# A Biased Topic Modeling Approach for Case Control Study from Health Related Social Media Postings

by

Jian Yang

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved October 2017 by the
Graduate Supervisory Committee:

Graciela Gonzalez, Co-Chair
Hasan Davulcu, Co-Chair
Huan Liu
Paolo Papotti

ARIZONA STATE UNIVERSITY

December 2017

# ABSTRACT

Online social networks are the hubs of social activity in cyberspace, and using them to exchange knowledge, experiences, and opinions is common. In this work, an advanced topic modeling framework is designed to analyse complex longitudinal health information from social media with minimal human annotation, and Adverse Drug Events and Reaction (ADR) information is extracted and automatically processed by using a biased topic modeling method. This framework improves and extends existing topic modelling algorithms that incorporate background knowledge. Using this approach, background knowledge such as ADR terms and other biomedical knowledge can be incorporated during the text mining process, with scores which indicate the presence of ADR being generated. A case control study has been performed on a data set of twitter timelines of women that announced their pregnancy, the goals of the study is to compare the ADR risk of medication usage from each medication category during the pregnancy.

In addition, to evaluate the prediction power of this approach, another important aspect of personalized medicine was addressed: the prediction of medication usage through the identification of risk groups. During the prediction process, the health information from Twitter timeline, such as diseases, symptoms, treatments, effects, and etc., is summarized by the topic modelling processes and the summarization results is used for prediction. Dimension reduction and topic similarity measurement are integrated into this framework for timeline classification and prediction. This work could be applied to provide guidelines for FDA drug risk categories. Currently, this process is done based on laboratory results and reported cases.

Finally, a multi-dimensional text data warehouse (MTD) to manage the output from the topic modelling is proposed. Some attempts have been also made to incorporate topic structure (ontology) and the MTD hierarchy. Results demonstrate that proposed methods show promise and this system represents a low-cost approach for drug safety early warning.

To my family for their love and support

ACKNOWLEDGMENTS

First and foremost, I would like to thank my advisor, Dr. Graciela Gonzalez, who has been a great mentor guiding me in my research, and helping me in the right direction for my research. She has contributed to this work in many ways. I am very grateful for her endless support over many years in my PhD program. Her help was always around, such as her critical thinking, technical writing, and presentation skills. Working toward my PhD. degree while doing full time job is not an easy task, but Dr. Gonzalez has inspired me every time I was frustrated. This dissertation would have been impossible without her support.

I also want to thank my co-advisor, Dr. Hasan Davulcu, who is a great mentor to follow. My research has benefitted from his insightful thoughts and words of wisdom. I am grateful to Dr. Huan Liu, who offered me great research opportunities, interesting projects, resources, and trust which allowed me to fully explore the data mining and social media computing area. I would also like to thank Dr. Paolo Papotti, who have provided valuable feedbacks, broadened my horizons, and helped me look beyond my current endeavors.

I would like to thank my colleagues at Arizona State University: Ryan Sullivan, Abeed Sarker, Davy Weissenbacher, Laura Wojtulewicz, for their constructive criticism, helpful suggestions, and support. They helped me a lot and greatly inspired me. I am also grateful to other lab members for helpful conversation and feedback on my work, particularly Siddhartha Jonnalagadda and Annie Akariah.

There are many friends and professors who have encouraged and helped me during my PhD study. The first name goes to Dr. Yi Chen, who has provided helpful suggestions for my research, I would like to thank my friends: Stephen Wu, Jim Celmer, Seow Lim, Yunzhong Liu and many others.

I would also like to thank my family for their wise counsel and sympathetic ear. You are always there for me, I am grateful for their love and understanding. Without their support, this thesis would not have been possible.

TABLE OF CONTENTS

CHAPTER Page

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

Figure Page

# 1 INTRODUCTION

## 1.1 Background and Motivation

There has been an increasing interest in performing biomedical text mining research in health-related social media in the recent years. The social media of today contains a treasure trove of information about subjects related to patients with disease, along with discussions and suggestions that may be of wide interest. Inquiring about health information from the internet or sharing experiences about drugs or disease is one of the most popular activities over the web. This results in an exponentially increasing amount of user generated biomedical textual data. Compared to traditional EMR data, this massive amount of health related textual data provides a better alternative for biomedical information retrieval, this is attributed to its diversity, fast response, and easy integration. This patient-contributed information is ever increasing in volume and provides a good opportunity for Heath Language Processing (HLP) to identify valuable information that can be used to monitor disease outbreaks, drug usage, treatments and Adverse Drug Reactions (ADR). ADR is a leading cause of death in the U.S. [1]. Traditionally small-scale patient surveys or voluntary report systems were used to monitor ADR, however, the online social network may provide larger scale information which is more timely and comprehensive.

"The FDA has classified OTC and prescription drugs into 5 categories in terms of safety for use during pregnancy (A, B, C, D, X)"[1]. This drug category system is used by the FDA to define what is safe and harmful during pregnancy. The problem with this system is that only animal data and minimal post-marketing human data from registries is available to define the risk category, and no ADR information can be obtained on pregnant women in the premarketing phase because pregnant women are excluded from clinical trials. Since the Information about adverse effects of medications in pregnancy is rare, social media can be a good source of information for this particular group, as well as many others. It has been proven by previous

[1] http://www.merckmanuals.com

research that social media discourse can, indeed, be used to analyze specific health-related behaviors in an automatic way.

In order to identify trustworthy knowledge from a social media, unsupervised text summarization techniques have their value in this case. These are more flexible, cheaper, faster, and more cost-effective as compared to supervised learning techniques, which require a lot of human annotation. Topic is defined as special concepts of interest within a document, and topics can play critical role for text mining in online social network user generated content. There has been increasing interest in performing topic modeling in health-related social media (Twitter, Dailystrength, PatientsLikeMe, and other forums) [2][3]. However, due to the nature of social network data, the traditional topic modeling approach has some challenges that affect their accuracy and performance, particularly due to the informal nature of biomedical social network data. In this study, we seek to integrate existing unsupervised text summarization techniques (Topic Modeling) with structured biomedical knowledge bases, the goal is to improve the quality of topic model by involving background knowledge, those widely available public biomedical dictionary and ontology data provides valuable resource as background knowledge for topic model.

In order to overcome the challenges of traditional topic modeling, this work improves and extends previously proposed topic modeling methods which incorporate background knowledge [4][5]. This method is a revised seed words approach, during the modeling process, a bag of seed words is provided to improve the modeling outcome (document/topic/word distribution) by biasing topics towards seed words. The intuition for this approach is that we believe that not all topics are equally important when case studies for ADR and medication usage are performed on social media. This is the reason why we want to “weight” health related topics more than “regular topics” like baby clothes or shopping activities Furthermore, the women taking the more dangerous medications would probably mention those adverse effects in greater volume or refer to more serious effects, it is crucial that the topics being generated would be more inclined towards health related information.

This study would improve and extend previous work in several substantial ways: (i) we modify the existing lexicon matching method for the background knowledge and replace it with a more accurate similarity scoring system; (ii) we propose a scoring system to score the ADR effects from the topic modeling result; (iii) we improve the performance of the existing approach by enabling the parallel execution of the algorithm in the Spark environment via data frames; and (IV) we extend this method by incorporating Ontology hierarchy as background knowledge.

In this research, we have applied this topic modeling approach on pregnant women twitter timelines, using background knowledge (seed words) from publicly available biomedical dictionary, we also calculate the ADR score based on their drug usage. Our approach generates topics that are biased towards adverse reactions from twitter timelines with different drug usages, those topics enable the comparison of specific cohorts for a case-controlled study by providing a quantitative model of the timelines. In biomedical domain, a case study is defined as "A study that compares patients who have a disease or outcome of interest (cases) with patients who do not have the disease or outcome (controls), and looks back retrospectively to compare how frequently the exposure to a risk factor is present in each group to determine the relationship between the risk factor and the disease."[2]. In our case, the health information summarized by biased topic modeling is utilized to compare the cohort of social media posting from different medication categories. Conclusively, we ought to relate the ADR information to the medication risk category.

This process is executed in HPC Spark cluster (52 nodes) by dividing the input timeline file into an RDD (Resilient Distributed Data Set), this way we can execute the training via a parallel framework. We performed the evaluation of this framework on a set of annotated data from domain expert, and compared the result with other different topic model methods. We also performed further evaluation by using the topic model as prediction model.

[2] https://himmelfarb.gwu.edu/tutorials/studydesign101/casecontrols.html

To evaluate the effectiveness of this proposed approach, we use the collected data to design and exam a prediction model, on which we can perform feature reduction to summarize social network user generated contents into a limited number of topics. A framework is proposed to take an unsupervised text mining approach to monitor social network contents for unsafe drug usage, such monitoring could be important indicators of potential adverse drug reactions. The proposed solution is built on top of a biased topic modeling which is optimized for biomedical text mining, it can also assist regulatory agencies like FDA or pharmaceutical companies to develop drug safety guideline and risk categories. The prediction process is optimized for handling massive amount of data by using semi supervised text mining, which means that limited amount of labeled and annotated data is needed to predict large corpus of unlabeled data with similar semantic features. We perform the evaluation of this framework on pregnancy timelines collected from twitter and perform a case control study. Currently, this framework can be used to predict drug usage category based on FDA risk category. We further hypothesize that this approach may eventually aid in categorizing given drugs into any risk classes based on the ADR risk. We verified its effectiveness via experimental evaluation with an official FDA pregnancy drug category.

This study also proposes an approach to adopt the MTD to manage the output from the topic modeling with background knowledge, in this approach, topic hierarchies (Ontology) from topic model can be incorporated with the MTD OLAP hierarchy. We have also designed an advanced star schema (diamond model) to represent probabilistic topic modeling approaches for text mining. We design an aggregation algorithm for the OLAP cube to aggregate the topic distribution from the topic levels. Given a pre-defined topic hierarchy based on ontology, this text multi-dimensional model can aggregate text semantic information based on topic hierarchy. Biomedical domain experts and power users can quickly digest the text information same way like a traditional relational database OLAP cube, those users can perform operations such as drilling down, slicing and dicing on top of aggregated data to view the text information from different dimensions.

## 1.2 Contribution

In this study, an advanced topic modeling framework to analyze complex longitudinal health information from social media with minimal human annotation has been designed. We propose an end-to-end framework to perform topic modeling on a dataset of twitter timelines of women that announced their pregnancy, and we sought to distinguish what kind of medications they were taking during pregnancy. This framework aims to enhance existing topic modeling methods to mine topics from large volumes of data and to try to minimize human intervention during the mining process. In this framework, novel uses of topic modeling that incorporate background knowledge including structured hierarchical knowledge have been explored. We chose biased topic modeling in order to obtain a “computable” model that “summarized” their discourse in social media. This framework is generic in that we can apply it to other data sets. This is a unique approach because it can greatly improve the accuracy and usability of the topic model for biomedical domains. This framework will allow researchers use background knowledge such as dictionary or ontology as knowledge to drive the topic modeling. The contributions of this research are summarized as follows:

- Developing and evaluating a topic-analysis framework that incorporates biomedical background knowledge which enables the comparison of specific cohorts for a case-control study by providing a quantitative model of the timelines. We also propose a scoring system to score the topic modeling output.
- Enhancing the topic modeling algorithm to incorporate background knowledge terms of different types, such as ADR, drug name library and disease symptoms; relationships between terms such as ontology can also be incorporated. Using this approach, we can output subset of topics which are semantically similar to background knowledge terms, this makes it possible for us to separate different types of health information into a different set of topics such as ADR and disease symptom.
- Validating the predictive power of the biased topic model for the pregnancy cohort by utilizing the quantitative scores based on the topics and comparing subgroups of the

cohort formed based on the categories of the medications they mention in their timelines. This classification approach combines both text mining and feature reduction techniques to summarize important text semantics features into a collection of topics, and those topics are adopted as the prediction models.

- Proposing a new approach to combine a traditional multidimensional database with probabilistic topic model with background knowledge. A revised star schema is proposed in order to store the topic distribution outcome from topic modeling, and incorporate the hierarchical PAM (Pachinko Allocation Model) Ontology with the multidimensional database Hierarchy. This approach will associate health information with the corresponding topic distributions to form patient semantic information units for effective query and knowledge discovery.

# 2 BIASED TOPIC MODELING FOR ADR MINING

## 2.1 Research Background

### 2.1.1 Topic Modeling

It is always a challenge to mine meaningful results from massive document collections. Topic modeling is one of the popular techniques used to provide an effective way to extract key topics and themes from a large corpus of documents. In particular, Latent Dirichlet Allocation (LDA), where each topic is a weighted collection of words that represent the meaning of the topic, is the most commonly used modeling method. During the process of topic modeling, a topic is defined as a distribution over a group of words with high co-occurrence. For example, a health topic has words about health, disease, drugs that often appear together, and the food topic has vocabularies about meat, vegetables which appear together with high probability. Those topics are the nature ways to represent semantic meaning of a document.

LDA can be considered as a cluster algorithm. Figure 2-1 shows the basic ouput process for LDA. Topic model is defined as “a type of statistics model for discovering the abstract latent topics that occur in document, which is distribution over words.” [6]. In this theory, both topic and word distribution are defined as multinomial distribution, and another dirichlet distribution(priori) decides those topic distributions and word distributions. Utimately, LDA is a dimensionality reduction tool that can help to reduce the number of dimensions of the feature vector to represent the data more accurately (removing redundancy and preserving the energy along the most probable direction).

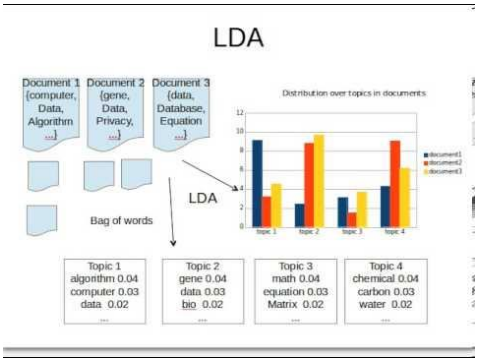



Figure 2-1: Topic modeling process.

For example, an LDA model might group words into topics such as DRUG_related and ADR_related. Different words like Amoxicillin, Lyrica, and Zanax can be classified as "DRUG_related." For ADR_related topics, there is a high probability that words such as vomiting, fatigue, alopecia, and drowsiness will appear. Words without any special relevance will be evenly distributed across different topics. The overall process can be described using the following steps:

a) Decide how many topics the output should contain, where the number of topics can be determined either by experience or by a performance evaluation. LDA will return the results based on the inter-class separation of the featured inputs.

b) A random topic is assigned to every word initally, then for every additional iternation, a new topic number will be assigned.

c) In next iternation, LDA will assign the topic assignment for current word is correct, then use information from other topic assignment to sample next topic assignment.

Represented graphically in Figure 2-2, the calculation is made for each tokens, then the topic assignment is updated based on: (i) how important that is word across topics and (ii) how important the topics are in the document.

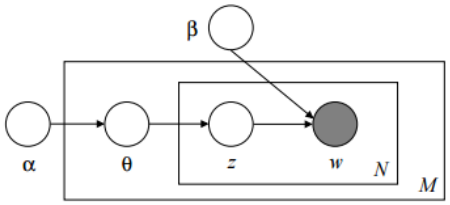



Figure 2-2: LDA Plate notation (from [8]).

In this study, the topic model is our fundamental method, utilized to uncover the hidden thematic structure from the Twitter timeline. We performed topic modeling on Twitter timelines and studied the presence of ADR using the topics from the topic modeling.

### 2.1.2 Hierarchical Topic Model

As previously described, LDA is a cluster algorithm that can help find correlations between words in a document. However, the output is just a cluster, and it does not describe the relationship between these clusters. Generally speaking, we know that there should be some relationship amid the topics within a document, and those topic relationship could be very useful to capture for biomedical research, for example, when we perform topic model on biomedical text, researchers are normally interested in discovering the relationship between drug related topics and ADR related topics, or distribution difference between different ADR related topics from different drugs. Some research has been done on hierarchical topic model to capture the relationship between topics.

