CONTEXT AWARE TOPIC MODELING FOR SHORT TEXT

A DISSERTATION

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENT

FOR THE AWARD OF THE DEGREE

OF

MASTER OF TECHNOLOGY

IN

COMPUTER SCIENCE AND ENGINEERING

Under the esteemed guidance of

Mr. MANOJ SETHI

Submitted by:

N MOGANA

Roll No-2K15/CSE/501

COMPUTER SCIENCE & ENGINEERING
DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi-110042

MAY 2018

DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi-110042

DECLARATION

I, N Mogana, 2K15/CSE/501. student of M.Tech (Computer Science & Engineering), hereby declare that the project dissertation titled "**Context Aware Topic Modeling For Short Text** " which is submitted by me to the Department of Computer Science & Engineering, Delhi Technological University, Delhi in partial fulfillment of the requirement for the award of the degree of Master of Technology, is original and not copied from any source without proper citation. This work has not previously formed the basis for the award of any degree, Diploma Associateship, Fellowship or other similar title or recognition.

Place: Delhi **N Mogana**

Date : (2K15/CSE/501)

Department of Computer Science and Engineering
DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi-110042

CERTIFICATE

I hereby certify that the Project Dissertation titled "**Context Aware Topic Modeling for Short Text**" submitted by **N Mogana, 2K15/CSE/501,** Computer Science & Engineering, Delhi Technological University, Delhi in partial fulfillment of the requirement for the award of degree of Master of Technology, is a record of the project work carried out by the student under my supervision. To the best of my knowledge this work has not been submitted in part or full for any Degree or Diploma to this University or elsewhere.

Place: Delhi

Date:

Mr. Manoj Sethi

Project Guide
Departement of CSE
Delhi Technological University

ACKNOWLEDGEMENT

I would like to thank the Almighty, who has always guided me to work on the right path of the life. My greatest thanks are to my family and all well wishers who bestowed ability and strength in me to complete this work.

I owe a profound gratitude to my project guide Mr..Manoj Sethi who has been a constant source of inspiration to me throughout the period of this project. I am also thankful to him for trusting my capabilities to develop this project under his guidance. I shall be indebted to him throughout my life.

I am very thankful to all the faculty members of the Computer Science Engineering Dept. of DTU. They all provided immense support and guidance for the completion of the project undertaken by me. I would also like to express my gratitude to the university for providing the laboratories, infrastructure, testing facilities and environment which allowed me to work without any obstructions.

I would also like to appreciate the support provided by our lab assistants, seniors and peer group who aided me with all the knowledge they had regarding various topics.

Date :

N MOGANA

(2K15/CSE/501)

ABSTRACT

With the advent of internet and advances in communications system, enormous amount of data is generated on day basis. The major portion of the text data contributed from the social media, blogs and emails, news forums are in short text form. This enormous data also called as big data, has potential to uncover hidden information which could be used for making business centric / concrete evident based decisions. In machine learning and natural language processing, topic modeling is a widely tool for discovering hidden semantic relations in a document corpus. It models a document as a distribution of topics and a topic as a probabilistic distribution of related words. State of the art methods like LDA, BTM are not considered suitable for short text due to data sparseness problem. In this paper, a novel method referred to as *Context Aware Topic Modeling* (CATM) for short text is proposed which extends previous *Bi-Term Pseudo- Document Topic Model (BPDTM) for short text*. The BPDTM constructs a manipulated corpus based on word co-occurrence network using bi-terms of the corpus for alleviating data sparseness problem. In the due process it includes several duplicate bi-terms and unwanted edges of the network into the pseudo-corpus, which drastically affects the coherence of the topics generated. In order to reduce the noise, the CATM algorithms prunes the word network by introducing an additional distribution for naturally eliminating the unwanted words during the learning process of the topic model. Also, a tool called Wordnet is used as a preliminary step to filter out totally unrelated words while constructing the word co-occurrence network. Besides, CATM naturally lengthens the documents, which alleviate the influence on performance exerted by data inadequacy issue. Experiments demonstrated that the proposed model outperformed baseline model- BPTDM, which proved its effectiveness on short text topic models.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

1. **BPDTM- Bi-term Pseudo-Document Topic Model**
2. **CATM – Context Aware Topic Modeling**
3. **PMI – Point-wise Mutual Information**
4. **NMI – Normalized Mutual Index**
5. **BTM – Bi-Term Topic Model**
6. **LDA- Latent Dirichlet Allocation**

CHAPTER 1

INTRODUCTION

1.1 OVERVIEW

The explosive growth of internet has resulted extensive use of web application and social media such as Twitter, Face book and others. Among these short texts has been predominantly used for communication and information sharing in Internet. For instance, the micro-blogging site (example of twitter which has maximum of 140 characters) has millions of terabytes of data in the form of short texts containing rich information. This vast collection of data is very important for retrieving important information from the knowledge domain for making evident based decision.

Topic modeling is one of the text mining techniques for extracting useful information from large collection of data. Originally it was used for mining data from books, newsgroups, literature etc. These topic modeling techniques does not provide accurate and precise results for the short text documents. This is because, the short texts documents which by nature are sparse contains too few words to infer any meaningful information from the document. Also, it becomes difficult to generate latent information for processing. Hence, discovering knowledge accurately, from short texts has been recognized as a demanding research area.
Many kinds of topic models exist to automatically extract topic related information from large collection of documents by deriving the latent topics. The latent topics are obtained by running the topic modeling techniques. The most popular technique is LDA model which has proven accuracy and efficiency. But it has resulted in mixed results when the same modeling technique was applied to short texts. Hence, to handle the short text to extract the latent features of the documents several literature works are being carried out. Among the solutions proposed, the technique to expand the short text and create pseudo-document is simpler and model independent. In this technique pseudo documents are created for the short text documents based on several factors like co-occurrences of words, auxiliary words, aggregating methods which are then given as input to existing modeling techniques without modifying it. One such method is Bi-Term Topic Model. Generally these kind of pseudo-document generated are based on methods that are dependent on the application or context for which they are designed, for example twitter-LDA, LF-LDA, Dual LDA etc. , There are several topic modeling techniques like LDA, LSA, PLSA, CoFE available to uncover

the topics from the regular sized text documents. However, the conventional methods of topic modelling are not considered apt for the short text because either there are too few documents, too many topics in a single document or documents are too short for identifying the pattern of words. This prevents the efficient and accurate usage of the methods to uncover the topics from short text. Intuitively, one such technique called Bi-Term Topic Model works on the principle of enlarging the short text documents to form a corpus, large enough to apply the conventional techniques to extract hidden content. This is achieved by identifying word to word (bi-term) relation within a document. In order to increase the coherence of words while constructing the corpus Context Aware Bi-Term(CATM) for short text is proposed.

1.2 TOPIC MODEL COMPONENTS

Topic models can sort out and offer bits of knowledge for us to see substantial accumulations of unstructured content bodies. Fig. 1.1 depicts the component of topic model. It has three components. Firstly the collection of documents acts as a input. The second component is the processing unit which acts like a black box and produces the required output. Topic modeling apportions the document into several topics for topic pattern analysis, summarizing, sub topic discovery, opinion comparison and understanding large archive of documents in brief time period. Unlike document summarization technique, it gives abstract overview of the text to the reader more clearly. In most cases, the documents are reduced to less than or equal to 20% of the original document [24].

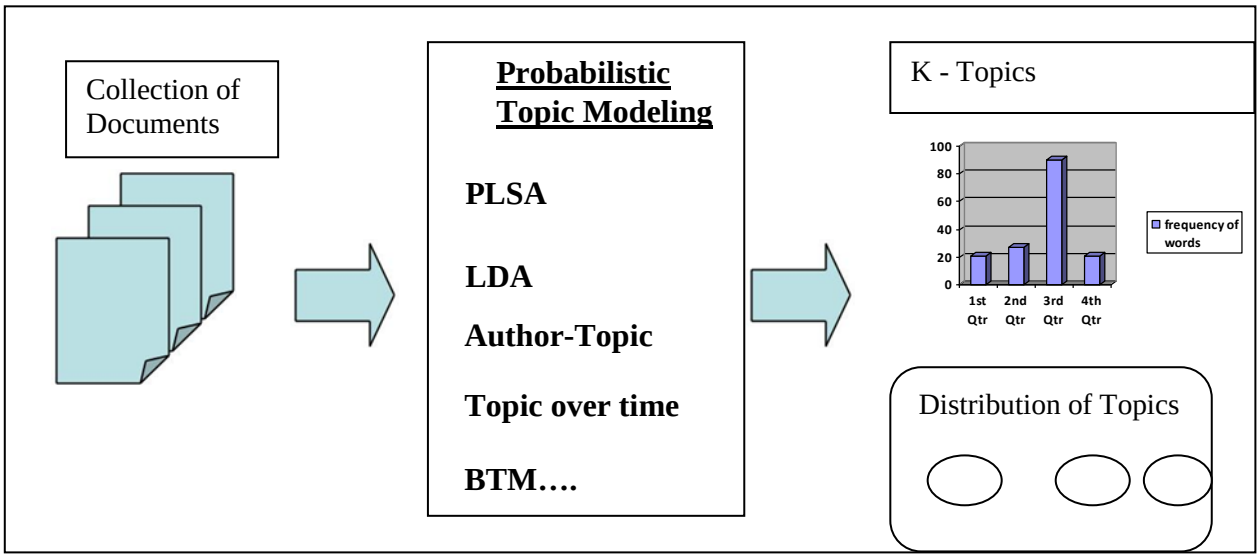

Figure 1.1 Components of Topic Model

There are several topic models for processing the collection of documents. The third component is output. It includes multinomial distribution of various topics document to topic distribution, frequency distribution of words for each topic, topic to word distribution and set of words related to each topic. Topic modeling is used in various areas. The various application of Topic Modeling is as listed below in Fig 1.2

