



# **Polynomial topic distribution with topic modeling for generic labeling**

By

**Md. Rezwan Ul-Hassan**

**151-35-917**

**Shadikur Rahman**

**151-35-988**

A thesis submitted in partial fulfillment of the requirement for the degree of  
Bachelor of Science in Software Engineering

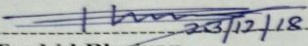

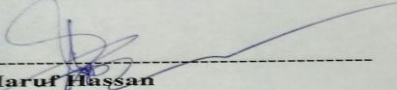
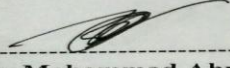
**Department of Software Engineering**  
**DAFFODIL INTERNATIONAL UNIVERSITY**

Fall – 2018

## APPROVAL

This thesis titled on “Polynomial topic distribution with topic modeling for generic labeling”, submitted by **Md. Rezwan Ul-Hassan, 151-35-917** and **Md. Shadikur Rahman, 151-35-988** to the Department of Software Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of Bachelor of Science in Software Engineering and approval as to its style and contents.

## BOARD OF EXAMINERS

 ----- <b>Dr. Touhid Bhatyan</b> Professor and Head Department of Software Engineering Faculty of Science and Information Technology Daffodil International University	<b>Chairman</b>
 ----- <b>Dr. Md. Asraf Ali</b> Associate Professor Department of Software Engineering Faculty of Science and Information Technology Daffodil International University	<b>Internal Examiner 1</b>
 ----- <b>Md. Maruf Hassan</b> Assistant Professor Department of Software Engineering Faculty of Science and Information Technology Daffodil International University	<b>Internal Examiner 2</b>
 ----- <b>Prof Dr. Mohammad Abul Kashem</b> Professor Department of Computer Science and Engineering Faculty of Electrical and Electronic Engineering Dhaka University of Engineering & Technology, Gazipur	<b>External Examiner</b>

## DECLARATION

It hereby declares that this thesis has been done by us under the supervision of **Ms. Syeda Sumbul Hossain**, Lecturer, Department of Software Engineering, Daffodil International University. It is also declared that neither this thesis nor any part of this has been submitted elsewhere for award of any degree.

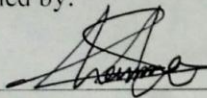
Rezwan  
24-12-18

Md. Rezwan Ul-Hassan  
Student ID: 151-35-917  
Batch: 16<sup>th</sup>  
Department of Software Engineering  
Faculty of Science & Information Technology  
Daffodil International University

Sadi  
24-12-18

Shadikur Rahman  
Student ID: 151-35-988  
Batch: 16<sup>th</sup>  
Department of Software Engineering  
Faculty of Science & Information Technology  
Daffodil International University

Certified by:

 24.12.2018

Ms. Syeda Sumbul Hossain  
Lecturer  
Department of Software Engineering  
Faculty of Science & Information Technology  
Daffodil International University

## ACKNOWLEDGEMENT

First of all, we are grateful to the Almighty Allah for giving us the ability to complete the final thesis.

We would like to express our gratitude to our supervisor **Ms. Syeda Sumbul Hossain** for the consistent help of my thesis and research work, through his understanding, inspiration, energy, and knowledge sharing. Her direction helped us to finding the solutions of research work and reach to our final theory.

We would like to express my extreme sincere gratitude and appreciation to all of our teachers of **Software Engineering** department for their kind help, generous advice and support during the study.

We are also express our gratitude to all of our friend's, senior, junior who, directly or indirectly, have lent their helping hand in this venture.

Last but not the least, we would like to thank our family for giving birth to us at the first place and supporting me spiritually throughout my life.

**Md. Rezwan Ul-Hassan**  
**151-35-917**

**Shadikur Rahman**  
**151-35-988**

## Table of Contents

<b>APPROVAL</b> .....	i
<b>DECLARATION</b> .....	ii
<b>ACKNOWLEDGEMENT</b> .....	iii
<b>LIST OF TABLES</b> .....	vii
<b>LIST OF FIGURES</b> .....	viii
<b>LIST OF NOMENCLATURE</b> .....	ix
<b>ABSTRACT</b> .....	x
<b>CHAPTER 1</b> .....	1
<b>INTRODUCTION</b> .....	1
1.1 Problem outline .....	1
1.2 Motivation .....	2
1.2.1 Elements of our study .....	2
1.2.2 Interest behind our research .....	2
1.2.3 Amusing to others .....	2
1.3 Research Questions .....	3
1.4 Research Objectives .....	3
1.5 Research Design.....	3
<b>CHAPTER 2</b> .....	5
<b>BACKGROUND</b> .....	5
2.1 Topic Model .....	5
2.2 Topic Modeling Algorithms.....	5
2.2.1 Latent Semantic Analysis (LSA).....	5
2.2.2 Probabilistic Latent Semantic Analysis (pLSA).....	6
2.2.3 Latent Dirichlet Allocation (LDA) .....	7
2.3 Reason for Using LDA in Our Research.....	8

2.4 Uses of LDA.....	9
2.4.1 Five impressive research using Latent Dirichlet Allocation in recent years in our aspect .....	10
2.4.1.1 Research 1 .....	10
2.4.1.2 Research 2.....	11
2.4.1.3 Research 3 .....	11
2.4.1.5 Research 4.....	12
2.4.1.6 Research 5 .....	12
<b>CHAPTER 3 .....</b>	<b>13</b>
<b>SYSTEMATIC MAPPING STUDY .....</b>	<b>13</b>
3.1 Search Strategy.....	13
3.1.1 Search String.....	13
3.1.2 Search Scope.....	13
3.1.3 Search Period .....	14
3.1.4 Search Method .....	14
3.1.5 Inclusion/Exclusion Criteria .....	15
3.2 Overview of Systematic Mapping Study .....	15
3.2.1 Classification Schema.....	15
3.2.2 Search Scope Results.....	23
3.2.3 Time Period Results.....	23
3.2.4 Publication Type Results .....	23
<b>CHAPTER 4.....</b>	<b>25</b>
<b>RESEARCH EXPERIMENT .....</b>	<b>25</b>
4.1 Text Preprocessing .....	27
4.1.1 Tokenizing Text.....	27
4.1.2 Stop words Removing .....	27
4.1.3 Lemmatizing words .....	28
4.2 Noun Phrase Choosing .....	29
4.3 Training LDA Model for Generating Topic Set.....	29
4.3.1 Contention Behind to Fix Pass Parameter Property Eight of LDA Model.....	31
4.3.2 Parameters We Have Used in LDA Model .....	33

4.3.3 Reasons Aft to Choose In between Two Thousand Words Document and Set as Topic Property Three and Word Property Two Behind Each Topic .....	33
4.4 Indexing Topic Set .....	35
4.5 Picking Top Weighted Words from Topic Sets .....	35
4.6 WordNet Processing.....	36
4.7 WUP Similarity Checking for Choosing Labels .....	37
<b>CHAPTER 5</b> .....	<b>40</b>
<b>RESULT ANALYSIS AND DISCUSSION</b> .....	<b>40</b>
5.1 Recall, Precision, F-measure .....	40
5.2 WUP Similarity .....	45
<b>CHAPTER 6</b> .....	<b>48</b>
<b>CONCLUSION</b> .....	<b>48</b>
6.1 Our Contribution .....	48
6.2 Future Works.....	48
<b>REFERENCES</b> .....	<b>49</b>
<b>Appendix A</b> .....	<b>51</b>
<b>Appendix B</b> .....	<b>52</b>

## LIST OF TABLES

Table 2. 1: Five Impressive Work Using LDA.....	10
Table 3. 1: Ordination of picked studies above electronic database.....	24
Table 4. 1: Result of an example document.....	30
Table 4. 2: Result of an example document with various pass.....	31
Table 4. 3: Result of an example document with topic four.....	34
Table 4. 4: Result of an example document with topic three.....	34
Table 5. 1: Confusion matrix classes and parameters.....	40
Table 5. 2: Label of selected 5 documents.....	43
Table 5. 3: Recall, Precision and F-measure of selected 5 documents.....	44
Table 5. 4: WUP similarity between topic and label.....	45



## LIST OF FIGURES

Figure 1. 1: Research Design.....	4
Figure 3. 1: Flowchart of the search and filtering process.....	16
Figure 3. 2: Algorithms used in our study.....	16
Figure 3. 3: Ordination of pertinent papers above time span each publicationspecies.....	24
Figure 4. 1: Overview of our research workflow.....	25
Figure 4. 3: Process of topic and label generation.....	25
Figure 4. 4: Training process of model.....	29
Figure 4. 5: WordNet Processing for label.....	37
Figure 4. 6: WUP similarity process for labeling.....	38
Figure 5. 1: F-measure score for topics of selected 5 documents.....	45
Figure 5. 2: WUP similarity between topic and label.....	46

## LIST OF NOMENCLATURE

Terms	Nomenclature
LDA	Latent Dirichlet Allocation
LSA	Latent Semantic Analysis
pLSA	Probabilistic Latent Semantic Analysis
SVD	Singular-Value Decomposition
TF-IDF	Term Frequency-Inverse Document Frequency
NMF	Non-Negative Matrix Factorization
HDP	Hierarchical Dirichlet process
NN	Noun
NNP	Proper Noun
VB	Verb
JJ	Adjective
POS	Parts-of-Speech
TP	True Positive
TN	True Negative
FP	False Positive
FN	False Negative
WUP	Wu-Palmer Similarity
SLR	Systematic Literature Review

## **ABSTRACT**

Topics generated by topic models are typically reproduced as a list of words. To decrease the cognitional overhead of understanding these topics for end-users, we have proposed labeling topics with a noun phrase that summarizes its theme or idea. Using the WordNet lexical database as candidate labels, we estimate natural labeling for documents with words to select the most relevant labels for topics. Compared to WUP similarity topic labeling system, our methodology is simpler, more effective, and obtains better topic labels.

*Keywords-Text mining, Topic model, Topic label, LDA, WordNet*

# CHAPTER 1

## INTRODUCTION

### 1.1 Problem outline

Statistical topic modeling plays vital roles in many research areas, such as text mining, language processing, and knowledge retrieval. Topic modeling techniques embrace Latent Dirichlet Allocation (Blei et al., 2003), Probabilistic Latent Semantic Analysis (Hofmann, 1999) and Latent Semantic Analysis (Deerwester et al., 1990). These techniques can automatically discover the abstract “topics” that occur in an exceeding assortment of documents. They model the documents as a mix of topics, and every topic is sculptural as a likelihood distribution over words. Though the discovered topics word distributions are typically intuitively significant, a serious challenge shared by all such topic models is to accurately interpret the means of every topic. The interpretation of every topic is incredibly necessary once people need to browse, perceive and leverage the topic. However, it's typically terribly exhausting for a user to grasp the discovered topics primarily based only on the polynomial distribution of words.

As an example, here are the highest terms for a discovered topic: {run, drive, car, speed, bike}. It is tough for a user to completely perceive this topic if the user is not terribly acquainted with the document assortment. The situation may deteriorate when the user faces with the variety of discovered topics and also the sets of top terms of the topics are usually overlapping with one another on several sensible document collections.

So as to deal with the above challenge, we design our method by extracting necessary phrase which gives higher tf-idf (Salton and McGill, 1983) value for given phrases and working with WordNet (Miller et al., 1995). For example, we may extract the phrase “car” if it provides the high value and then working with our process model for generic labeling for this exact phrase. The topic labels will facilitate the user to grasp the topics to some extent.

If we choose the word as the label which provides higher value by training model it gives a result however the case will deteriorate when some ambiguous phrase is employed or multiple distinct phrases with poor coherence are used for a topic. To address the drawbacks of the above labels, we need to provide additional contextual data and think about employing the natural label to represent the topics.

## **1.2 Motivation**

The overall perspective of this research is: Topic modeling is employed to extract latent topics for documents. It provides an advantage to the reader to present an easy touch what's occurring over the document. But for polynomial topic it's terribly exhausting to grasp for a user. If it's properly labeled it'll be far better for understanding.

We developed following queries, to spot specific studies and analysis queries from overall analysis topic, goal.

### **1.2.1 Elements of our study**

We studied “Natural language processing” for preprocess of given text datasets, statistical topic models in search of latent distributed topics over the documents. WordNet for context collection and selection process of accurate phrase.

### **1.2.2 Interest behind our research**

Because for a reader who doesn't want to read the full document or a specific page or a specific content, topic models give a glance of that criterion which user or reader choose. Then when it's come through as polynomial it seems very confusing to the user what is very irritating. If we can label the polynomial topic it will more comfortable.

### **1.2.3 Amusing to others**

Because typically one has not got a lot of time or not interested to read the whole document. However, he needed to grasp what's the main keywords of a specific page, content or whole document. Then if one gets the topics then conception will come but if one gets confused for polynomial topics besides a topic our labeling provides a certain concept on that polynomial topics.

### **1.3 Research Questions**

RQ1: What are the topic modeling algorithms existing in the current-state-of-art?

RQ2: How to set the effective/ appropriate number of topics, words and passes parameter for LDA training for best topic pulling out to obtaining the most effective result?

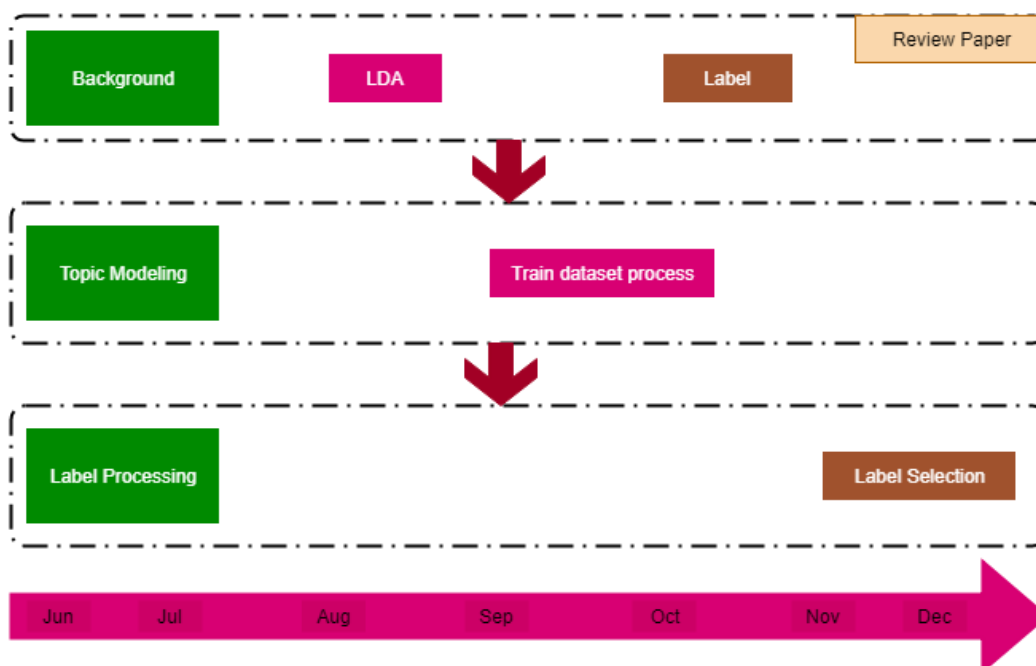
RQ3: How to figure out the most covering topics and label for polynomial topics?