Pachinko Allocation Models (PAM) [7] and Hierarchical PAM [8] are two variations of hierarchical topic models. In hierarchical topic model, tree structure is used to define the topic relationships, this tree structure is also defined as directed acyclic graph (DAG). in Hierarchical PAM [8], every node is mapped to a word distribution, since it is possible for Hierarchical PAM to support model tree structure, and this kind of tree structure implementation shares some similarity with OLAP hierarchical structure, it would be interesting to leverage Hierarchical PAM with the OLAP cube so that we can navigate the topic distribution at different levels (Figure 2-3).

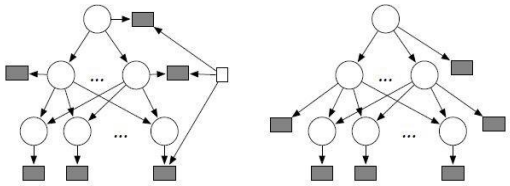


Figure 2-3: Example of structures of hPAM [10].

In this study, we propose to generate topic relationships from hieratical topic modeling automatically, and therefore, to model this hierarchy into the data warehouse hierarchy so that we can cross-reference the existing ontology structure.

### 2.1.3 Research Opportunities

The existing topic modeling methods have the following challenges when mining social network data.

- Accuracy: When we perform topic modeling on a document, a document level co-occurrence probability is used to group semantically-related words into a single topic. Since the objective of these models is to maximize the co-occurrence of the observed data, they have a tendency to explain only the most obvious and superficial aspects of a corpus. However, the information on the web can be sparse. This leads to a situation where the topic output from LDA is not in accordance with the subjects in which we are really interested. For example, if we perform topic modeling on a web site such as dailystrength.com, we may be able to identify meaningful results with topics related to drugs, ADR, and outcomes. However, for generic web sites like Twitter, where people do not focus on health-related issues, topic modeling has a tendency to explain only the most obvious and superficial aspects of a corpus. It effectively sacrifices performance with rare topics in order to do a better job of modeling frequently occurring words. In this case, we end up with a large number of unrelated topics and a skewed impression of the corpus; consequently, this approach will not perform well with extrinsic tasks.
- Performance: We need a scalable and efficient way of distributing the computation across multiple machines and an effective way to store the results where the results can be retrieved effectively. Most topic modeling techniques run out of memory when applied to even a modest number of documents (50,000 to 100,000 documents), and the process to build the model and calculate probability is long. Consequently, a Big Data based approach is needed so that we can linearly scale the computation and spread the computation task to multiple servers.

The output from topic modeling is a collection of document/topic distribution and topic/word distribution, which can be used to analyze UGC. Users can manually go through all topics to identify information such as the frequency of topic, related vocabularies, whether there

was a good or bad outcome, etc. However, in order to make the results from topic modeling more useful, we need the ability to exploit the structure to aid in discovery and exploration, as there is a disconnection between the output of topic modeling and its link to automatic information retrieval. There are many exciting new directions for research to enhance topic modeling, including the following [9]:

- Visualization and user interfaces: Results from topic model are normally difficult to be used by human directly, one research direction is to make the topic model result more user friendly, by visualizing topics and frequent words. This kind of visualization can help us to exploit text structure to aid in better information discovery from document or social media user generated content.
- Topic models for data discovery: Although topic model is an unsupervised text mining approach, it is always desirable to involve domain knowledge to explore those topics. As topic modeling is a text mining model, it will be interesting for users from other background to should be able to use the topic modeling output directly for problems like case control study about the data. In general, this problem is best addressed by providing domain experts such as biomedical researchers with a tool that they can use topic models output to help resolve real word problems or hypotheses from text semantic information without too much computer science knowledge. If this is possible, topic modeling can be a real useful interdisciplinary computational methodology for proposing and drawing conclusions [9].

### 2.1.4 Early Work on ADR Mining

Some early work was done to mine the ADR from social media using lexical approach. DailyStrength data is used in this study. DailyStrength is a health related social media, where various disease-related support groups are provided. Users can join the supporting group by creating profiles. It is a useful resource that allows people to connect with other patients with similar conditions, and share with their experience. “As of September 2007, DailyStrength had 14,000 average daily visitors, each spending 82 minutes on the site and each viewing

approximately 145 pages “[13]. We propose using the patient-submitted data contained in health social network to study adverse drug effects.

For this study, we used an automatic web crawler to gather data from the social network. Because of the massive amount of data (over 500 supporting groups with millions of pages), a highly parallelized web crawler was designed to ensure that data collection could be done in a reasonable timeframe. The system was run from two Dell Precision servers and was controlled by a central database server with the crawling jobs dynamically assigned so that the crawling of the whole social network can be finished within 24 hours. All information collected by the crawler was saved into a central database system as raw HTML. After the raw data were collected in a database, they were preprocessed for text mining by removing all HTML tags and filtering comments about alternative treatments. Only comments about drug names that exist in FDA drug library were kept in the database. For each comment, the following were extracted: user comment, date, disease name, drug name, and comment text. For each user, we collected age, gender, and location.

Figure 2-4 contains an illustrative sample data flow process for the DailyStrength data. For data analysis, we created a subset the comments text associated with ten focus drugs known to cause adverse events. Overall, our data set contains 25,311 comment records for ten drugs that related to over 100 different diseases (Table 2-1) posted by 24,210 different users. The extracted data is saved in SQL Server 2005 relational database. The data structure is shown in Figure 2-5.

After the data cleaning on DailyStrength data is done, the ADR extraction was performed in the following steps.

a) Reported reaction vocabulary generation: the reaction terms from the Canada ADR database are converted into two sets of vocabularies (bag of words). The first set is called known ADR vocabularies, which includes drugs for disease Autism / Autism Spectrum, Schizophrenia, Epilepsy & Seizures, and Alzheimer. The second set is called

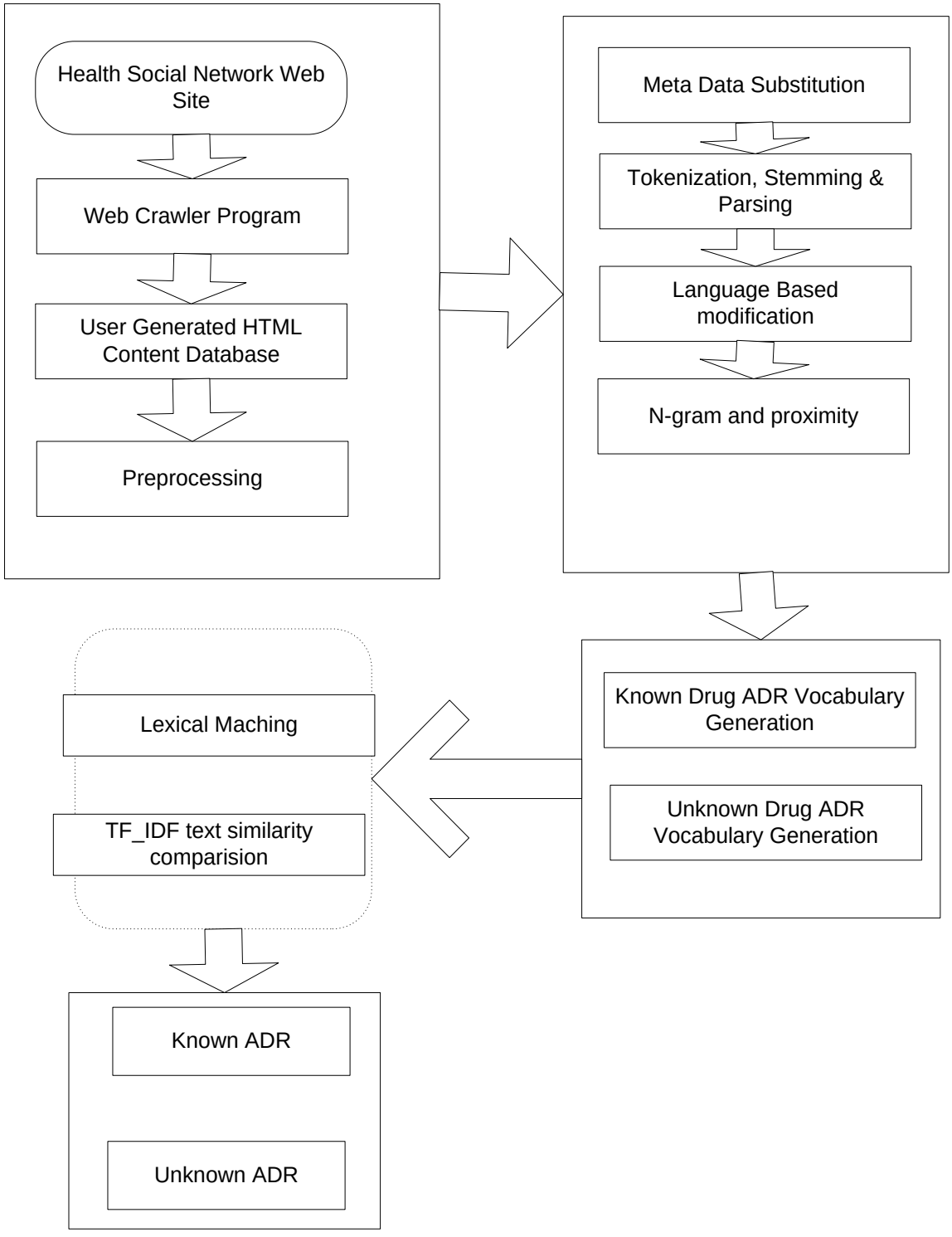



Figure 2-4: DS data collection.

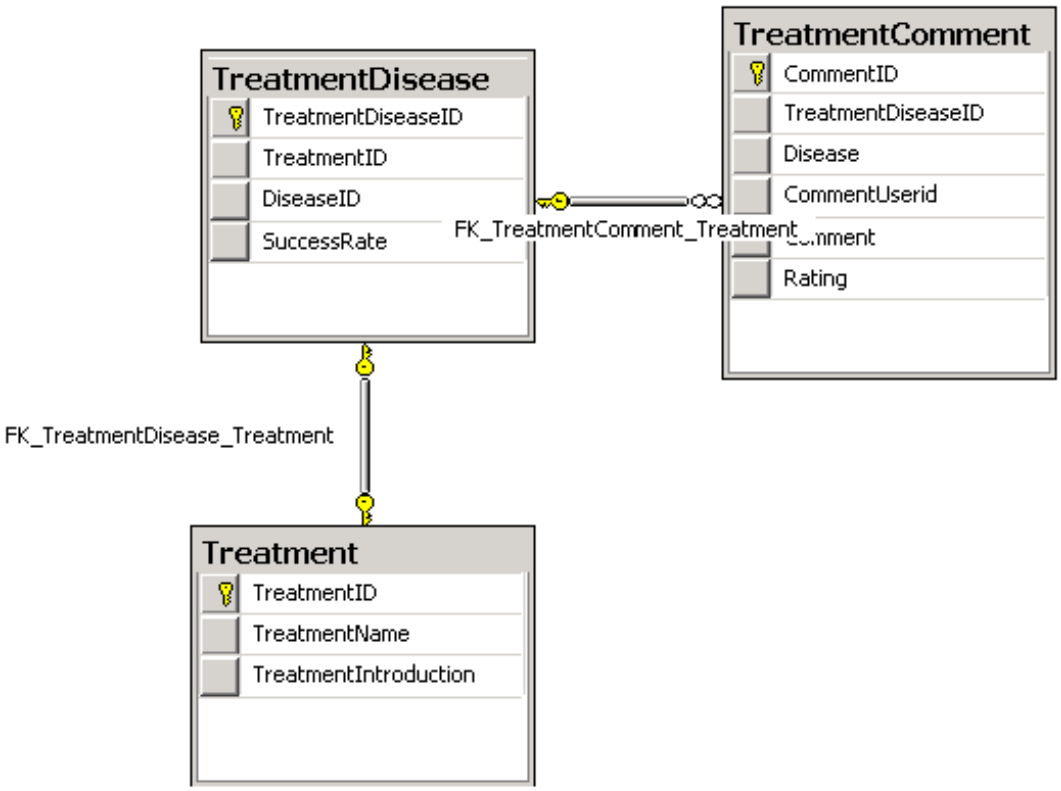



Figure 2-5: Database schema for DS data.

| Total Disease | Total Drugs | Total Comments |
|---|---|---|
| 458 | 1057 | 571,689 |

Table 2-1: DS data summary.

unknown ADR vocabulary, which is the set of ADR reaction terms that do not appear in the first vocabulary set.

b) Known ADR reaction extraction: For each of the drug reaction terms in the known vocabulary set, we check whether it is contained in the DailyStrength user comment data on the same drug. This process is implemented as text similarity. If the reaction term is unigram, then an exact lexical match is used. If the reaction term is n-gram, then Frequency/Inverse Document frequency (TF-IDF) is used. The TF-IDF is a text statistical-based technique that has been widely used in text mining applications. We capture the similarity of the comment and reaction terms using a cosine similarity measurement. The cosine similarity is calculated by measuring the cosine of the angle between two strings (the reaction term and the user comment). The text similarity threshold we used in this project is 0.5. If either exact lexical match or text similarity on the n-gram or stem words exceeding the threshold, the comment will be marked for extraction.

c) Unknown ADR reaction extract: For each of the ADR reaction terms in the unknown reaction vocabulary, we checked all comments for DailyStrength comments on drugs for Autism / Autism Spectrum, Schizophrenia, Epilepsy & Seizures, and Alzheimer. The text similarity checking uses the same method as the known ADR reaction extraction method.

Some sample ADR mentioned from text mining is shown inTable 2-2.

From the result, we can see the significant difficulty of this domain, particularly due to the mismatch between the informal language employed by the users and the formal language used clinically especially when the reaction vocabulary contains more than one word. Also it is common to find a significant number of orthographic and grammatical errors in the user comments. A closer analysis reveals that while many comments include terms which could mention an adverse reaction, the filtering and TF-IDF method employed did not profitably trade-off between precision and recall, which make it difficult to improve system performance. This

| Sample ADR |
| --- |
| I took it & it worked but I lost a lot of weight & didn't have an energy because I was allergic to it.<br>On 175 mg BID, side effects include insomnia, speech issues, word retrieval & vocabulary is NOT what it used to be, immediate and short term memory problems as well as my balance being off somewhat. Since increasing dosage by 50 mg. last November have been seizure free, but continue to have "non-specific" (as neuro calls it) sensations in my head, feels as if my head is being squeezed. So, I'm not sure the Lamictal is working for me.<br>Extremely bad alergic reaction. Seizures increased a dangerous amount! Had to stop taking even at the lowest dose! memory loss dizziness insomnia and fatigue sleeping all the time (was used for muscle spasms caused by other meds) made him lethargic and absent minded lack of appetite worked on the seizures but gained 30 pounds in a few months and developed hormonal problems Terrible side effects,drunk feeling, dizziness and felt disconnected from world, no seizure control Really bad side effects, terrible headaches blurred vicion nausea. |

Table 2-2: Sample comments from health-related social networking website.

appears to be related to not being able to correctly handle situations where multiple terms combine to form the meaning of another single term. For example, the phrase "hair loss" and term "alopecia" have the same meaning, but the semantic similarity between the two will be low using the measure we implemented.

Another challenge for this approach is its application it to generic social media site like Twitter: DS is a health-focused site that most comments are related to personal health experience, Twitter is a general networking site where the information about a particular drug and its side effects is very spare, which make it even more difficult to detect ADR.

We expect future work to greatly improve the performance of our system, for instance, by taking a topic modeling approach utilizing a relatively simple feature set, topic modeling may result in providing common topics that may summarize the semantics meaning of the document.

