ISSN 2348 – 9928
IJAICT Volume 3, Issue 11, March 2017 Doi:01.0401/ijaict.2017.03.10 Published on 05 (04) 2017

TOPIC MODELING IN CLINICAL REPORTS - A SURVEY

Ms. S. Ponmalar
UG Scholar

Ms. D. Ponnarasi
UG Scholar

Ms. A. Sangeetha
UG Scholar

Prof. R. Kingsy Grace
Associate Professor

Department of Computer Science and Engineering
Sri Ramakrishna Engineering College
Coimbatore, Tamilnadu, India

Abstract* — *Text mining is a process of converting unstructured data into meaningful data. It may be loosely characterized as the process of analyzing text to extract information that is useful for particular purposes. Topic modeling is a form of text mining, a way of identifying patterns in a corpus. The topics produced by topic modeling techniques are clusters of similar words that are frequently occur together. Topic modeling is also a frequently used text-mining tool for discovery of hidden semantic structures in a text body. Intuitively, a document is about a particular topic, one would expect particular words to appear in the document more or less frequently. This paper, presents a survey on topic modeling in clinical documents.

Keywords* — *Topic Modeling, Text Mining, Clinical Documents, Prediction, Data Mining.

I. INTRODUCTION

Data mining [1, 12] refers to extracting useful information from vast amounts of data. The other terms which are used to name data mining are knowledge mining from databases, knowledge extraction, analyzing data, and data archaeology. Nowadays, it is commonly agreed that data mining is an essential process of Knowledge Discovery in Databases, or KDD. Data mining is the process of discovering knowledge from large amount of data stored either in databases, data warehouses, or other information repositories. Data mining is used in many different sectors of business to both predicting and discovering trends. It is a proactive solution for business looking to gain a competitive edge. In the past, it is only able to analyze what a customers or clients HAD DONE, but now, with the help of Data Mining, one can predict what client WILL DO. With Data Mining, companies can make better and more effective business decisions such as advertising, marketing, etc decisions that will help these companies grow. In finance and banking, data mining is used to create accurate risk models for loans and mortgages. It is also helpful for detecting fraudulent transactions. Data mining techniques are used to improve conversions, increased customer satisfaction and created targeted advertising campaigns in marketing. This is done by looking at historical sales and customer data and creating powerful prediction models.

Text mining [11, 12] is also referred as text data mining, roughly equivalent to text analytics, is the process of deriving high-quality information. High-quality information is typically derived through the devise of patterns and trends through means such as statistical pattern learning. Text mining usually involves the process of structuring the given input text, deriving patterns within the structured data, evaluation and interpretation of the output. 'High quality' in text mining usually refers to some combination of application, novelty, and interestingness.

Typical text mining tasks includes text categorization, text clustering, and concept extraction, production of granular taxonomy, document summarization, and entity relation modeling. Text analysis involves information recovery, word frequency analysis, and pattern recognition and information extraction. Text mining techniques such as association analysis, visualization, and predictive analytics are used to turn text into data for analysis, via application of Natural Language Processing (NLP) and analytical methods.

The field of text mining seeks to extract useful information from unstructured textual data all the way through the identification and exploration of interesting patterns. The techniques employed usually do not involve deep analysis or parsing, but rely on simple “bag-of-words” text representations based on vector space. Usually text mining helps an organization to derive valuable business insight from text-based content for example word documents, emails and postings on social media stream like Facebook, Twitter and LinkedIn. Mining unstructured data with Natural Language Processing (NLP), numerical modeling and machine learning techniques can be challenging, however, because natural language text is

often inconsistent. It contains ambiguity caused by inconsistent syntax and semantics, as well as slang, language specific to vertical industries and age groups, double entendres and sarcasm. Vast amounts of new records and data are generated each day through economic, academic and social activities with large potential economic and societal value. Techniques such as text and data mining and analytics are required to exploit large amount of data [12].

Topic Modeling [4, 7, 9, 10] provides a convenient way to analyze big unclassified text. A topic contains a cluster of words that frequently occurs together. A topic modeling can connect words with similar meaning and distinguish between uses of words with multiple meaning. Topic modeling describes a topic as a recurring pattern of co occurring words. Text analytics gives a lot of remuneration to business, governments and the individual. However, privacy, security, and misuse of information are the big problems if they are not addressed and resolved properly. Topic model provides a mean for indexing large and unstructured corpora with inferred semantics, but incorporating these methods in clinical text has begun recently.