### **1.4 Research Objectives**

- To present the topic modeling algorithms existing in the current-state-of-art.
- Find out the effective number of topics, words and passes parameter for LDA training to best topic pulling out for obtaining the most effective result.
- To figure out the most covering topics and label for polynomial topics.

### **1.5 Research Design**

This research has been divided into 2 segments. Segment 1 is pulling out topics by using LDA topic modeling and another segment is labeling based on topic pulling result. We have conjointly developed two background studies associated with the topic model and labeling process. Later, that helped us to ascertain our study.



*Figure 1. 1: Research Design*

# **CHAPTER 2**

## **BACKGROUND**

### **2.1 Topic Model**

One similar technique in the field of text mining is Topic Modelling. Topic model is a method to automatically recognize topics in any datasets and to get out hidden patterns shown by a text corpus in our datasets (Gildea et al., 1999).

I have put your diary and that I have only 2 minutes to feel your inmost secrets! However, about studying it from the scrape? In 2 minutes? Not possible! However, I have a text mining robot who will method and explore the total diary in but 2 minutes and by topic modeling. Topic model techniques will lightly get out valuable data and insights from diary.

With an example, it is easier to hold out remember you are reading newspaper or book and you have a gathering of colored highlighters in your hand. It is previous fashion? I catch lately only a few people read newspapers or books in print, everything is digital and highlighters are so yesterday. Make up you are your father or your mother thus, as you are study the newspaper or books you are highlighting the fascination keywords. An extra assumption! You utilize a special color for highlighting the keywords of various themes. You cluster the keywords cover the assigned color and themes. Every list of words known by an individual color is the list of keywords for a topic. The count of distinct colors your usage represents the count of topics. This is the most primary topic modeling concept.

### **2.2 Topic Modeling Algorithms**

There are several algorithms for doing topic modeling. The top most popular ones include LSA (Deerwester et al., 1990), pLSA (Hofmann and Thomas, 1999), LDA (Blei et al., 2003).

#### **2.2.1 Latent Semantic Analysis (LSA)**

Latent semantic Analysis is also known as Latent Semantic Indexing (LSI) (Deerwester et al., 1990). Latent semantic Analysis is an absolutely automatic statistical method for pulling



out and assume rapport of awaited contextual usage of words in passages of an essay. It is not a conventional language processing or AI program; it uses no humanly made dictionaries, information bases, semantic networks, grammars, syntactic parsers, or morphologies, etc. And takes as its input solely raw text parsed into words defined as distinctive character strings and separated into significant passages or samples like sentences or paragraphs.

The primary mode is to represent the text as a matrix in which every row stands for a novel word and every column stands for a text passage or alternative context. Every cell holds the oftenness with that the word of its row seems within the passage designated by its column. The cell ingress is exposed to an initial conversion in that every cell oftenness is weighted through a function that discloses each the words significance within the specific passage and therefore the degree to which the word sort carries data within the area of discourse usually.

LSA (Deerwester et al.,1990) embed singular value decomposition (SVD) (Golub et al.,1970) to the matrix. This is often a type of factor exploration, or additional properly the mathematical generalization of that factor exploration might be a special case. In SVD (Golub et al.,1970) an oblong matrix is decomposed into the outcome of 3 different matrices. One element matrix narrates the main row entities as vectors of formed perpendicular factor costs, another narrates the main column entities within the equivalent manner, and therefore the third is a diagonal matrix hold scaling costs like when the three components are matrix-multiplied, the main matrix is rebuilt. There is a mathematical evidence that any matrix is thus decomposed altogether, using no additional factors than the tiniest dimension of the main matrix. When fewer than the mandatory number of factors are used, the rebuilt matrix is a least-squares best match. One will cut back the dimensionality of the analysis just by defacing coefficients within the diagonal matrix, normally beginning with the tiniest.

### **2.2.2 Probabilistic Latent Semantic Analysis (pLSA)**

Probabilistic Latent Semantic Analysis (pLSA) (Hofmann and Thomas,1999) is a method of topic models. Its principal target is to model co-occurrence data below a probabilistic framework to find the underlying linguistics structure of the information. It was created in 1999 by Th. Hofmann and at the start used for text-based applications such as indexing, retrieval, clustering. But its use shortly unfolds in different fields like computer vision or

audio process. pLSA (Hofmann and Thomas,1999) is often considered in two apparent different ways:

- Latent variable model: The probabilistic form of pLSA (Hofmann and Thomas,1999) relies on a statistical model, referred to as the facet model. The latent variables stated by topics are related to the discovered variables which stated by documents and words, for the text region.
- Matrix factorization: Likewise, Latent semantic analysis (LSA) (Deerwester et al.,1990), pLSA (Hofmann and Thomas,1999) aims to factorize the distributed co-occurrence matrix so as to scale back its dimensionality. However, pLSA is sometimes viewed as an additional sound technique because it provides a probabilistic explanation, as LSI acquires the factorization by using solely mathematical base. More exactly, pLSA (Hofmann and Thomas,1999) uses the singular value decomposition procedure.

### **2.2.3 Latent Dirichlet Allocation (LDA)**

Latent Dirichlet Allocation (LDA) (Blei et al., 2003) model, an unsupervised, statistical procedure is introduced as modeling document corpora through finding latent semantic topics within extensive collections of text documents. The chief insight within LDA (Blei et al., 2003) is the assumption that words hold powerful semantic knowledge regarding the document. Hence, it is logical to expect that documents on pretty related topics will use the equivalent genus of words. Latent topics are in this way identified by identifying the genus of words in the corpus that commonly befall together inside documents. Learning in this model is unsupervised because the input data is unfinished: the corpus gives only the words inside documents; there is no training set with topic or subject vaccines. In LDA (Blei et al., 2003) the worths of the latent random variables (topics) are supposed by conditioning over the audited random variables (words) applying Bayes provision.

LDA (Blei et al., 2003) assumes that the generative manner as every document in a corpus: for every word  $w_{d,i}$ , in the corpus, it forms a topic  $z$  relied on the blend  $\theta$  attached to the document  $d$  and later it produces a word from the topic  $z$ . To clarify this fundamental model, the volume of the Dirichlet frequency  $k$ (amount of topics  $z$ ) is supposed to be acquainted and stable. The Dirichlet prior is used because it has various compatible

characteristics that simplify guess and parameter determination algorithms for LDA (Blei et al., 2003).

Later on we will describe the process of LDA (Blei et al., 2003) briefly as we focused on LDA (Blei et al., 2003) and use it in our process.

### **2.3 Reason for Using LDA in Our Research**

There are several approaches for getting topics from a text like – Term Frequency and Inverse Document Frequency. Non-negative matrix resolution techniques. Latent Dirichlet Allocation is that the preferred topic modeling technique (Blei et al., 2003). LDA could be a Bayesian version of pLSA (Hofmann and Thomas,1999).

If read the count of topics as count of clusters and therefore the probabilities because the ratio of cluster membership then exploitation LDA (Blei et al., 2003) could be a method of sentimental clustering your mixed and elements. Distinction this with say k-means wherever every existence will only enter to at least one cluster.

If select the amount of topics to be less than train datasets exploitation LDA (Blei et al., 2003) could be a method of minimizing the dimensionality of the first composite versus half knowledge set. With the datasets currently retail to a lower dimensional latent topic area, you will be able to currently attach different machine learning algorithms which can like the smaller variety of dimensions. For instance, you will run your documents through LDA (Blei et al., 2003) so exhausting cluster them exploitation Density-based special clustering. Of course the most argument you'd usage latent Dirichlet allocation is to uncover the themes lurking in your knowledge. By exploitation LDA (Blei et al., 2003) on burger orders, you would possibly infer burger topping themes as spicy, salty, savory, and sweet.

Probabilistic Latent Semantic Analysis (Hofmann-UAI99) approach is more principled than Latent Semantic Analysis, since it possesses a sound statistical foundation (Hofmann-UAI99). LDA (Blei et al., 2003) is similar to pLSA (Hofmann and Thomas,1999), but with dirichlet priors for the document-topic and topic-word distributions. This prevents overfitting, and gives better results.

## **2.4 Uses of LDA**

LDA is known as a generative probabilistic model of a corpus of some document. The initial notion is that the documents are illustrated as random blends upon latent topics, where a topic is distinguished by an ordination upon words. Latent Dirichlet allocation (LDA) (Blei et al., 2003), foremost proposed by Blei, Ng and Jordan in 2003 (Blei et al., 2003), is one among most popular procedure in topic modeling. LDA (Blei et al., 2003) illustrates topics by word likelihoods. The words with maximum likelihoods in every topic generally impart an excellent notion of what the topic is can word likelihoods from LDA (Blei et al., 2003).

### 2.4.1 Five impressive research using Latent Dirichlet Allocation in recent years in our aspect

Table 2. 1: Five Impressive Work Using LDA

No.	Author	Used Model	Years	Problem Domain
1	Valle et al.,2018	LDA	2018	Birds breeding and biogeographical shifts
2	Guo et al.,2017	LDA	2017	Online ratings and reviews analysis
3	Feuerriegel et al.,2016	LDA	2016	Financial news and stock prices
4	Pinoli et al.,2014	LDA	2014	Gibbs sampling and gene function
5	Lienou et al.,2010	LDA	2010	Satellite Images

#### 2.4.1.1 Research 1

**Extending the Latent Dirichlet Allocation model to presence/absence data: A case study on North American breeding birds and biogeographical shifts expected from climate change (2018)**

They represent this method with the datum from the North American Breeding Bird Survey (BBS). All-inclusive, their model recognized 18 foremost bird combinations, exposing attractive spatial forms for every combination, multiple of which were almost correlated with temperature and rainfall gradients. Moreover, by distinguishing the predicted ratio of every combination for two periods of time (1997-2002 and 2010-2015). Their outcomes designate that 9 (out of 18) breeding bird combinations manifested an expansion northward and diminution southward of their ranges, exposing astute but significant community-level biodiversity shifts at a continental measure that are apt with those prospected below climate transition.

#### **2.4.1.2 Research 2**

##### **Mining meaning from online ratings and reviews: Tourist satisfaction analysis using latent dirichlet allocation (2017)**

They introduce a novel procedure to excerpt hidden dimensions of customer gratification from rich online consumer reviews. For dimension excerption, the LDA (Blei et al., 2003) exploration of consumer reviews exposes significant dimensions that are not found in conventional ways. The relevant tenor of the excerpted dimensions is classified following to the intensity of the conversations for everyone. They also determine the heterogeneity of appreciation across various demographic profiles of customers practicing the dimensions. The research takes a nearly large sample of 25,670 hotels placed in 16 countries, allowing them to make more positive generalizations than earlier studies using conventional study designs. This study moreover presents a stepwise regression and sensorial analysis for TripAdvisor's five customer ratings for hotels and in total customer gratification. Room experience including service standard are recognized as the multiple significant dimensions in our investigation.

#### **2.4.1.3 Research 3**

##### **Analysis of How Underlying Topics in Financial News Affect Stock Prices Using Latent Dirichlet Allocation (2016)**

They practice a Bayesian framework to choose the fundamental latent topics. Once they recognize the topic model, they can investigate how stock market interests rely upon the particular topic groups. More correctly, they investigate the influence of topics found in ad hoc publications on the German stock market. To move out such a large review, they discover the topics of each specific confusion relief and carry them in our model. Topic measurement is performed via the usage of Latent Dirichlet Allocation (Blei et al., 2003). They attach a significance to any of the latent topics by reviewing the most familiar words compared with them. In order to predict the stock market effect, they estimate the unusual gains for each stock on the notification date to reduce embarrass impacts. Our rating knows topics with no resulting control on unusual records of stocks, whereas different topics, such as drug experiment, show a huge impact.

#### **2.4.1.5 Research 4**

##### **Latent Dirichlet Allocation based on Gibbs Sampling for Gene Function Prediction (2014)**

This is the paper where they proposed to use two versions of Latent Dirichlet Allocation (Blei et al., 2003) approach based on deteriorated Gibbs Sampling, as substitutes to truncated Singular Value Decomposition, to forecast lost bimolecular vaccines on the base of the at present alive ones. They operated multiple operations on datasets of Homo sapiens and Rattus norvegicus genes annotated to Gene Ontology property terms split into three sub-ontologies (Biological Processes, Molecular Functions, and Cellular Components). Outcomes parade that their methods out do the tSVD (Hansen et al.,1992) one, forecasting a larger amount of annotations that were observed verified by computational or human curated annotations in an updated variant of the fundamental datasets granted for the forecast.

#### **2.4.1.6 Research 5**

##### **Semantic Annotation of Satellite Images Using Latent Dirichlet Allocation (2010)**

They proposed to utilize text analysis tools in order to semantically annotate large high-resolution satellite pictures, applying ideas defined by the user. The method proposed here

joins the LDA (Blei et al., 2003) model that enables one to classify the applications of the huge picture into the given semantic classes, acknowledgments to its latent topics, and the spatial knowledge between these patches, which raises the annotation review.

## **CHAPTER 3**

### **SYSTEMATIC MAPPING STUDY**

#### **3.1 Search Strategy**

The research procedure approaches the research string we declare, the scope of the research, the limit of search, the research method, and the inclusion/exclusion measures we place. Appropriate has comprehensive coverage of the state of the art, the research string should be carefully Determined.