### 2.2 Biased Topic Modeling System

The goal of this research is to learn the ADR term distribution for different drug usage patterns from the Twitter timelines of pregnant women. Formally, given a timeline $X_n$ with m posts $\{p_0, p_1, ..., p_{m-1}\}$, our task is to predict the ADR distribution for each timeline in $X_n$ .To mine the patient health information units from the Twitter timeline, we need to identify the drug usage pattern from the tweets and analyze the ADR distribution from the outcome. Our approach starts with an LDA model that has integrated background knowledge. Then, we can calculate the ADR-weighted score based on the ADR topics distributed at the drug and drug category levels.

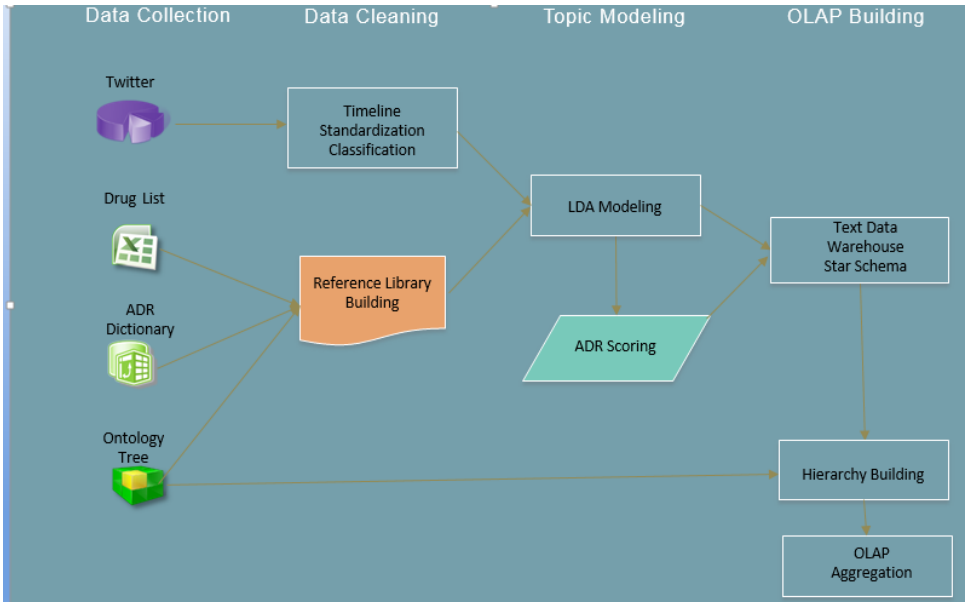



Figure 2-6: Biased topic modeling process flow.

### 2.2.1 Data Prepossessing

The raw data crawled from the website needs to be preprocessed before we do any text mining, and the following steps were performed on the user-generated data:

- HTML tag clean up; remove all HTML tag element like <TR>, <TD>
- Tokenization: The user comment is broken into a bag of words (token).
- Removing stop words: Stop words need to be removed before topic modeling, because those words are normally insignificant and always have high frequency, their existence could potentially skew the topic modeling result, as topic model is highly based on words occurrence.
- Stemming word: In this process, words are conflated into their original root (distraction -> distract). This step is necessary for searching and matching.

### 2.2.2 LDA with Background Knowledge

The unbiased LDA model will build a generic model for the data, where the generic models will generate random topics. In order to incorporate the background knowledge into the LDA process, we need to modify the classic LDA approach. In this new process, we still maintain the original process of generating the per-document Dirichlet distribution, which is unbiased and random. In addition, we also need to generate background related Dirichlet distributions by using the background knowledge bag of words. In this process, multiple background knowledge libraries (dictionary terms) can be used as input. The background related distribution is a subset of the generic distribution, and it will return a multinomial distribution over a sub-set of topics. This sub-set of topics is generated towards terms semantically related to the dictionary terms, which we called biased topic modeling. The output of the process will be multiple models. One model will contain all the topics, and each background knowledge set will generate a subset of topics, this method is built by extending existing topic modeling methods with knowledge [4][5].

During the model training, a semantic distance check will happen before we compute a word's distribution probability. If the word is not semantically related to a background term (dictionary term), it will be used to train the full generative model; if the word is semantically

related to the dictionary term, then it will be used to train the background related topics. However, both generic distribution and biased distribution will be updated at the same time.

The overall process for this method is described as follows:

Let D be a collection of documents, ad S be a collection of background knowledge dictionary (seed). For each d ∈ D of length N , let fd : {1, . . . , N } →{0, 1} be an indicator function, this function will return either 0 or 1, where 1 mean that the current word is a seed word(from background dictionary), and 0 means that the current word is not a seed word. To generate the collection D:

---

**for** k = 1 to T **do**

Draw $\Phi_{k,\cdot} \sim$ Dirichlet(β)

**end for**

**for** each seed set s = 1· · · S,

Choose topic distribution $\Psi s \sim$ Dirichlet ($\alpha^s$)

**end for**

Draw Θi ∼ Dirichlet(αyi,·)

**for** each word $0 \leq i < N$ in document $d$

**for** each s = 1 to S

**if** $f_d(i) = 1$ draw topic $z_i$ from ψs,

**else** draw topic $z_i$ from $\theta$,

draw $w_i$ from $\varphi_{z_i}$.

**end for**

**end for**

**end for**

---

Figure 2-7: Biased LDA process.

Using this method, we can input multiple dictionary seed words (such as the ADR disease symptom) to separate different entities into different topics.

### 2.2.3 Similarity Measurement

During the look up process to map background words to the dictionary data, it is critical that we reduce the significant percentage of spelling errors we found in the user comments, so we implemented a measurement of string similarity based on the Jaro-Winkler edit distance [10][11]. Jaro-Winkler was chosen because, in this string similarity metric, more weight is put on matching of the beginning of the strings, and it can be assumed that, even with a misspelling, the first few letters will be correct. This situation commonly happens in social media.

There are two options this score is used, option A: This scoring is used in F(n) as a similarity score between the tokens in the dictionary and the documents on which we perform topic modeling. F(1) will return 1 if the final score (d(n)) is greater than a configurable threshold.

$$f(n) = \begin{cases} 1, & d(n) \geq threshold \\ 0, & d(n) < threshold \end{cases}$$

If f(n) returns 1, we draw topic z from seed topic distribution, if returns 0, we draw topic from general topic distribution.

In option B, the distance score is mapping to a probability to choose between seed topic distribution and general topic distribution, if distance score is high, then we have more probability to draw topic from seed distribution, else, we have more probability to draw from general distribution.

### 2.2.4 Term Mapping

Informal language is widely used in social network user generated content. It is normal to see the gap between the terms used in social media and those professional terms in official media, such as the FDAs. In this study, we adopt mapping tools to map the professional terms to

the consumer terms. Those tools include Metamap[3], which maps the Unified Medical Language System (UMLS) to a unique identifier; the unique identifier can also be mapped to a series of terms. ADRmine[12] is used for ADR mapping. For mapping drug names, we use the rxNorm database. RxNorm [13] is a widely-used drug database that contains drug name variations and relationships from various source. There are multiple relationships defined in RxNorm, including "contains," "trademark of," and "brand name." During the mapping process, we start with the FDA drug category, then search those drug names in RxNorm for all variation of drug names that have relationships, then use all the name variations for mining purposes.

### 2.2.5 ADR Distribution Weight Calculation

In order to assign a weighted score to each ADR term to show the importance of the ADR token within the topic, we use a weighted score that we sum the weight of the known ADR tokens within each topic. Finally, we multiply the score with the topic distribution probability, this approach extends an existing ADR score calculation method[4].

Given an observed ADR term in a topic $Z_n$, the conditional distribution weight of the ADR term is defined as follows:

$$\frac{1}{Z}\sum_{k=1} P(w_j|z_k)P(z_k|d_i))\, Weight(d_i)$$

The process of calculating ADR score is shown in Figure 2-8. For each topic within each group, we check every word to determine if it is an ADR term or not. For every ADR term, we calculate the ADR score, then this score is aggregated on document level then normalized by number of topics to get the document level ADR score.

.

[3] https://metamap.nlm.nih.gov/

### 2.2.6 Big Data Implementation

This framework is implemented on Spark as parallel LDA to archive the best performance for a massive amount of data. There are some benefits to the parallel execution of the LDA model. First, we prefer this framework to processing massive amount of training data

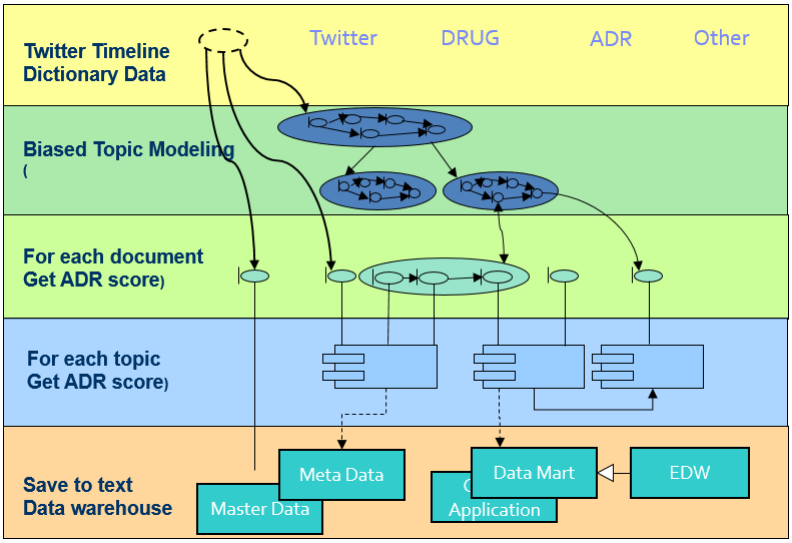



Figure 2-8: ADR distribution score calculation.

because more data means that more hidden semantic values can be revealed, and the model can be more accurate and achieve better performance. Secondly, we can train a “larger” model. We prefer that the model summarize more detail and long-tail topic models. For example, if we have 1 million topics, a larger model means the matrix for $N_{wt}$ can be large enough for our purpose. The size of $N_{wt}$ is decided by $V_{xK}$. V means the number of words and V is the number of topics. A large $N_{wt}$ will not easily fit into the memory unless we implement a parallel distributed execution. The algorithm is implemented in following major steps:

1. Transform RDD into bag-of-words.
2. Suppose we have multiple processes, each one of which will initiate a sampling process, and every process has an a copy of topic/word matrix RDD ($N_{wt}$).
3. Train (W,T) and $N_{td}$ in parallel, refresh $N_{wt,}$

4. Divide the matrix ($N_{wt}$) cross RDD to get the rows corresponding to the words in the mini batch.
5. MapReduce
    a. Map phase: Gibbs sampling is executed in mapper class to run sampling process in parallel.
    b. Reduce phase: Summarize sampling result from map phase for next iteration.
6. Update the rows of the topic/word matrix RDD corresponding to the words in the mini batch.

Figure 2-9 illustrates the overall flow of the distributed process.

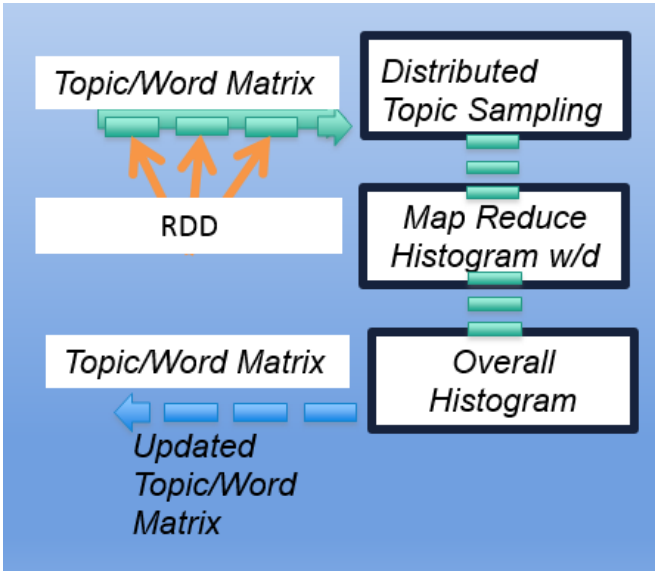



Figure 2-9: Distributed LDA on spark.

## 2.3 Experiment: Biased Model

### 2.3.1 Use Case

A study from 2012 has shown that 26% of online adults discuss health information using social media [14], with approximately 90% of women using online media for healthcare information and 60% using pregnancy-related apps for support. These statistics suggest that social media sources play a critical role for women as they contain key information and can be

used to share drug and ADR experience. Compared to clinical reports from health care organizations, the data from social networks is more generic and is not limited to geographical locations.

We collected user posts about drugs from Twitter. Twitter has been widely studied in recent years in the Heath Language Processing domain. It is a microblogging site that is actively used by over 320 million users. Their real-time tweets help health monitoring services and researchers in multiple ways. For example, by tracking tweeted first-hand reports of disease outbreaks, interested agencies can observe patterns of their spread and take appropriate actions to minimize the effects. One advantage of Twitter over other social networks is its high frequency of user tweets, which makes it easier to find drug mentions and user reactions compared to other social media venues. Hence, Twitter has been a widely-used source of social media data in pharmacovigilance research [15]. However, it comes with its own challenges in terms of information extraction due to its use of abbreviations, informal language, and colloquial terms.

It would be interesting to test if it is possible to elicit common public concerns, issues, and misconceptions about the disease from social media sources. In this paper, we study a subset of Twitter data, which is pregnant women's timelines. We queried the Twitter timelines of pregnant women, followed their activities on Twitter, and used their timelines as longitudinal networks, which revealed a wide range of information about users, including their drug use patterns. Then, we performed topic modeling with background knowledge to assess the prevalence of the use of different drugs among these pregnant women, obtained a distribution of ADR topics among those topics, and fed the output into the OLAP cube.

### 2.3.2 Data Preparation

We use several sets of data in this research: tweets pregnancy timelines, drugs lists, and ADR term. We also annotated some timeline data manually to evaluate the performance of the system. Reference libraries for formal biomedical dictionary data (ADR, Drug and Disease)

have been built, some data cleaning and data qualities techniques were utilized during the reference library building process[16], [17].

#### 2.3.2.1 Tweets Timeline

A total of 35,355 tweets during a one-year time period starting January 2014 and ending January 2015 were collected using Twitter's search API with a list of search queries. Some of the search queries used for collecting these tweets were "i am weeks pregnant lang:en since:2015-01-01 until:2015-07-31" and "i am months pregnant lang:en since:2014- until:2015-01-01." Once a user is determined as pregnant, all of the postings from that user are retrieved from the timeline. After those timelines are obtained, some post processing is then performed to identify true pregnancy timelines.

#### 2.3.2.2 Drug List Reference Library

The FDA has established some guidelines to category drug usage for pregnancy, this category indicates the risk of a drug to cause birth defects. Because of the difficulty to involve pregnant women in clinical trial, those categories are determined by animal study and post market reporting, so limited information is provided to define this category. There are five pregnancy categories—A, B, C, D, X—with A being the safest and X being the riskiest state to take the drug in during pregnancy. We divided the twitter timeline with the drug usage category, based on the following criteria.

1. For all timelines, label the timeline based on drug mentions extracted from the timeline. (A/B/C/D/X/Non drug usage)
2. Because multiple drug mentions from one timeline are frequently seen, the following criteria are adopted to group the timeline to the drug category:
    a. Any timeline that contains a category X drug mentions will be removed from the whole collection and moved to category X.
    b. From the rest timelines, any category that contains group D drug mentions will be removed and moved to category D.

c. From the rest timelines, any timeline that contains category C drug mentions will be removed and moved to category C.
d. From the rest timelines, any timeline that contains category B drug mentions will be removed and moved to timeline category B.
e. From the rest timelines, any timeline that contains a category A drug mentions will be removed and moved to timeline category A.
f. All remaining timelines are classified as N/A group.

