IJCSE International Journal of Computer Sciences and Engineering Open Access
Survey Paper Vol.-6, Issue-5, May 2018 E-ISSN: 2347-2693

A Survey on Advanced Algorithms in Topic Modeling

Padmaja Ch V R[1*], Lakshmi Narayana S[2], Divakar Ch[3]

[1*]Dept. of CSE, Raghu Engineering College, Visakhapatnam, AP, India
[2] Former Principal Scientist, NIO, Visakhapatnam, AP, India
[3] Dept. of IT, SRKR Engineering College, Bhimavaram, AP, India

Corresponding Author: cvrpadmaja@gmail.com, Tel.: +91-9989346633

Available online at: www.ijcseonline.org

Accepted: 16/May/2018, Published: 31/May/2018

***Abstract*—** In this paper, Survey of various topic modeling algorithms is presented. Introduced classification differs from earlier efforts, providing a complementary view of the field. This survey provides a brief overview of the existing probabilistic topic models and gives motivation for future research.

***Keywords*—**Topic modeling, pLSI, LDA, Dynamical Topic Model, Supervised LDA.

I. INTRODUCTION

The latent topical structures in data can be discovered using probabilistic topic models, which are a collection of machine learning algorithms. Although many applications are found in various data mining areas such as image annotation, audio and video analysis, they are primarily invented for use in finding topics in textual data. Profiling and modeling knowledge from scientific papers is one area of research that benefits most.

Few surveys of topic models already exist; among most significant are [1], [2] and [3]. Among three, the first one [1] is more prominent in topic modeling survey and gives a classification of directed probabilistic topic models and a border view on graphical models. The other two papers [2] and [3] gives a reasonable overview and summary of the topic modeling domain. In paper [4] discusses the classification of probabilistic topic modeling algorithms

Main criterion of classification in [1] is functionality, and models are presented in a chronological order in a systematic evolution-based fashion. This is not the purpose of this survey; functionality is not of a primary interest for us. Criteria of presented classification are chosen as to highlight fundamental approaches and assumptions used in topic modeling. Also, [1] focuses on directed probabilistic topic models while we impose no such restriction. Introducing general ideas and formal definition has also been done in [2] and [3] so this is not our primary goal either. In [1] models are also classified according to their original problem domain.

As many of those problems, such as topic discovery, topic evolution, document classification and many others, present a subproblem to the modeling and profiling of the knowledge from scientific papers, such distinction is not made in this survey. Survey of topic models is presented with emphasis on different approaches used.

II. CLASSIFICATION

Topic models are classified according to three orthogonal criteria. First criterion is based on word ordering and a document representation. Two distinct approaches are possible. The first one is bag-of-words, which is very simple one. In this approach, the word ordering is ignored while representing the document. This makes us to focus on semantic structure rather than modeling the word order dependencies. Other approach, that doesn't neglect word ordering will be referred to as a sequence of words.

Although first approach is appreciated for its simplicity and is often sufficient, second approach bear more information which can supposedly lead to better results in some problem domains.

Second criterion is taking external knowledge into consideration. First approach where no such knowledge is provided is simpler and for many purposes sufficient. Second approach is based on using in-domain knowledge for the target problem, yielding more specific and human interpretable topics.

Third and final criterion is dependability on labeled data. Main idea behind topic models is unsupervised clustering of topics which renders them applicable to a broad range of real life problems where there are no data labels and cannot be provided.

Most of the topic models are fully unsupervised. Some models can be used in a supervised or semi-supervised manner to be applicable to classification tasks or simply to yield better results if labeled data for domain is already present. The classification tree is shown in Figure 1.

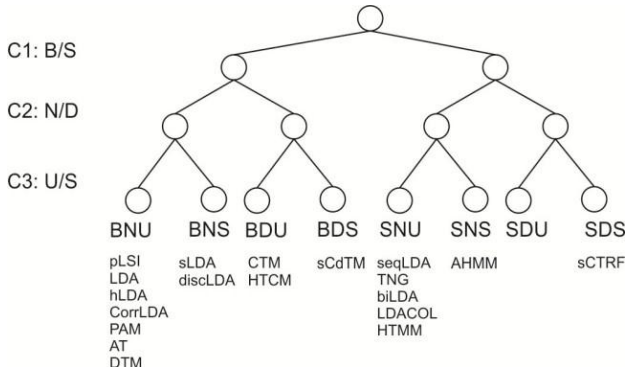

Figure 1. The classification three of probabilistic topic models. **Legend**: B/S – bag of words vs sequence of words; N/D – no in-domain vs in-domain; U/S – unsupervised vs supervised. **Description:** The classification three obtained by successive application of the chosen criteria. **Implication:** The class of unsupervised sequence of word models with in-domain knowledge requirements have no known implementations.