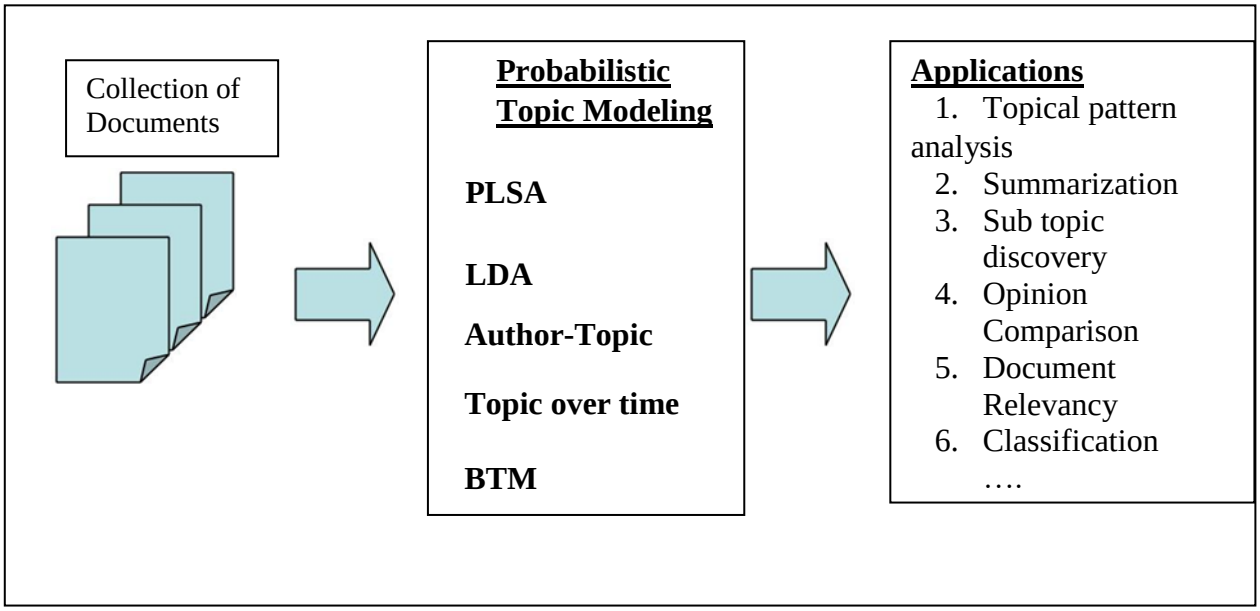

Figure 1.2. Application of Topic Model

1.3 PROBLEM STATEMENT

The study on various topic models reveals that the conventional topic models are not suitable for short text documents which work on word co-occurrence patterns. In order to use the proven conventional models, one of the simpler solutions is to enlarge the corpus to regular sized documents called pseudo-documents. For enlarging the short text documents into large corpus, several techniques are proposed in literature, like using the auxiliary information, aggregating of words etc. Jiang et al [7] in his work creates a large document from bi-term of individual documents using BTM technique. To increase the coherence of the topic model a word network based on triangular relation between bi-term is proposed which are then converted into a manipulated pseudo-document. However, the pseudo-document generation has shown several unwanted and duplicate edges being induced in to the psuedo-

corpus. Hence, to overcome the noise that is introduced because of spurious edges, a modified version of the algorithm called as Context Aware Topic Modeling is proposed.

1.4 THESIS ORGANIZATION:

The thesis is organized in various chapters as follows:

Chapter 2 gives an overview of the related work of the study, that is, the various research works that have been done in this area and how all those work helped in evolution of my study.

Chapter 3 summarized research methodologies used in this thesis including overview of the recommended framework.

Chapter 4 describes the algorithm developed as an enhancement to existing BPDT Model

Chapter 5 states the implementation of algorithm with the dataset and parameter setting for algorithm.

Chapter 6 explains the result and analysis of the program output.

chapter 7 concludes the research work carried out with suggestion for future scope of work.

CHAPTER 2
LITERATURE REVIEW

2.1 TOPIC MODEL INTRODUCTION

Topic modeling is a text mining tool for extracting hidden information from a large document collection. Unlike supervised classification technique, in this technique the input data is unlabelled and unclassified. It does not fully fall under unsupervised learning technique as well because in unsupervised the data are grouped based on clustering. There is no predefined format or structure in the input document. The document collection is totally amorphous. It tries to find the hidden topics based on the word pattern that occurs across the documents. For example, given a document it is more likely that the related words (or words belonging to same topic) co-occur more often than those that are unrelated.

2.2 CONVENTIONAL TOPIC MODELS

Discovering patterns by using the words and detecting relationship between documents are some of the important functions of topic modeling. This is achieved by understanding the way in which documents are generated. Over last few years various topic models are designed and improved to enhance the accuracy and coherence of topic classification. Some of the popular techniques are:

(a) Latent Semantic Analysis (LSA)
(b) Probabilistic Latent Semantic Analysis (PLSA)
(c) Latent Dirichlet Allocation (LDA)
(d) Correlated Topic Model (CTM)

Topics Modeled based on time factor is called topic evolution models(TEM). Some of TEMs are as listed below:

(a) Topic Over Time (TOT)
(b) Dynamic Topic Models (DTM)
(c) Multi-scale topic tomography
(d) Dynamic topic correlation detection

Over the past few decades, topic modeling has turned out to be exceptionally prevalent as a totally unsupervised procedure for topic revelation in extensive archive accumulations.

Example for such a model is LDA. LDA is data mining tool that works on the principal of probability theory(Bayesian) it is considered most popular technique. LDA models the generative process of the document that endeavors to learn what the composition procedure is. So it endeavors to create a document apportioned with set of topics. There are several models which have inherited the base idea of LDA model. Some of the Example are: temporal text mining, author- topic analysis, supervised topic models, latent Dirichlet co-clustering and LDA based bioinformatics. The essential thought of the procedure is, each archive is demonstrated as a blend of topics, and every topic is a discrete probability that characterizes how likely each word is to show up in a given topic. These subject probabilities give a compact portrayal of a document. Here, a "document" is a "bag of words" with no structure past the point and word insights. Similar efforts to find an efficient topic modeling technique for short texts have been put. Example for such models are Self Aggregation Topic Model (SATM), Pseudo-Document based TM(PTM), Sparcity-enhanced PTM (SPTM). Probabilistic topic models have been broadly used to consequently remove topical data from extensive file of archives which included content accumulations like news articles, look into papers and online journals.

2.2.1 Comparative Analysis Of Conventional Topic Models

There are several topic models existing for different type of data. The topic model for regular documents and short text documents are presented separately, each with its features and limitations in table 2.1 and table 2.2 respectively.

Table 2.1 Topic model for regular documents

MODEL	FEATURES	LIMITATION
Latent Semantic Analysis (LSA)	* Derives topics from words having similar meaning * Does not have strong statistics to prove	*difficult to identify topics. * Difficult to operate
Probabilistic Latent Semantic Analysis (PLSA)	* Capable of detecting multiple topics from single document. * PLSA handles polysemy.	*Not a probabilistic model at document level. *not easy to detect relationship among topics
Latent Dirichelet Allocation (LDA)	* very effective for regular sized documents	* depicting relationship among topics

		*manual pre-processing of documents
Correlated Topic Model (CTM)	* Uses logistic normal distribution to create relations among topics. *Allows the occurrences of words from other topic labels and representing graphs.	*Statistical operation driven *uses common words for topics

2.3 TOPIC MODELS FOR SHORT TEXT

The quick development of client produced content in smaller scale blogging destinations like (twitter, face-book, online journals) has raised the enthusiasm of analysts in growing more proper techniques to separate valuable data in short content situations. The short text poses the problem of feature extraction due to following reasons

(i) The document has too few words to infer a topic

(ii) Too few document are available

(iii) Too many topics exists within a single document

The problem of data sparseness of short text can be handled in following ways.

(i) By creating new modeling techniques for short text

(ii) By modifying existing modeling technique available for long text to make it suitable for document containing less number of words

(iii) By constructing larger pseudo-documents from short text documents

Numerous research works has been carried out in proposing a solution to the problem of handling short text. The last has the primary preferred standpoint of being more straightforward and technique free, since it just changes the information. Hence, the research is restricted to topics related to third approach of solving the problem. The most typical probabilistic topic modeling technique is LDA which has made incredible progress in displaying content accumulations like news articles, explore papers and websites. Be that as it may, the outcomes are blended when LDA is connected straightforwardly to short messages, for example, tweets, texts and gathering messages. The reason is principally because of the absence of word co-occurrence in small documents when contrasted with consistent estimated records. For instance, tweets contain printed content as well as logical data, for example, initiation, hash-labels, time and areas.