#### 2.3.2.3 ADR Dictionary Data

The ADR dictionary term used for the background knowledge is the list of adverse reactions found in the SIDER database [18] and the MEDDRA ontology of adverse events [19].

### 2.3.3 Evaluation

"The unsupervised nature of topic models makes model selection difficult" [20]. Because of the unsupervised nature of topic model, the manner in which topic models are evaluated and their expectation has some disconnection. The basic approach to modelling validation is as follows. First, select a random subset from testing document, then, select a set of different topic model techniques to the rest of the corpus and approximate a measure of model fit (probability) for each trained model on the test set, at the end, the method which has the best held-out performance is selected [9]. Some validation and verification enhancement on top of this basic approach has been developed, those approaches can evaluation performance by some secondary tasks, which include document topic inference, unknown document estimation based on training document, information retrieval, social media response prediction, et al. We will explain these approaches below.

- Intrinsic evaluation metric: One widely-used topic modelling evaluation metric is Perplexity ([6] and [21]). Perplexity is an information retrieval measurement, simply speaking, it is the uncertainty of predicting a word, with lower number indicating higher performance. However, research has indicated that, in some cases, a good perplexity score does not always produce semantically meaningful topics[22].

- Semantic Coherent Score for topic quality: Another commonly-used method to validate and verify topic quality is the topic coherent score, which is either manually carried out by a human or automatically by a program. In [22], a human scoring approach was proposed to compute the word and topic “Intrusion score” and test the concept of “word intrusion” and “topic intrusion.” A group of intruder words and topics was manually selected (those words and topics that had low portability on the same topic or document) and annotators were asked to identify the intrusion words and topics. When the authors validated three models (LDA, PLSI and CTM), they found that the CTM has the worst score using this scoring system. In [23], a graph mining approach for the external evaluation of topic models was developed.
- Topic Likelihoods Estimating – Another approach to performing topic model comparisons is to use the likelihood of unseen documents to perform testing or model comparison. Several IR (information retrieval) methods have been proposed, such as harmonic mean [21] [24], Empirical Likelihood [7], and unbiased perplexity scores for held-out words for model validation [25]. However, those approaches that were based on traditional IR techniques were proven to be biased. The harmonic mean method often significantly overestimates, and simple IS methods tend to underestimate. In [26] and [20], several new methods for likelihood evaluation were proposed based on the left-to-right algorithm and the A Chib-style method [27]. In [28], a method for comparing the predictive performance of any topic model relative to a baseline model was proposed. This strategy is based on the annealed importance sampling (AIS) method as applied to the topic model evaluation. All of these methods focused on the statistical (or quantitative) evaluation of topic models [7]. However, the interpretability of topics are not measured in those approaches. [8] has some interesting finding that that there could be a negative correlation between a human and automatic evaluation of topic models[23].
- Automatic Topic Labelling – Although automatic topic modelling [29]–[31] cannot be applied to topic evaluation directly, it could be useful to find a single most representative

phrase (i.e., topic label or name) for each topic, then the summary of the topic modelling can be generated and compared and a user can interpret the discovered topic. This can, then, be compared with other models for validation and verification [31].

- Evaluation for LDA variance – There has been some work done to evaluate LDA variance, which is part of my research and which includes LDA topic models, hierarchical topic models, LDA topic models with knowledge priori, and hierarchical topic models with ontologies as background knowledge. In [32] the authors evaluated their system for topic modelling based on fuzzy set theory using the results of a document classification task. In [33], the authors showed that LDA effectively captures relevant topics in biomedical text by showing significant improvement in classification through adding LDA-based topic modelling. This evaluation methodology can be used to evaluate our LDA model. In [34], an interesting approach to automate the evaluation of topic modelling by using an external concept hierarchy has been developed. Topic coherency is used to define a conceptual topic score based on a concept hierarchy (ontology). This is a relevance-based approach, which inspirited the evaluation method in our research as well.

Overall, the evaluation method used is determined by the purpose of the topic modelling, particularly whether it is for prediction or just finding hidden semantic information. If prediction is preferred, then likelihood works better; otherwise, the score measure works better.

We have evaluated our results by using the topic coherent measurement(CM) [35], this method can help us to understand the coherency of the topic. We compare the topic words with human annotation and background knowledge dictionary to judge if those words are health issues related (disease, drug, ADR). We used four methods—a. classic LDA, b.LDA with lexical prior knowledge [5], c. biased LDA with option A , d. biased LDA with option B –and evaluated the results of those four methods using the topic coherent score. The score is measured by number of intuition words within ADR topic. We calculated the number of intrusion words by

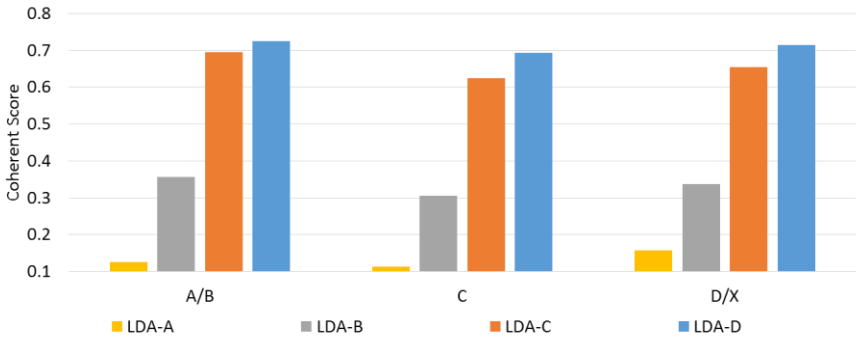
comparing the topics to the annotations from the Twitter Adverse Drug Reaction corpus from Ginn [3].

Two evaluation cases have been done, four drugs (serequel, trazodone, verlafaxine and vyvanse) were selected. Then drugs are selected based on FDA medication category (A-X), and timelines from those categories are aggregated then rerun the evaluation.

The result shows that our method has significant improvement in topic coherency over classic and LDA with lexical priories. In all cases, option A always outperforms option B.

A corpus of previously-annotated tweets was used to validate our results. Experiments were performed to assess the accuracy of the approach based on the ADR score. After that, we will translate our ADR score results into a classification model by different drug categories based on the ADR model.

Using the annotated data set[3], we also compared the ADR scores of the timeline with drug usage against timelines without any drug usage. We found that with our topic modeling method, timelines with drug usages had a minimum ADR score 2.23 times higher than the timelines without any drug usage. This provides evidence that our method creates topics that capture adverse drug reactions better.



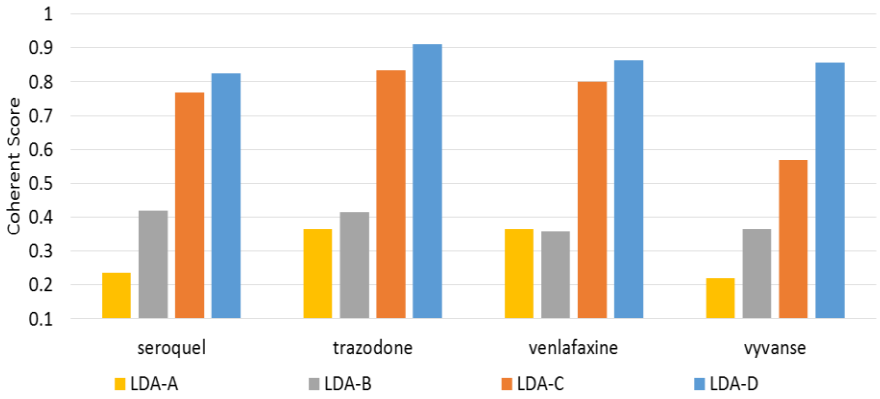



Figure 2-10: Topic coherent scores.

### 2.3.4 Result Analysis and Conclusion

To score each drug, every the timelines mentioning a particular drug were considered one document, then topic estimation is performed, the estimation will give approximate distribution of the topics and discover the portion of the document which could be covered by a given topic. We next calculated the score by combining the topic percentage with the per-topic ADR weight. Table 2-3 is an example of score by sampling timelines from each drug category.

From the initial results, shown in Table 2-3, we can see that the ADR score from the drug category is significantly higher than the timelines without drug usage; however, the ADR scores seem to be somewhat off from the A-X medications, where A should be the safest and X the most dangerous. The scores for the B category are significantly higher than for other categories, and X is lower than expected.

| Drug Category | ADR Score |
|---|---|
| A | 105.2 |
| B | 856.3 |
| C | 623.2 |
| D | 956.3 |
| X | 632.5 |
| N/A | 89.7 |

Table 2-3: ADR score for different drug categories.

To resolve the first issue, we undertook topic modeling for drugs in every category, and then we tried to average the drug ADR score within each category. Finally, we classified the drugs based on the ADR score into three categories (low ADR, medium ADR, and high ADR). The results are shown in Figure 2-11, with the y-axis showing the percentage of drugs within each category labeled as low, medium, or high ADR.

From Figure 2-11, we can see that the majority of category A and B drugs are classified as low ADR drugs, while for category D and X drugs, although some of these drugs are still classified as low ADR drugs, the proportion of drugs classified as medium or high ADR drugs is much higher. This implies that we are able to capture evidence from the results of topic modeling that during, pregnancy D-X cause more adverse effects.

C and D are ideal examples to indicate why we need to incorporate topic modeling with the multi-dimensional OLAP cube, which is also part of my thesis topic. It is highly desirable to aggregate text information in different dimension to for sophisticated analyses, where a topic hierarchy with roll up/drill down relationships can be performed.

There are some other reasons why the ADR score is not precisely matching the risk category, firstly, mentions of the presence of the ADR drug in Twitter posts are not evenly distributed, and with category B drugs being mentioned the most. It appears that Twitter users have taken category B drugs much more often than drugs from the other categories, secondly, on Twitter, recommendation is mixed with experience, for example, “If you felt sick this morning, better take an aspirin. Other reasons is that the occurrence of ADR does not necessarily related to the severity of ADR. For example, women taking D and X category could have ADRs like “miscarriage” or “suicidal thoughts”, this kind of ADR is much more severe than ADR from A and B groups, like “sleepy” or “nausea”, however, when we calculate the ADR score, only the number is occurrence is calculated, so it does not reflect the risk level.

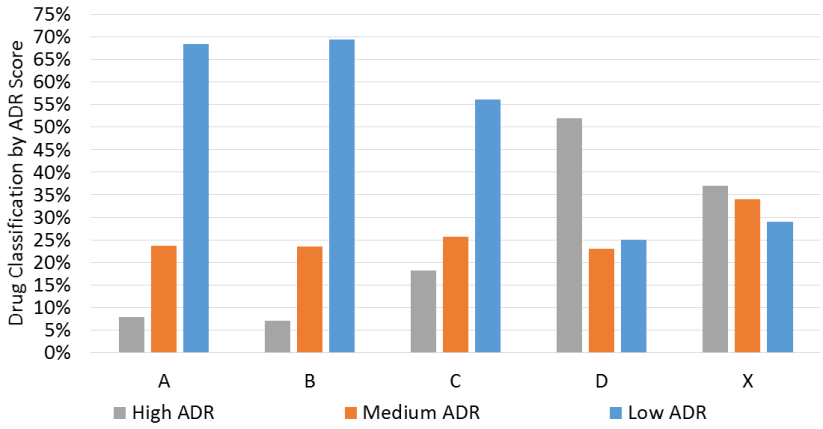



Figure 2-11: Drug categorization by ADR score.

## 2.4 Future Work

Some future work could be undertaken to improve the performance of this model.

- Data cleaning: There are existing online resources to map drug names and symptoms, and these can be used to remove symptoms from the ADR candidate set before running the topic modeling.
- In the data warehouse OLAP cube, when modeling the topics using the time dimension, the time dimension is based on the Tweeter timeline. Through this approach, we can distinguish the ADR topics that occurred before the date when the drug was mentioned, which we consider as disease symptoms after or on the same day of as the drug was mentioned. This provides us with what we consider to be the true ADR.
- Associating ADRs with pregnancy outcomes, we can categorize the outcomes of all of the valid cases of pregnancy into two categories: good and bad. A pregnancy is categorized as having a bad outcome if there is evidence of a miscarriage, stillbirth, or other birth complications. In this way, we can map the ADR presence with teratogenic effects.

# 3 TOPIC MODELLING PREDICTION APPLIED TO ONLINE SOCIAL MEDIA

## 3.1 Introduction

In this research, we adopt the biased topic modelling approach described in the previous chapter to accurately predict the safety quotient of drugs used by pregnant women. There are several goals: (a) Design an unsupervised topic modelling approach to automatically summarize ADRs based on UGC, and (b) automatically infer the semantic information from ADR events and use this semantic information to detect unsafe drug usage. The objective of our research is to build an efficient prediction system to accurately monitor unsafe drug usage.

The basis of drug usage prediction is through co-occurrence topic modelling for text summarization. The rationale behind such analysis is that if a drug can cause an adverse reaction, then the drug and the adverse reaction should be frequently mentioned together, where the frequency of co-occurrences can be summarized into certain topics. There are some challenges to summarize heath related (ADR/Drug) information from a general social network, and using it for prediction purposes. Varying the degree of information quality is one of the biggest challenges, as the drug mention and its corresponding ADR could be mentioned in different messages. For example, one woman could mention the use of drug Alomide in one tweet, while in another tweet, she mentions a miscarriage. Although the ADR and drug usage are in different tweets, it is possible that the miscarriage was caused by the drug usage. In addition, a social network will include a mixture of ADR information along with other noisy information such as hearsay or advice, and this could have a negative impact on the ADR mining performance. ADR and disease symptoms are similar and it is very challenging to distinguish between the two during the mining process. Since the twitter timeline contains large number of vocabularies with different variation, it is crucial for us to reduce the high dimensionality into limited number of latent topics, those topics can help to summarize the semantic structure of original document, and make it possible for us to extract useful ADR information from other noise data such as hearsay and disease symptoms.

Traditionally, a supervised learning classifier requires a large amount of labelled training data, and this labelled training data needs to be high quality in order to produce high quality classification. However, in a social network, high quality labelled data is sparse and highly distributed. For example, when we perform consumer ADR text mining, ADR related information is only a very small portion of the information in a social network UGC, so it requires tremendous manual work from subject experts to extract this sparse and highly distributed information. In addition, due to the fast growth of social networks, the labels may be outdated very soon. This small number of positive examples adds up to an additional challenge in terms of text summarization because the featured distribution of all the positive examples may not be easily represented.

The biased topic modelling method we propose provides an opportunity to address these challenges. Using timelines collected from a pregnant women's twitter, our proposed system integrates both biased topic modelling and dimension deduction techniques to automatically perform text summarization by extracting semantic topic features, and then predict the drug usage group by distance measurement. The unknown timeline can be classified into different drug usage groups based on the semantic similarity from existing labelled timelines.

We use several sets of data in this research, which includes tweets of pregnancy timelines, drugs lists, disease symptoms and ADR terms. By using timelines as longitudinal networks which reveal a wide range of information about the subjects including their drug usage patterns, this framework can be used to identify users of a specific drug usage category from social media (in this case, pregnant women). Secondly, collected data was used to design and test a prediction model which can perform feature reduction to summarize social network user generated contents into a limited number of topics. This framework is built top of unsupervised text mining classification and can be used to monitor and identify social media user generated contents for unsafe drug usage, by text summarization and monitoring important signal of potential adverse drug reactions. The proposed solution is to build on top of a specialized topic modelling approach which is customized for biomedical text mining, this can also assist agencies

like FDA or pharmaceutical companies to build an early warning system to prevent future ADRs. The prediction process is optimized to handle massive amounts of data by using a partially supervised learning method, which means that we only need a small sized positive data set to predict a large corpus of unlabelled data with similar test features.