While these techniques have yielded promise results, they have generally been limited to basic methods that have not incorporated more recent advance in the topic modeling field. Development of new topic modeling methods that incorporate various clinical information and structure have the potential to unlock the information contained in clinical reports for use in developing clinical tools. Topic modeling in clinical reports where the less studied is a problem of analyzing the text as a whole to create temporal indices that capture relationship between learned the clinical events. This paper deals with a survey on topic modeling used on clinical documents. The rest of the paper is organized as follows: Section II presents different topic modeling techniques in the literature. The comparison of different topic modeling techniques is discussed in Section III. Section IV concludes the paper.

II. TOPIC MODELING TECHNIQUES

Williams Speier, et al. [2] have proposed a topic modeling based automatic summarization system for clinical documents. The proposed model is based on informative prior probabilities. The clinical reports are represented by two forms such as i). Chained n- grams ii). Dirichlet hype parameter. The topic modeling proposed in this paper used Latent Dirichlet Allocation models (LDA) for testing huge amount of clinical documents. Clinical document such as glioblastoma multiforme (GBM), lung cancer and acute ischemic stroke affected patients were collected from disease –coded research database which was approved by Institutional Review Board (IRB). Around 936 patients medical reports were collected. The total counts of medical report were 84,201. The proposed model is evaluated using Empirical likelihood method which was proposed by Wei and McCallum [15] .The proposed model Dirichlet parameter is used for tailoring the words which are frequently occurring in the clinical document. The proposed model classifies the clinical documents based on disease and document author. The achievement of disease classification accuracy is 99.8% and the accuracy for document author is 83.2%. The proposed model is combined with Metadata and Phrase driven Topic model (MPT), Dirichlet Multinomial Regression (DMR) and Topical N-Grams (TNG) is proved to be better for both disease and author classification.

Philip Resnik, et al. [3] have discussed a supervised topic modeling approach for Twitter depression related data. The proposed model is based on language analysis method and it includes words and n-grams and manually defined word categories. Supervised Latent Dirichlet Allocation (LDA) and supervised anchor topic modeling techniques are used in this proposed model to analyze linguistic signal for detecting depression. The dataset is collected from Twitter collection which contains 3 million tweets from about 2,000 twitter users of which around 600 users are clinically diagnosed with depression. The proposed model uses three features such as LDA, Supervised Nested Latent Dirichlet Allocation (SLDA), and Supervised Anchor (SANCHOR) LDA. The proposed model provides the automatic identification of depression in a sophisticated manner. The weekly grouping provided by SANCHOR improved precision by R=0.5 to 74% and precision by R=0.75 to 62%.

Rubayyi Alghamdi and Khalid Alfalqi [4] have proposed topic modeling in text mining. The proposed model contains two categories. First one is methods of topic modeling which contains four methods such as Latent Semantic Analysis (LSA), Probabilistic Latent Semantic Analysis (PLSA), Correlated Topic Model (CTM), and Latent Dirichlet Allocation (LDA). The second is Topic Evolution Model (TEM) which uses the models such as Dynamic Topic Models (DTM), Multiscale Topic Tomography, Topic Over Time (TOT), Detecting Topic Evolution in scientific literatures and Dynamic Topic

Correlation Detection models to find the time. Latent Semantic Analysis (LSA) is a method of Natural Language Processing (NLP) and is mainly used to compute similarity between texts. Probabilistic Latent Semantic Analysis (PLSA) is a method that used to automate document indexing and also used to identify different contexts of word usage. Latent Dirichlet Allocation is used for huge number of electronic document collections and it generates the document on the given topic. Correlated Topic Model (CTM) is used to find the topics in a group of documents. The proposed model explains the difference in terms of limitations and characteristics between LSA, PLSA, LDA and CTM.