III. EXISTING METHODS

For each class defined in previous section most prominent examples are presented, if such solutions exist.

A. Unsupervised bag of words topic models with no in-domain knowledge requirements

This class of models reside on word exchangeability assumption, i.e. discards information on word position within documents. Such models are often used regarding problems such as information retrieval, document clustering and summarization due to their simplicity introduced with bag of words approach and greater real-world problem applicability based on their unsupervised nature. Most of the probabilistic topic models, including the earliest ones, fall into this category. There are numerous extensions from the baseline approach (Latent Dirichlet Allocation) that introduce additional abilities beside modeling word-topic and document-topic distributions, some of which are presented here.

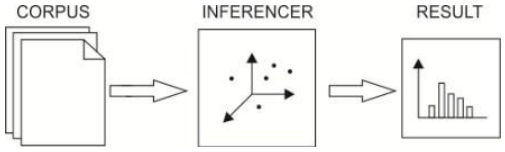

Figure 2. Outline of the unsupervised bag of words topic model with no in-domain knowledge requirements. **Description**: Appropriate inference equations stipulated by the particular model are applied to the textual corpus after tokenizing and pre-processing. Word ordering is neglected.

A.1. Probabilistic Latent Semantic Indexing

T. Hoffman invented Probabilistic Latent Semantic Indexing (pLSI) in 1999 as variant of Latent Semantic Analysis [5]. It has a sound statistical basis and describes a proper generative data model.

pLSI is a generative model, in which each word of the documents is sampled from multinomial distributions that can be taken as topics. The proportions corresponding to mixture weights are sampled from a separate multinomial distribution for each document in the corpus.

Based on generative model, an inference algorithm is defined as a method for inferring topic-word distributions, as well as document-topic distributions, from textual corpora.

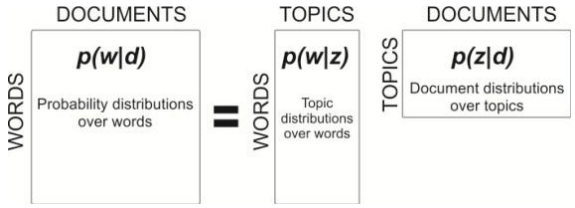

Figure 3. Probabilistic Latent Semantic Indexing viewed as a matrix factorization

There are several methods for computing word-topic and topic-document distributions, one widely accepted is Expectation Maximization algorithm. Equations for E and M steps are inferred directly from the generative model (Fig. 4).

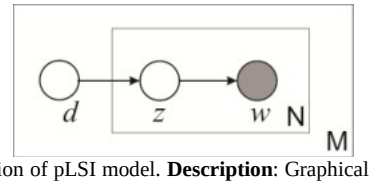

Figure 4. Plate notation of pLSI model. **Description**: Graphical presentation of a Bayesian network corresponding to pLSI model. For interpretation of this as well as other graphical models presented in the survey, reader is encouraged to read [1] and [2].

pLSI efficiently resolve several issues of Latent Semantic Analysis (LSA) [6], it's non-probabilistic predecessor, such as capturing polysemy. Also, as opposed to LSA, this generative model has a strong theoretical justification. Problem that pLSI is often confronted to is large number of estimation parameters that depends on corpus size which can create problems with overfitting as number of documents increases, as well as inability to be applied incrementally to unseen documents due to its offline nature.

A.2. Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA), an extension to pLSI is a generative process that introduces priors on document-topic distributions proposed by Blei in 2003 [7]. In later years, this method became a basis for various latent structure discovery algorithms, known as probabilistic topic models.

By using Dirichlet prior distribution, LDA resolves the issues of pLSI. With these prior distributions the number of estimation parameters were reduced and overcome the failure of the model to be applied incrementally to unnoticed documents.

Because of its increased complexity in comparison to pLSI, exact inference is intractable from the generative model (Figure 5. **LDA model in plate notation**). Many approximation inference algorithms are derived like Variational Inference, and various Markov Chain Monte Carlo algorithms such as Gibbs Sampling [8].