We performed the evaluation of this framework on pregnancy timelines we collected from twitter and the preliminary results are encouraging. Currently, this framework can be used to predict drug usage categories based on the FDA risk categories. We hypothesize that this approach may eventually aid in categorizing given drugs into any risk classes based on the ADR risk. We verified its effectiveness via experimental evaluation with an official FDA pregnancy drug category.

### 3.2 System Overview – Biased Topic Modelling Prediction (BTMP)

In this section, we briefly introduce the background, including some underlying techniques used in existing work for prediction in an online health social network.

We use biased topic modelling as the basic prediction model for the proposed framework to classify and predict the drug usage group based on a semi-supervised learning classification approach. This classification approach is based on standard topic model inference and combines both text mining and features reduction techniques to summarize important text semantic features into a collection of topics. It starts with a limited amount of labelled data, and this limited labelled data can be used to predict the drug usage group of a twitter timeline with similar text features from a large amount of unlabelled data. The overall framework is shown in Figure 3-1, including the major components and data processing flow, the input and output for each component, and tools or techniques used by each component.

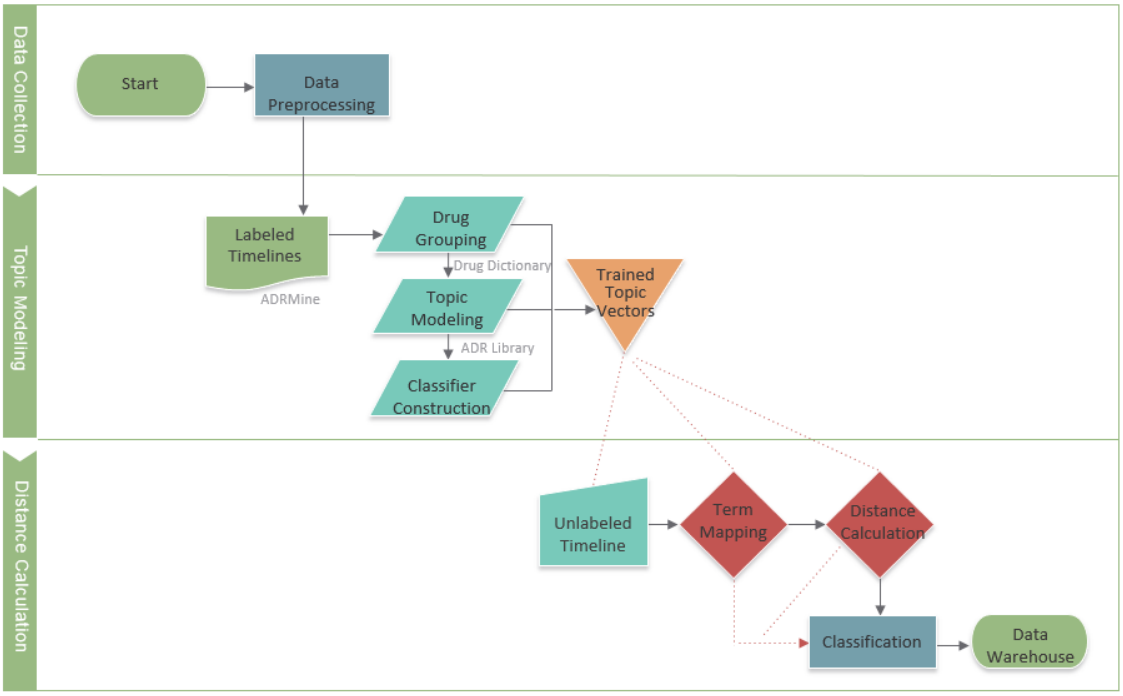



Figure 3-1: BTMP system overview.

After we perform the data collection and pre-processing, our prediction system applies the biased topic modelling to summarize the collection of twitter timelines into a collection of topics. The biased topic model can be used to represent documents/timelines in the form of latent topics, instead of the bag of words, this is useful for text summarization. By reducing the document dimensions into topics, the document semantic meaning can be summarized, and we can use this collection of topic space to calculate the distance between the incoming unlaced timeline and existing models. Some major steps of this process are described as follows:

1. Label tweets with the drug usage as training data
2. Train topic model from labelled timelines
3. Infer the topic distribution from the unknown timeline
4. Relate unknown timeline topics to each of the training labelled topics, and calculate a similarity score
5. Classification/prediction based on similarity score and voting algorithm

Finally, we make a prediction for each unknown timeline, and summarize the drug usage categorization via a voting algorithm, with this information being saved into a health data warehouse for information retrieval.

### 3.2.1 Biased Topic Modelling Training

During this process, we perform feature reduction so that we can summarize twitter timelines with high term based dimensionality into a low dimensional topic collection. We designed and implemented a biased topic modelling process which was used in this study to perform the feature reduction and text summarization, this way we were able to preserve as much of the semantic structure of the original document as possible.

Our biased topic modelling can produce a collection of topics that summarize a large number of timelines. During the topic modelling process, each timeline can be treated as a document, which can be modelled as a collection of topic distribution. A topic is modelled as a collection of word distribution. The outcome of this process is the topic association of each of the words from the timeline, which is the document-topic distribution and topic-word association distribution.

We group together timelines from the same medication category, then perform topic modeling on each group, this way, every group of timelines can produce a trained topic model, which is a collection of document/topic distribution and topic/word distribution, We can then use those trained models to infer hidden model from an unknown timeline using standard LDA inference and calculate the distance, the process is described in the section below.

### 3.2.2 Distance Score Calculation

Using the word-topic association from the previous step, we can automatically infer the topic structure of an unknown timeline, and then use this inferred topic structure to measure the similarity distance between the new timeline to the existing topic structure. Specifically, we can use a subset of labelled timelines to find other timelines with the same class label. This is distance based prediction is based on standard topic inference technique[37].

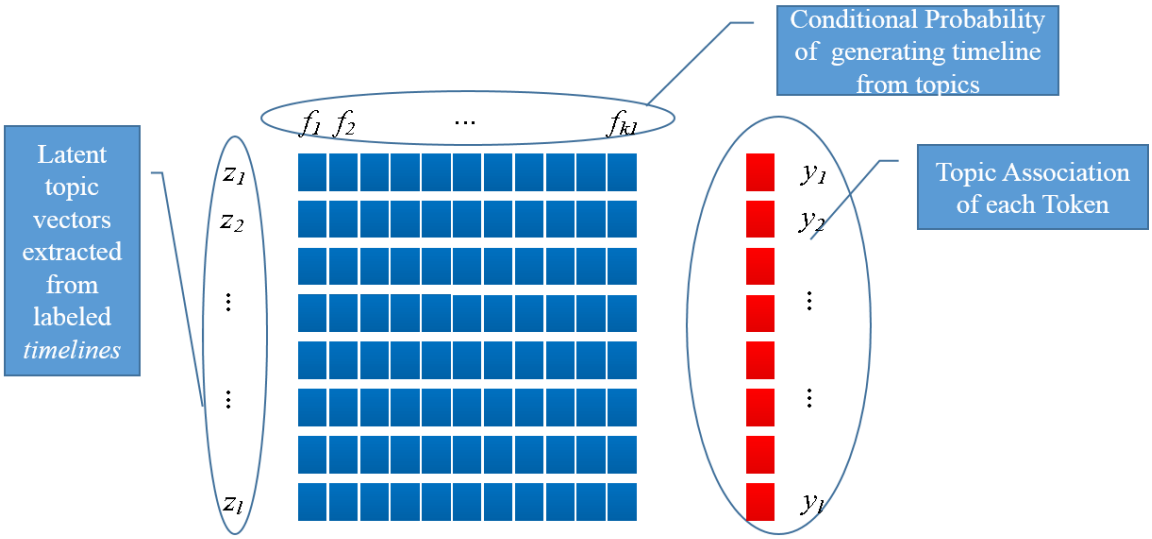



Figure 3-2: Similarity distance calculation process.

Figure 3-2 illustrates the similarity calculation process, during this process, two scores were calculated, one is the distribution score, and the other one is the ADR score. For each unknown timeline, we assign each of the tokens from the timeline to one of the trained topics using inference, then we can access the score from the topic/word distribution, this score can be normalized into a vector of probability, we then sum the weighted score of each token. Finally, we multiply the score by the distribution probability to get the weight score. This score can indicate the association between the document and the topics. For terms in the ADR dictionary, we give them extra weight, which is the ADR score. Conclusively, we can aggregate the conditional distribution score and ADR score to measure the distance between a new timeline with the labelled data, and in this way we can measure the semantic distance between an unlabelled timeline to the labelled timelines with medication category information.

For similarity distance calculation, we use a similar approach as in [37], where documents $\{d_m\}_{m=1}^{m}$ can be summarized into latent topic $\{Z_j\}_{j=1}^{T}$, and the conditional probability of generating document $d_m$ from unknown timeline can be represented as

$$p(z_j|x) = \frac{P(x|z_j)p(Z_{j)}}{\sum_{j=1}^{T} P(d_i|z_j)P(z_j)} = \frac{1}{z}p(z_j)\prod_{w\in d} p(w|z_j)$$

This is based on a standard LDA inference. Suppose we have a set of training timelines, Xn = {$p_0$,$p_1$,….,$P_{m-1}$}, label R = {$Y_0$,$Y_1$,….,$Y_{m-1}$}, during the training process, the conditional weighted distribution of the topic association of the words can be represented as

$$P(d_m|x) = \sum_j P(d_m|x_j^l)\, P(z_j|x) Similarity(x)$$

With the weight of each token, the aggregated score can be used to measure how close the words from the unknown timeline to the word-topic association we trained from the topic modelling process.

After we calculate the weight score, we calculate the ADR score, Given an observed ADR term in a topic $X_n = \{p_0, p_1, ..., p_{m-1}\}$, the aggregated conditional distribution weight of the ADR term is defined as follows:

$$P(d_m|D^L) = \frac{1}{Z}\sum_{x_j^l \epsilon d^L} P(d_m|x_j^l)\, Similarity(x)$$

### 3.3 Experiment Evaluation – BTMP

#### 3.3.1 Problem Statement

In this section, we describe experiments which sought to answer the following two questions - (1) can the proposed framework BTMP help to predict the drug usage? In our case, given an existing topic model S, and an unknown Twitter timeline T, can we predict the drug category usage? (2) How does background information affect the performance of BTMP? We begin by introducing experimental settings.

We performed our proposed method on user generated twitter timelines. Given a collection of user generated content from the social network, our goal was to predict the potential unsafe drug usage from unknown user generated content. The FDA group pregnancy related drugs into five categories, namely Categories A, B, C, D, and X. Pre-market clinical trials

assess the safety of drugs in limited settings, however the effects of those drugs on particular patient groups (e.g., pregnant women) cannot be assessed. In addition, spontaneous reporting systems that are in place for post-market surveillance, suffer from problems such as under-reporting [38]. It would be interesting to run a drug usage category classification using our method, and then compare it to the official categorization to see the difference. In this research, firstly, we developed a text summarization system, which can identify the health related topics from user generated data, then associated the ADR information with the corresponding drug category. Second, we have developed an automatic drug categorization usage system to predict the drug usage category based on sematic information. We integrated them to build a drug early warning system, which can be used for drug misuse identification. We verified its effectiveness via experimental evaluation with an official ADR knowledge base as well as human-annotated grounded truth.

The input of this process is a collection of twitter posts with labelled drug usage. The output is the predicted drug usage category. By using a small number of labelled data, we can run a prediction for a large amount of data, which makes it possible to identify the drug usage on the social network. In this research, every timeline from Twitter is considered as a document for topic modelling, which is represented as $D0 = \{d_1,...,d_n\}$, all the drugs can be represented as $M = \{m_1,...,m_p\}$, and the topics can be denoted as $S = \{s_1,...,s_q\}$.

### 3.3.2 Data Pre-processing

**Timeline of Tweets:** A total of 255,355 tweets during a one-year time-period from Jan 2014 to Jan 2015 were collected using Twitter search API with a list of search queries. Some of the search queries used for collecting these tweets are “I am weeks pregnant lang:en since:2015-01-01 until:2015-07-31”, “I am months pregnant lang:en since:2014- until:2015-01-01”. Once a user was determined to be pregnant, all the postings from that user were retrieved across the timeline.

The tweets were obtained from 15,523 different users, thus we retrieved 15,523 timelines[26]. Timeline statistics and distributions are shown in Table 3-1.

| Range | <100 | (100,1000) | (1000-2000) | (2000,3000) | (3000-4000) | >4000 |
|---|---|---|---|---|---|---|
| Distribution of number of tweets per timeline | | | | | | |
| #of tweets | 424 | 1852 | 2251 | 4193 | 6601 | 208 |
| Distribution of number of tokens per timeline | | | | | | |
| Range | <10000 | (10000,20000) | (20000-30000) | (30000,40000) | (40000-50000) | >50000 |
| # of tokens | 1959 | 2132 | 3157 | 4667 | 2731 | 915 |

Table 3-1: Statistical analysis of timelines.

**Drug List:** The FDA has established some guidelines to category drug usage for pregnancy, this category indicates the risk of a drug to cause birth defects. Because of the difficulty to involve pregnant women in clinical trial, those categories are determined by animal study and post market reporting, so limited information is provided to define this category. There are five pregnancy categories: A, B, C, D, X, with A being the safest and X being the riskiest type of drug to take during pregnancy.

**ADR dictionary data:** The ADR dictionary terms used for the background knowledge was the list of adverse reactions found on the SIDER database [18], and the MEDDRA ontology of adverse events [19].

#### 3.3.2.1 Drug Category Grouping

Here, each timeline is mapped to the drug usage. To categorize the drug usage on the timeline, we start with a lexicon-based approach to categorize all the timelines into five drug usage categories: A, B, C, D, and X. A timeline is categorized as having a certain drug usage if there is evidence of drug usage from a certain drug category. For example, phrases like “adrenaline”, “ovidel”, “aspirin” etc. were searched for using the lexicon.

Although lexicon based categorization is a straight forward way to determine the drug usage, we can see that not all tweets with drug mentions are based on legitimate drug usage. There are variety of reasons for a drug mention, for example, the tweets could talk about drug

usage by a friend or family members, or it could be a general recommendation from friends. In order to detect true drug usage, we adopted a mixture of methods to eliminate tweets which were not related to legitimate drug usage. Some manual annotation and supervised classifications were adopted to determine the true drug usage timeline. We adopted a Support Vector Machine [39] to classify true drug usage from negative drug usage. SVM is a supervised classification technique used for both regression and classification problems. Given training data, SVM finds an optimal hyperplane from which it classifies new data. SVM works best for two-label (binary) classification. In our case, SVM was employed to perform the initial classification to split the timelines into “Drug Usage” and “Non Drug Usage” groups, by various factors such as the presence of ADR, drug name, and symptom. After we split the timeline into “Drug Usage” and “Non Drug Usage” groups, we use a quantitative method for determining the approximate category of the drug, based on its frequency of discussion in social media. We posit that this method in combination with other classification techniques will help to achieve higher accuracy rates.

During the study, we found that that it was normal for a timeline to contain drug usage across categories. To resolve conflicts, if a duplicate was found in two categories, we removed the drug from the category with the lower severity. This technique gives the drug a category with a higher risk. For each drug across the 5 categories, we record the count of unique users who have mentioned the drug.

- Total of 7,387 Drugs from 5 FDA categories [40]
- For each drug, use the first word to build a mapping table using the rxNorm database[13].

  A summary of the results from the drug search is given in table 3-2.

| Drug Category | # of Drugs | # of Timelines | # of Mentions |
|---|---|---|---|
| A | 79 | 1,520 | 7,951 |
| B | 312 | 6,523 | 14,252 |
| C | 123 | 963 | 6,827 |
| D | 67 | 419 | 1,420 |
| X | 156 | 2,128 | 13,342 |
| N | 33 | 215 | 231 |
| N/A | 0 | 3,755 | 0 |
| **Total** | **770** | **15,523** | **44,023** |

Table 3-2: Statistical analysis of drug mentions.