##### **3.1.1 Search String**

We made the research string following to the leading goal and the research question what we have set. The strings should be lite, so as to achieve several results and include the topics specifically. We have used the OR Boolean operatives to attach the leading terms and their equivalents. The terminal research string is:

("Topic Model" OR "Topic-Model" OR "Topic Modeling" OR "Topic-Modeling" OR "Topic Label" OR "Topic-Label" OR "Topic Labeling" OR "Topic-Labeling")

In order to use the string formed with the AND/OR Boolean operators, we have used the accessible advanced search in all database.

##### **3.1.2 Search Scope**

For gaining a dominant coverage of obtaining the pertinent studies and publications, six electronic databases were included in the search scope. Pursuant to Dyba et al. (2007), Laguna and Crespo (2013), Novais et al. (2013), and Vasconcellos et al. (2017) in their



systematic mapping studies, these electronic databases are the traditional and effective to convey systematic studies in the behalf of software engineering, re-engineering and technical studies for our research. They are highly suggested for searching past publications widely. The databases names are

ScienceDirect([www.sciencedirect.com](http://www.sciencedirect.com)),IEEEExplore([www.ieeexplore.ieee.org/Xplore/home.jsp](http://www.ieeexplore.ieee.org/Xplore/home.jsp)),ACMDigitalLibrary([www.dl.acm.org](http://www.dl.acm.org)), Springer ([www.link.springer.com](http://www.link.springer.com)), conferences, Journal of Machine Learning Research ([www.jmlr.org](http://www.jmlr.org)) and Semantic Scholar ([www.semanticscholar.org](http://www.semanticscholar.org)). To obtaining more helpful studies that do not appear in the standard search manner, the Google Scholar ([www.scholar.google.com](http://www.scholar.google.com)) database was admitted, nevertheless the fact that the search returns tend to be iterative with the search outcomes of the chosen databases.

### **3.1.3 Search Period**

This time span coated all similar papers published in books, book chapters, journals, conferences, magazines, articles from almost 2003 up to 2018. The topic modeling became popular around the beginning of 2003 when LDA come out at first. So, we choose that time as the beginning period. June 2018 is the time that we began operating on this research of topic modeling with our labeling criteria.

### **3.1.4 Search Method**

We managed both automated and usual researches in the study. In the automated search, all electric database, provided by a search engine, analyzes the search titles covering the metadata, mutually with the title, inclusive, and keywords of per paper in the database. In the usual search, on the one hand, we saw inside per conference, books, articles, magazine, and journal noted in the database that was linked to the search string for the papers specifically associated to our mapping study, but which did not look in the automatic search. On another hand, we used the snow-balling technique (Wohlin 2014), which allows the reference list of the final accepted studies from the electronic query to be examined. Therefore, we have introduced them to our mapping study.

### **3.1.5 Inclusion/Exclusion Criteria**

The selection standards goal to get all pertinent papers in our issue of systematic mapping as bellows.

#### **Inclusion criteria:**

Those papers which published since 2003 to 2018 and few bare from before, the entire paper issued in the conference, journal, articles, and chapter in a book, papers where the search string appears in the caption, abstract, and main keywords.

#### **Exclusion criteria:**

Those papers that are not correlated to search strings. Project reports, research proposal are also eliminated. Copied papers of the alike study in several variants, journals, conferences, and books chapter excluded. Those papers which are not having the complete lesson gainable and not issued in English are similarly excluded.

## **3.2 Overview of Systematic Mapping Study**

In this sections, we describe the results regarding study scope, study selection, and the demographic directions.

### **3.2.1 Classification Schema**

Pursuant to (Kitchenham and Charters 2007), in stage two of the regular literature review (SLR), the situation appraisal manner of the studies is followed for examining and evaluating the selected papers to be involved in the data removal and reporting manner. Filing the papers in faces is a nice base for explaining research questions. Per face is determined by means of various suitable keywords. As an outcome of the fast reading, a set of facets can be settled

into which the papers can be classified. In our case, the aspects were also inspired by the classification structure upon judgments suggested by Pérez et al., 2011.

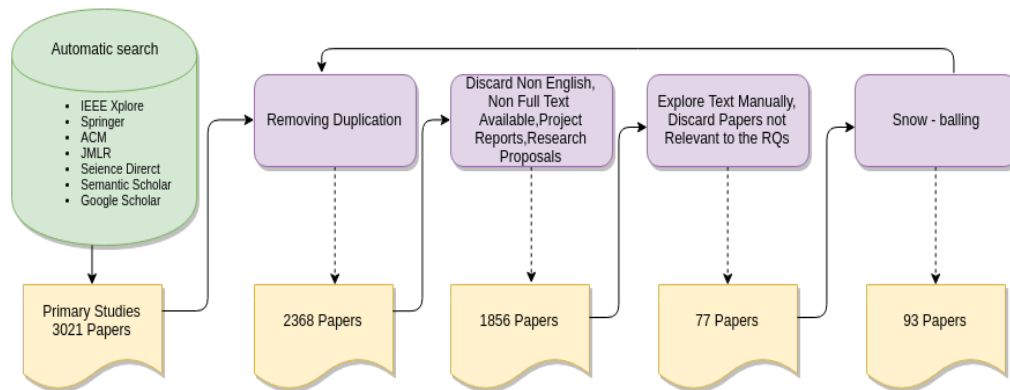


Figure 3. 1: Flow diagram of search and filtering procedure

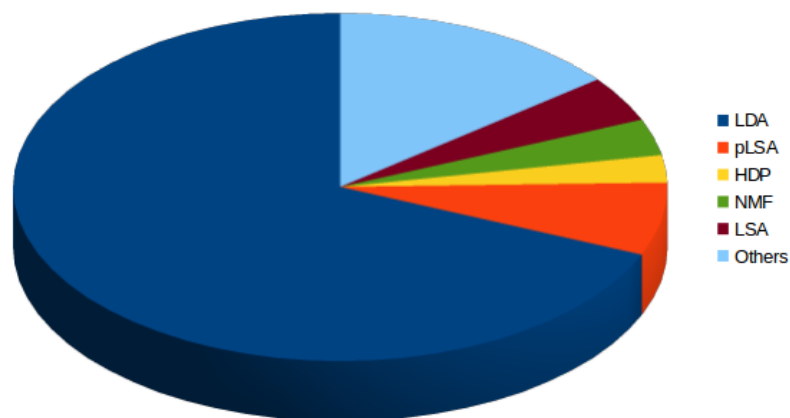


Figure 3. 2: Algorithms used in our study

We have seen on our study that 67% used LDA, 7% used pLSA, 3% used HDP, 4% used NMF, 5% used LSA and 14% used other algorithms which are not much popular.

The selected 93 papers were published as journals, book, book chapters, articles, or conferences. Where 53% were conference papers, 35% were from journals, 8% were articles, and 4% were book and book chapters.

The authors worked with AT model in [S1] which comes from LDA and in this paper, they worked for extracting topics from large text documents. They had used AT model with Gibbs sampling for assigning extracted topics to the author.

In [S2], they had built a novel model using LDA which they called MME-LDA. They had introduced it for giving image annotations. Natural language processing (NLP) is also used therewith topic modeling method. Finally, they made it and get a better result from other versions of the LDA model. This was kind of text analysis process. The study [S82] in addition use NLP and LDA for sectional preferences.

The authors of study [S3] developed a new geometric method using NMF algorithm for latent topic discovery and then they had shown the result under some label but manually select it as human aspect.

Study [S4, S20, S4] introduced an important factor based clustering algorithm (IFCA) which was a graph-based clustering method. They also used NLP as [S2] and then cluster their dataset on behalf of an importance level. Study [S20] is a cluster topic over a document with semantic similarity like [S10] where they also work with that kind of similarity.

We can correlate study [S5, S16, S72, S91] at the same set because of their approaches. The authors of [S5] just presented four kinds of topic modeling algorithm VSM, LSI, pLSI and LDA. Last of the paper, they were talking about some important tools that were already implemented in different languages like Stanford topic modeling toolbox, Gensim, Mallet and BIGartm. [S16] is kind of [S5] but the author only talked about LDA probabilistic topic modeling method. In [S72] the author of the paper had surveyed of topic modeling four methods like [S5]. [S91] was about topic modeling applications and all generic of LDA with implications.

Study [S6, S17, S18, S74, S75, S76, S77, S79, S80, S81, S89, S93] are correlate themselves because most of them are regarding ontology concept, WordNet and labeling. [S6, S89] had proposed OntoLDA an ontology-based topic modeling approach, [S6] along with a graph-based topic labeling method for the task of topic labeling. They generated labels for different words. They had used the LDA model by integrating ontological concepts for generating

labels for every word. [S17, S18] are also both impressive work for topic label generation but for the group of words not like [S6]. [S74] was also about labeling topic through text summaries. [S75, S79] both were very close to [S6, S17, S18, S74] but these were for multinomial topics. [S76] was similar to these [S6, S17, S18, S74] and use WordNet like [S10]. Though [S77] about labeling but they had induced L-LDA for that task. [S80, S93] is slightly unique for labeling it use Wikipedia titles. [S81] is task about conceptual labeling and similar [S80] for Wikipedia concept.

They had proposed an LDA based model and named as Biterm Topic Model (BTM) in [S7]. BTM can well capture the topics within short texts by explicitly modeling word co-occurrence patterns in the whole corpus. BTM a promising tool for content analysis on short texts for various applications, such as recommendation, event tracking, and text retrieval, etc.

Study [S8, S21] are similar type work we have found. The author of [S8] used LDA for clustering user of microblogging platform like Twitter. Here they were differentiated users by their same activity and interest. They had used user tweets like [S7] where they use tweets and they also used NLP like [S2, S4]. In [S21] they as well worked for microblogging platform Twitter for tweet pulling and labeling via topic modeling.

Study [S9] was about social emotion classification they had proposed contextual sentiment topic model (CSTM) for adaptive social emotion classification. It was like sentiment analysis over social media emoticon. Though [S31] has not direct connectivity with [S9] but they are all talked about sentiment analysis by their own procedure. The authors of [S31] had introduced a new semi-supervised approach for sentiment analysis, which consists of the following steps: domain identification and sentiment analysis based on LDA background topic labeling, additionally, they present an automatic modification of the last one. Study [S92] has a little bit connection with [S31] because it was also about sentiment analysis with the grace of LDA over short text.

Study [S10] where they had used LDA and also used WordNet. In this paper word, according to the sentences generative process, calculating topic-importance and the topic-distribution of sentences, they proposed a novel sentence-ranking method to get the salience of sentences. They calculated the semantic distance of the sentences using WordNet. If we brief generally it was a document summarization process. [S26] is very similar with [S10]. In [S26], the paper had presented a contextual topic model for multi-document summarization in which per sentence is observed as hierarchical topics with regard to contextual knowledge, and has

shown how this model (hLDA) can be utilized to gain insight into various documents and use it to decide the sentence similarity. Though [S27] is similar to [S10, S26] but the greatest thing that they had introduced an extractive summarization procedure for novel documents. The approach was developed based on the LDA topic modeling algorithm, where under the demands of high concentration ratio and topic difference, the importance evaluation function of candidate sentences was invented to extract a machine summary for a novel document.

In [S11], they had worked for labeling newsgroup dataset. It was closer to [S6] for Ontology-based labeling method. Both of the studies have used LDA. The study also used NLP processes like [S2, S4, S8].

Similar type of work is also traced in study [S12, S73]. The authors of [S12] had introduced the clustering common topic from asynchronous text sequences. It was kind of work like [S1, S4]. [S73] is quite similar work of text categorization like [S12] topic clustering.

From our study where we find big data term firstly in [S13]. They used LDA topic modeling process for big data in social science. Also introduced some process for visualizing topics.

In [S14], this paper they had proposed the Constrained Latent Space Model (CLSM), which employs a multi-modal paradigm to simultaneously describe social network information and user behavioral data using a latent space representation. The latent space is inferred via Latent Dirichlet Allocation (LDA).

They had introduced a new method for topic modeling in [S15] called Regularized Latent Semantic Indexing (RLSI). In [S10] they also work with semantic relativity.

Study [S19] they had proposed a novel inexpensive document classification algorithm which requires minimal supervision. The algorithm was based on the generative property of LDA. It was slightly related to [S2, S4, S8, S11]. Study [S83] task is although same to [S19] but here the task was cluster scientific documents.

The authors of the paper had used topic modeling for mining Wikipedia data in [S22]. In this paper, they had proposed a novel approach, ML-LDA, to mine multilingual topics from Wikipedia.

An overview to topic modeling and its applications in bioinformatics had covered in [S79]. The studies had shown that a topic model can accomplish the task of clustering and classification of biological data.

In [S84], they had introduced a unique task for bibliometric analysis. They measured scholarly impact for publications.

The authors of the paper had introduced a new approach for vocabulary reduction in bag of words in [S86]. It was based on filtering words in the topic feature space instead of directly in the original word space.

Study [S87] had presented a set of algorithms for automatic annotation of metadata. The problem was discussed by two different methods. They improve poorly annotated metadata and recommend tag.

In [S88], the authors in this study worked with corpus and analyze the topic number, similarity and topics stability and predicted how many topics will be more appropriate and stable for the taken corpus.

In a set of study [S24, S34, S44, S68, S29, S70, S73], they had introduced a weakly supervised text classification algorithm founded on the generative field of unsupervised LDA. In their algorithm supervision comes in the kind of labeling of a few LDA topics.

Both study [S28, S41] were give an empirical implication that weighting words can improve topic models, where all term weighting schemes improve the basic models in some degree. Our CEW uses word co-occurrences to detect informative words, and it performs the best in most settings.

Study [S31] had introduced a new semi-supervised approach for sentiment analysis, which grow of the following steps: domain identification and sentiment analysis based on LDA background topic labeling, additionally, they present an automatic modification of the last one. [S92] have a little bit connection with [S31] because was also about sentiment analysis.

While topic modeling methods such as LDA, PLSA, and NMF are generally implemented in a variety of domains to separate unstructured text. In [S32, S39, S46], the papers they have described that for both methods, this can result in significant differences in the topics provided over multiple runs over the same corpus.