Some of the methods for short term expansion are as follows:

(i) Self-Term Expansion: Paulo Bicalho et al. [1] proposed a document enlarging method without utilizing any outer external information. STE works by supplanting documents words with an arrangement of corresponded terms, where the score of correlation is registered by examining co-occurrence of word. The technique utilizes score of PMI, which catches semantic relationship between word sets, to make a positioned connection list for each word of the dataset, which is utilized to extend the corpus' archives.

(ii) One clear technique, generally embraced for short messages in web based life, is to use external context related data called auxiliary data, relevant to total short messages into customary estimated pseudo archives applying prior to a standard topic model. For instance, tweets include printed content as well as relevant data, for example, origin, hash-labels, time and areas. This context related information can be utilized to total tweets while performing topic modeling. Other than combing short messages, a few experiments include these external context related data for conducting document generation process of the models. They regulate trained models as per some auxiliary data. Case for such a model is Dual–LDA. The issue lies in that external data is not generally accessible or just too exorbitant for arrangement.

(iii) Yuan Zuo et al [9], proposed a Pseudo-document-based Topic Model (PTM) , yet another method, for short texts that does not include auxiliary data for topic modeling. Their work is among the most punctual investigations this fascinating way. The base idea of PTM is the presentation of pseudo collection of documents for understood conglomeration of short messages against information inadequacy issue. Along these lines, the modeling of topic of colossal short messages is changed into the topic modeling of substantially less pseudo archives, which could be gainful for parameter assessment as far as both exactness and productivity is concerned. To additionally dispense with undesired relations between pseudo archives and inert topics, they likewise propose a Sparsity-enhanced PTM (SPTM) by applying Spike

(iv) Yan et al [22]. Proposed a Bi-term based model called BTM to extract from short texts which is current model selected for implementing on the metric space. It works on the entire document corpus for retrieving the latent features.

(v) Lin al. [20] proposed the dual sparse topic model, which trains centered topics of an archive and centered terms of a topic by supplanting symmetric Dirichlet priors called priors of Slab and Spike.

(vi) Among recent contributions, Quan et al[6], proposed a self-aggregated model named SATM, which can total small messages into pseudo records belonging to same topics as opposed to assistant data. Notwithstanding, the quantity of parameters of SATM develops with the measure of information, which makes it inclined to over-fitting. In addition, the time intricacy of SATM is additionally inadmissibly high. The two shortcomings keep it from being generally connected.

(vii) Gabriel and Picaso[16] proposed a method for document expansion called Co-occurrence Frequency expansion (CoFE), which is does not include context information and the maximum sized of the pseudo-document is made user-specific as per requirement. The core idea of CoFE is to calculate the number of time the words appear together across the documents. The higher the number, the higher is their chances of belonging to the same topic. Such word pairs are earmarked for inclusion into the pseudo document resulting in increase of word count in the words X doc matrix. This technique is tested on a generic framework and its capabilities for latent feature extraction are analyzed.

(viii) With respect to development techniques already proposed particularly for short content topic modeling, most methodologies concentrated on growing Twitter information. Two diverse tweet conglomeration plans are proposed: one in light of the tweets creators and another in view of each expression of the vocabulary collection. Their objective was to gather a topic distribution for the two messages and creators in the document collection. They assessed four tweet pooling plans to enhance the aftereffects of LDA, including merging of tweets composed by a similar creator, published around the same time, having common hash labels and by latent topics. They discovered those merging of tweets by hash labels produces better execution when thought about than the other proposed approaches.

(ix) As opposed to above proposals, a word co-occurrence based model, named WNTM [6] to create phony archive of documents was proposed. An undirected weighted graph is drawn from the archive which is based on co-occurrence of words. Words that co-occur at any rate once in an archive are connected, and the edge is represented by the word co-occurrence recurrence. Every word w I of this chart create a completely new phony document, which is formed by the nearby words to w I in the diagram. WNTM investigates the way that, even in short-content situations, the word-word space is somewhat thick, putting together the calculation less touchy to record length or diverseness of the topic dissemination.

(x) Regarding the works that produce or adjust prevailing techniques for topic extraction, Zhao[10] altered LDA to devise a more reasonable techniques for Twitter.

(a) Twitter-LDA, accept that every solitary tweet is more often than not about a solitary topic, and each Twitter client has a topic circulation that characterizes their likelihood of composing a tweet identified with every topic.

(b) Jin [13], thus, advocated Dual LDA (DLDA), which improves topic modeling for short messages by means of exchange of words from an assistant dataset of lengthy messages. They utilize URLs display in the short instant messages that reference lengthy reports to deliver the assistant dataset, and subsequently expect most records in the archive do have connections to lengthy archives. Past research have taken after a comparable stand, yet they overlooked the hidden and semantic inconsistencies between the objective and helper information.

(c) The current cutting edge strategies for short content are LDA, Latent Feature (LF-LDA) and Bi-term Topic Modeling (BTM).

2.3.1 COMPARATIVE ANALYSIS OF TOPIC MODELS FOR SHORT TEXT

Table 2.2. Topic Models for short text documents

METHOD	FEATURES	LIMITATIONS
BTM Bi-term Topic Model	* directly model word pairs *works on the entire document corpus	*document level distribution is not available
Co-occurrence Frequency Expansion (CoFE)	* context-independent *user defined document size	*if topic is spread over far then does not yield better result
Pseudo-document-based Topic Model(PTM)	* generation of phony documents for implicit combing of short texts	* requires extensive training on model
Self-aggregated topic model named (SATM)	*one document for each topic	* time complexity is high * the number of parameters is directly proportional to size of the data
Word co-occurrence network-based model named (WNTM)	*undirected graph based *insensitive to topic distribution	*creates new graph for every new co-occurring word.

CHAPTER 3
RESEARCH PROCESS

3.1 RESEARCH BACKGROUND AND MOTIVATION

Co-occurrence of words generates high-dimensional representations of words. Words co-occurrence measurements portrays how words occur together that readily catches the connections between words. Words co-occurrence insights are computed basically by counting how at least two words occur together in a given corpus.

In practice, the co-occurrence counts are converted to probabilities. A bi-term is not ordered in a short context. It includes a small window to cover the number of words to form bi-terms. The concept of investigating words co-occurrences can be stretched out from various perspectives. For instance, the count of how frequently a grouping of three words occurs together to produces trigram frequencies. The count of how frequently a couple of words occurs together in sentences regardless of their situations in sentences can likewise be figured. Such occurrences are called skip-bigram frequencies. Due to such varieties in how co-occurrences are indicated, these strategies as a rule are known as n-gram techniques. The term context window is frequently used to determine the co-occurrence relationship. For bigrams, the context window is asymmetrical single word long to one side of the present word in co-occurrence counting. For trigrams, it is asymmetrical and two words in length. In words to vector conversion approach by means of co-occurrence, things being what they are a symmetrical context window taking a gander at one going before word and one after word for computing bigram frequencies gives better word vectors.

3. 2 BI-TERM TOPIC MODEL

In BTM, a topic in a document is understood have contain all terms which are inter-related. It presents that the document is composed of words that are correlated and thus appear together. In traditional topic models, the frequency of words are used for extracting the topic information which may grossly misinterpret as it may contain latent (hidden) meaning in the text document. Secondly, learning of documents requires considerable amount of text words. If the text document does not contain sufficient words for leaning the method may fail with respect to quality of the topics generated. Rather, if

every one of the words of the document contributes to the learning process by expressing its relation with every other word of the text then the hidden meaning of the topic can be revealed by the correlation between the bi-terms. The bi-term topic model deals with this thought and alleviates the data sparseness problem.

3. 2.1 BI-TERM EXTRACTION:

The term "bi-term" indicates unordered set of words co-occurring within a small context window. In short messages with constrained length, for example, tweets and instant messages, each document is considered to a context window, where every word is combined with other word. For instance, an archive with three unique words will create three bi-terms:

(w1; w2; w3) => {(w1, w2); (w2, w3);(w1, w3); }

Bi-terms are unordered. In the wake of extricating bi-terms in each record, the entire corpus presently transforms into a bi-term set. The process of bi-term extraction is completed by thorough scan over entire corpus and at the same forming bi-terms by combing that words that appear together.

3.2.2 BI-TERM MODEL DESCRIPTION:

Generally the topic models functions by modeling the process in which the document are generated. But in case of BTM, the models functions by modeling the generation of bi-terms in the corpus. Hence, the learning process takes place at corpus level. The key thought is that if two words co-occur all the more every now and again, they will probably have a place with a same topic. Two words of the bi-term are assigned independently from a topic where the topic is picked from set of topics of the whole corpus.