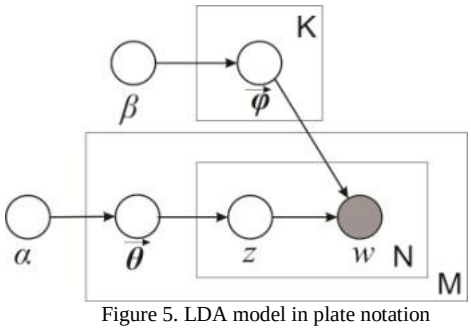

Figure 5. LDA model in plate notation

LDA serves as a basis for many topic models, some of which are presented in further text. LDA is computationally more expensive than earlier models such as pLSI and LSA. So many alternatives or extensions of LDA came into existence to model the relationship among topics [9], [10], and [11]. Some of them are modeling evolution of topics over time based on document metadata [12] and [13], modeling authorship [14], modeling arbitrary document metadata [15] and others. To resolve the computational time requirements several implementations by means of protentional parallelism are made [16].

A.3. Hierarchical Topic model

As an extension to LDA, in 2003 Blei introduced Hierarchical LDA, which can model a tree of topics rather than horizontal topic structure [9].

To model topic hierarchies, HLDA uses non-parametric Bayesian approach. Tree of topics is defined procedurally by an algorithm that constructs hierarchy as data are made available. Each node in the tree characterizes a random variable and has a word-topic distribution. The generation of the document can be done by traversing the tree from the root to one of its leaves while sampling topics along the path. Graphical model for hLDA can be seen in Figure 6.

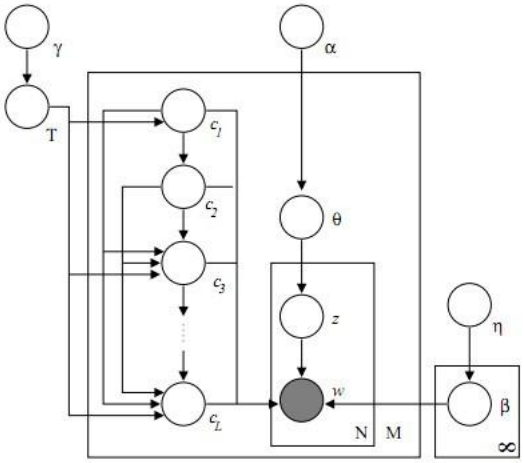

Figure 6. Hierarchical Latent Dirichlet Distribution

A.4. Dynamic Topic Model

Dynamical Topic Model (DTM) are introduced by D. Blei and J. Laferty in Proceedings of the 23rd international conference on Machine learning in New York, USA 2006., as an enhancement to Latent Dirichlet Allocation which enabled modeling of topic evolution in time [12].

Dynamical topic model includes notion of time in topic modeling using document metadata and therefore can describe evolution of word-topic distributions. Using this approach topic trends can be observed.

Dynamic topic model, an extension of LDA, gives more complicated inference as shown in Figure 7. Variational Kalman Filtering or Variational Wavelet Regression are used [12] came into exitance due to non-conjugacy, sampling methods are more difficult to infer.

The ability to track the topics through time makes DTM more advantage than the previous probabilistic topic models. But, there are some limitations in DTM. The most significant are fixed number of topics and distinct notion of time. The complexity of DTM variational inference grows fast with increase in time generality. This leads to a problem of

determining the appropriate resolution because of memory and computational time requirements.

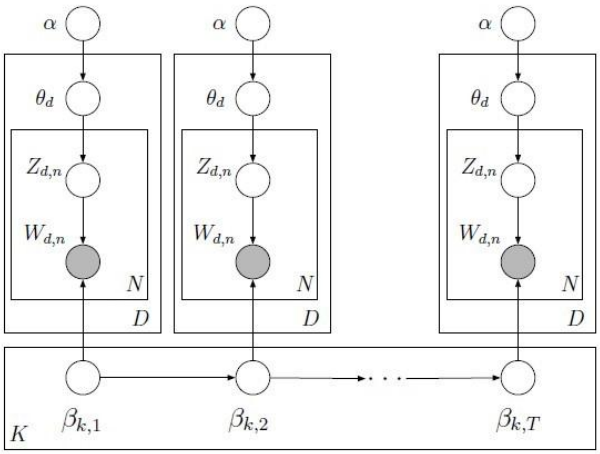

Figure 7. Dynamical Topic Model in plate notation

A.5. Correlated topic Model

Correlated Topic Model (CorrLDA) is a probabilistic topic model that improves base LDA with modeling of correlations between topics and is introduced by D. Blei and J. Laferty [10]. CorrLDA is more expressive model as it provides a graph representation of topic relationships. Generative model is represented in plate notation in Figure 8, upon which appropriate mean-field variational inference algorithm can be based [10].