Study [S33, S47] they had studied two administered topic models, called FLDA and DF-LDA, for the multi-label report categorization task. They had estimated the offered models on the multi-label Yahoo. FLDA and DFLDA Dependency-LDA.

Both in study [S35, S45] they had introduced a weakly executed sentiment-topic model for short text representations. In this model LDA, they have shown a novel method for combined modeling sentiments and topic. [S90] is in extension work with short text by including word embeddings.

They had introduced a tag-topic model for blog mining based on the author-topic model AT model in [S36] and the model was already discussed on [S1]. In this model LDA, each is tag denoted by a probability distribution over topics.

Study [S37, S23, S54, S59] were similar to themselves. They had introduced an approach for topic modeling demanded to social media. The text preparing with 2 gram, tf-idf weights is more effective. The lemmatization does not give enough improvement since there are many misspellings in social media.

The authors introduced LDA guidelines for computer-assisted content analysis of e-suplications which involves close human supervision throughout the training and evaluating the process in [S38]. They inquired to automatically recognize latent topics from WtP suplications.

GPLSA was a unique model we have found in [S40], They had presented useful algorithms for graph regularized PLSA (GPLSA) to probabilistic topic analysis of both single- and multiple- modality data description. In GPLSA, topic proportions of a data entity are mapped to a graph and this comparison between topic creations on the graph are measured with divergences between discrete probabilities.

Study [S42] they had investigated a generalized topic model CSTM for short texts. We have already investigated that [S7, S9, S35, S45] are based on short text processings. Their target was capture the background noise words by prefacing a new letter of topic, namely common topic.

The article [S43] they had introduced a content based tag instruction model Similar Word. The common idea of Similar Word is to build use of the tag-content importance phenomenon that they have empirically verified. Some special cases of Similar Word are also studied.

In [S48], this study they had introduced an application of two well organized unsupervised topic models (LDA and MM) for the task of text segmentation. Out of the two strategies proposed in this document, the LDA based method was able to measure the segment frames with higher efficiency as related to the limits expected by the MM based method. Study [S85]



also a paper based on text segmentation this paper described a system which uses entity and topic coherence for improved Text Segmentation (TS) accuracy. Linear Dirichlet Allocation (LDA) algorithm was used to obtain topics for sentences in the document.

Study [S56, S52, S59, S49, S50, S51, S57, S60, S65, S71, S61, S25, S66, S67] were very similar where they had expressed latent Dirichlet allocation, a bending generative probabilistic model for sets of discrete data. LDA is founded on a simplistic exchangeability opinion for the words and topics in a document; it is since obtained by a candid application of the Finetti's description theorem. They had view LDA as a dimensionality reducing technique, in the quality of LSI, but with individual underlying generative probabilistic semantics that complete sense for the type of data that it models.

They had performed A-EPSGLD, an attempt of integrating two distributed counting paradigm, i.e. parameter server (PS) and embarrassingly parallel (EP), for large-scale LDA inference in [S53].

In [S55], they had proposed the Word-Topic Mixture (WTM) model trains word embeddings and topic model together based on LDA and word embeddings. WTM links the ideas of TWE and LFLDA. It first uses LDA to catch the word-topic responsibility and includes external corpus as the words semantic supplement into TWE to determine topic vectors and word embeddings.

Both in study [S62, S30], they had introduced a massive amount of unstructured data in the kind of documents, blogs, tweets etc., is continuously existing produced in the world. As a result, today the problem is not about the unavailability of data, instead, it is about abundant data, which is giving the task of obtaining specific information involved.

In [S64], Generative models for text, such as the topic model, have the potential to make significant augmentations to the statistical analysis of massive document sets, and the evolution of an extensive perception of human language learning and processing. Topic models demonstrate how using a separate description can provide new insights into the statistical modeling of language, including several of the key assumptions behind LSA.

They had presented a comparison between topic coherence scores and human topic ranking when creating LDA topics from abstract and full-text data in [S63].

The recommended system means to collect news data from such different sources, capture the different studies, summarize, and present the news. In [S69] involves knowing topic from

real-time news removals, then make clustering of the news documents based on the topics. Previous procedures, like LDA, know topics efficiently for long news texts.

### **3.2.2 Search Scope Results**

Table 3.1 confers a total of 3021 papers as the outcome of the initial search manner, the number of papers as each electronic database, and including the percentage that it signifies. It can be mentioned that Springer, Semantic Scholar, Science Direct and ACM DL delivered the highest collections of papers. In these databases, we found various papers from different fields that were not associated with our study. The IEEE Xplore database delivered just 319 papers, Springer database returned 412 papers but a majority of them were relevant to the study, so this was as a result more efficient as compared to the remaining databases. Concerning the Google Scholar database, the number of relevant papers was 1237, in which most of the delivered papers were repeated in the six main databases.

### **3.2.3 Time Period Results**

Figure 3.3 shows the frequency of associated papers upon the time period from 2003 to 2018. During the years 2004,2005 and 2008 no papers were published in journals and proceedings, book chapter. From 2003 to 2008, the number of published papers is less than 3 per year, increasing slowly. As Figure 3.3 shows that, from 2009 to 2015 the increase was much high. From 2010 until the end date of this study (December 2018), there has been a huge dap compared with the years earlier 2015, with a peak in the last 3 years, specifically in 2018.

### **3.2.4 Publication Type Results**

The picked 93 papers were issued as a journal, conferences, articles, books, or book chapters. Figure 3.1 shows the distribution of selected papers above the publication species and days.

Table 3. 1: Ordination of picked studies above electronic database

DB source	No. of papers	Ratio	Relevant papers	Effectiveness
Science Direct	292	9.7%	57	19.5%
IEEE Xplore	319	10.6%	113	35.4%
ACM DL	473	15.7%	157	33.2%
JMLR	158	5.2%	43	27.2%
Springer	412	13.6%	16	3.9%
Semantic Scholar	130	4.3%	27	20.8%
Google Scholar	1237	40.9%	236	19.1%
Total	3021	100%		

Published in proceedings as follows:

(51) conferences, (31) journals, (6) articles, (1) book, (4) book chapters and 92% (82 studies) of the selected papers were published in prominent journals and conferences while hardly 1% were publications as book.

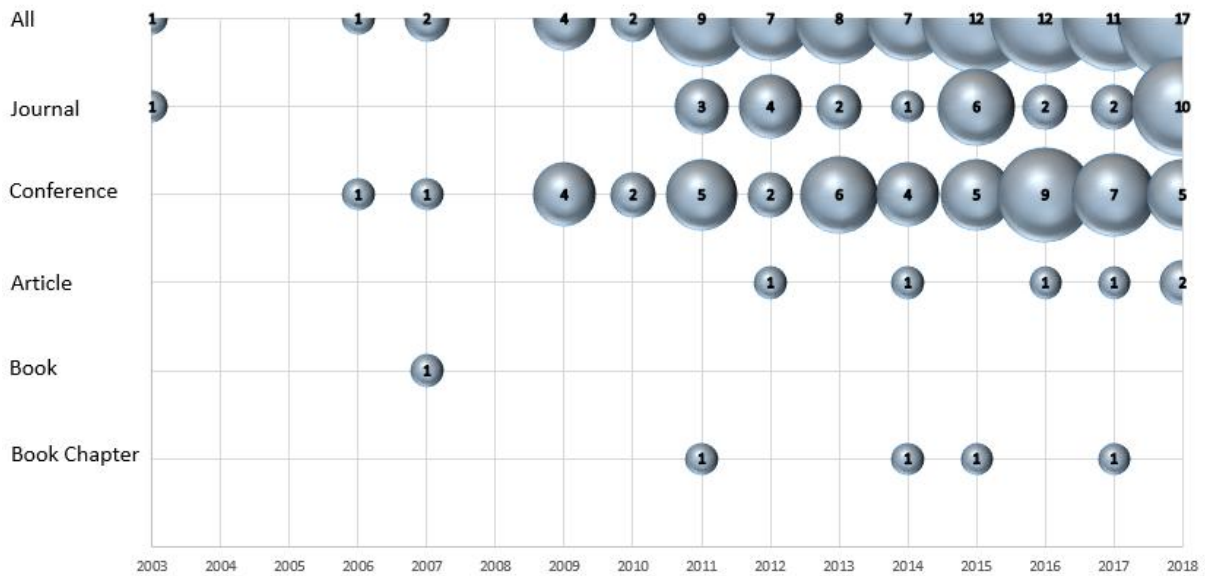


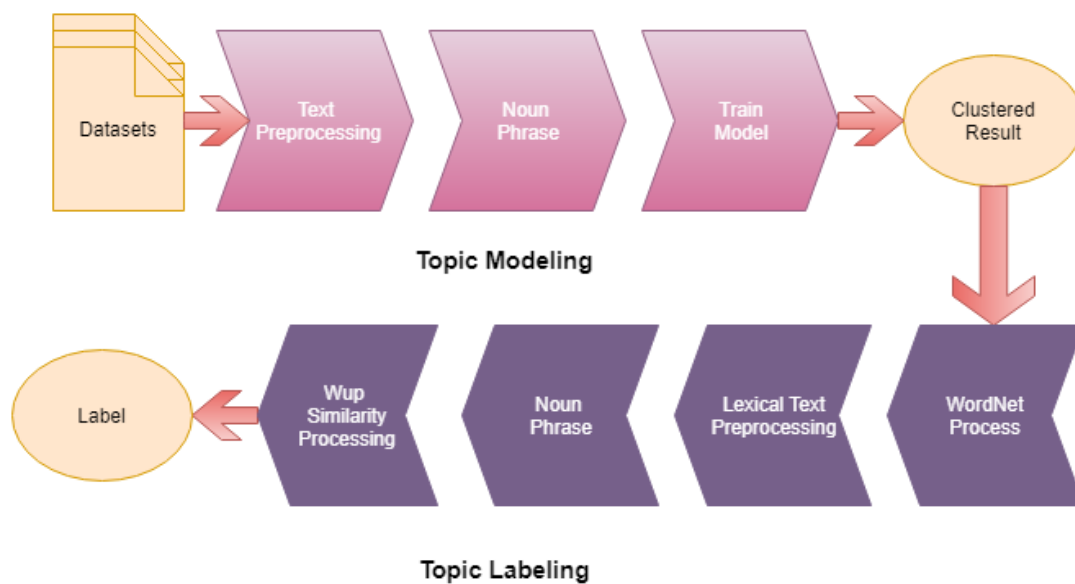
Figure 3. 3: Ordination of pertinent papers above time span each publication species

## CHAPTER 4

### RESEARCH EXPERIMENT

In this section, we have described the overall process of our research work.

First of all, we have set select our dataset. For dataset, we have chosen some online document to complete our experiment process, we have done work we have to cross a lot of process for example step by step pre-processing, noun phrase separation, training model, label processing with the help of WordNet. Then we acquire to find out topic label based on our topic model result.



*Figure 4 1: Overview of our research workflow*

#### Overall process of topic labeling

In This Section, The whole process is described in this Figure 4 2: Process of topic and label generation.

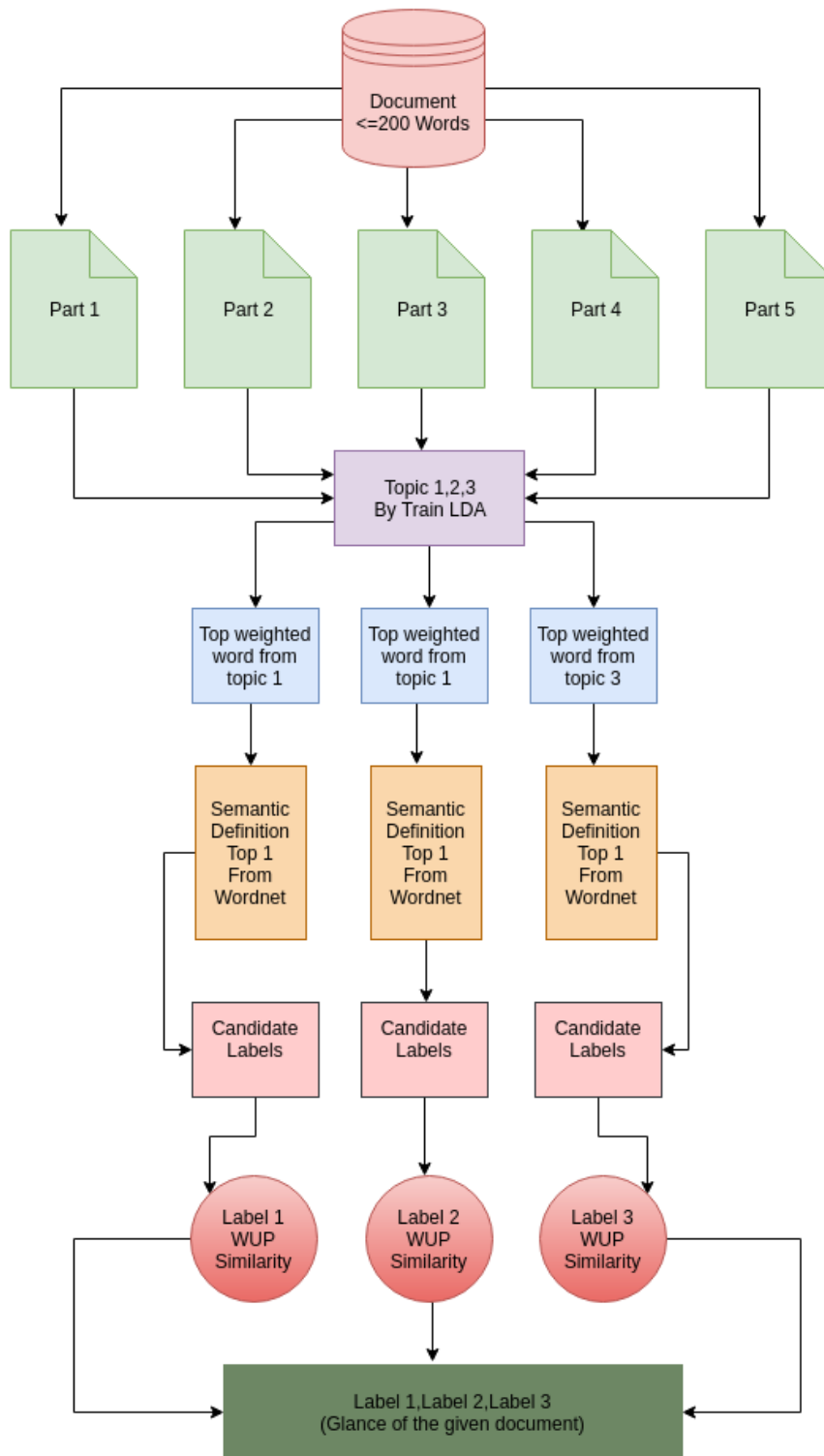


Figure 4 3: Process of topic and label generation

## 4.1 Text Preprocessing

Quantitative resolution demands that we modify our documents within numerical data. Allowing the reason that word sequence may be influenced of with minimum costs for thought (see Grimmer and Stewart, 2013, for discussion) and a 'bag of words' description employed. Research is typical practice (some subset of) any more binary preprocessing (Denny et al., 2018) moves in constructing the appropriate document-term matrix. We now explain these in few details, since certain are the focus of our research work.