Let us say there is a corpus with D number of documents,

In total there are B number of bi-terms given as **B ={** $\boldsymbol{b_{i=1}^{nb}}$**}** with $b_i=(w_{i,\ 1},\ w_{i,\ 2})$

with K topics expressed over W unique words in the vocabulary.

The distribution of topic over the corpus represented as P(z)= ɵ= $\theta_{k=1}^{k}$, by k-dimensional multinomial distribution ɵ= $\theta_{k=1}^{k}$

If the topic indicator variable is given as Z[1, K] . The word distribution for topics P(w/z) is given as

Φ = K X W matrix

where the kth row Φ_k is a W-dimensional multinomial distribution with entry and

$\Phi_{k,\ w}$=P(w/z=k) and $\sum_{w=1}^{W} \Phi$=1

Following the convention of LDA [6], symmetric Dirichlet priors for ɵ and Φ with single-valued hyper parameters α and β, respectively can be obtained.

3.2.3 ALGORITHM FOR BTM

As explained above in the previous section, the generative process of BTM is nothing but the generation of the bi-term on the whole corpus, which is as given by Cheng et al[6] illustrated is below:

(i) Draw θ~Dirichlet (α)

(ii) For each topic k belongs to [1,k]

a. Draw Φ_k ~ Dirichlet (b)

b. Draw $w_{i,1}, w_{i,2}$ ~ multinomial (Φ_z)

Its graphical representation is shown in Fig 1.1 Following the above procedure, the probability of bi-term bi can be determined by calculating the parameters ɵ and Φ using the hyper parameters α and β. The probability of bi-term can be obtained by integrating over ɵ and Φ. The product of probability of all single bi-terms in the whole corpus can be used to calculate the likelihood of the whole corpus.

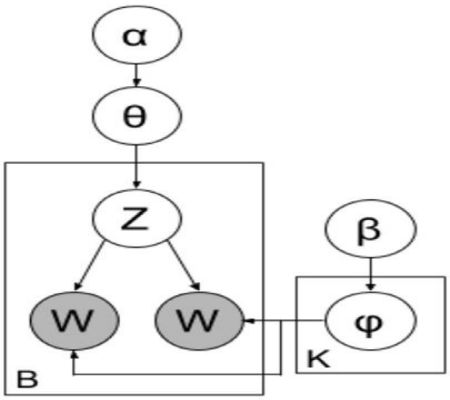

Figure 3.1 Plate Notation for BTM

3.3 BI-TERM PSUEDO-DOCUMENT TOPIC MODEL(BPDTM) FOR SHORT TEXT

BPDTM was proposed by Jiang et al.[7]. BPDTM involves following steps:

3.3.1 BI-TERM REPRESENTATION

Bi-term topic model learns topics at corpus level based on the aggregated information generated using bi-terms of individual documents. Since the learning takes place at corpus level the problem of data sparseness is alleviated. The whole corpus is considered as collection of topics generated using bi-terms. The bi-term is assigned a topic independently. The probability of a bi-term drawn belonging to a particular topic is additionally depended on the probability of each word of the bi-term belonging to the topic. Assume α and β are the Dirichlet priors. BTM overcomes the information sparsity issue by drawing topic assignment z from the corpus-level topic distribution θ. The records consider the bi-terms rather than a solitary word topic assignment like in unigram and LDA by breaking documents into bi-terms. Thusly, BTM can keep the correlation between words, as well as can catch numerous topic angles in a record, since the topic assignments of various bi-terms in an archive are free. The plate notation given above in Fig. 3.1 explains the same.

3.3.2 CONSTRUCTING WORD-NETWORK

A Word – network is constructed by taking all words of the corpus and represented as undirected but weighted graph as in Fig 3.2. Each word in the corpus is represented as a node and the relation between them as edges. The edge weight is nothing but the number of times the two words co-occur in the corpus.

3.3.3 PSEUDO-CORPUS GENERATION FROM BI-TERM

Pseudo-corpus generation is done for every node of the word network. The algorithm is as given below:

For each node of the network, say A_0

- take is adjacent sub- network.
- for every two nodes x, y in sub- network
  - check if it is connected to A_0. If it is connected then include x, y in the pseudo document for A_0

The pseudo-documents are created by using the triangular relations between the nodes. If any two nodes are related, then the third node is automatically added into the corpus irrespective of the actual semantic relation. This is done with the most compelling intuition that if B and C are related to A, then it is most likely that B and C are related. Refer the Fig. 3.2 the list of pseudo-documents generated for each node of the corpus.

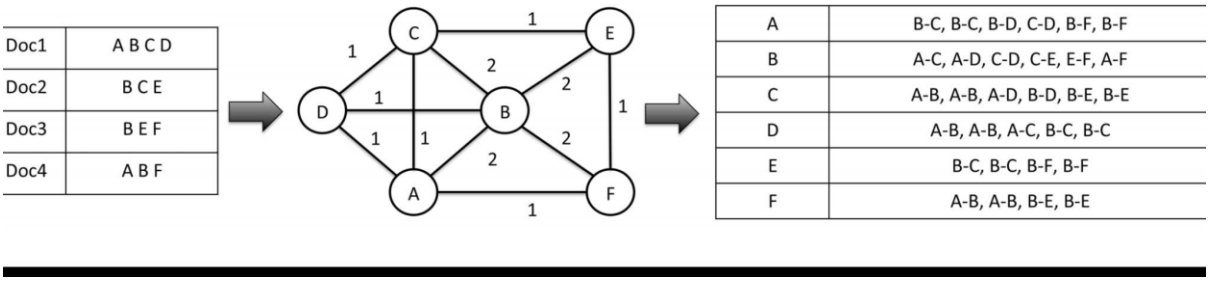

Figure 3.2. Procedure for BPDTM

Observation: It can be seen in adjacent network of nodes, the bi-terms are repetitive

3.3.4 GENERATIVE PROCESS FOR BPDTM

The process by which the model learns to generate the pseudo-document and topic-word distribution is given below. Unlike in BTM where the generative process learns the document to topic word distribution, here in BPDTM, the process generates the pseudo-document.

Output: Φ_z, Θ_j

1: for each topic z do
2: draw a topic-pseudoword distribution $\Phi_z \sim Dir(\beta)$
3: end for
4: for each pseudo document d_j in the pseudo corpus do
5: draw a pseudodocument-topic distribution $\Theta_j \sim Dir(\alpha)$
6: end for
7: for each biterm b in pseudo document d_j do
8: (a) draw a biterm-topic distribution $z_b \sim Multi(\Theta_j)$
9: (b) for b draw two words $w_i, w_j \sim Multi(\Phi_z)$
10: end for

CHAPTER 4

THE PROPOSED MODEL

CONTEXT- AWARE TOPIC MODELING FOR SHORT TEXT

4.1 INTRODUCTION

As explained in the BPDTM model above, it can be observed that BPDTM alleviates the data sparseness problem by constructing Pseudo-corpus using bi-terms of the documents. The model, in order to maintain coherence of the topics uses the triangular relations between the bi-terms which in due process results in repetitive bi-terms in pseudo-corpus as illustrated in Fig 3.2. Also, it introduces spurious edges into the pseudo-document. This in turn adds noise into the documents and hence coherence of the topics is affected drastically. This problem can be addressed in two ways:

(a) Firstly, using some technique to calculate the semantic similarity index of two words of bi-terms selected. Example technique for calculating the similarity index may include Point-wise Mutual Information (PMI) or using certain tool like Wordnet. In this thesis, Wordnet is used as a preliminary step for filtering out unwanted bi-terms. In Wordnet, the smallest similarity index value 0 indicates, words are totally unrelated and as we move towards the maximum value of 1, the closeness rate also increases. These indices may help in estimating the semantic closeness but imposes bigger problem of setting threshold or margin which could delimit the range for inclusion or exclusion of bi-terms. However, to address this issue it is proposed in this thesis, that all the bi-terms with semantic index equal to 0 are excluded from pseudo-corpus. It is done so because, determining an appropriate margin for similarity index is non-trivial and cumbersome. Hence, the second solution is proposed for enhancing coherence at learning phase automatically.

(b) As explained, Wordnet still does not solve the major issue of selecting appropriate bi-terms to exclude the spurious bi-terms into the pseudo-corpus. Hence, a novel technique to filter out spurious bi-terms on its own is proposed. This is done by introducing a special distribution for every node of the word- network while selecting its adjacent sub-network. This is explained in detail in sections 4.3.

4.2 WORDNET

4.2.1 OVERVIEW

WordNet is a large lexical database of English that are connected with each other by semantic structure. It is a combination of dictionary and thesaurus. For example: Nouns, adjectives, verbs, and adverbs. In Wordnet, words are arranged by synsets. A synset represents an abstract and unique concept. They are interlinked with each other conceptually and lexical values. It forms a network of concepts. In data mining, it is used for computational linguistics and natural language processing by forming a network of words. The major difference between Wordnet and thesaurus is that the WordNet is based on senses of word and links the words together. Secondly, WordNet while forming synset considers the semantic relation not just their synonyms. (i. e. , concepts that speakers have adopted word forms to express). The Wordnet synset relates with other synsets in hierarchical form to explain a concept from a more generic to specific concept. The important features of Wordnet is that it has hierarchical representation of relations- uses hypernym (general) or hypernym (specific)

Structure

The structure of words in Wordnet is unordered set of synonym words which belong to a concept. This group of words are called as synset and it is unordered set. Words with different meaning are represented by different synset based on their inherent meaning.