#### 3.3.2.2 FDA Non Categorized Drug List Generation

Not all FDA approved drugs are placed into an FDA category." Less than 10% of medications approved by the U.S. FDA since 1980 have enough information to determine their risk of causing birth defects" [10]. In this research, we run through the FDA un-categorized drug list to predict which drug category (A-X) the uncategorized drugs should belong to. In this way we can validate the effectiveness of our system.

To generate a list of un-categorized drugs, a master list of drugs from FDA database was compiled. This list was then compared with the FDA categorized drug list and a list of non-categorized drugs (2,423 drugs) was compiled.

### 3.3.3 Biased Topic Modelling

In this step, we use training data to train a probabilistic topic model for timelines for each drug usage category (A, B, C, D, X), and then build a weight map for our topic model. This process is based on a text summarization process which is on top of the biased topic modelling and co-occurrence analysis. It assumes that if a drug and an adverse reaction co-occur frequently in the same information unit, then the drug is considered likely to cause an adverse

reaction. After we build the topic models, we use topic models as a feature set to build prediction system where we attempt to identify the drug usage group based on clinical text.

This case involves learning topic models from labelled timelines, estimating the topic distribution in the timeline, and then generating the document/topic and topic/word distribution. This way we have a weight map for our topic models, and then we can use this weight map to assign a score to each of the unknown timelines using the following main steps:

1. Label tweets with the drug usage as training data
2. Infer topic modelling from labelled timelines
3. Infer topic distribution from the unknown timeline
4. Relate unknown timeline topics to each of the training labelled topics, and then calculate the similarity score
5. Classification based on the similarity score

In the previous step, we summarized the timelines into a set of topics. In this step we will generate a similarity score from those topics. The basic concept is that since each topic contains the distribution of tokens, we can use this distribution to measure the weight of a token within a topic using a score, and this weight becomes the main component of our scoring system.

In the scoring process, we first train topic models for each medication groups we wish to use in the timelines to run the prediction, plus the non-drug usage group. We then estimate the distribution of topics, the purpose of the estimation is to get idea on the portion of text which can be represented by the topics, which can be used for parameter tuning. ADR score is calculated from each unknown timeline by combining topic and word distribution of ADR terms and then normalization on top of drug usage group.

Using biased topic modelling as the basic model for the prediction system (BTMP) to classify drug usage, the proposed framework exploits background knowledge to improve the performance of the drug usage prediction. In the next sections, we will go over the method and evaluations.

#### 3.3.3.1 BTMP-A

The basic mode is based on standard topic model inference, on top of the basic prediction model, we develop two variances to improve the performance. In this variant (Option A), we go through different steps to categorize the timelines into different grouping. First, we divide the timelines into Non Drug usage group and drug usage group, generating topics for those two groups. The reason why we have this step is that because the N/A group is very similar to A and B group. By having two groups first, we can better classify NA group from drug usage group by training a NA topic model and drug usage group topic model. Then we future divide the medication usage group into three sub groups for training purpose, which are A/B, C and D/X group. We then summarize the timelines from each of three groups into topics, finally we use the conditional probability to train the labelling timelines, and calculate the distance between an unlabelled timeline to the topic models from each category. Distance calculation is based on standard topic model inference. For every unknown timeline, related tokens are assigned to one of the topics from each category and then the topic distribution can be represented as a vector. For example, after we train the topic modelling using a labelled timeline, we have a collection of topics from each drug category: (Drug, ADR, Disease symptom…). Each topic represents a group of topic/word distribution. When we have an unknown timeline, e.g. “I took Zyrtec yesterday and it caused back pain for me and I could not sleep”, each of the tokens from this timeline would be assigned to one of the trained topics, which returns both a post-topic association and a word-topic association. After the mapping, we have a vector which represents the topic distribution of all the tokens like [0.6, 0.1, 0.1], and then we can aggregate the conditional probability [37]. Aggregated scores from different drug usage groups are ranked and the unknown timeline is classified as the corresponding group with the highest similarity score. When we have an unknown timeline, it will be compared with Non-drug and drug usage group first, if the timeline is identified as drug usage groups, then we repeat this process within three subgroups, A/B, C and D/X until the final groups is decided.

#### 3.3.3.2 BTMP-B

For option B, we took a slightly different approach to the build the background knowledge. We still trained the topic models for the four groups – Non drug usage, A/B, C and D/X. We used Non drug usage group topics as the background model, calculate the difference between the background topic and drug category topics. During similarity calculation from timeline to each models special weight is added to those words appearing in drug usage topics with less similarity to background words.

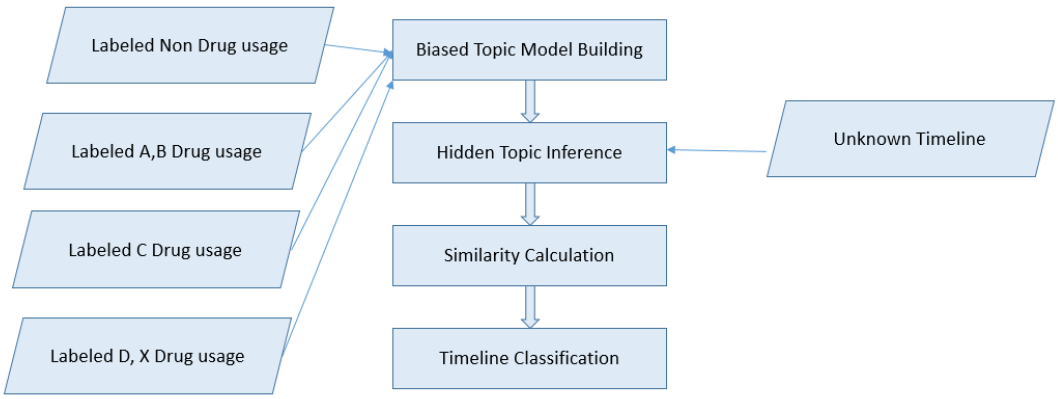



Figure 3-3: Distance calculation option A.

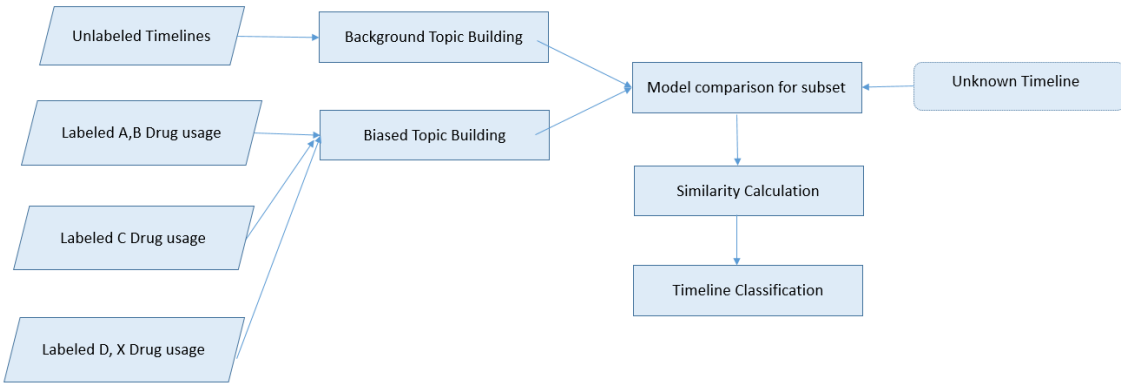



Figure 3-4: Distance calculation option B.

### 3.4 Result and Evaluation

In this section, we perform extensive experiments on real-world online twitter to evaluate the effectiveness of our model. We perform both a qualitative evaluation and quantitative evaluation. The evaluations were conducted by comparing the predicted drug usage group from the knowledge base or a human-annotated ground truth.

#### 3.4.1 Ground Truth

We use two types of ground truth to evaluate the performance of the different methods. The first type of ground truth was the official FDA Pregnancy Categories. The FDA Pregnancy Categories have multiple sources and are updated every year. This data set allowed us to evaluate the proposed methods in a large-scale fashion and with negligible human effort. The second type of ground truth was the pregnancy category from other countries such as Australia. This data source provided us with some reference on drug risk when a drug was not found or conflicting information was found from FDA data.

#### 3.4.2 Comparison Method

We tested our proposed methods with different options and compared them with existing supervised or semi-supervised learning methods. We compared the proposed approach in this study with an existing supervised learning method, the Multi-Class Supervised LDA Prediction [41][42]. We also compared our method with LDA with lexical priors [5]. Multi-Class Supervised LDA is chosen because it is a widely used prediction method on top of LDA. We also choose LDA with lexical priors because it is an approach which is similar to ours, which utilizes background knowledge.

As our goal is to predict the ADR drug usage group, this can be evaluated with the set-based evaluation measures: precision (P), recall (R), and f-measure (F). The ground-truth pairs are the sampled data set. For a given drug group, the precision is defined as the ratio of the number of correct predicted timelines to the total number of timelines within that group; the recall is the ratio of the number of correct predicted timelines to the total number of timelines in that data set; and the f-measure is the harmonic mean of the precision and recall.

$$p = \frac{tp}{tp+fp} \qquad r = \frac{tp}{tp+fn} \qquad f = \frac{2*p*r}{p+r}$$

### 3.4.3 Evaluation

Because of the sparse distribution of ADR posts in social media, and the unbalanced radio between positive and negative data set, we tested our methods using a different number of timelines. Performance was evaluated by setting different number of topics. To compare the performance of all four methods, we performed evaluations by cross-validation. Training data was divided into k-sized equal subsets. In one iteration of the cross validation, out of these k subsets, k-1 subsets were given to the classifier for training and the remaining subset was used as a testing set where the classifier assigns the label for each data in the testing set. Similarly, k different iterations were run using a different testing set each time and the overall performance was based on the mean performance of each iteration. Testing sets are generated by random sampling from overall data set . Finally, the evaluation result was measured by averaging the score from all the test runs.

During the cross-validation process, the timeline for the same drug could result in a different prediction result. In this case we adopted a voting algorithm which means that the prediction result is decided by the result from the majority of the timelines.

#### 3.4.3.1 Parameter Analysis

Before we perform the final evaluation comparison, we used a subset of data to determine the optimal value of parameter iteratively. During the topic modelling process, the number of topics (T) is one of the most important parameters affecting the performance of the topic modelling, because the dimensionalities of the topics are critical for the model training. It is very useful to explore how the different numbers of topics affect the overall performance.

To determine the optimal number of topics, we performed several rounds of experiments with different parameters for the number of topics by using the chosen data set. For this parameter analysis, we randomly sampled a subset of timelines. We use this same data set to run multiple topic modelling by changing topic number from 50 to 500. Figure 3-5 shows

that the performance of the model with different number of topics. We observe that the performance of the model reaches the maximum score when the topic number is between 100 and 200.

After we determine that the T should be between 100 to 200, we tried another round of different topic numbers T {10,20… ,190,200}. Starting from 10 and increasing in steps of 10 up to a maximum of 200. This result shows that the performance of the topic modelling increases when the number of topics is less than 160, and then decreases when T > 160. Based on this result, we chose T =160 as the number of topics. Figure 3-5 shows the ADR result using different numbers of number of topics.

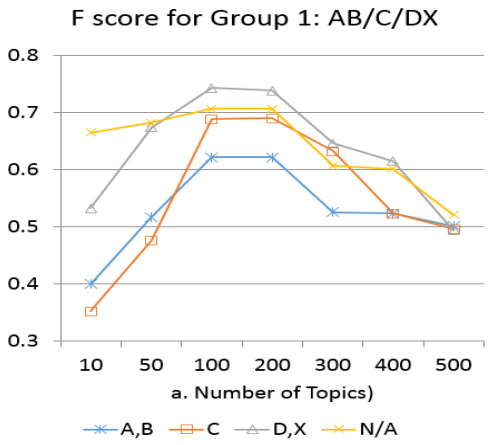



Figure 3-5: Effect of drug grouping on F score for different topic sizes.

During the process of parameters tuning, we noted that the properly setting parameters are crucial for the result, there is normally a range of parameters which can help to archive ideal result, it is important to perform different parameter setting to find this range.

Table 3-3 shows the notation of the two comparison methods, LDA-MC[41][42] and LDA-LP[5], and two variants of our proposed methods. We chose the comparison methods from two different approaches, the supervised LDA and the semi-supervised LDA, both of which are popular methods for topic modelling. For all four of those methods, we first determine the optimal topic numbers for each method, then we randomly sample the training and testing data,

finally we predicted the timeline to a drug usage category. The prediction accuracy was measured based on whether the predicted category matches the FDA category. When there were multiple timelines for one drug, a voting algorithm was used for the overall prediction.

#### 3.4.3.2 Performance Comparison

Table 3-4 shows the performance comparison of the four methods. By comparing the results, we can see that:

- The proposed biased topic modelling can significantly improve the prediction performance. From our study, compared with the performance of the other method, BTMP can improve the performance by 10-20% with a small amount of labelled data. Using the ADR dictionary as background information, the semantic meaning of the text is well preserved and sparse information is weighted higher during the modelling process. This result demonstrates the importance of background knowledge in topic modelling prediction.
- BTMP-A has better performance than BTMP-B, the difference between that options A and B are the distance measurement options. Where Option A measures the distance between an unknown timeline and the probability vectors derived from each drug category, and each token from the unknown timeline is assigned a weight and this weight is aggregated to obtain the overall score. And Option B focused on the generated topics by comparing drug/ADR related topics and the background topics which come from the non-drug related topics. Option A is observed to achieve a better performance in measuring the similarity between an unknown timeline with the existing topic model.
- The multi-class supervised LDA uses a similar method for the similarity measures as BTMP with the difference being how the topic model is trained. Multi-class supervised LDA uses a supervised approach with labelled data, while the BTMP is semi-supervised. From the results we can see that BTMP has better or comparable score than the multi-class supervised LDA for every group. Considering that this is a purely

| | | |
|---|---|---|
| Biased Topic Modelling | BTMP-A | Biased Topic Modelling Option A |
| | BTMP-B | Biased Topic Modelling Option B |
| Supervised | LDA-MC | Multi-Class Supervised LDA |
| Semi Supervised | LDA-LP | LDA with Lexical Priors |

Table 3-3: Notation of the comparison methods.

| Method | Non Drug | | A,B | | C | | D,X | |
|---|---|---|---|---|---|---|---|---|
| | Precision | Recall | Precision | Recall | Precision | Recall | Precision | Recall |
| BTMP-A | 0.8109 | 0.7106 | 0.6896 | 0.5935 | 0.7449 | 0.7812 | 0.8504 | 0.7660 |
| BTMP-B | 0.7659 | 0.6985 | 0.6015 | 0.5518 | 0.6844 | 0.5906 | 0.4802 | 0.7710 |
| LDA-MC | 0.7623 | 0.6652 | 0.7454 | 0.5017 | 0.6954 | 0.6835 | 0.6501 | 0.7955 |
| LDA-LP | 0.5623 | 0.5158 | 0.4401 | 0.5617 | 0.4102 | 0.3752 | 0.3562 | 0.5834 |

Table 3-4: Result of comparing BTMP again supervised and unsupervised methods.

supervised method, this requires large effort in the labelling data process, and the feature set is much smaller. This demonstrates that the proposed semi-supervised method can achieve similar or better performance with supervised methods.

- BTMP has a good trade-off between the precision and recall compared to other methods. When the training size is relatively small. The multi-class supervised LDA achieves a similar to higher precision and recall to BTMP, and when the training size is large enough the BTMP can achieve better precision than the multi-class supervised LDA. The reason for this is that the BTMP method requires a larger amount of training data for summarization with background knowledge, which can result in a more robust performance with larger amounts of data.