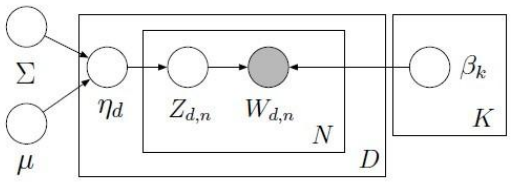

Figure 8. Correlated Topic Model, plate notation

Topic visualization and exploration can be done more effectively through CorrLDA. It suits textual corpora in a better way by modeling the relations between topics.

A.6. Pachinko Allocation Model

Pachinko Allocation Model was first introduced by Wei Li and Andrew McCallum in 2006 as a flexible alternative to Correlated Topic Model [11]. The correlations among topics can be modelled by PAM. In CorrLDA, the topic correlations are modelled using covariance matrix representing pairwise correlations between topics. The PAM improves the concept of topic distribution over words and other topics rather than the distribution over words.

Like Correlated Topic Model, PAM can model correlations between topics. As opposed to CorrLDA where topic correlations are modelled using covariance matrix representing pairwise correlations between topics, PAM redefines the concept of topic as a distribution not only over words, but as a distribution over words and other topics also. This approach enables modeling arbitrary DAG topic structure that cannot be modelled using CorrLDA. Generative model is represented in plate notation in Figure 9, and appropriate inference can be done using Gibbs Sampler.

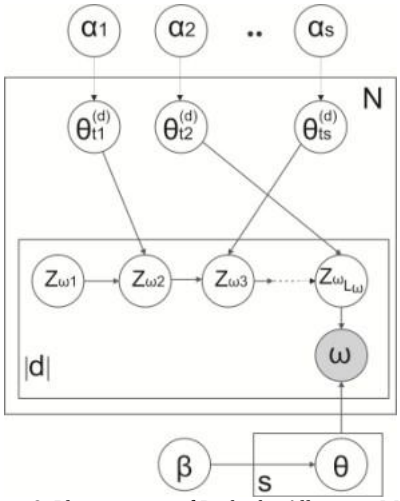

Figure 9. Plate notation of Pachinko Allocation Model

Using different approach but with the same objective, Pachinko Allocation Model provides several benefits over Correlated Topic Model. PAM can capture nested and n-ary correlations and the choice of underlying distribution is not restricted to logistic normal distribution.

A.7. Author topic Model

Author Topic Model (ATM) is a generative probabilistic topic model introduced by M. R. Zvi et al., in in 2010, derived from LDA as a model for detecting topics distribution corresponding to each author in textual corpora, based on metadata [14]. Instead of modeling only document-topic and topic-word distributions, ATM models author-topic distributions which is dependent on metadata associated with each document in corpus. The generative model is shown in Figure 10.

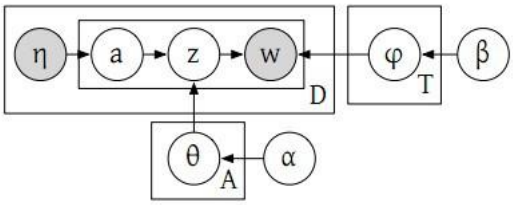

Figure 10. Plate notation of Author Topic Model

B. Supervised bag of words models with no in-domain knowledge requirements

This class of models stems from unsupervised bag of words models with no in-domain knowledge requirements, as a group of models used for classification instead of clustering. Due to their supervised nature, on some tasks these models can exhibit better modeling results.

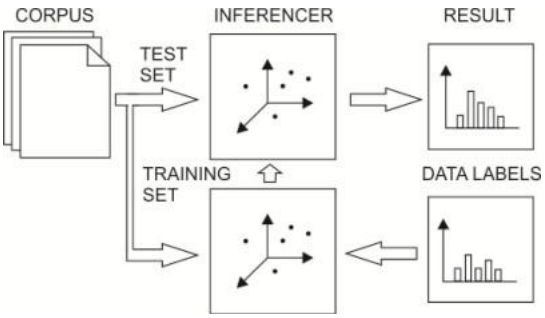

Figure 11. Outline of the supervised bag of words topic model with no in-domain knowledge requirements. **Description**: Textual corpus is divided into training and test set (to be evaluated on) and tokenizing and preprocessing is applied. Appropriate inference equations stipulated by the model are applied to training set given appropriate set of data labels effectively learning latent parameters. Finally, model with inferred parameters can be used for evaluation on test set or completely new set of unknown, unlabeled data. Word ordering is neglected.