Relation

Following are the relations that are represented in a wordnet:

(a) **Hypernym:** It is the most frequently encoded form of relation. It has parent to child relation between entities. That is it represents IS-A relation. The parent in the relation is the hypernym. It is more generic in nature. Example- chair is a type of furniture.

(b) **Hyponyms:** In the parent child relationship of words, the more specific entity(child) is a hyponym. Hyponyms are transitive in nature.

(c) **Meronymy**: It represents a part of the whole relation. Example – {chair}- {legs, arm rest}. It is not inherited upward.

4. 2. 2 LEXICAL GRAPH

A lexical graph represents the words and their alternate forms, that is synonyms which includes the semantic meaning of the words. It indicates three important features of the synset- minimality, coverage and replaceability. For achieving coverage, the words in synset are ordered as per the frequency. Example-{ family, household, house, home}. Replaceability property replaces the most frequent word with the closest synonym. The wornet has 117595 synsets in total containing several strings for noun, verb, adverb and adjectives. It excludes all stopwords grammar classes.

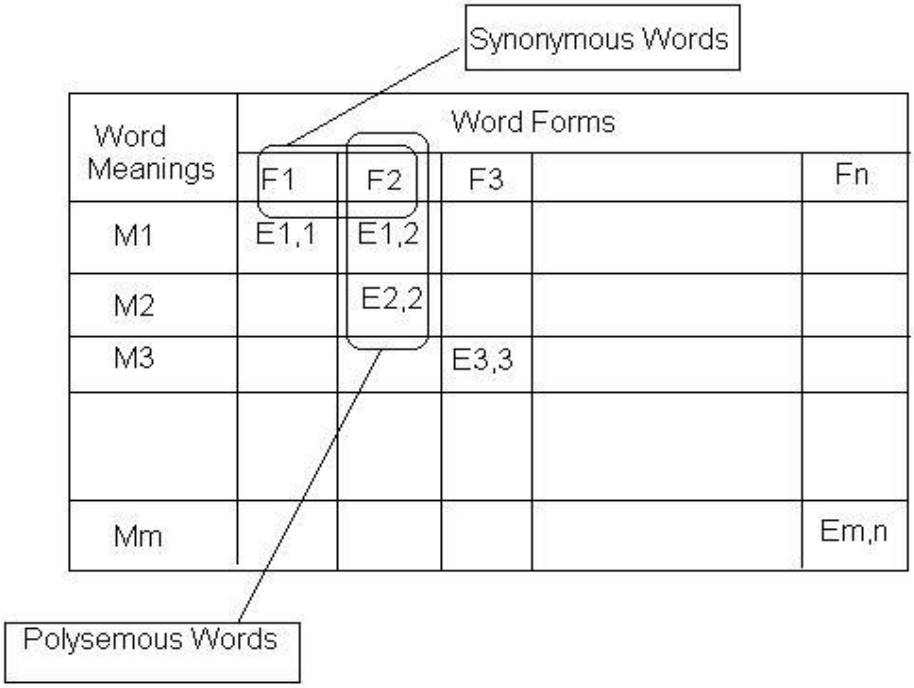

Figure 4. 1. Lexical Graph

An on-going research stream is to incorporate meta-data variables into topic modeling. The meta-data may include the auxiliary information of the document corpus. This is generally done to enhance the model in terms of accuracy and topic extraction. Also this information is helpful in estimating the results. This study aims at incorporating Wordnet synset information into topic models. In wordnet analogy, a Topic may be the combination of Wordnet synsets, or/and the hidden co-occurrence structure. The wordnet synset affects the topic inference at the token level. The BPDTM algorithm works on topic

inference calculation at corpus level learning process by enlarging the set of input documents into a large document. The enhanced model incorporates wordnet synset into the topic model. Inference is then done using Gibbs Sampling algorithm.

4. 2. 3 SEMANTIC SIMILARITY

The relations between Synsets which are, collection of literally and semantically related words, are represented by graph. From a graph, the value of similarity of synsets based on the shortest path between them can be calculated . Its values range between 0. 0- 1. 0. Lower the value, lesser is the similarity and higher the value more identical are the synsets. This value is called the **path similarity**, and it is given as:

Path Similarity Index = `1 / ((smallest distance between synset1 and synset2) + 1).`

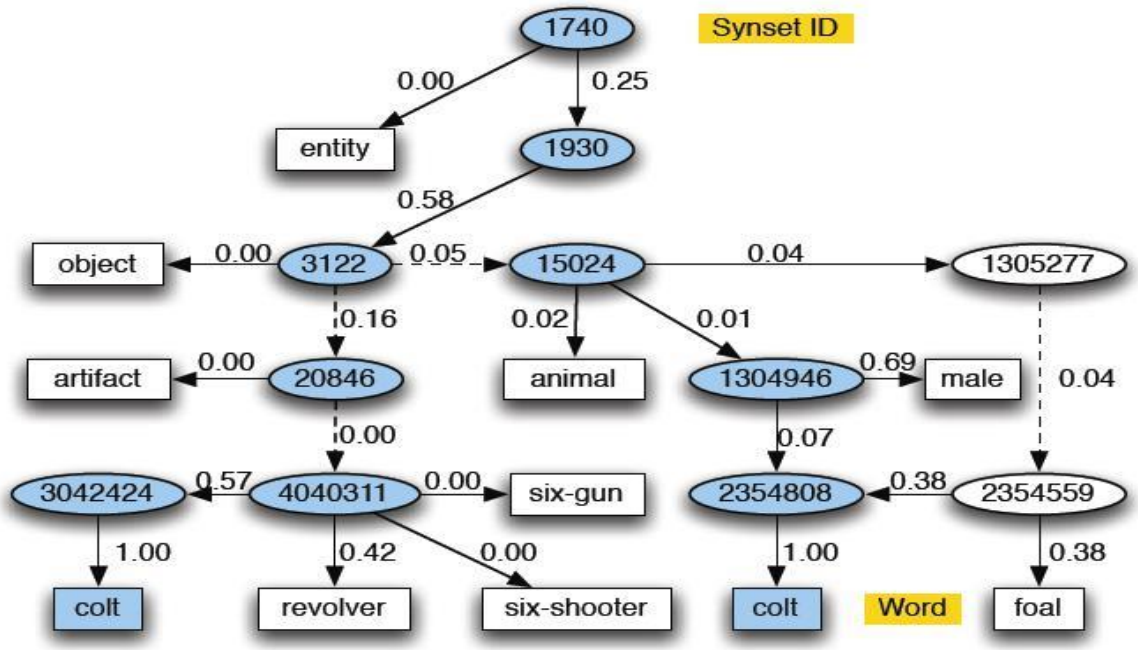

Figure 4. 2. Similarity Index of words

4. 3 CATM PROCEDURE

The steps involved in CATM topic modeling are as given underneath:

4. 3. 1 BI-TERM GENERATION

For enlarging the documents to alleviate the data sparseness problem, BTM constructs bi-term corpus. Based on the co-occurrence of the words, the BTM builds a pseudo-document for every document in the corpus.

4. 3. 2 PRELIMINARY SETUP- WORDNET EMBEDDING

The Wordnet is embedded into the Bi-term corpus during the construction of bi-term corpus as a preliminary setup for building word co-occurrence network. During the construction of the pseudo-documents the similarity index of the words are calculated as discussed above and only those bi-terms are used for inclusion into the manipulated corpus whose index value is not zero. Thus the bi-terms that are totally unrelated are removed from the corpus before the construction of word co-occurrence network.

4. 3. 3 WORD CO-OCCURRENCE NETWORK GENERATION

As in BPDTM, a word co-occurrence network is created. It is an undirected and weighted graph where each node represents a word from the vocabulary and the weighted edges between any two nodes represents the co-occurrence frequency between the words. Using this graph, the pseudo – corpus is generated as given in Fig 3. 2. As a result, for every node of the graph an adjacent list is formed using the bi-terms pseudo-documents.

4. 3. 4 THE PROPOSED ALGORITHM - GENERATIVE PROCEDURE FOR CATM

In order to filter out the unrelated words during the learning process by itself and enhance the coherence of topics, a separate distribution is added along with existing distribution. The semantically similar words are generated from the topics where as the unrelated words are generated from the separate distribution added into BPDTM method. The plate

notation explains the same in the Fig. 4. 4 given below. Also, the differences between the BPDTM and the proposed method can be noted.