Overall, by introducing background knowledge into topic modelling process, the prediction and inference performance can be improved significantly.

### 3.4.4 FDA Non Categorized Drug Prediction

In this task, we established some comparisons of the timelines for women taking different kinds of medication. And then we examined if the timelines of those that mentioned taking uncategorized medication matched the timelines of those taking A & B or those taking D&X or C, thus suggesting a potential categorization for the future.

To generate the set of un-categorized medication, we performed the steps.

- Compile a full drug list from the FDA drug database: drug name/brand name/ingredient:
- Compile a FDA pregnancy category list: (A/B/C/D/X):
- Compare data from steps A and B, and generate a non-categorized drug list

Once we generated the list of non-categorized drugs, we captured the timeline with the drugs on this list and then run through the prediction process.

| Predicted Medication Group | Percentage |
|---|---|
| Non Drug Usage | 64.1% |
| A, B | 17.9% |
| C | 15.3% |
| D, X | 2.7% |

Table 3-5: FDA Non categorized drug prediction.

### 3.4.5 FDA Multiple Categorized Drug Prediction

Alternately, instead of running predictions on drugs which are not in a FDA category, we also ran the prediction for drugs which appear across categories. Currently, we there are around 120 drugs that can be found in multiple categories. There are several reasons why they exist in multiple categories:

- The drugs come from three different sources, thus there is conflicting information about these drugs.
- FDA has moved drugs into different categories over the years, which has usually involved moving C category drugs to the AB or CD groups.

By running predictions on these drugs and comparing the prediction result to existing categories, it is possible to determine whether the existing categorization is more valid. This can help us to validate the accuracy of the drug category by using social media.

### 3.4.6 Topic Modelling Observations

To score each drug, we performed topic modelling on all four groups, where all the timelines that mentioned a particular category were considered as one document. Then we ran the topic estimation to estimate the distribution of the topics, and what percentage of the document could be explained by a given topic.

| Predicted Medication Group | Percentage |
|---|---|
| Non Usage | 5.4% |
| A, B | 19.2% |
| C | 48.3% |
| D, X | 17.1% |

Table 3-6: FDA multiple categorized drug prediction.

We undertook visualization of the topics from each drug category. This visualization could help to discover the topics and the underlining timeline structures, and then we could use this kind of visualization to identify the corpus structure from different drug categories.

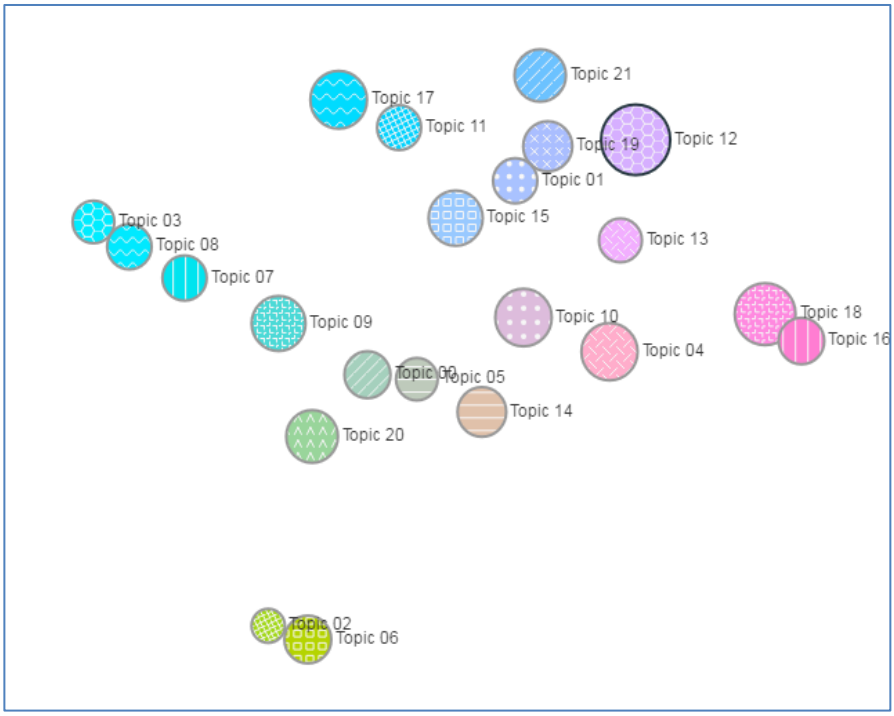



Figure 3-6: Non medication usage group topic visualization.

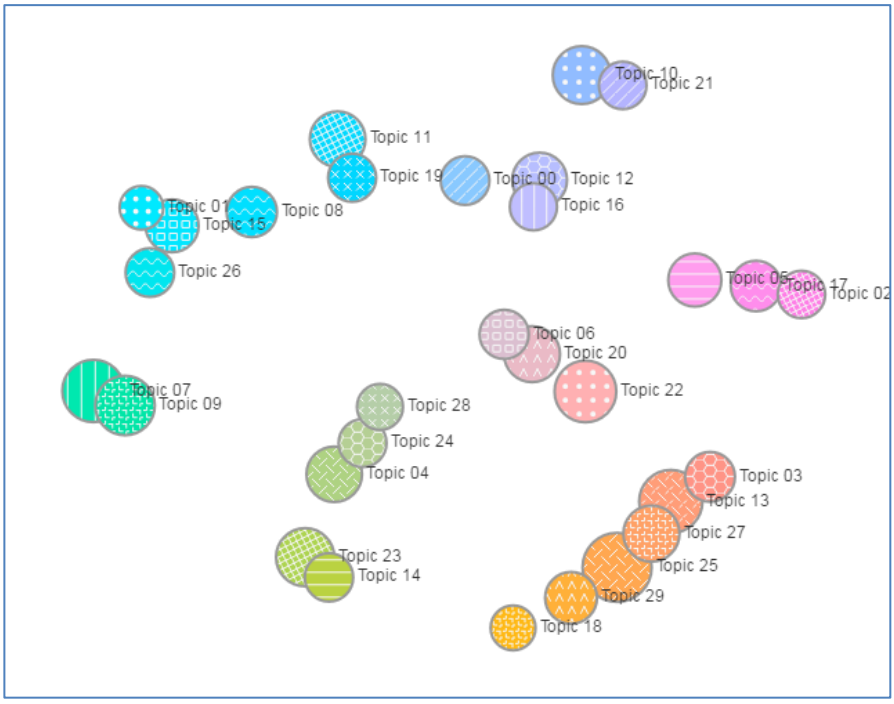



Figure 3-7: A/B group topic visualization.

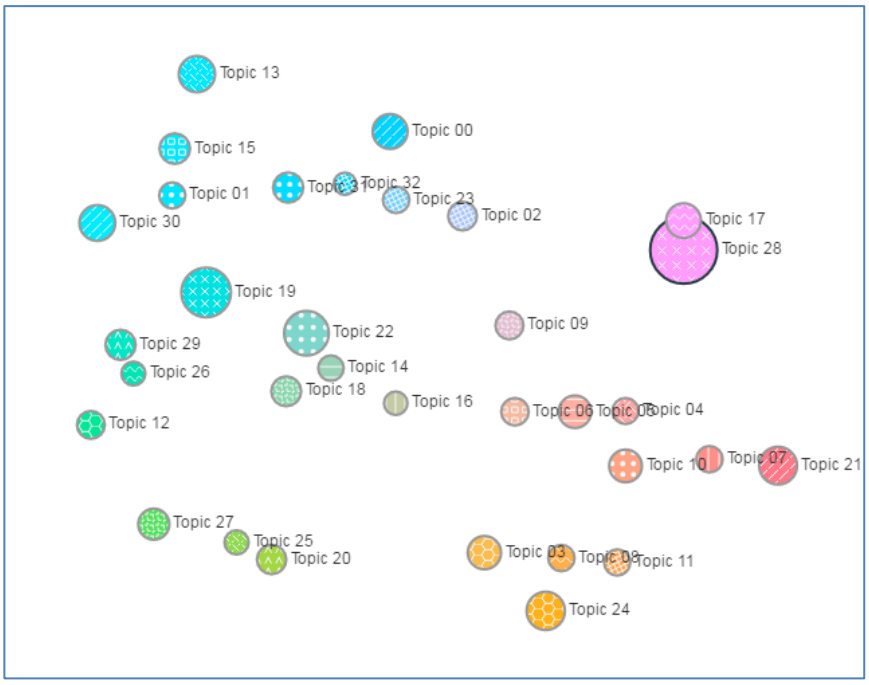



Figure 3-8: Group C topic visualization.

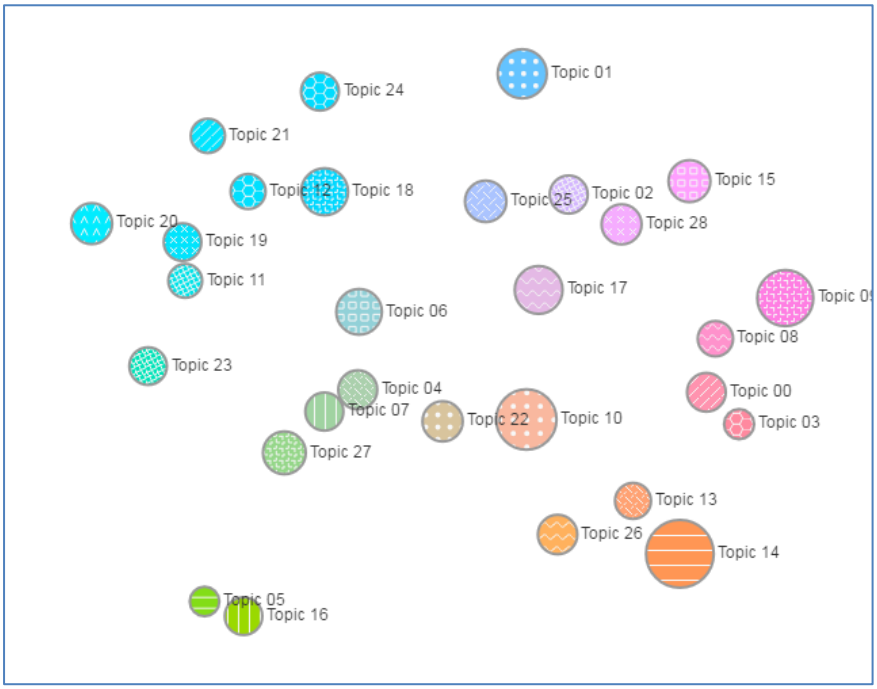



Figure 3-9: Group D/X topic visualization.

In the visualization, the topics summarize the timelines where each topic is visualized as a shape. The probability of the terms is defined based on the different colour schema in this visualization, with the yellow colour as ADR related vocabularies, and related topics being connected by sharing the same colour scheme. As we can see from this visualization, the N/A group contains almost no ADR related topics except for 1 or 2 with a very light yellow colour scheme. From the A/B group, we have 4-5 different ADR related groups, and those topics are almost the same size, which means that they have a similar distribution probability. However, for the D/X group, we have ADR distributed in one or two large shapes. This visualization matches with our observation, that the A/B group contains the most commonly used drugs and has been mentioned large amount of times in the timelines. The D/X group has relatively limited ADR mentions, however, these mentions always come with severe consequences such as a miscarriage.

We are able to infer some meaningful observations from these results. We performed some analysis on the predicted results and compared them with the documented category, and then the additional information could be used to improve the accuracy in identifying ADRs from the social network. By running prediction on FDA un-approved drug list, the prediction results show that most drugs in this list are predicted to be non-drug usage group, some of them goes to group A/B and C. For prediction on FDA multiple-categorised drug list, most them goes to C group.

### 3.4.7 Error Analysis

Figure 3-10 shows the drug category prediction across different groups. It can be seen that for timelines with D/X groups, these had the highest accuracy, while the accuracy was considerably low for the A/B group. For drug group C, most missed predictions went to A/B, which means that for drug usage group C, high risk drugs were easier to predict than low risk drugs.

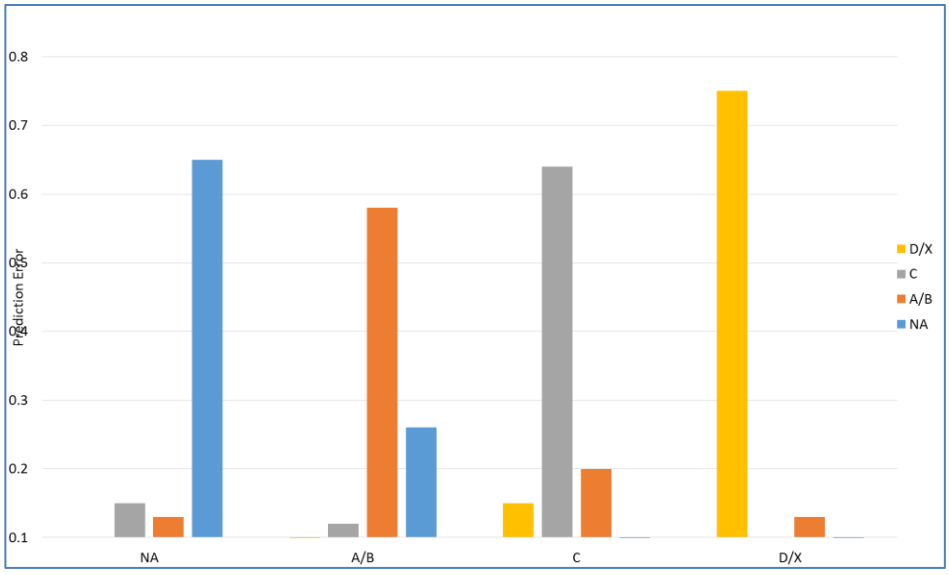



Figure 3-10: Drug category prediction distribution.

Some error analysis on the reasons which affect the prediction performance was first conducted. The greatest challenge was to differentiate the health related terms from other information in the user generated data where informal language is widely used. The prediction performance was decreased due to the variation in the medical terms such as the drug name, disease symptom, and ADR, and because of this, some social network acronyms were not identified by the text mining system because of the informal language being widely used on the social network. Building a medical term variation library can help to improve the performance.

The second factor which affects the performance is the quality of the data. Not all the terms from ADR vocabularies are really ADR, sometimes they are recommendations, common feelings, symptoms, or somebody else's experiences. We sought to resolve this quality issue in the topic modelling building phase by introducing a bag of words describing symptoms, so the ADR topics could be separated from the symptom topic. Another solution that could be introduced is to utilize context information so that we can better differentiate personal experience from recommendations and other people's experiences.

As we can see from the comparison result, BTMP-A outperformed BTMP-B, which indicates the importance of preserving the whole picture of the topic model structure for distance measurements.

The parameters in these variants were determined via cross-validation and the result shows that different parameters should be chosen for different methods.

### 3.5 Conclusion and Future Work

In this study, we have proposed a novel framework to predict drug usage from a social network. This framework contains two important modules: the biased topic modelling for text summarization, and the classification process for the drug usage group prediction.

Biased topic modelling serves as an important text summarization method in this research, which can be used to filter the limited information we need from the social network. One major contribution of this research is that the proposed approach can capture the low density ADR mentions from high dimension social network content, and preserve semantic meaning in topic models. Moreover, it also avoids performance downgrading when we perform text summarization from the training data.

We performed the evaluation by using data previously annotated by experts. First, we conducted experiments to determine the parameters, such as the number of topics, and we also compared the performance of our methods with existing methods. The results suggest that the proposed method outperforms the existing method, and thus the outcome of this research can be used to act as early warning system to detect new ADRs.