### 4.1.1 Tokenizing Text

Tokenization (Manning et al., 2014) is a measure which divides larger strings of text into smaller parts or tokens. Massive chunks of text can be tokenized within sentences; sentences can be tokenized within words. Similarly, as processing is usually achieved behind a portion of text has been properly tokenized. Sometimes segmentation is used to transfer to the categorization of a large chunk of text within pieces greater than words, while tokenization is held for the analysis process which results particularly in words.

Tokenize process

```
tokenizer = RegexpTokenizer(r'\w+')  
raw = i.lower()  
tokens = tokenizer.tokenize(raw)
```

### 4.1.2 Stop words Removing

After complete tokenization, we need to remove stop words as because we do not need all words at all.

Stop word removing process

```
en_stop = get_stop_words('en')
stopped_tokens = [i for i in tokens ifnot i in en_stop]
```

Stop words (Denny et al., 2018) are mean a set of commonly used words in any language, not just English. The purpose of processing stop words are significant to various applications is that we separate the words that are very regularly applied in a given language, we can adjust on the relevant words instead. For example, if we search any query as “how to develop topic modeling with labeling”, If the search engine decides to find web pages that received the articles “how”, “to” “develop”, “topic”, “modeling”, “with”, “labeling” the search engine is working to obtain a lot of extra pages that contain the articles “how”, “to”, “with” than pages that receive information around topic modeling and labeling because the articles “how”, “to” and “with” are so regularly used in the English language. So, if we ignore these two articles, the search engine can really focus on topic modeling pages that receive the keywords: “develop” “topic” “modeling” “labeling” – which would more nearly bring up pages that are actually of interest. This is just the basic concern for inspiration stop words.

### 4.1.3 Lemmatizing words

Lemmatization (Toman et al., 2006) is the classification mutually of various classes of the same word.

Lemmatizing process

```
l_lemma = WordNetLemmatizer()
lemmatize_tokens = [l_lemma.lemmatize(i) for i in stopped_tokens]
```

In exploration queries, lemmatization (Toman et al., 2006) enables end users to query each variant of a base word and get appropriate results. Because search engine algorithms utilize lemmatization (Toman et al., 2006), the user is easy to query each inflexional form of a word and get appropriate results. For example, if the user queries the plural custom of a word

(works, cars, cats, better), the search engine identifies and return appropriate content that relates the singular form of the identical word (work, car, cat, good).

## 4.2 Noun Phrase Choosing

After preprocessing, we only pick the noun and proper noun from the preprocessing result. Using this strategy, the topic is obtained by close top nouns with the largest frequency.

Noun phrase choosing phrase

```
nn_tagged = [(word,tag) for word, tag in a if tag.startswith('NN') or tag.startswith('NNP')]
```

The moves involved are as follows: First, tokenization of text is implemented to lemma out the words. The tokenized text is then tagged with parts of speech (NN (nouns), NNP (proper nouns), VB (verbs), JJ (adjectives) etc.) before lemmatize and stop-words removal as Parts-of-Speech(POS) tagging is flow labeling process and trust on word order. Therefore, removing stop-words results in equivocality and will lose the necessary information expected by POS tagger. The stop-words are removed after POS tagging. In the final stage, words including with their tags and frequencies are put in a hash table and most solid nouns (Sajid et al., 2017) are extracted from those to create a heading for a text.

## 4.3 Training LDA Model for Generating Topic Set

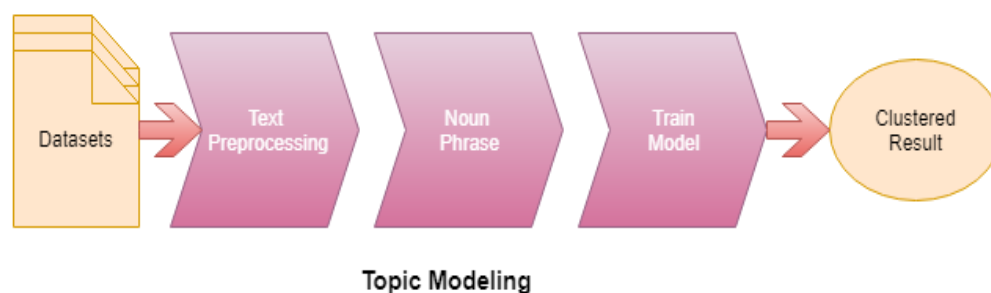


Figure 4 4: Training process of model



While training an LDA (Blei et al., 2003) model, we need to start with a set of documents and any of these is expressed by a fixed-length vector (bag-of-words). LDA (Blei et al., 2003) is a common Machine Learning (ML) technique, which indicates that it can also be practiced for other unsupervised ML problems where the input is a set of fixed-length vectors and the aim is to traverse the building of this data.

Latent Dirichlet Analysis is a probabilistic model, and to obtain cluster assignments, it uses two probability values:  $P(\text{word} | \text{topics})$  and  $P(\text{topics} | \text{documents})$ . These values are determined based on an initial random distribution, after which they are reproduced for the specific word in the specific document, to determine their topic distribution. In an iterative method, these probabilities are determined multiple times, until the convergence of the algorithm.

To perform an LDA (Blei et al., 2003) model, you first begin by determining the number of 'topics' that are started in your set of documents. Now we will show the model output below:

Here we take number of topics =3 and Number of Words =2

*Table 4. 1: Result of an example document*

<b>Topic 1</b>	<b>Topic 2</b>	<b>Topic 3</b>
0.033*"sweet"	0.094*"brother"	0.075*"health"
0.033*"brother"	0.094*"sweet"	0.043*"may"

Example document set:

“Sweet dangerous to eat. My brother likes to have sweet, but not my mother. My mother consumes a lot of time pushing my brother about to dance exercise. Doctors recommend that driving may produce improved stress and blood pressure. My father never seems to drive my brother to do better. Health experts say that sweet is bad for your health.”

### 4.3.1 Contention Behind to Fix Pass Parameter Property Eight of LDA Model

When we have trained our documents with our LDA (Blei et al., 2003) model we go through some experiences.

We see that passing parameter which bypasses document through the model and it has a big contribution for extracting topics and words behind topics.

It exactly makes changes in our output every time. When we were started to getting our topics and words then it makes sense to us. It affects every dataset we have.

For an experiment, we choose here document 1.

*Table 4. 2: Result of an example document with various pass*

<b>Passing times</b>	<b>Topics</b>		<b>Execution time</b>
1	Topic 1	jam, traffic	5.537 sec
	Topic 2	traffic, cause	
	Topic 3	cause, traffic	
5	Topic 1	road, jam	8.491 sec
	Topic 2	traffic, rule	
	Topic 3	cause, vehicle	
6	Topic 1	road, jam	7.382 sec
	Topic 2	traffic, rule	
	Topic 3	cause, vehicle	

7	Topic 1	road, jam	6.940 sec
	Topic 2	traffic, rule	
	Topic 3	cause, vehicle	
8	Topic 1	road, jam	6.530 sec
	Topic 2	traffic, rule	
	Topic 3	vehicle, cause	
10	Topic 1	road , jam	7.090 sec
	Topic 2	traffic, rule	
	Topic 3	vehicle, cause	
5000	Topic 1	road , jam	25.583 sec
	Topic 2	traffic, rule	
	Topic 3	vehicle, cause	
10000	Topic 1	road , jam	78.450 sec
	Topic 2	traffic, rule	
	Topic 3	vehicle, cause	

Training process of model

```
dictionary = corpora.Dictionary(clean words)
corpus = [dictionary.doc2bow(text) for text in l]
ldamodel = gensim.models.ldamodel.LdaModel(corpus, num_topics=3, id2word =
dictionary, passes=8,random_state=1)
```

Dictionary makes a single list of cleaned words for build corpus which is used to make the bag of words. Bag of words creates a unique id for each word for every individual topic set and count the frequency of that word within the topic.

When we have fixed passes=8 maximum number of iterations allowed to LDA (Blei et al., 2003) algorithm and it pulling out better result with proper topics and words distribution and also run time and result are better and meaningful than before, like passes =1/5/6/7. Because we have experienced passes=8 is the pick point of our required document like 200 words documents. If we set passes =8 then we get standard topics with belonging words and also we get our result within sort execution time. If we use more than passes=8 like passes=10/5000/1000 then we get the same result but takes unnecessarily long execution time where there is no need for that. That is why we have chosen passes=8.

We also see that LDA (Blei et al., 2003) model randomly pick frequent words and make topics. So, if we do not set random\_state=1 it makes changes model output every time for our document. But this is not our headache this time. So, random\_state=1 is enough for our required document.

#### **4.3.2 Parameters We Have Used in LDA Model**

- **corpus**–Stream of text vectors or rare matrix of the pattern.
- **num\_topics**– The number of inquired latent topics to be selected from the training corpus.
- **id2word** – Charting from word IDs to words. It is used to define the vocabulary size, as well as for debugging and topic print.
- **passes** – The Highest number of iterations passed to LDA algorithm for convergence.
- **random\_state**– Either a random State object or an each to create one. Beneficial for reproducibility.

#### **4.3.3 Reasons Aft to Choose In between Two Thousand Words Document and Set as Topic Property Three and Word Property Two Behind Each Topic**

We know the very short text is sparse and noisy and very bigger are tough. It also requires high hardware computer to execute the model that we do not have at this time. But if we choose short article approximately around 200 words which words are very relevant with each other it is very exquisite to find proper topics and words behind. From our understanding, the heart of topic modeling techniques is co-occurrence of terms like we have

used LDA (Blei et al.,2003) model for that. For our experiment, here we only work in between 200 words document and choose only three topics and at first, we take two words under each topic. Because if we take more topics for as our document requirements we get same topics and words again that is called redundant problem.

Suppose look for an example for document 2.

When we train our model by passing our dataset with four topics and three words we get output like below-

*Table 4. 3: Result of an example document with topic four*

Topic 1	love, yes, life
Topic 2	year, life, nonsmoker
Topic 3	cigarette, brush, change
Topic 4	life, year, yes

But as we propose our work as passing 3 topics and two words with the dataset we get output like below-

*Table 4. 4: Result of an example document with topic three*

Topic 1	love, yes
Topic 2	life, year
Topic 3	cigarette, brush

Here, for first approach we see that “yes”, “life”, “year” are redundant but for second approach we do not get any redundancy as before. We could also train model with two topics but for getting adequate topics and words we set parameter topic as three and word as two. It clearly clarifies that, within 200 words documents we need to assure topic and word

parameter not over three topics and two words because, it creates topic and word redundancy as we see above for first approach of document 2.

When we get topics from the documents we only take top weighted word from the topic because its impact is robust for its topic set. Then we search the semantic definition from the lexical database for English parts of speeches which is called WordNet (Miller et al.,1995). We select initial definition because that is most appropriate within its terms. After preprocessing the definition, we get the candidate labels and from these candidate labels we measure the candidate labels with the main topic word for conceptual semantic relatedness measurement by WUP (Wu et al., 1994) similarity. Then actual generic label for each topic come out.

#### 4.4 Indexing Topic Set

Indexing topic set process

```
output = ldamodel.print_topics(num_topics=3, num_words=2)
topic1 = output[0][1]
topic2 = output[1][1]
topic3 = output[2][1]
```

In this part, we separate each topic and then measure the top weighted value and set these words behind their topics through 2d matrix. Cause model is not able to provide polite result what we need. Utmost we get topic 1, topic 2 and topic 3 as our expectation.

#### 4.5 Picking Top Weighted Words from Topic Sets

In this section, we have picking up our top valuable word for each topic. But this section has one major problem. Trained model provide output which are in complex string format in a lists. Like-

```
[("0.033*sweet+0.033*brother"), .....]
```

Then we split the string value from the list-

```
"0.033*sweet', '+', '0.033*brother"
```

Then we select the largest weighted words from our topic dataset but we do not need the score with we just need largest weighted word. That's why we use regular expression in our when processing. Then we find out the top weighted words as its root topic.

Picking word process

```
tv2 = t1.split()[0]
tv3 = " ".join(re.findall("[a-zA-Z]+", tv2))
tv4 = re.split("^[a-zA-Z]*", tv3)
best_word = ""

for item in tv4:
    best_word = str(item)
```

#### 4.6 WordNet Processing

WordNet (Miller et al.,1995) is a great lexical database of English. Nouns, verbs, adjectives and adverbs are classified into sets of cognitional synonyms (synsets), each meaning a distinct idea. Synonyms are interlinked by means of conceptual-semantic and lexical relationships. The turn network of meaningfully similar words and ideas can be operated with the browser. WordNet is also easily and openly accessible for download. WordNet is building makes it a valuable tool for computational philology and natural language processing.

WordNet (Miller et al.,1995) partially relates a dictionary, in that it classifies words mutually based on their suggestions. However, there are any significant differences. First, WordNet (Miller et al.,1995) interlinks not just word makes strings of words but special functions of words. As a result, words that are seen in near concurrence to one extra in the system are semantically disambiguated. Second, WordNet (Miller et al.,1995) specifies the semantic relationships between words, whereas the classification of words in a dictionary does not match any specific decoration other than determining the identity.