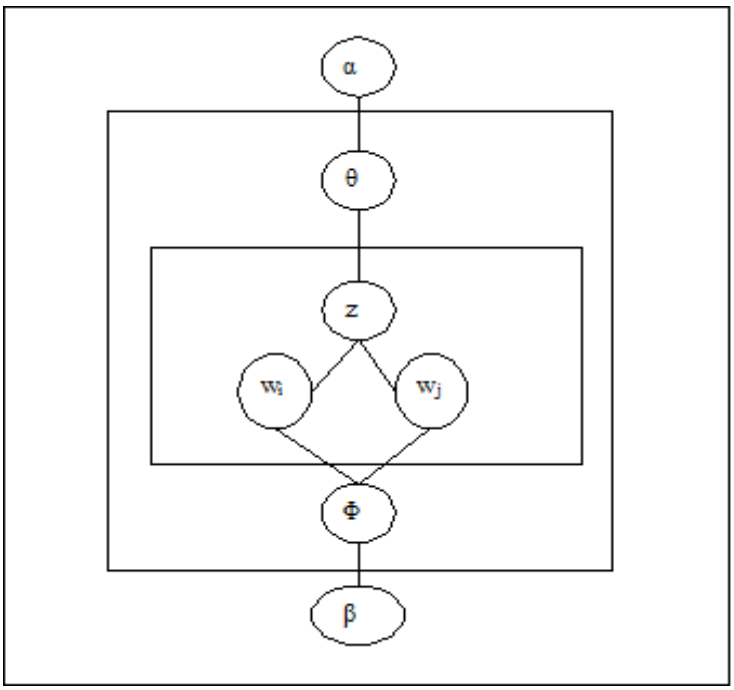

Figure 4. 3 . Plate notation of BPDTM

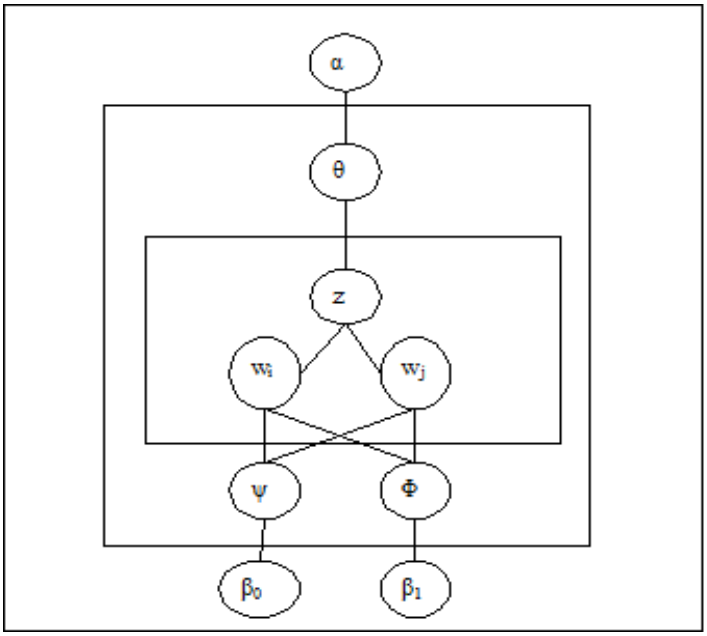

Figure 4. 4. Plate Notion of CATM

In the plate notation of CATM, the α , β_0 are dirichlet priors representing per-document topic distribution and per topic-word distribution respectively and correspond to K multinomial distribution Φ and θ. The parameter β_1 represents dirichlet prior for multinomial distribution ψ for L number of topics. This is the additional distribution added to represent the adjacent sub-network of the word co-occurrence network. Also, during the learning phase a word can be generated either from the topic distribution Φ or newly added distribution ψ. If the word is generated from the Φ distribution then a separate variable, say p is set to 0 else it is set to 1. Probability of generation of value of p is drawn from Bernoulli formula using λ which is drawn from symmetric Beta(γ).

THE PROPOSED GENERATIVE PROCESS FOR CATM IS AS GIVEN BELOW:

Input : topic count K, bi-term pseudo-corpus

Output : Φ, ψ, θ

1. **For every latent topic z do**
    a. **Draw a latent topic pseudo-word distribution $\Phi z \sim Dir(\beta_0)$**

    End for
2. **For every adjacent sub network A_0 of the node n_0 in the corpus network**
    a. **Draw $\theta_0 \sim Dir(\alpha)$**
    b. **Draw $\lambda_0 \sim Beta(\gamma)$**
    c. **Draw $\psi_0 \sim Dir(\beta_0)$**

    End for
3. **For every bi-term b in the psuedo-document d_i**
    a. **Draw $P_b \sim \lambda_i$**
    b. **If $P_b=0$ then draw $z_b \sim Multi(\theta z)$ and $b \sim Multi(\Phi_i)$**
    c. **If $P_b=1$ then draw $b \sim Multi(\psi_i)$**

    End for

4.3.5 TOPIC INFERENCE

The posterior inference is done by learning the model using Gibbs Sampling technique. The topic inference is done as given in the BPDTM technique. As given in Lan Jiang et al[7], the following equation gives the topic to word distribution within a document

$$P(z|d) = \sum_{w_i} P(z|w_i)P(w_i|d)$$

$$= \sum P(z|pd_i)P(w_i|d)$$

Where $P(z|pd_i)$ is the topic-document pd_i that equals topic-word distribution of word w_i and $P(w_i|d)$ is the normalized word count in the document

$$P(w_i, d) = |W_{i,d}|/|d|$$

where $w_{i,d}$ stands for the frequency of word i in document d.

4.3 EVALUATION METRICES

(a) QUALITY OF TOPIC

Topic coherence is proposed by Mimno et al. [26] as a way to measure how closely the classes are related. Coherence value is used to measure the semantic similarity of topics in a document. Topic coherence calculations are done by most representative words of a topic and then overall semantic similarities formula is applied. The formula for coherence is given as:

$$Co(k,W) = \sum_{m=2}^{M} \sum_{l=1}^{m-1} \log \frac{T(w_m, w_l)+1}{T(w_l)}$$

In the formula, k and W mean that in topic k, the total words set is W. T(w_m, w_l) is the number of documents containing both word m and word l while T(w_l) is the count of documents containing word l. In the work proposed by Cheng et al[6], it is said the term T(w_m, w_l) in the above equation, could be added with value 1 or any smaller value as it has been introduced to avoid calculating log of zero in the result, which might occurs when the similarity index gives 0 for the words. Where as Zuo et al.[26] suggests a value of 10^{-12}. In the experiment carried out above, it has been chosen to stick with value 1 as suggested by Cheng et al[6]. The Table 6.1 lists the average coherence value of the two models calculated as per the input parameters.

(b) NORMALIZED MUTUAL INDEX(NMI)

The formula to calculate the NMI is as given below (Mehrotra et al [15]).

$$NMI(X,Y) = \frac{2I(X;Y)}{H(X)+H(Y)}$$

The set X is set of topics labeled as i, represented as X= {X1, X2, ...Xn} while Yj is the set of words in topic j represented as Y = {Y1, Y2, ...Yn}. I (X ; Y) is mutual information between X and Y and the formula for calculating its value is given as

$$I(X;Y) = \sum_{y\in Y}\sum_{x\in X} p(x,y)\log(\frac{p(x,y)}{p(x)p(y)})$$

In the formula, p (x_i) represents probability of being labeled to topic i, p (y_j) means probability of classified to topic j while p(x_i, y_j) means probability of being classified to cluster i but actually labeled to cluster j. NMI was calculated for K value equal to 10, 35, 60 and 85 and the average NMI for each of K value is given in Table 6.2.

(C) ENTROPY

Mehrotra et al. [15] proposed the entropy formula. Entropy is another matrices to measure the document clustering. Entropy H(X) of X is given as:

$$H(X) = -\sum_{i=1}^{n} p(x_i)\log_2 p(x_i)$$

In the formula, p (x_i) indicates probability of topic i belonging to topic i, p (y_j) represents probability of topic i labeled as topic j, while p(x_i, y_j) means probability of being classified to cluster i but actually labeled to cluster j. Entropy was calculated using the above formula for the given set. The values obtained are listed in Table 6.3.

(d) PARAMETER ESTIMATION

The parameter estimation for Θ and Φ are done using Gibbs Sampling technique. As per Gibbs Sampling Technique each of the bi-terms of the network are assigned a topic randomly. In each iteration, the topics are conditioned on the model parameters. When the end of the iteration is reached, the proportion of bi-terms in each topic k denoted by n_k,

and the number of times that each word w assigned to topic k, denoted by $n_{w|k}$ is calculated. These counts are used to estimate Θ and Φ as mentioned by Xueqi Cheng et al[5] as given below

. $\Phi_{k,\ w}=(\ n_{w|k}+\beta)\ /(\ n_{w|k}+W\ \beta)$

$\theta_k=(n_k+\alpha)/(n_b+k\ \alpha)$

CHAPTER 5
IMPLEMENTATION DETAIL

5.1 DATA SET DESCRIPTION

Text documents from 20 Usenet newsgroups collection[1] containing different domain like sports, politics, and entertainment are used as input. The dataset set contains a set of 20 folders having 11347 documents of real-time data. Out of which 8575 is used for training the documents and remaining 2825 documents are used for testing purpose. The ratio of testing to training of documents is taken as 4:1as suggested by Checn et al.[19] in his work on short text. The newsgroup names are given in Tables 5.1.