Sungrae Park, et al. have proposed an associative topic modeling which finds the cluster of numerical time-series text data. Associative Topic Model (ATM) is used to predict the higher level of accuracy of numerical time series of data. Associative probabilistic topic model gives the prior knowledge of word distribution and it generates words in a corpus of documents. The topic modeling method Latent Dirichlet Allocation is used the Bayesian approach to modeling of generation process. The proposed model integrates text and numerical data at a time and identifies the representation of events which are not directly recognizable from the corpus and the numbers. The proposed model introduces an Associative Topic Model which integrates time-series data of texts and numerical values. The proposed model applies ATM to a financial news corpus and stock indexes. For the financial news corpus, the articles were collected from Bloomberg. The 60,500 articles were selected and from which 500 articles were selected for every week. For the stock indexes, datasets were collected from DJIA. The proposed model is applicable to product reviews and sales records over time.

Jian Tang, et al have [6] proposed the limiting factors of topic modeling in posterior contraction analysis. The machine learning tool Latent Dirichlet Allocation (LDA) is used for topic modeling. The proposed model provides a semantic analysis of LDA in various things and it also identifies limiting factors in performance analysis of LDA. The proposed model validates two types of datasets such as synthetic and real world dataset and the datasets were collected from Wikipedia articles, news articles from New York Times (NYT) and short messages from Twitter and these datasets are validated using minimum matching Euclidean algorithm. The proposed model is based on mild geometric assumptions and is independent of inference algorithms and datasets were used for different evaluation measures.

Corey Arnold, et al [7] have proposed topic model tailored to the clinical reporting environment. The proposed model is able to identifying predefined concepts from clinical documents. The proposed model is based on Latent Dirichlet Allocation (LDA) for topic modeling and that provides a method for indexing large unstructured corpora and it identifies the use of topic model in a patient's clinical document such as glioblastoma multiforme (GBM), an aggressive brain cancer. The temporal patterns were reviewed by a neuroradiologist and were found to associate with valid sequences of clinical events. The proposed model is able to identifying patients based on patterns of temporal topic and it is also used for case based reasoning.

Jordan Boyd-Graber and David M. Blei [8] have proposed multilingual topic models. The proposed model uses topic modeling method for unaligned text and is also use Latent Dirichlet Allocation for topic modeling. Topic modeling is a powerful technique for unsupervised analysis of corpora. The proposed model collects data and incorporates prior information to find the exact matches. The performance of multilingual topic modeling is examined based on three criteria. First is the qualitative logic of learned topics and the second is the accuracy of the learned matching. The proposed model able is distinguish the consequence of the learned matching from the information already available through the matching prior and it improves the quality of translation.

Chong Wang and David M. Blei [9] have proposed collaborative topic modeling for scientific articles. The proposed model incorporates collaborate filtering and probabilistic topic modeling and if forms recommendation about both existing and newly published articles. This model helps to access scientific information quickly but it limits researchers to particular reference communities. The proposed model describes two types of recommendation problems such as i). Recommendation task and ii). Recommendation by matrix factorization. To combine collaborative filtering with topic modeling, Collaborative Topic Regression (CTR) technique is used. CTR explains the user latent space using the topics learned from the data. Datasets are users and their libraries of articles obtained from CiteULike and removed empty articles, merged duplicated articles, and removed users with fewer than

10 articles to obtain a data set of 5, 551 users and 16,980 articles with 204, 986 observed user-item pairs. The proposed

Table 1 : Comparison of topic modeling techniques

S.NO	*TECHNIQUE*	*TOOL USED*	*PERFORMANCE PARAMETERS*	*MERITS/DEMERITS*
1.	*Topic modeling in clinical reports using LDA [2]*	*Latent Dirichlet Allocation models (LDA)*	*accuracy*	*Handling large amount of dataset*
2.	*Supervised topic modeling in Twitter [3]*	*Supervised Nested Latent Dirichlet Allocation (SLDA), and Supervised Anchor (SANCHOR) LDA*	*precision*	*Easier than LDA*
3.	*Topic modeling in Text mining [4]*	*Latent Semantic Analysis(LSA)*	*information retrieval tasking*	*Improved data access*
4.	*Associative topic models with numerical time series [5]*	*Associative Topic Model (ATM)*	*accuracy*	*Increases higher level of accuracy*
5.	*Limiting factors of topic modeling [6]*	*Latent Dirichlet Allocation models (LDA)*	*accuracy*	*Increase performance*
6.	*Topic models of clinical reports [7]*	*Latent Dirichlet Allocation models (LDA)*	*accuracy*	*Convenient method for large unstructured data*
7.	*Multilingual Topic Models for Unaligned Text [8]*	*Latent Dirichlet Allocation models (LDA)*	*quality*	*Improves quality of translation*
8.	*Collaborative Topic Modeling [9]*	*Collaborative Topic Regression (CTR)*	*accuracy*	*Useful for large set of data*
9.	*A Topic Modeling Toolbox Using Belief propagation [10]*	*Variational Bayes (VB) and collapsed Gibbs Sampling (GS)*	*accuracy*	*Improves performance*