B.1. Supervised LDA

Supervised Latent Dirichlet Allocation is first introduced by Blei and McAuliffe in 2007 as a supervised extension to LDA [17].

As opposed to other probabilistic topic models that work in purely unsupervised fashion, sLDA extends on LDA by introducing a observable response variable in the model for each document. This extension enables sLDA to fit latent topics that will best predict future unlabelled documents.

Most appropriate approximate inference method used for estimating the unknown parameters is Mean Field variational inference and can be derived from graphical model in Figure **12**.

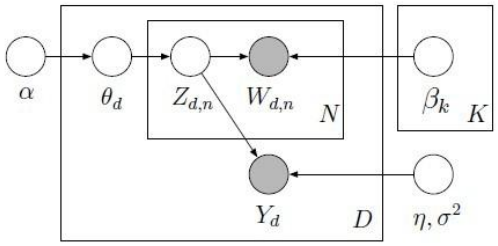

Figure 12. Supervised Topic Model

B.2. Dirichlet Multinomial Regression

Dirichlet Multinomial Regression is presented by Mimno and McCallum in 2008 [15] as an extension to LDA that can incorporate various document metadata.

As opposed to previous probabilistic topic models that account for document metadata, DMR can incorporate arbitrary types of document metadata without additional coding. This is achieved by conditioning on metadata, rather than generating metadata or estimating metadata topical densities.

Gibbs sampler for this model can be derived based on graphical model in Figure 13.

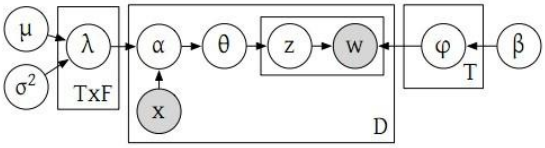

Figure 13. Multinomial Dirichlet Regression

C. Unsupervised bag of words models with in-domain knowledge requirements

Models that belong to this category make abundant use of in-domain knowledge while retaining unsupervised learning strategy. This approach is used to increase the human interpretability of topics. For instance, if modeling of a

biology corpus is required, additional constraints induced by a biological ontology are expected to yield better results.

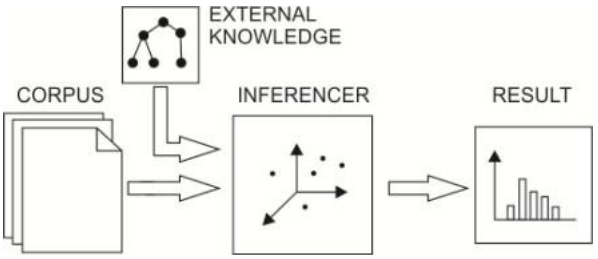

Figure 14. Outline of the unsupervised bag of words topic model with in-domain knowledge requirements. **Description**: Appropriate inference equations stipulated by the model are applied to the textual corpus after tokenizing and pre-processing. Additional in-domain knowledge is supplied, usually in form of ontology or thesaurus. Word ordering is neglected.

C.1. Concept Topic Model

Concept Topic Model [18] is an attempt to introduce semantically rich concepts into the probabilistic model.

CTM is an extension to LDA where beside ordinary learned topics also exists several constrained topics where non-zero probabilities can be assigned only to words representing human defined concepts that are provided along textual data.

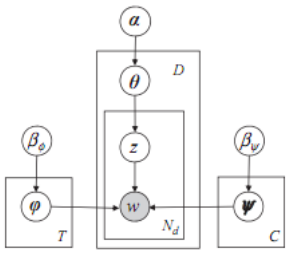

Figure 15. Concept Topic Model

D. Supervised bag of words models with in-domain knowledge requirements

This group of topic models attempt to employ additional constraints from domain of interest in classification tasks, while retaining simplicity of the bag of words assumption.