Table 5.1. Newsgroup Dataset

No.	Newsgroup Name	No.	Newsgroup Name
1	alt.atheism	11	rec.sport.hockey
2	comp.graphics	12	sci.crypt
3	comp.os.ms-windows.misc	13	sci.electronics
4	comp.sys.ibm.pc.hardware	14	sci.med
5	comp.sys.mac.hardware	15	sci.space
6	comp.windows.x	16	soc.religion.christian
7	misc.forsale	17	talk.politics.guns
8	rec.autos	18	talk.politics.mideast
9	rec.motorcycles	19	talk.politics.misc
10	rec.sport.baseball	20	talk.religion.misc

5.2 PARAMETER SETTING AND CONFIGURATION

All parameters and the values used for the parameters for implementing the model are explained below. There are four main parameters specified as input to the model:

(i) The number of topics (k) which were initially given as 10, 20, 30, 40 as specified in the base paper Yan[22]. As the matrices -NMI and Entropy for measuring topic coherence did not display significant difference when the

[1] http://qwone.com/jason/20Newsgroups

topic count was lower, the topic count were increased till 85 to demonstrate the significant improvement.

(ii) The hyper- parameters α and β_0 are the Dirichlet Priors given as 50/k and 0.01 respectively as given in the base paper Yan[22]. Similarly, the hyper-parameter β_1 was set to 0.0001, γ=0.3 as suggested by Wang[25] for the additional distribution introduced in the generative process of CATM.

(iii) The number of sampling iterations given as 100 as suggested in the base paper (Yan[22]).

The parameter estimation for Θ and Φ are done using Gibbs Sampling technique. As per Gibbs Sampling Technique each of the bi-terms of the network are assigned a topic randomly. In each iteration, the topics are conditioned on the model parameters. When the end of the iteration is reached, the proportion of bi-terms in each topic k denoted by n_k, and the number of times that each word w assigned to topic k, denoted by $n_{w|k}$ is calculated. These counts are used to estimate Θ and Φ as mentioned by Xueqi Cheng et al[5] as given below

. $\Phi_{k,\ w}=(\ n_{w|k}+\beta)\ /(\ n_{w|k}+W\ \beta)$

$\theta_k=(n_k+\alpha)/(n_b+k\ \alpha)$

5.3 PRE-PROCESSING OF DOCUMENTS

The details of the dataset are listed below in Table 5.2. The total of number of documents in the collection and the average number of words in the collection are listed in the Table 5.2.

Table 5.2. Data Set Information

DATA SET	NEWS GROUP
NO.OF DOCUMENTS	11347
AVERAGE NO.OF WORDS IN EACH DOCUMENT	2000

The data from above collection are pre-processed before applying topic models. The datasets are pre-processed in following steps:

(i) Converted all words to lower case

(ii) Removed non-alphabetic characters and stop words

(iii) Stemming of the words has been carried out

5.4 ENVIRONMENT AND MODEL

The implementation were conducted on Windows-7 environment on Intel Core machine with 1.67 GHz speed and 2GB RAM. The language used for implementation is Java. The base algorithm and enhanced version was executed as per the parameters specified above.

Observation: As the size of the data is too large to execute in limited environment, the time taken to complete the process is high.

CHAPTER 6
RESULT AND ANALYSIS

6.1 TOPIC COHERENCE

Topic coherence is proposed by Mimno et al. [26] as a way to measure how closely the classes are related. Coherence value is used to measure the semantic similarity of topics in a document. Topic coherence calculations are done by most representative words of a topic and then overall semantic similarities formula is applied. The Table 6.1 lists the average coherence value of the two models calculated as per the input parameters.

Table 6.1 . Average Topic Coherence Score

Topic Models	Average Coherence
BPDTM	-481.00
CATM	-362.00

The coherence value has been calculated from the theta and phi values for every document and topic count after running it for 100 iterations. The values are then averaged out by dividing them by total number of documents and the number of topic counts. As coherence value approaches to zero, it represents a greater semantic similarity. From the above table, it can be inferred the enhanced version (CATM) outperforms the base version (BPDTM) for the given dataset.

6.2 DOCUMENT CLUSTERING

Normalized Mutual Index (NMI) and Entropy Index are the two matrices used for measuring the document clustering. Document clustering refers to degree to which the clusters within the document are related.

6.2.1 NORMALIZED MUTUAL INDEX(NMI)

NMI was calculated for the given data set from the theta and phi values. NMI was calculated for K value equal to 10, 35, 60 and 85 and the average NMI for each of K value is given in Table 6.2.

Table 6.2 Comparison of NMI values on base and enhanced versions

Topic Models	Average NMI			
	K=10	K=35	K=60	K=85
BPDTM	0.0966	0.1034	0.1088	0.1106
CATM	0.1638	0.1534	0.1548	0.1632

When NMI = 0 implies, the topics are least related, while NMI = 1 implies topics are closely related to each other. It can be observed that the number of topics is critical to deciding the values of output. The experiments were run with different values for k-number of topics staring from 10 to 85 with an interval of 25. It is seen that as the number of topics (K) increases, the NMI values for the CATM improves significantly. Following graph – Fig 6.1 depicts the same.

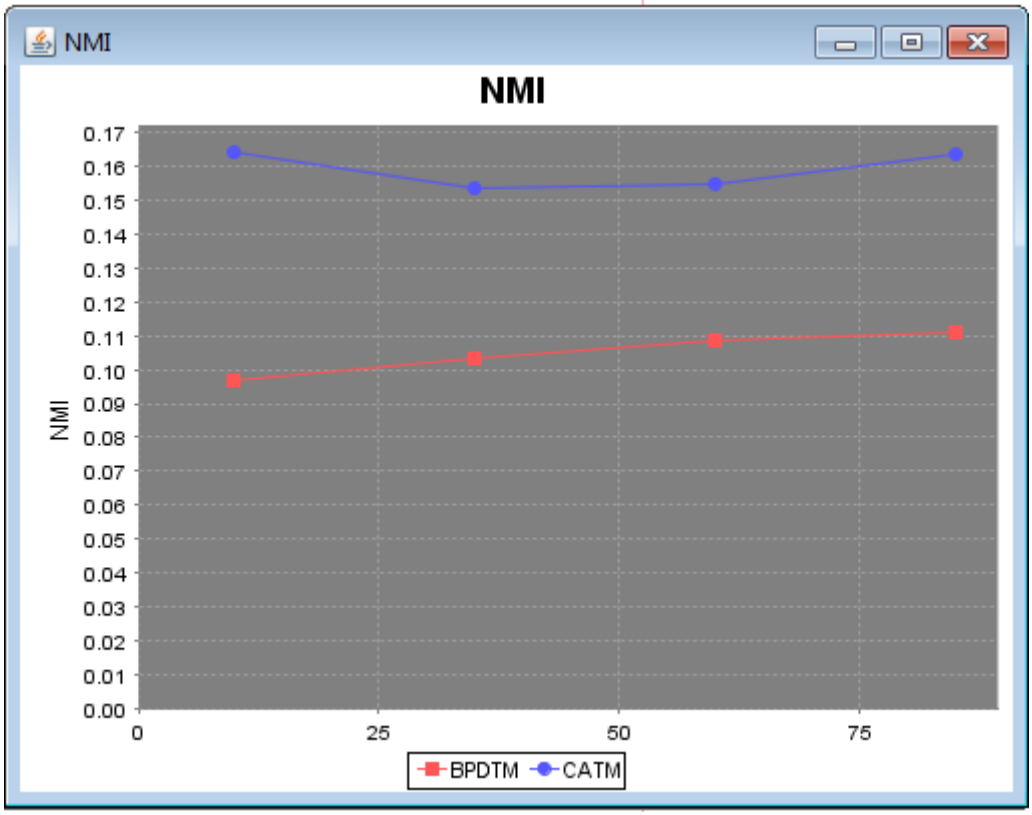

Figure 6.1 . Comparison of NMI values on base and enhanced versions

6.2.2 ENTROPY

Entropy was calculated using the above formula for the given data set. The values obtained are listed in Table 6.3. The input for calculating entropy was derived from theta and phi values.

Table 6.3. Comparison of Entropy values on base and enhanced versions

Topic Models	Average Entropy			
	K=10	K=35	K=60	K=85
BPDTM	0.0995	0.1034	0.1072	0.1065
CATM	0.1269	0.1583	0.1601	0.1709

From the table 6.3, it can be inferred that the value for entropy is closer when the topic number is less. As the topic number increases the gap between the entropy of two methods

CATM and BPDTM widens significantly, thereby proving that CATM outperforms the baseline value and indicating the fact that words in topics generated are better correlated. The graph given in Fig.6.2 reflects the same.

