligence and neuroscience 2016.

[8] Lops, Pasquale., Gemmis, Marco De., Semeraro, Giovanni. (2011). Content-based recommender systems: State of the art and trends, Recommender systems handbook. Springer US, 2011. 73-105.

[9] Khan, M., Taimoor, Durrani M., Khalid S., Aziz, F. (2016). Lifelong aspect extraction from big data: knowledge engineering, Complex and Adaptive Systems Modeling 4 (1).1.

[10] Khan M.T., Durrani M., Ali A., Inayat I., Khalid S., Khan K. (2016) Sentiment Analysis and the complex Natural Language, Complex and Adaptive Systems Modeling, 4 (1).

[11] Khan M.T., Durrani M., Ali A., Khan K., Khalid S. (2015) Aspect-based Sentiment Analysis on Large-scale data: Topic Models are the Preferred Solution, BUJICT, 8 (2).

[12] Xue, Gui-Rong, et al. (2008). Topic-bridged PLSA for crossdomain text classification, *In:* Proceedings of the 31st annual international ACM SIGIR Conference on Research and development in Information Retrieval. ACM.

[13] Chen, Zhiyuan., Liu, Bing. (2014). Mining topics in documents: standing on the shoulders of big data, *In:*Proceedings of the 20th ACM SIGKDD International Conference on Knowledge discovery and data mining. ACM, 2014.

[14] Mimno, D. . Wallach, H. M Talley, E. . Leenders, M. McCallum, A(2011) Optimizing semantic coherence in topic models, In EMNLP, p. 262-272.

[15] Khan M.T., Khalid S. (2015). Sentiment Analysis for health care, *International Journal of Privacy and Health Information Systems*, 3 (2) 80-94.

[16] Agarwal, Apoorv, et al (2011). Sentiment analysis of twitter data, *In:*Proceedings of the workshop on languages in Social media. Association for Computational Linguistics.

[17] Kumar, Akshi., Sebastian, Teeja Mary.(2012). Sentiment analysis on twitter, *IJCSI International Journal of Computer Science* Issues 9 (4) 372-373.

[18] Grubmller, Verena., Gtsch, Katharina., Krieger, Bernhard . (2013), Social media analytics for future oriented policy making, *European Journal of Futures Research* 1 (1) 1-9.

[19] Khan M.T., Durrani M., Ali A., Khan K., Khalid S., (2015) Aspect-based Sentiment Analysis on Large-scale data: Topic Models are the Preferred Solution, BUJICT, 8 (2).

[20] Khan M.T., Khalid S. (2015). Sentiment Analysis for health care, *International Journal of Privacy and Health Information Systems*, 3 (2) 80-94.

[21] Mohammad, Zuber, (2014). A Survey of Data Mining Techniques for Social Network Analysis, *International Journal of Research in Computer Engineering and Electronics* 3.6, 2014.

[22] Blei, David., M., Ng, Andrew Y., Jordan, Michael I. (2003). Latent dirichlet allocation, *The Journal of Machine Learning Research* 3. 993-1022.

[23] Mukherjee, Arjun., Liu, Bing (2012). Aspect extraction through semi-supervised modeling, *In:*Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Vol. 1. Association for Computational Linguistics.

[24] Bing, Xiang., Zhou, Liang., Reuters, Thomson. (2014). Improving Twitter Sentiment Analysis with Topic-Based Mixture Modeling and Semi-Supervised Training, ACL (2).