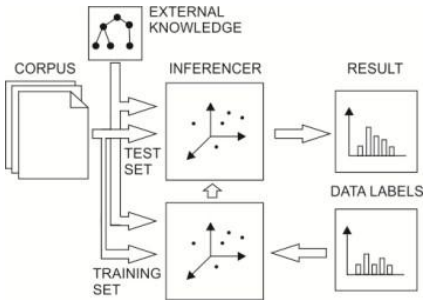

Figure 16. Outline of the supervised bag of words topic model with in-domain knowledge requirements. **Description**: Textual corpus is divided into training and test set (to be evaluated on) and tokenizing and preprocessing is applied. Appropriate inference equations stipulated by the particular model are applied to training set given appropriate set of data labels effectively learning latent parameters. Additional in-domain knowledge is supplied, usually in form of ontology or thesaurus. Finally, model with inferred parameters can be used for evaluation on test set or completely new set of unknown, unlabeled data. Word ordering is neglected.

D.1. *Supervised Conditional Topic Model*

Supervised Conditional Topic Model (sCdTM) [19] is proposed by J.Xu and E.Xing in 2010 as an attempt to utilize nontrivial input features to improve performance.

As opposed to Dirichlet Multinomial Regression [15], that can utilize arbitrary document-level metadata, Supervised Conditional Topic Model can utilize metadata at word level which enables use of retch feature such as POS tags and ontologies in modeling. This is accomplished through conditioning on metadata instead of a generative approach.

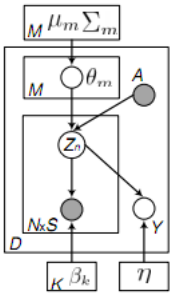

Figure 17. Conditional Topic Model in plate notation

E. Unsupervised sequence of words models with no in-domain knowledge requirements

Models belonging to this group go beyond bag of words model and account for sequential nature of textual data. Unsupervised nature of these models makes them applicable to many real-world problems where data labels aren't at

disposal. Lack of in-domain knowledge requirements makes them simpler and more applicable to some problems with presumably less domain-specific and humanly interpretable results.

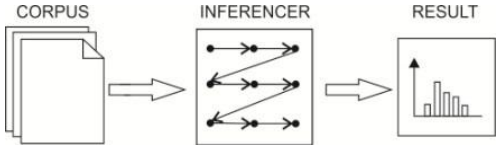

Figure 18. Outline of the unsupervised sequence of words topic model with no in-domain knowledge requirements. **Description**: Appropriate inference equations stipulated by the particular model are applied to the textual corpus after tokenizing and preprocessing. Word order is not neglected.

E.1. Topical N-Grams

Topical N-Grams (TNG) is defined by X. Wang et al in 2007 [20], as a generative probabilistic model that attempts to relieve bag of words assumption made by Latent Dirichlet Allocation [10]. As opposed to LDA, which relies on bag of words assumption and models only unigrams, TNG also models N-grams up to arbitrary N. Using this approach, although still relying on bag of words assumption, TNG attempts to account for sequential nature of text and enable modeling of complex phrases as well as unigrams. Inference is slightly more complicated than in LDA, but similar approximate inference algorithms are still applicable. Structure of TNG generative model is given in Figure 18.

Benefits of Topical N-Grams model are semantically richer topic representations, enabling modeling of concepts made of multiple words which was impossible by earlier probabilistic topic models, but such benefits come at a greater computational cost.

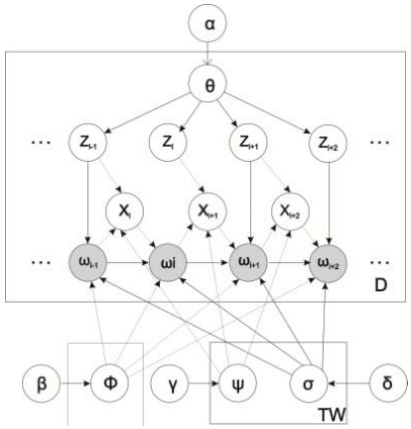

Figure 19. Plate notation of Topical N-Grams model

F. Supervised sequence of words models with no in-domain knowledge requirements

Analog to supervised variants of bag of words models, these models are intended for use in classification tasks, i.e. tasks where labels corresponding to training data are provided. These models pose no in-domain knowledge requirements.