Jia Zeng [10] has proposed topic modeling tool box based on belief propagation. The proposed model uses Bayesian model for probabilistic topic modeling. Topic modeling toolbox is implemented by MEC C++ in Matlab. This model uses Latent Dirichlet Allocation for solving topic modeling and it also uses two approximate inference methods such as Variational Bayes (VB) and collapsed Gibbs Sampling (GS). This model developed LDA topic modeling for Author Topic Models (ATM) and Relational Topic Models (RTM).

III. SUMMARY

The comparison of various topic modeling techniques are done based on the aspects such as technique, algorithm used, performance parameters and merits/demerits. The detailed comparison is shown in Table 1.

IV. CONCLUSION

Incorporating patient and document metadata, as well as capturing expressions in clinical text, enormously enhances the topic representation of clinical reports. Topical n-grams are combined into the model catches common anatomical concepts, tests, and ailments while additionally separating words that are equivocal in a pack of words model. Including patient and document level information makes a more instructive earlier on the topics in a document, bringing about points that better speak to the contained text. This paper presents a detailed survey on the literature having different topic modeling techniques and its merits/demerits. Most of the work is the survey uses LDA and is preferred for topic modeling. Most of the proposed model checks the accuracy of the prediction using topic modeling. Our future work incorporates using the topics found in this study to drive a web application that naturally outlines a patient's medical records, including concept, source, and time situated perspectives.

References

[1] Yihao Li, "Data Mining: Concepts, Background And Methods Of Integrating Uncertainty In Data Mining", Vol. 6, No. 1, 2009.

[2] William Speier, Michael K. Ong, Corey W. Arnold, "Using phrases and document metadata to improve topic modeling of clinical reports", Journal of Biomedical Informatics , Vol . 61, PP. 260–266, 2016.

[3] Philip Resnik, William Armstrong, Leonardo Claudino, Thang Nguyen, Viet-An Nguyen, and Jordan Boyd-Graber, "Beyond LDA: Exploring Supervised Topic Modeling for Depression- Related Language in Twitter", Resnik et al 2015.

[4] Rubayyi Alghamdi, Khalid Alfalqi, "A Survey of Topic Modeling in Text Mining", (IJACSA) International Journal of Advanced Computer Science And Applications, Vol. 6, No. 1, 2015.

[5] Sungrae Park, Wonsung Lee, Il-Chul Moon, "Associative topic models with numerical time series", Vol.51, PP.737-755, 2015.

[6] Jian Tang, Zhaoshi Meng, Xuan Long Nguyen, QiaozhuMei, MingZhang, "Understanding the Limiting Factors of Topic Modeling via Posterior Contraction Analysis", Proceedings of the 31st International Conference on Machine Learning,Vol.32,2014.

[7] Corey Arnold and William Speier, "A Topic Model of Clinical Reports", August 12-16,2012.

[8] Jordan Boyd-Graber, David M. Blei, "Multilingual Topic Models for Unaligned Text", Vol. 6, No. 1, 2009.

[9] http://observationbaltimore.com/blog/2011/09/data-mining %E2%80%94why-is-it-important/.

[10] C.Uma, S.Krithika, C.Kalaivani, "A Survey Paper on Text Mining Techniques", International Journal of Engineering Trends and Technology (IJETT) – Volume-40 Number-4 - October 2016.

[11] https://books.google.co.in/books?isbn=3319091441.

[12] L. Wei, A. McCallum, Pachinko allocation: DAG-structured mixture models of topic correlations, ICML '06 Proc. 23rd Int. Conf. Mach. Learn. (2006) 577–584, http://dx.doi.org/10.1145/1143844.1143917.