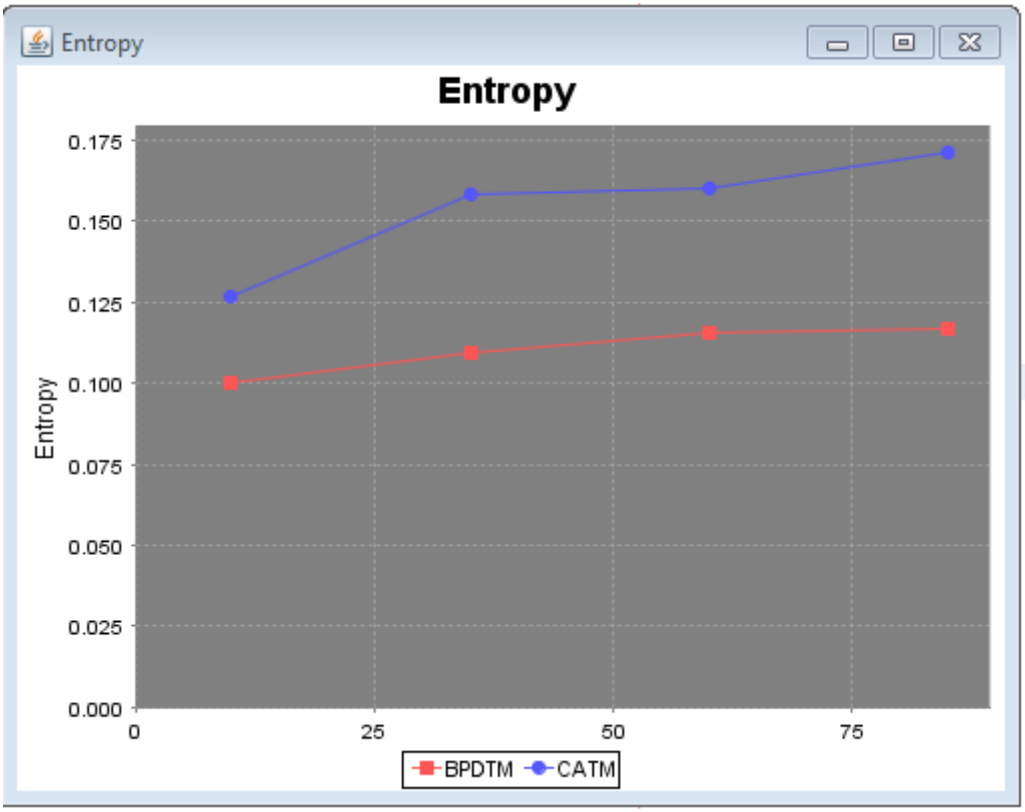

Figure 6.2. Comparison of entropy values on base and enhanced versions

CHAPTER 7
CONCLUSION AND FUTURE WORK

7.1 CONCLUSION

In this thesis, a novel method referred to as *Context Aware Topic Modeling* (CATM) for short text is proposed which extends previous *Bi-Term Pseudo- Document Topic Model (BPDTM) for short text* . The model *CATM* for sort text proposes to enhance the coherence of words in a document corpus by altering the learning process of the model so that it inherently detects the spurious words during the learning process and eliminates them. Experiments conducted on dataset showed that proposed model outperformed the baseline in terms of two topic evaluating matrices Topic coherence and NMI.

7.2 FUTURE WORK

The experiments were conducted on document corpus of short text documents, the proposed algorithm may be ascertained on regular documents for measuring its effectiveness.

REFERENCES

[1] Alghamdi, Rubayyi, and Khalid Alfalqi. "A survey of topic modeling in text mining." *Int. J. Adv. Comput. Sci. Appl.(IJACSA)* 6, no. 1 (2015).

[2] Bicalho, Paulo, Marcelo Pita, Gabriel Pedrosa, Anisio Lacerda, and Gisele L. Pappa. "A general framework to expand short text for topic modeling." *Information Sciences*393 (2017): 66-81.

[3] Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent dirichlet allocation." *Journal of machine Learning research* 3, no. Jan (2003): 993-1022.

[4] Aletras, Nikolaos, Timothy Baldwin, Jey Han Lau, and Mark Stevenson. "Evaluating topic representations for exploring document collections." *Journal of the Association for Information Science and Technology* 68, no. 1 (2017): 154-167.

[5] Cheng, Xueqi, Xiaohui Yan, Yanyan Lan, and Jiafeng Guo. "Btm: Topic modeling over short texts." *IEEE Transactions on Knowledge and Data Engineering* 26, no. 12 (2014): 2928-2941.

[6] Chen, Qiuxing, Lixiu Yao, and Jie Yang. "Short text classification based on LDA topic model." In *Audio, Language and Image Processing (ICALIP), 2016 International Conference on*, pp. 749-753. IEEE, 2016.

[7] Chen, Weizheng, Jinpeng Wang, Yan Zhang, Hongfei Yan, and Xiaoming Li. "User based aggregation for biterm topic model." In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, vol. 2, pp. 489-494. 2015.

[8] Chen, Guan-Bin, and Hung-Yu Kao. "Word co-occurrence augmented topic model in short text." *International Journal of Computational Linguistics & Chinese Language*

Processing, Volume 20, Number 2, December 2015-Special Issue on Selected Papers from ROCLING XXVII 20, no. 2 (2015).

[9] Elkan, Charles. "Text mining and topic models." *Lecture notes*(2010).

[10] Gandomi, Amir, and Murtaza Haider. "Beyond the hype: Big data concepts, methods, and analytics." *International Journal of Information Management* 35, no. 2 (2015): 137-144.

[11] Jiang, Lan, Hengyang Lu, Ming Xu, and Chongjun Wang. "Biterm Pseudo Document Topic Model for Short Text." In *Tools with Artificial Intelligence (ICTAI), 2016 IEEE 28th International Conference on*, pp. 865-872. IEEE, 2016.

[12] Kamath, Krishna Y., and James Caverlee. "Expert-driven topical classification of short message streams." In *Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on*, pp. 388-393. IEEE, 2011.

[13] Larkey, Leah S., Lisa Ballesteros, and Margaret E. Connell. "Improving stemming for Arabic information retrieval: light stemming and co-occurrence analysis." In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 275-282. ACM, 2002.

[14] Mazarura, Jocelyn, and Alta de Waal. "A comparison of the performance of latent Dirichlet allocation and the Dirichlet multinomial mixture model on short text." In *Pattern Recognition Association of South Africa and Robotics and Mechatronics International Conference (PRASA-RobMech), 2016*, pp. 1-6. IEEE, 2016.

[15] Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. "Distributed representations of words and phrases and their compositionality." In *Advances in neural information processing systems*, pp. 3111-3119. 2013.

[16] Matwin, S. (2013) Institute for Big Data Analytics: Message from the Director, https://bigdata.cs.dal.ca/about, retrieved July 28, 2016.

[17] Mimno, David, Hanna M. Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. "Optimizing semantic coherence in topic models." In *Proceedings of the conference on empirical methods in natural language processing*, pp. 262-272. Association for Computational Linguistics, 2011.

[18] Miller, George A. "WordNet: a lexical database for English." *Communications of the ACM* 38, no. 11 (1995): 39-41.

[19] Pedrosa, Gabriel, Marcelo Pita, Paulo Bicalho, Anisio Lacerda, and Gisele L. Pappa. "Topic modeling for short texts with co-occurrence frequency-based expansion." In *Intelligent Systems (BRACIS), 2016 5th Brazilian Conference on*, pp. 277-282. IEEE, 2016.

[20] Sajid, Anamta, Sadaqat Jan, and Ibrar A. Shah. "Automatic Topic Modeling for Single Document Short Texts." In *2017 International Conference on Frontiers of Information Technology (FIT)*, pp. 70-75. IEEE, 2017.

[21] Wang, Chan, Caixia Yuan, Xiaojie Wang, and Wenwei Xue. "Dirichlet process mixture models based topic identification for short text streams." In *Natural Language Processing andKnowledge Engineering (NLP-KE), 2011 7th International Conference on*, pp. 80-87. IEEE, 2011.

[22] Wang, Fei, Rui Liu, Yuan Zuo, Hui Zhang, He Zhang, and Junjie Wu. "Robust Word-Network Topic Model for Short Texts." In *Tools with Artificial Intelligence (ICTAI), 2016 IEEE 28th International Conference on*, pp. 852-856. IEEE, 2016.

[23] Yan, Xiaohui, Jiafeng Guo, Yanyan Lan, and Xueqi Cheng. "A biterm topic model for short texts." In *Proceedings of the 22nd international conference on World Wide Web*, pp. 1445-1456. ACM, 2013.

[24] Y. Zuo, J. Zhao, and K. Xu, "Word network topic model: a simple but general solution for short and imbalanced texts," k*nowledge and Information Systems*, vol. 48, no. 2, pp. 379–398, 2016.

[25] Yan, Xiaohui, Jiafeng Guo, Yanyan Lan, and Xueqi Cheng. "A biterm topic model for short texts." In *Proceedings of the 22nd international conference on World Wide Web*, pp. 1445-1456. ACM, 2013.

[26] Zuo, Yuan, Junjie Wu, Hui Zhang, Hao Lin, Fei Wang, Ke Xu, and Hui Xiong. "Topic modeling of short texts: A pseudo-document view." In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 2105-2114. ACM, 2016.

[27] Zuo, Yuan, Junjie Wu, Hui Zhang, Hao Lin, Fei Wang, Ke Xu, and Hui Xiong. "Topic modeling of short texts: A pseudo-document view." In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 2105-2114. ACM, 2016.