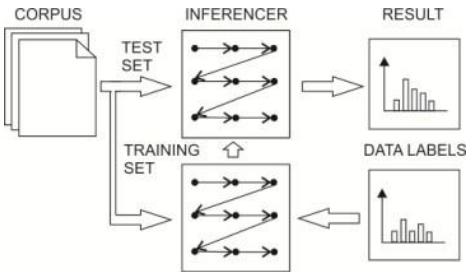

Figure 20. Outline of the supervised sequence of words topic model with no in-domain knowledge requirements. **Description**: Textual corpus is divided into training and test set (to be evaluated on) and tokenizing and preprocessing is applied. Appropriate inference equations stipulated by the particular model are applied to training set given appropriate set of data labels effectively learning latent parameters. Finally, model with inferred parameters can be used for evaluation on test set or completely new set of unknown, unlabeled data. Word order is not neglected.

F.1. Aspect Hidden Markov Model

Aspect Hidden Markov Model (AHMM) [21] is invented by D.Blei and P. Moreno in 2001. as an attempt to use Hidden Markov Models for topic modeling. AHMM is based on segmenting Hidden Markov Model and providing intuitive topical dependency between words and cohesive segmentation model.

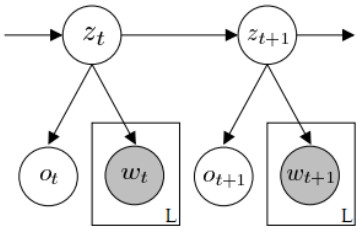

Figure 21. Plate notation of Aspect Hidden Markov Model

G. Supervised sequence of words models with in-domain knowledge requirements

In this model, the data is classified using preassigned labels for training set. The labels are assigned using some domain knowledge.

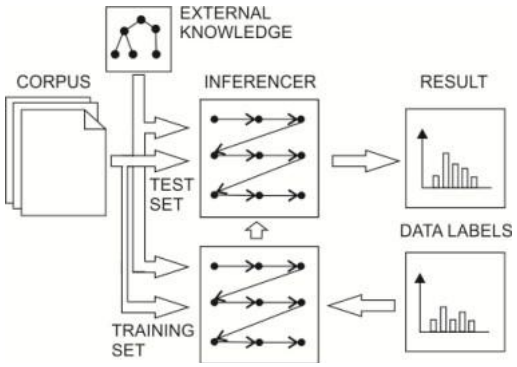

Figure 22. Outline of the supervised sequence of words topic model with in-domain knowledge requirements.

G.1. Supervised Conditional Topic Random Field Model

Supervised Conditional Topic Random Field Model is created by J. Xu and E. Xing in 2010 [19] as an attempt to utilize nontrivial input features to improve performance and to incorporate Markov dependency between topics assigned to neighbouring words.

This model presents further enhancement over Dirichlet Multinomial Regression and Conditional Topic Models in modeling using feature rich metadata, by employing a Markov dependency between topics thus accounting for sequential nature of textual data. This is accomplished through use of Conditional Random Field, a type of undirected graphical model.

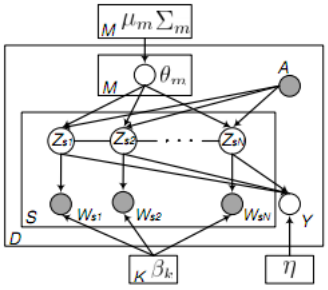

Figure 23. Plate notation of Conditional Random Field Model

H. Unsupervised sequence of words models with in-domain knowledge requirements

This class of topic models [22] make use of word-order information, while attempting to increase applicability and interpretability of results to domain of interest with additionally supplied in-domain knowledge. This class is especially interesting because of lack of instances; there are few if any models that fall into this category. Authors are non-aware of such solutions.

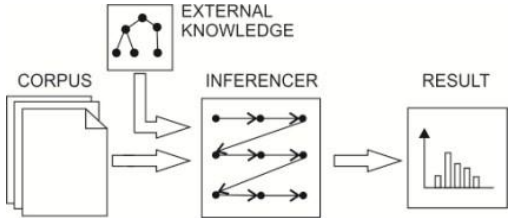

Figure 23. Outline of the unsupervised sequence of words topic model with in-domain knowledge requirements. **Description**: Appropriate inference equations stipulated by the model are applied to the textual corpus after tokenizing and preprocessing. Additional in-domain knowledge is supplied, usually in form of ontology or thesaurus. Word order is not neglected.

IV. CONCLUSION

In this study, an attempt was made to present the most significant probabilistic models which serves as a motivation to the budding researchers in selecting the appropriate model for their research work. After introducing LDA, many more probabilistic models came into existence to capture topics over a time, author-topic models, supervised LDA and so on. The advantages and their limitations are also discussed in a broad way.