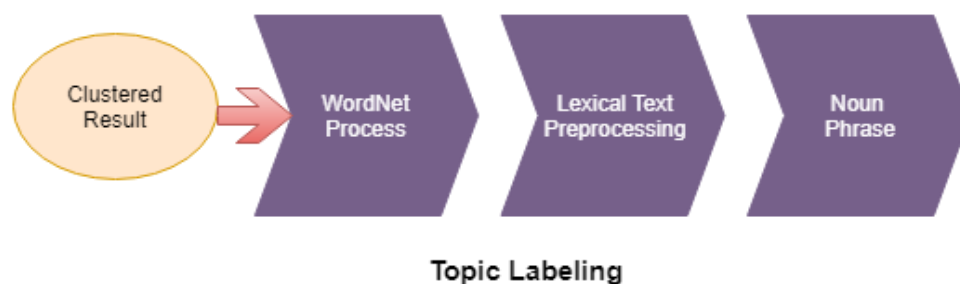
Get definition from WordNet

```
syns = wordnet.synsets("best_word")  
word_def = syns[i].definition()
```

In our topic modeling cluster result in our largest valuable words and next works in WordNet term. This WordNet term gives a word definition in our selected word. Suppose our selected word is “Sweet” then WordNet synset gives a definition are below:

S: (n) dessert, sweet, after (a dish served as the last course of a meal)

Then we started again preprocessing in our WordNet definition and also pick up the noun and proper noun phrase.

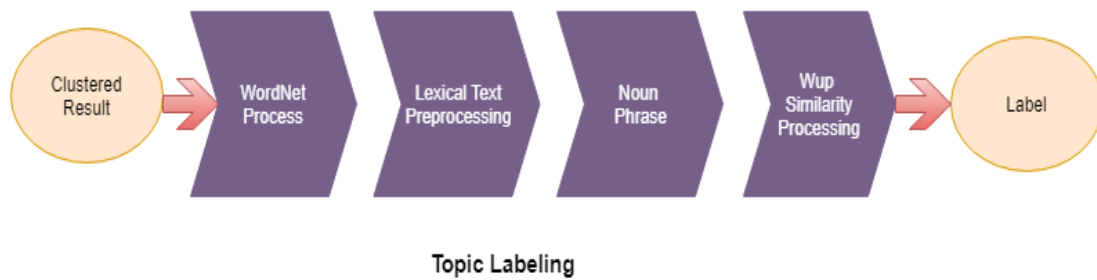


*Figure 4 5: WordNet Processing for label*

#### **4.7 WUP Similarity Checking for Choosing Labels**

When the WordNet (Miller et al.,1995) process is finished. Then we started in our WUP similarity process for labeling. In this section, WUP (Wu et al., 1994) similarity process gives labels for our topics.





*Figure 4 6: WUP similarity process for labeling*

Wu & Palmer (WUP) – Words Similarity:

Wu & Palmer (WUP) is a module for computing semantic relatedness of word senses by applying the edge calculating process of the Wu & Palmer (Wu et al., 1994).

The WUP measures the relation by considering the depths of two synsets behind the WordNet taxonomies, along together with the depth of the least common subsumer(lcs).

$$\text{The formula score} = \frac{2 * \text{depth}(lcs)}{(\text{depth}(s1) + \text{depth}(s2))} \quad (4.1)$$

That forecasts it as  $0 < \text{score} \leq 1$ . The score can not be zero as the depth of the lcs is nowise zero hither in the depth of the root of taxonomy is always one. The score exhibit one when the two different input assumptions are aspects like same.

### Processes of WUP:

1. Initialize the two things, one is WordNet Database and another WuPalmer object.
2. Fix MFS as true. It does Most Frequent Sense(MFS). MFS extends calculation to speed above.
3. Acquire the synsets because of input words as per their POS.
4. Repeat two synsets to measure relatedness score of synsets between them.
5. Deliver maximum score for the synsets.

#### *WUP similarity check*

```
word1 = wordnet.synset(best_word)
word2 = wordnet.synset(candidate_label)
score = word1.wup_similarity(word2)
```

Which word gives the top score in WUP (Wu et al., 1994) similarity we chose it as label belonging our query word.

## CHAPTER 5

### RESULT ANALYSIS AND DISCUSSION

#### 5.1 Recall, Precision, F-measure

A confusion matrix is needed for our predicted result analysis. So now we are going to discuss on recall, precision, and f-measure (Makhoul et al.,1999).

A confusion matrix is one kind of table what is usually practiced to represent the appearance of a classification model upon a set of dataset set for which the true values always remain known. All the measures are assumed by using leftmost four parameters. Thus, let's discuss concerning these four parameters.

Table 5. 1: Confusion matrix classes and parameters

	Predicted class		
Actual Class		Class = Yes	Class = No
	Class = Yes	True Positive	False Negative
	Class = No	False Positive	True Negative

True positive and true negatives are the two parameter which measurements that are precisely calculated and hence presented in green. We desire to decrease false positives and false negatives. Therefore, they are presented in red color. Those terms are a bit baffling. So let's get every term one after one and catch on it entirely.

True Positives (TP)  $\Rightarrow$  Those are precisely predicted positive values as means that the value of the exact class is yes and the value of the predicted class is also yes. For example, if the exact class value symbolizes that this will occurred and predicted class mentions the very same information.

True Negatives (TN)  $\Rightarrow$  Those are precisely predicted all negative values as it means that the value for the exact class is no and for the predicted class is also no. For example, if exact class value symbolizes that this will not occur and predicted class mentions the very same information.

False positives and false negatives, those values happen when original class denies with the predicted class.

False Positives (FP)  $\Rightarrow$  While the original class is no and predicted class is yes. For example, if original class value symbolizes that this will not occur and predicted class mentions that will occur.

False Negatives (FN)  $\Rightarrow$  While the original class is yes but predicted class in no. For example, if original class value symbolizes that this will occur and predicted class mentions that will not occur.

From this four parameters, we can measure Recall, Precision and F-measure for our prediction.

**Precision**  $\Rightarrow$  It dictate the proportion of perfectly predicted positive audit to the entire predicted positive audit. The precision is the proportion of the quantity of relevant items discovered over the entire quantity of items obtained.

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

$$= \frac{True\ Positive}{Total\ Predicated\ Positive}$$

So the term is actually looks like -

$$Precision = \frac{| \{Relevant\ Items\} \cap \{Retrieved\ Items\} |}{| \{Retrieved\ Items\} |} \quad (5.1)$$

**Recall**  $\Rightarrow$  It dictate the proportion of precisely predicted positive audits to all audits in actual class - yes. The recall is the proportion of the amount of relevant items obtained across the entire amount of associated items.

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

$$= \frac{True\ Positive}{Total\ Actual\ Positive}$$

So the term is actually looks like-

$$Recall = \frac{|{\{Relevant\ Items\}} \cap |{\{Retrieved\ Items\}}|}{|{\{Relevant\ Items\}}|} \quad (5.2)$$

**F-Measure**⇒ F-measure is the calculation behind the weighted average of Precision and Recall. Accordingly, the aforementioned score tackles both false positives and false negatives inside account. Intuitively that is not so facile to catch as accuracy, but F-measure is habitually very useful than accuracy, wonderfully in case you have a rough class allocation. Accuracy runs great if false positives and false negatives have comparable cost. In case the cost of false positives and false negatives are very several, it is good to eye at both Precision and Recall.

$$F\text{-measure} = \frac{2(Precision * Recall)}{Precision + Recall} \quad (5.3)$$

So, as we build a model, this measure helps us to find out what these parameters actually mean and how well our model has attained.

Table 5. 2: Label of selected 5 documents

Documents	Topics		Top Weighted Word	Candidate Labels	Label
Document 1	Topic 1	road, jam	road	way, travel, transportation	transportation
	Topic 2	traffic, rule	traffic	aggregation, thing, vehicle, locality, period, time	aggregation
	Topic 3	vehicle, cause	vehicle	conveyance, transport, people	transport
Document 2	Topic 1	love, yes	love	emotion, regard, affection	emotion
	Topic 2	life, year	life	state, mode, living	mode
	Topic 3	cigarette, brush	cigarette	ground, tobacco, paper, smoking	tobacco
Document 3	Topic 1	child, childhood	child	person, sex	person
	Topic 2	work, event	work	activity, something	activity
	Topic 3	labour, life	labour	class, labor, work, wage	labor
Document 4	Topic 1	role, lesson	role	action, activity, person, group	activity
	Topic 2	mother, heart	mother	woman, birth, child, term, address, mother	mother
	Topic 3	child, love	child	person, sex	person
Document 5	Topic 1	school, activity	school	institution	institution
	Topic 2	work, excursion	work	activity, something	activity
	Topic 3	student, thing	student	learner, institution	learner

In Table 5.2 we have labeled the documents. First we select candidate key. Then select final label measure by WUP.

*Table 5. 3: Recall, Precision and F-measure of selected 5 documents*

Documents	Topics	Recall	Precision	F-measure
Document 1	Topic 1	75%	50%	.6
	Topic 2	86.714	54.54%	.667
	Topic 3	75%	75%	.75
Document 2	Topic 1	75%	60%	.667
	Topic 2	75%	75%	.75
	Topic 3	80%	66.66%	.727
Document 3	Topic 1	66.667%	50%	.571
	Topic 2	66.66%	40%	.5
	Topic 3	80%	57.143%	.667
Document 4	Topic 1	80%	57.143%	.667
	Topic 2	100%	66.667%	.80
	Topic 3	66.667%	50%	.571
Document 5	Topic 1	50%	50%	.5
	Topic 2	66.667%	40%	.53
	Topic 3	50%	50%	.5

Meanwhile, we have already described recall, precision, and f-measure. Here for calculation of recall, precision and f-measure we take words after preprocessing as reference dataset behind each topic. And we need the second dataset to compare with it. So we take candidate

labels what we get after lemmatization of reference dataset and given topic word because they have nexus with each other.

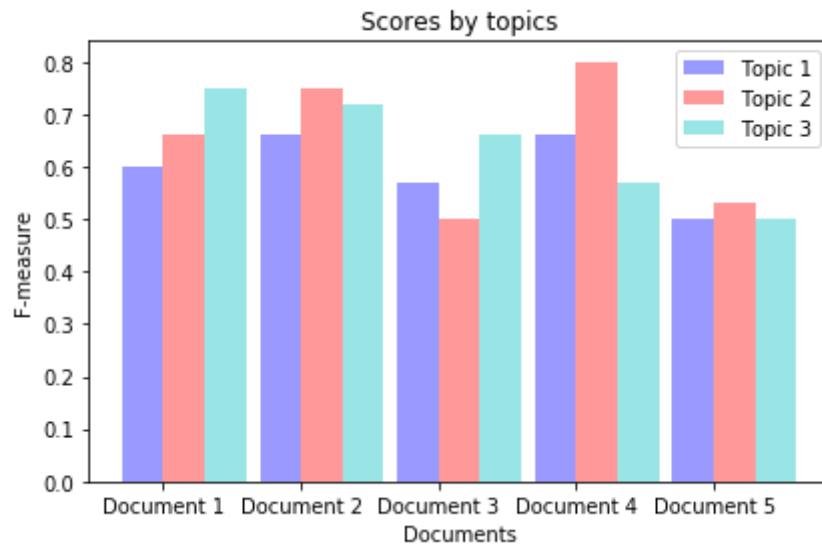


Figure 5. 1: F-measure score for topics of selected 5 documents

## 5.2 WUP Similarity

Here we show the score between our topic and label on basis of lexical semantic WUP (Wu et al., 1994) similarity. This table explain matching score between topics with label. WUP present good accuracy.

Table 5. 4: WUP similarity between topic and label

Documents	Topics	Label	WUP similarity	Average WUP
Document 1	Road	Transportation	0.714	0.845
	Traffic	Aggregation	0.888	
	Vehicle	Transport	0.933	
Document 2	Love	Emotion	0.923	0.789
	Life	Mode	0.545	
	Cigarette	Tobacco	0.9	
Document 3	Child	Person	0.75	0.891
	Work	Activity	0.923	
	Labour	Labor	1.0	
	Role	Activity	0.8	



Document 4	Mother	Mother	1.0	0.85
	Child	Person	0.75	
Document 5	School	Institution	0.857	0.828
	Work	Activity	0.923	
	Student	Learner	0.705	

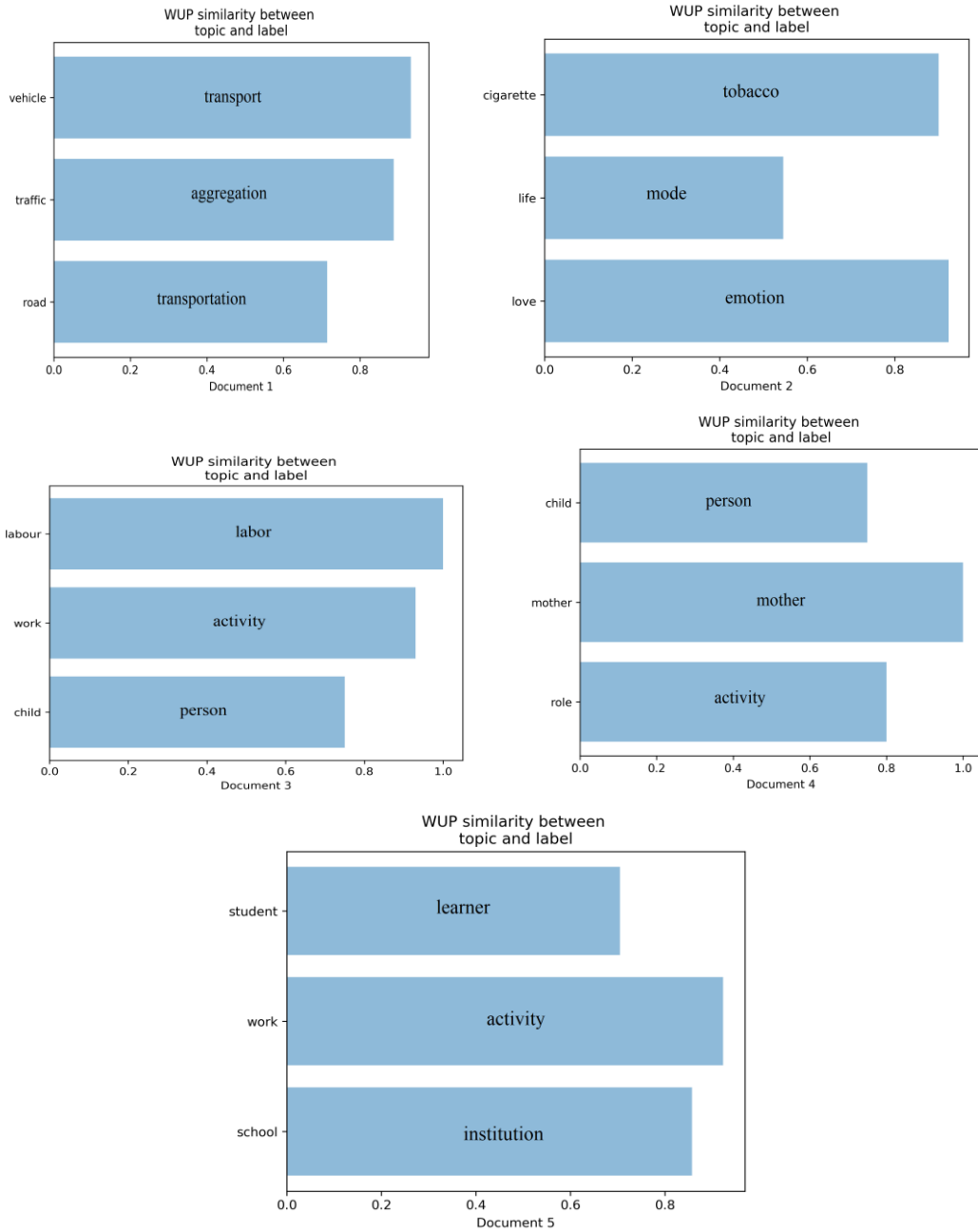


Figure 5. 2: WUP similarity between topic and label

In figure 5. 2, WUP similarity between topic and label shown the score through the horizontal graph.

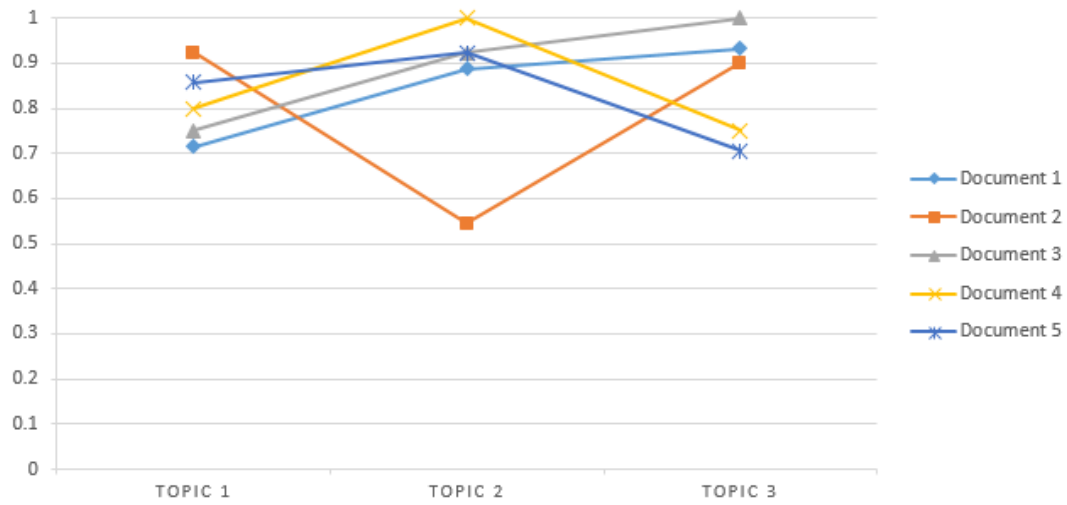


Figure 5. 3: WUP similarity score of Documents

In figure 5. 3, WUP similarity between Score of document is shown through line graphs.

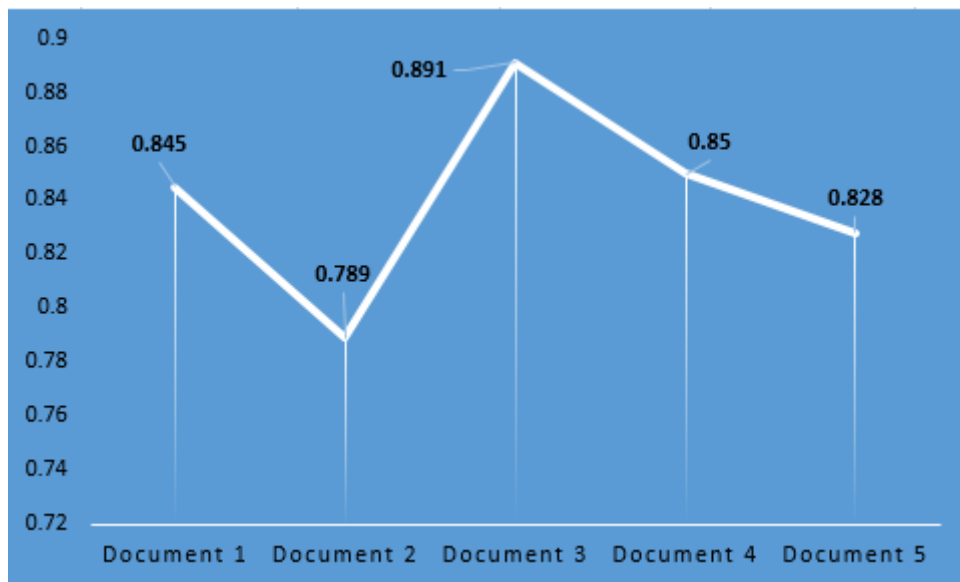


Figure 5. 4: Average WUP similarity score of documents

In figure 5. 4, Average WUP similarity between Score of document is shown through line graphs.

## **CHAPTER 6**

### **CONCLUSION**

#### **6.1 Our Contribution**

This work has suggested a unique mechanism for genetic labeling detection the topic of the text document by high weight words result and WordNet WUP similarity scores. It has been observed that our proposed method approaches are application to find out the appropriate topic label for the polynomial topic over text in between 200 words and extract the high weight topic word with description using WordNet that can concisely convey the generic label for a document. The results shown that the Noun phrase approach is better within the unique mechanisms as it gives the most relevant candidate labels. It can be concluding that the most relevant and suitable word are Nouns for WUP similarity description for choosing word as a label.

#### **6.2 Future Works**

We have done this experiment on short document having word count approximate 200 .In future we will do this for large data set.

## REFERENCES

- Approaches to strategic alignment of software process improvement: a systematic literature review. *Journal of Systems and Software*, 123, 45–63.
- Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent dirichlet allocation." *Journal of machine Learning research* 3.Jan (2003): 993-1022.
- Deerwester, Scott, et al. "Indexing by latent semantic analysis." *Journal of the American society for information science* 41.6 (1990): 391-407.
- Denny, Matthew J., and Arthur Spirling. "Text preprocessing for unsupervised learning: why it matters, when it misleads, and what to do about it." *Political Analysis* 26.2 (2018): 168-189.
- Dyba, Tore, Torgeir Dingsoyr, and Geir K. Hanssen. "Applying systematic reviews to diverse study types: An experience report." *Empirical Software Engineering and Measurement*, 2007. ESEM 2007. First International Symposium on. IEEE, 2007.
- Feuerriegel, Stefan, Antal Ratku, and Dirk Neumann. "Analysis of how underlying topics in financial news affect stock prices using latent dirichlet allocation." 2016 49th Hawaii International Conference on System Sciences (HICSS). IEEE, 2016.
- Gildea, Daniel, and Thomas Hofmann. "Topic-based language models using EM." *Sixth European Conference on Speech Communication and Technology*. 1999.
- Golub, Gene H., and Christian Reinsch. "Singular value decomposition and least squares solutions." *Numerische mathematik* 14.5 (1970): 403-420.
- Guo, Yue, Stuart J. Barnes, and Qiong Jia. "Mining meaning from online ratings and reviews: Tourist satisfaction analysis using latent dirichlet allocation." *Tourism Management* 59 (2017): 467-483.
- Hansen, Per Christian, Takashi Sekii, and Hiromoto Shibahashi. "The modified truncated SVD method for regularization in general form." *SIAM Journal on Scientific and Statistical Computing* 13.5 (1992): 1142-1150.
- Hofmann, Thomas. "Probabilistic latent semantic analysis." *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc., 1999.
- Kitchenham, B., Charters, S. (2007). Guidelines for performing systematic literature reviews in software engineering. In *Tecnical. Report EBSE-2007-01*, Keele University.
- Laguna, M. A., & Crespo, Y. (2013). A systematic mapping study on software product line evolution: from legacy system reengineering to product line refactoring. *Journal of Science of Computer Programming*, 78(8), 1010–1034.
- Lienou, Marie, Henri Maitre, and Mihai Datcu. "Semantic annotation of satellite images using latent Dirichlet allocation." *IEEE Geoscience and Remote Sensing Letters* 7.1 (2010): 28-32.
- Makhoul, John, et al. "Performance measures for information extraction." *Proceedings of DARPA broadcast news workshop*. 1999.

- Manning, Christopher, et al. "The Stanford CoreNLP natural language processing toolkit." Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations. 2014
- Manning, Christopher, et al. "The Stanford CoreNLP natural language processing toolkit." Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations. 2014.
- Miller, George A. "WordNet: a lexical database for English." Communications of the ACM 38.11 (1995): 39-41.
- Novais, R. L., Torres, A., Mendes, T. S., Mendonça, M. G., & Zazworka, N. (2013). Software evolution visualization: a systematic mapping study. Information & Software Technology, 55(11), 1860–1883.
- Pérez, J., Moha, N., Mens, T. (2011). A classification framework and survey for Design Smell management. In Technical report. 2011/01, GIRO Research Group, Departamento de Informática, Universidad de Valladolid.
- Pinoli, Pietro, Davide Chicco, and Marco Masseroli. "Latent Dirichlet allocation based on Gibbs sampling for gene function prediction." Computational Intelligence in Bioinformatics and Computational Biology, 2014 IEEE Conference on. IEEE, 2014.
- Sajid, Anamta, Sadaqat Jan, and Ibrar A. Shah. "Automatic Topic Modeling for Single Document Short Texts." Frontiers of Information Technology (FIT), 2017 International Conference on. IEEE, 2017.
- Salton, Gerard, and J. Michael. "McGill. 1983." Introduction to modern information retrieval (1983).
- Toman, Michal, Roman Tesar, and Karel Jezek. "Influence of word normalization on text classification." Proceedings of InSciT 4 (2006): 354-358.
- Valle, Denis, et al. "Extending the Latent Dirichlet Allocation model to presence/absence data: A case study on North American breeding birds and biogeographical shifts expected from climate change." Global change biology 24.11 (2018): 5560-5572.
- Vasconcellos, F. J., Landre, G. B., Cunha, J. A. O., Oliveira, J. L., Ferreira, R. A., & Vincenzi, A. M. (2017).
- Wohlin, C. (2014). Guidelines for snowballing in systematic literature studies and a replication in software engineering. In In Proceedings of the 18th international conference on evaluation and assessment in software engineering (p. 38). London, England: EASE.
- Wu, Zhibiao, and Martha Palmer. "Verbs semantics and lexical selection." Proceedings of the 32nd annual meeting on Association for Computational Linguistics. Association for Computational Linguistics, 1994.

# Appendix A

## Datasets:

Document 1: <https://goo.gl/Cz2H1q>

Document 2: <https://goo.gl/2C1uRc>

Document 3: <https://goo.gl/gafH2E>

Document 4: <https://goo.gl/bMMv46>

Document 5: <https://goo.gl/8RNwew>

## Appendix B

### Relevant work included in systematic mapping-

[S1] Shalinie, S. Mercy, K. Sundarakantham, and S. Pushparathi. "A author topic model based unsupervised algorithm for learning topics from large text collections." Recent Trends in Information Technology (ICRTIT), 2011 International Conference on. IEEE, 2011.

[S2] Tian, Jing, et al. "A Multi-Modal Topic Model for Image Annotation Using Text Analysis." IEEE Signal Process. Lett. 22.7 (2015): 886-890.

[S3] Ding, Weicong, et al. "A new geometric approach to latent topic modeling and discovery." Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on. IEEE, 2013.

[S4] Chanda, Prateek, and Asit Kumar Das. "A Novel Graph Based Clustering Approach to Document Topic Modeling." 2018 9th International Conference on Computing, Communication and Networking Technologies (ICCCNT). IEEE, 2018.

[S5] Barde, Bhagyashree Vyankatrao, and Anant Madhavrao Bainwad. "An overview of topic modeling methods and tools." Intelligent Computing and Control Systems (ICICCS), 2017 International Conference on. IEEE, 2017.

[S6] Allahyari, Mehdi, and Krys Kochut. "Automatic topic labeling using ontology-based topic models." Machine Learning and Applications (ICMLA), 2015 IEEE 14th International Conference on. IEEE, 2015.

[S7] Cheng, Xueqi, et al. "Btm: Topic modeling over short texts." IEEE Transactions on Knowledge & Data Engineering 1 (2014): 1-1.

[S8] Dolatabadi, Hossein, et al. "Clustering Users in Micro Blogging Social Networks Using Probabilistic Topic Modeling-A Framework." 2012 12th International Conference on Computational Science and Its Applications. IEEE, 2012.

[S9] Rao, Yanghui. "Contextual sentiment topic model for adaptive social emotion classification." IEEE Intelligent Systems 1 (2016): 41-47.

[S10] Bian, Jinqiang, Zengru Jiang, and Qian Chen. "Research on Multi-document Summarization Based on LDA Topic Model." Intelligent Human-Machine Systems and Cybernetics (IHMSC), 2014 Sixth International Conference on. Vol. 2. IEEE, 2014.

[S11] Adhitama, Rifki, Retno Kusumaningrum, and Rahmat Gernowo. "Topic labeling towards news document collection based on Latent Dirichlet Allocation and ontology." Informatics and Computational Sciences (ICICoS), 2017 1st International Conference on. IEEE, 2017.