REFERENCES

[1] A. Daud, J. Li, L. Zhou, and F. Muhammad, "*Knowledge discovery through directed probabilistic topic models: a survey*," Frontiers of Computer Science in China, vol. 4, no. 2, pp. 280–301, Jun. 2010.

[2] David M. Blei. *Introduction to Probabilistic Topic Models.* Communications of the ACM, 2011

[3] Steyvers, M. and Griffiths, T., *Probabilistic Topic Models*. In T. Landauer, D. S. McNamara, S. Dennis, & W. Kintsch (Eds.), handbook of Latent Semantic Analysis. Hillsdale, NJ: Erlbaum, 2007

[4] Jelisavcic, V., Furlan, B., Protic, J., & Milutinovic, V. M., "*Topic Models and Advanced Algorithms for Profiling of Knowledge in Scientific Papers*", 35th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO'2012), 1030–1035.

[5] Hofmann, T., *Probabilistic Latent Semantic Indexing.* In Proceedings of the 22nd ACM SIGIR Conference on Research & Development on Information Retrieval, Berkeley, CA, USA, 1999.

[6] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "I*ndexing by latent semantic analysis",* Journal of the American Society for Information Science, vol. 41, pp. 391–407, 1990.

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

[8] T. L. Griffiths and M. Steyvers, "*Finding scientific topics*", Proceedings of the National Academy of Sciences of the United States of America, vol. 101, no. Suppl 1, pp. 5228–5235, Apr. 2004.

[9] D. Blei, T. Gri, M. Jordan, and J. Tenenbaum, "*Hierarchical topic models and the nested chinese restaurant process*", 2003.

[10] D. M. Blei and J. D. Lafferty, "*Dynamic topic models*", in Proceedings of the 23rd international conference on Machine learning, ser. ICML '06. New York, NY, USA: ACM, 2006, pp. 113–120.

[11] W. Li and A. McCallum, "P*achinko allocation: DAG-structured mixture models of topic correlations*", in Proceedings of the 23rd international conference on Machine learning, ser. ICML '06. New York, NY, USA: ACM, 2006, pp. 577–58.

[12] X. Wang and A. McCallum, "*Topics over time: a non-Markov continuous-time model of topical trends*", in Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, ser. KDD '06. New York, NY, USA: ACM, 2006, pp. 424–433.

[13] D. M. Blei and J. D. Lafferty, "*Dynamic topic models*", in Proceedings of the 23rd international conference on Machine learning, ser. ICML '06. New York, NY, USA: ACM, 2006, pp. 113–120.

[14] M. R. Zvi, C. Chemudugunta, T. Griffiths, P. Smyth, and M. Steyvers, "*Learning author-topic models from text corpora*", ACM Trans. Inf. Syst., vol. 28, no. 1, pp. 1–38, Jan. 2010.

[15] D. Mimno and A. McCallum, "*Topic Models Conditioned on Arbitrary Features with Dirichlet-multinomial Regression*", in Proceedings of the 24th Conference on Uncertainty in Artificial Intelligence (UAI '08), 2008.

[16] J.P. Yamron, I. Carp, L. Gillick, S. Lowe, and P. van Mulbregt, "*A Hidden Markov Model Approach to Text Segmentation and Event Tracking*", *Proceedings ICASSP-98, Seattle, May 1998.*

[17] D. M. Blei and J. D. Mcauliffe, "*Supervised topic models*", in Proceedings of th Neural Information Processing Systems – NIPS, 2007.

[18] Steyvers, Mark, Padhraic Smyth, and Chaitanya Chemuduganta. "*Combining background knowledge and learned topics*", Topics in Cognitive Science 3, no. 1 (2011): 18-47.

[19] Zhu, J., Xing, E.P., "*Conditional topic random fields*", Proc. 27th Int. Conf. Mach. Learn. 2010, 1239–1246.

[20] Wang, Xuerui, Andrew McCallum, and Xing Wei. "*Topical n-grams: Phrase and topic discovery, with an application to information retrieval*" In Data Mining, 2007. ICDM 2007. Seventh IEEE International Conference on, pp. 697-702. IEEE, 2007.

[21] Blei, David M., and Pedro J. Moreno. "*Topic segmentation with an aspect hidden Markov model*" In Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 343-348. ACM, 2001.

[22] Bisgin, Halil, Zhichao Liu, Hong Fang, Xiaowei Xu, and Weida Tong. "*Mining FDA drug labels using an unsupervised learning technique-topic modeling*" BMC bioinformatics 12, no. 10 (2011): S11.