[S12] Wang, Xiang, et al. "Topic mining over asynchronous text sequences." IEEE Transactions on Knowledge and Data Engineering 24.1 (2012): 156-169.

[S13] Sukhija, Nitin, et al. "Topic Modeling and Visualization for Big Data in Social Sciences." Ubiquitous Intelligence & Computing, Advanced and Trusted Computing,

Scalable Computing and Communications, Cloud and Big Data Computing, Internet of People, and Smart World Congress (UIC/ATC/ScalCom/CBDCCom/IoP/SmartWorld), 2016 Intl IEEE Conference. IEEE, 2016.

[S14] Cho, Yoon-Sik, et al. "Latent space model for multi-modal social data." Proceedings of the 25th International Conference on World Wide Web. International World Wide Web Conference Steering Committee, 2016.

[S15] Wang, Quan, et al. "Regularized latent semantic indexing." Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval. ACM, 2011.

[S16] Blei, David M. "Probabilistic topic models." Communications of the ACM 55.4 (2012): 77-84.

[S17] Mei, Qiaozhu, Xuehua Shen, and ChengXiang Zhai. "Automatic labeling of multinomial topic models." Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2007.

[S18] Lau, Jey Han, et al. "Best topic word selection for topic labelling." Proceedings of the 23rd International Conference on Computational Linguistics: Posters. Association for Computational Linguistics, 2010.

[S19] Hingmire, Swapnil, et al. "Document classification by topic labeling." Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval. ACM, 2013.

[S20] Badenes-Olmedo, Carlos, José Luis Redondo-García, and Oscar Corcho. "Efficient clustering from distributions over topics." Proceedings of the Knowledge Capture Conference. ACM, 2017.

[S21] Mehrotra, Rishabh, et al. "Improving lda topic models for microblogs via tweet pooling and automatic labeling." Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval. ACM, 2013.

[S22] Ni, Xiaochuan, et al. "Mining multilingual topics from wikipedia." Proceedings of the 18th international conference on World wide web. ACM, 2009.

[S23] Xue, Zijun. "Scalable Text Analysis." Proceedings of the Tenth ACM International Conference on Web Search and Data Mining. ACM, 2017.

[S24] Hingmire, Swapnil, and Sutanu Chakraborti. "Topic labeled text classification: a weakly supervised approach." Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval. ACM, 2014.

[S25] Kling, Christoph Carl, et al. "Topic model tutorial: A basic introduction on latent dirichlet allocation and extensions for web scientists." Proceedings of the 8th ACM Conference on Web Science. ACM, 2016.

[S26] Yang, Guangbing, et al. "A novel contextual topic model for multi-document summarization." Expert Systems with Applications 42.3 (2015): 1340-1352.



- [S27] Wu, Zongda, et al. "A topic modeling based approach to novel document automatic summarization." *Expert Systems with Applications* 84 (2017): 12-23.
- [S28] Wei, Xing, and W. Bruce Croft. "LDA-based document models for ad-hoc retrieval." *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2006.
- [S29] Song, Yangqiu, et al. "Topic and keyword re-ranking for LDA-based topic modeling." *Proceedings of the 18th ACM conference on Information and knowledge management*. ACM, 2009.
- [S30] Mohr, John W., and Petko Bogdanov. "Introduction—Topic models: What they are and why they matter." (2013): 545-569.
- [S31] Sokhin, Timur, and Nikolay Butakov. "Semi-automatic sentiment analysis based on topic modeling." *Procedia Computer Science* 136 (2018): 284-292.
- [S32] Belford, Mark, Brian Mac Namee, and Derek Greene. "Stability of topic modeling via matrix factorization." *Expert Systems with Applications* 91 (2018): 159-169.
- [S33] Li, Ximing, Jihong Ouyang, and Xiaotang Zhou. "Supervised topic models for multi-label classification." *Neurocomputing* 149 (2015): 811-819.
- [S34] Pavlinek, Miha, and Vili Podgorelec. "Text classification method based on self-training and LDA topic models." *Expert Systems with Applications* 80 (2017): 83-93.
- [S35] Xiong, Shufeng, et al. "A short text sentiment-topic model for product reviews." *Neurocomputing* 297 (2018): 94-102.
- [S36] Tsai, Flora S. "A tag-topic model for blog mining." *Expert Systems with Applications* 38.5 (2011): 5330-5335.
- [S37] Uteuov, Amir, and Anna Kalyuzhnaya. "Combined document embedding and hierarchical topic model for social media texts analysis." *Procedia Computer Science* 136 (2018): 293-303.
- [S38] Hagen, Loni. "Content analysis of e-petitions with topic modeling: How to train and evaluate LDA models?." *Information Processing & Management* (2018).
- [S39] Xiao, Ding, et al. "Coupled matrix factorization and topic modeling for aspect mining." *Information Processing & Management* 54.6 (2018): 861-873.
- [S40] Wang, Xin, et al. "Efficient algorithms for graph regularized PLSA for probabilistic topic modeling." *Pattern Recognition* 86 (2019): 236-247.
- [S41] Li, Ximing, et al. "Exploring coherent topics by topic modeling with term weighting." *Information Processing & Management*(2018).
- [S42] Li, Ximing, et al. "Filtering out the noise in short text topic modeling." *Information Sciences* 456 (2018): 83-96.
- [S43] Wu, Yong, et al. "Guiding supervised topic modeling for content based tag recommendation." *Neurocomputing* 314 (2018): 479-489.

- [S44] Vo, Duc-Thuan, and Cheol-Young Ock. "Learning to classify short text from scientific documents using topic models with various types of knowledge." *Expert Systems with Applications* 42.3 (2015): 1684-1698.
- [S45] Selvi, M., et al. "Classification of Medical Dataset Along with Topic Modeling Using LDA." *Nanoelectronics, Circuits and Communication Systems*. Springer, Singapore, 2019. 1-11.
- [S46] Belford, Mark, Brian Mac Namee, and Derek Greene. "Stability of topic modeling via matrix factorization." *Expert Systems with Applications* 91 (2018): 159-169.
- [S47] Bittermann, André, and Andreas Fischer. "How to identify hot topics in psychology using topic modeling." *Zeitschrift für Psychologie* (2018).
- [S48] Misra, Hemant, et al. "Text segmentation: A topic modeling perspective." *Information Processing & Management* 47.4 (2011): 528-544.
- [S49] Ramirez, Eduardo H., et al. "Topic model validation." *Neurocomputing* 76.1 (2012): 125-133.
- [S50] Yan, Jian-Feng, et al. "Towards big topic modeling." *arXiv preprint arXiv:1311.4150* (2013).
- [S51] Agrawal, Amritanshu, Wei Fu, and Tim Menzies. "What is wrong with topic modeling? And how to fix it using search-based software engineering." *Information and Software Technology* 98 (2018): 74-88.
- [S52] Zeng, Jia. "A topic modeling toolbox using belief propagation." *Journal of Machine Learning Research* 13.Jul (2012): 2233-2236.
- [S53] Newman, David, et al. "Distributed algorithms for topic models." *Journal of Machine Learning Research* 10.Aug (2009): 1801-1828.
- [S54] Yang, Xiaoyan, et al. "Enhancing Topic Modeling on Short Texts with Crowdsourcing." *Asian Conference on Machine Learning*. 2016.
- [S55] Fu, Xianghua, et al. "Improving distributed word representation and topic model by word-topic mixture model." *Asian Conference on Machine Learning*. 2016.
- [S56] Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent dirichlet allocation." *Journal of machine Learning research* 3.Jan (2003): 993-1022.
- [S57] Weinshall, Daphna, Gal Levi, and Dmitri Hanukaev. "LDA topic model with soft assignment of descriptors to words." *International Conference on Machine Learning*. 2013.
- [S58] Cheng, Dehua, Xinran He, and Yan Liu. "Model selection for topic models via spectral decomposition." *Artificial Intelligence and Statistics*. 2015.
- [S59] Tang, Jian, et al. "Understanding the limiting factors of topic modeling via posterior contraction analysis." *International Conference on Machine Learning*. 2014.
- [S60] Shubankar, Kumar, AdityaPratap Singh, and Vikram Pudi. "A frequent keyword-set based algorithm for topic modeling and clustering of research papers." *Data Mining and Optimization (DMO), 2011 3rd Conference on*. IEEE, 2011.

- [S61] Morstatter, Fred, and Huan Liu. "A novel measure for coherence in statistical topic models." Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Vol. 2. 2016.
- [S62] Minhaj, Mohamed. "Clustering of conference papers using LDA based topic modelling."
- [S63] Syed, Shaheen, and Marco Spruit. "Full-text or abstract? Examining topic coherence scores using latent dirichlet allocation." Data Science and Advanced Analytics (DSAA), 2017 IEEE International Conference on. IEEE, 2017.
- [S64] Steyvers, Mark, and Tom Griffiths. "Probabilistic topic models." Handbook of latent semantic analysis 427.7 (2007): 424-440.
- [S65] Chundi, Parvathi, and Susannah Go. "Latent dirichlet allocation approach for analyzing text documents." Encyclopedia of Information Science and Technology, Third Edition. IGI Global, 2015. 1819-1824.
- [S66] Zhang, Jianwei, et al. "LDA Revisited: Entropy, Prior and Convergence." Proceedings of the 25th ACM International on Conference on Information and Knowledge Management. ACM, 2016.
- [S67] Hansen, Joshua A., Eric K. Ringger, and Kevin D. Seppi. "Probabilistic explicit topic modeling using wikipedia." Language Processing and Knowledge in the Web. Springer, Berlin, Heidelberg, 2013. 69-82.
- [S68] Zou, Yue-peng, Ji-hong Ouyang, and Xi-ming Li. "Supervised topic models with weighted words: multi-label document classification." Frontiers of Information Technology & Electronic Engineering 19.4 (2018): 513-523.
- [S69] Sahni, Aashka, and Sushila Palwe. "Topic Modeling on Online News Extraction." Intelligent Computing and Information and Communication. Springer, Singapore, 2018. 611-622.
- [S70] Kavvadias, Spyridon, George Drosatos, and Eleni Kaldoudi. "An Online Service for Topics and Trends Analysis in Medical Literature." World Congress on Medical Physics and Biomedical Engineering 2018. Springer, Singapore, 2019.
- [S71] Drosatos, George, Spiros E. Kavvadias, and Eleni Kaldoudi. "Topics and Trends Analysis in eHealth Literature." EMBEC & NBC 2017. Springer, Singapore, 2017. 563-566.
- [S72] Alghamdi, Rubayyi, and Khalid Alfalqi. "A survey of topic modeling in text mining." Int. J. Adv. Comput. Sci. Appl.(IJACSA) 6.1 (2015).
- [S73] Sriurai, Wongkot. "Improving text categorization by using a topic model." Advanced Computing 2.6 (2011): 21.
- [S74] Wan, Xiaojun, and Tianming Wang. "Automatic labeling of topic models using text summaries." Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Vol. 1. 2016.

- [S75] Prier, Kyle W., et al. "Identifying health-related topics on twitter." International conference on social computing, behavioral-cultural modeling, and prediction. Springer, Berlin, Heidelberg, 2011.
- [S76] Musat, Claudiu, et al. "Concept-based topic model improvement." Emerging Intelligent Technologies in Industry. Springer, Berlin, Heidelberg, 2011. 133-142.
- [S77] Ramage, Daniel, et al. "Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora." Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1. Association for Computational Linguistics, 2009.
- [S78] Liu, Lin, et al. "An overview of topic modeling and its current applications in bioinformatics." SpringerPlus 5.1 (2016): 1608.
- [S79] Magatti, Davide, et al. "Automatic labeling of topics." Intelligent Systems Design and Applications, 2009. ISDA'09. Ninth International Conference on. IEEE, 2009.
- [S80] Bhatia, Shraey, Jey Han Lau, and Timothy Baldwin. "Automatic labelling of topics with neural embeddings." arXiv preprint arXiv:1612.05340 (2016).
- [S81] Sun, Xiangyan, et al. ""On Conceptual Labeling of a Bag of Words."" IJCAI. 2015.
- [S82] Ritter, Alan, and Oren Etzioni. "A latent dirichlet allocation method for selectional preferences." Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2010.
- [S83] Yau, Chyi-Kwei, et al. "Clustering scientific documents with topic modeling." Scientometrics 100.3 (2014): 767-786.
- [S84] Song, Min, and Ying Ding. "Topic modeling: Measuring scholarly impact using a topical lens." Measuring Scholarly Impact. Springer, Cham, 2014. 235-257.
- [S85] John, Adebayo Kolawole, Luigi Di Caro, and Guido Boella. "Text Segmentation with Topic Modeling and Entity Coherence." International Conference on Hybrid Intelligent Systems. Springer, Cham, 2016.
- [S86] Fernández-Beltran, Rubén, Raul Montoliu, and Filiberto Pla. "Vocabulary reduction in bow representing by topic modeling." Iberian Conference on Pattern Recognition and Image Analysis. Springer, Berlin, Heidelberg, 2013.
- [S87] Tuarob, Suppawong, et al. "A generalized topic modeling approach for automatic document annotation." International Journal on Digital Libraries 16.2 (2015): 111-128.
- [S88] Greene, Derek, Derek O'Callaghan, and Pádraig Cunningham. "How many topics? stability analysis for topic models." Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer, Berlin, Heidelberg, 2014.
- [S89] Davoudi, Heidar, and Aijun An. "Ontology-Based Topic Labeling and Quality Prediction." International Symposium on Methodologies for Intelligent Systems. Springer, Cham, 2015.

[S90] Qiang, Jipeng, et al. "Topic modeling over short texts by incorporating word embeddings." Pacific-Asia Conference on Knowledge Discovery and Data Mining. Springer, Cham, 2017.

[S91] Jelodar, Hamed, et al. "Latent Dirichlet Allocation (LDA) and Topic modeling: models, applications, a survey." Multimedia Tools and Applications (2017): 1-43.

[S92] Wu, Zewei, et al. "Sentiment Detection of Short Text via Probabilistic Topic Modeling." International Conference on Database Systems for Advanced Applications. Springer, Cham, 2015.

[S93] Yokomoto, Daisuke, et al. "Lda-based topic modeling in labeling blog posts with wikipedia entries." Asia-Pacific Web Conference. Springer, Berlin, Heidelberg, 2012.