In future research, we plan to incorporate more types of user generated content. Twitter is a general social network where users can discuss any topic, in contrast, in some forums such as Dailystrength, users tend to be more focused on health related issues. However, the contents of these more specialized forums tend to be far less than Twitter. It would be interesting to combine Twitter data with health related social network data. Secondly, the topic modelling/training/prediction process needs to be retrained every time there is an update in the

reference data, which is a time consuming process. We plan to incorporate the concept of the transfer learning model, where knowledge can be transferred from one domain to another.

# 4 TOPIC MODEL TEXT DATA WAREHOUSE

## 4.1 Introduction

Topic models can help summarize a large collection of text into topics. However, in order to make the results from topic modeling useful, we need the ability to exploit the structure to aid in discovery and exploration. For example, we have the following challenges when exploring the result from the topic model.

- The visualization of the topics: The output from topic model is always hard to be visualized by a normal user. This is attributed to the fact that there is a large collection of topic and words, and not all topics can be interpreted easily. A common way is to display the top words from each topic to the user, however, this method has limitations as well because not all topic words are related.
- The interpretation and labeling of topics: Another challenge is to interpret and label certain topics from a document level, even though topics come with hidden information on how those words are grouped together. Automatically labelling the topic is beneficial, so that the user can quickly identity the meaning of topics.
- Topic connection: It is always interesting to infer how a topic is connected to the original document, this can help to explain how the document is formed.

Those challenges provide opportunities for us to further utilize topic model to provide meaningful information for text summarizing. This meaningful information always comes with different dimensions, such as age, gender, locations, and other medical related information. Deep diving into the topic information from different dimensions can further expand the insights into the useful hidden semantics information for researchers. In the biomedical domain, we can use this information to answer the following questions:

- How do we develop a tool to examine if there are any differences in the ADR distribution for known and unknown ADR for a particular drug?
- How do we develop a tool which can be used by the user to see the ADR distribution from a hierarchical view by drilling down and rolling up (Drug category - > drug -> ADR)

- How do we compare the topic distribution from different level of cube cells so that the user can compare the ADR across different drugs

The multidimensional database is widely used in industry to help answer these kinds of questions for a traditional database/data warehouse. An OLAP cube is a common way to accomplish this goal. OLAP is built on top of the multidimensional database, and it is aggregated from data warehouse to provide online analytical processing (OLAP) power (Figure 4-1). It is normally built from atomic detail data from a star schema database and provides a multi-dimensional, aggregated data view. In a multidimensional database, dimensions are used to aggregate measurement, then this aggregation across different dimensions can be used by OLAP cubes to improve query performance over relational databases. Although the data for OLAP is directly coming from a traditional database, OLAP cube manages data differently from traditional database, normally in an aggregated way.

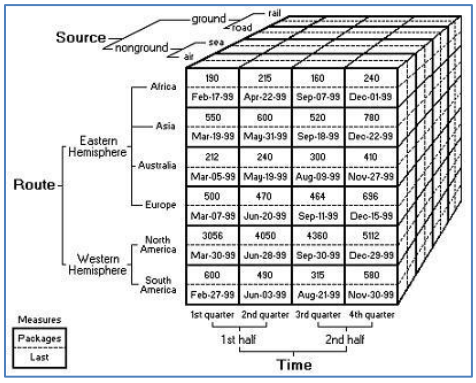



Figure 4-1: Sample multidimensional database.

An OLAP cube is optimized for read-only analytical purposes. Typical operation are slicing and dicing, drill down and roll up, those operations can be performed on millions of records at a short time. OLAP is also normally user friendly so business users without too much technique background can query a multidimensional database. There has been substantial research on multidimensional data warehousing, with the star schema and the OLAP cube taking center stage. The use of the star schema has implications at the logical and physical levels, and is designed for querying using SQL. Adopting OLAP model for the topic model has

been challenging because of the complex features inherent in topic and word distribution and demanding user needs and expectations. It will be very useful for a biomedical professional user to be able to slice and dice measurement through a pre-defined topic model relationship structure, this structure can be either Ontology or any predefined parent-child relationship. Imagine that a biomedical professional user is interested in analyzing the ADR distribution of FDA category B drugs in month 2 of a pregnancy, with the traditional data cube, this user can only query to return all the tweets in this condition, which is a collection of text documents, those text documents would have to be extracted and processed using other applications. Moreover, the basic OLAP cube drill down and roll up operation is based on a hierarchical design as shown in Figure 4-1: Sample multidimensional database where aggregation on top of the topic distribution won't be done easily. A more advanced data warehouse model, as compared to the star schema, is required to better respond to the challenges in the health language processing domain, and a unified model that adequately satisfies these needs is not currently available.

In this study, we adopt the multi-dimensional text data warehouse to manage the topic distribution from the topic modeling process and propose to model the Ontology structure using the OLAP hierarchy.

## 4.2 Research Background

Managing topic modeling output using Multi-Dimensional Text data warehouse (MDT) is now a hot research theme to address the above challenges and automate the topic model data discovery. An MDT is a special type of data warehouse, with mixture types of dimensions, those dimensions can be traditional type of structured dimensions, such as time, age and gender. In addition to structured dimensions, we also have text dimension or topic dimensions, those dimensions include dimensions which are related to text information. Text dimensions are generated from the topic modeling process, and those dimensions are useful to arrange detail topic information in a structured way so that the knowledge can be discovered. After running topic modeling on documents, a wide range of heterogeneous concepts and dimension are automatically discovered. For example, when we perform ADR analysis on an amazon product

review, we assign each ADR mention within the review a score to measure the risk of the product. In order to summarize this detail information from the topic model, this can be best managed within the MDT. In our case, those ADR scores need to be aggregated in different dimensions, such as on medication category, age, or gender. Because ADR score cannot be simply aggregated by normal sum operation, we need to design special aggregation algorithms to aggregate those scores in different levels.

With this kind of advanced multi-dimensional data warehouse cube, case control study can be done easily. For example, if a biomedical researcher is interested in finding all the amazon product reviews in the year 2016, for product category “Vitamins&DietarySupplements”, which talks about ADR for “abdominal pain”, a simple query e.g. (“Vitamins&DietarySupplements”, “Jan. 2016”, t=“abdominal pain”) can be utilized to obtain all the relevant narratives for this topic. In addition, many of the related topics that have been mentioned that use other words for the same concept can be identified by using a query like (“Vitamins&DietarySupplements”, “Jan. 2016”, t=“?”).

In biomedical research world, sophisticated analyses are always needed, which require more complex queries than previous example, this kinds of queries normally requires a topic hierarchy that roll up/drill down relationships can be performed to enable the aggregation of topic measures at different levels, in those cases, topics with hierarchically structure can be user friendly to determine group profiling at different levels [43] and the compare the semantic meaning of documents [44][45].

There are different ways to generate a topic hierarchy, it can be manually defined, such as medication hierarchy, or it can be automatically derived from the topic modeling process. Topic hierarchy automatic generation has been proposed in different approaches. Some research has shown that topic structures can be generated from an ontological point of view [46], in this approach, by feeding ontology information in to the topic model, the topic relationships can be generated. Some other studies have demonstrate the relationship between hierarchical topic models and ontologies, where the ontology can be automatically generated

from a topic modeling system [47]. In our research, we propose a method to automatically define an ontology as a possible hierarchy in MDT. Some attention has been given to formally define the ontology hierarchy. “MedDRA® is the Medical Dictionary for Regulatory Activities terminology and is the international medical terminology developed under the auspices of the International Conference on the Harmonisation of Technical Requirements for the Registration of Pharmaceuticals for Human Use (ICH)”[4]. In this study, we use MedDRA ontology information to model our text dimensions, and also use this ontology to guide the topic modeling process so that topic relationships can be generated. We also use MedDRA to evaluate the output topic structure.

In this study, an approach is proposed to incorporate the multi-dimensional text data warehouse with results from biased topic modeling, this text data warehouse model provides topic dimensions by extending the traditional data cube standard dimension. A topic hierarchy can be further defined on top of topic dimension, in addition, distribution from topic model can be stored in the cell level of text data warehouse.

This research extends the existing text cube by adapting background knowledge and a topic hierarchy (Ontology) in topic modeling and data warehouse aggregation process, those background information are effective in capturing the rich features of biomedical texts, and help to model the data in the data warehouse.

By conducting query evaluation, the goal is to show that the proposed topic text data warehouse can outperform the existing one in terms of performance and the effectiveness of analyzing and searching a large set of text. The proposed OLAP model allows users to study and analyze the characteristics of unstructured text data contained in social network user generated data. This approach can fully take advantage of all meaningful information. We present several use cases through which the analyst can use the tool to study ADR topic distributions alongside pregnancy twitter timelines.

## 4.3 Method

### 4.3.1 Overview

In this study, we propose a new approach to extend traditional OLAP to handle distribution from topic modeling and enable distribution from topic model to be stored in the cell level of text data warehouse. This approach provides a topic hierarchy structure which is defined on top of topic dimension. The main challenge for the text cube instead of traditional cube is the process of aggregation. In the traditional cube, the aggregation is pre-calculated at the lowest level, then simply aggregated at a higher level, however, this is not possible in text topic distribution, because the topic distribution at the lower level won't be summed up to get the distribution at the upper level.

We defined our cube as follows: A topic cube is a special OLAP cube which is built from a text multidimensional data warehouse. Each cell from this cube stores the measurement from topic modeling, which is the topic/word distribution. Then those measures are aggregated by predefined hierarchy. As with the traditional cube, we also define the following measurements and dimensions:

- Dimension Hierarchy: This is the topic hierarchy we can define manually, which is primarily for the end user to drill down and roll up. An example of the hierarchy is the drug category hierarchy.
- Measurements: Distribution score for contextual factors of the analyzed documents, this is the summarization of text documents in a cell, in the form of a numeric value.
- Dimensions: Topics, domains, metadata such as date, location, drug, etc.

### 4.3.2 Star Schema

An example of star schema for a topic cube is given in Figure 4-3. There are six different dimensions. Three kinds of measures are stored in a topic cube cell, word distribution, topic distribution, and ADR score.

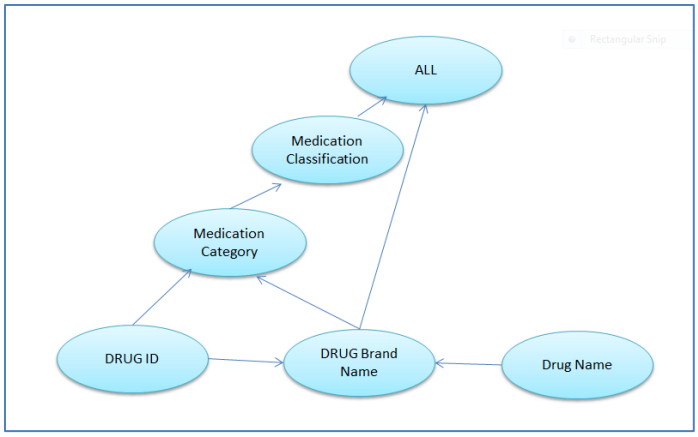



Figure 4-2: Medication hierarchy.

### 4.3.3 ADR Score Aggregation

To materialize topic cubes efficiently, we have three different levels of aggregation. The lowest level of aggregation is that of the database table level, corresponding to the detail topic level distribution. The purpose of which is to help the user to access detailed information for each topic if needed. There is no summarization in this layer, but it is important to provide this direct access when the user needs conduct data analysis on an individual topic or event transaction. Although there is no additional aggregation, database tables are de-normalized into a logical star schema view. As described in the previous section, we have all the three measurements saved in fact table. These measures from fact tables are aggregated during the run time at the OLAP layer. The ADR score and distribution are aggregated using the similar approach in Oukid et. [48]. Instead of using the term probability, we use ADR score to calculate the weight.

- Step 1: Compute the weights of the words which exist in the concept hierarchy by using the ADR score- computed by the following formula

$$ADR(n_k, n_{fi}) = ADR(n_k) + ADR(n_k, n_{fi})$$

- Step 2: The ADR score of its ancestors from the topic structure are computed by the formula below [48]

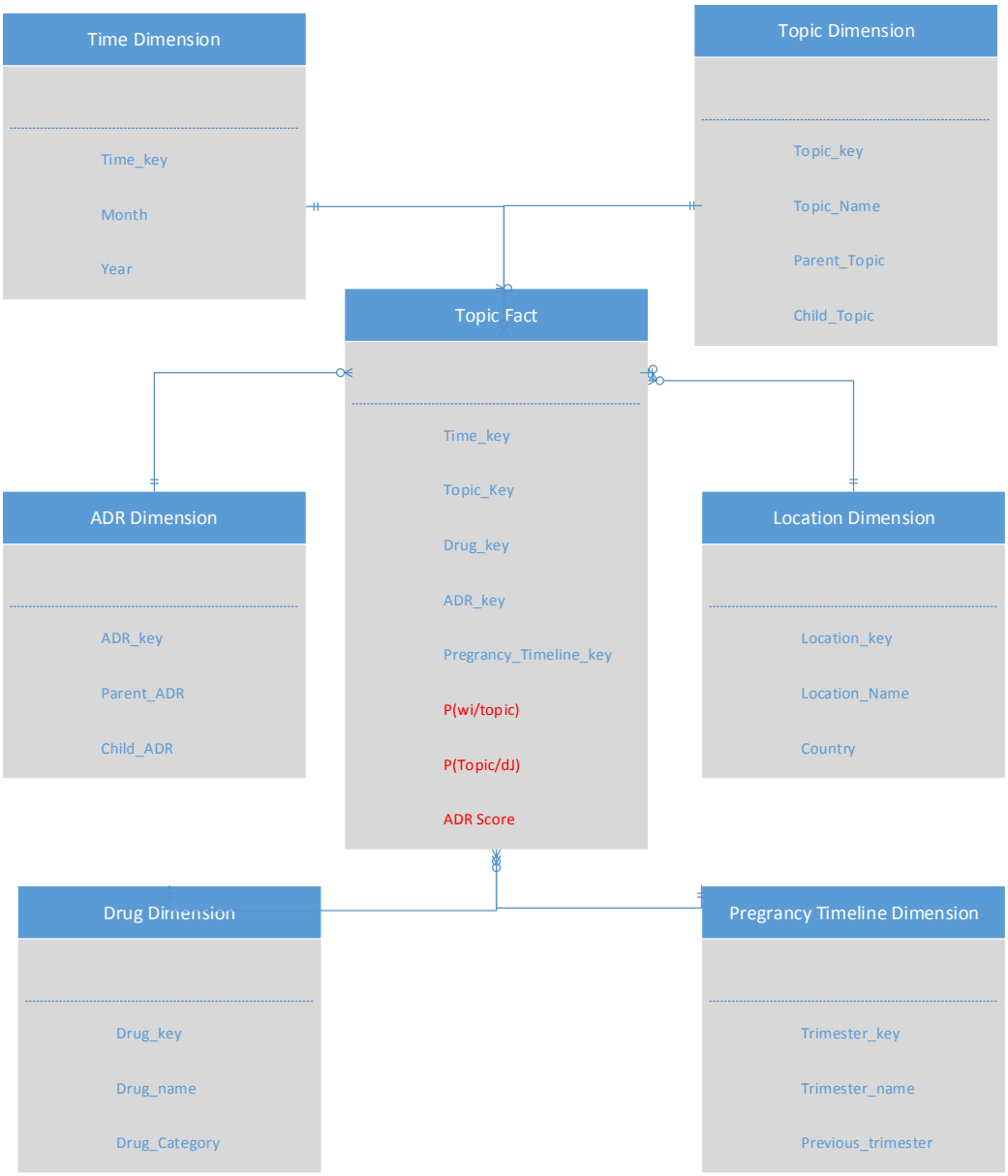



Figure 4-3: Star schema for text cube.

$$P(c|w_1, \ldots\ldots, w_n) \propto P(C) \prod_{i=1}^{n} P(w_i|c)$$

The OLAP performance benefits from the revised data schema design, with three levels of granularity, which satisfy different types of queries. During the cube construction, the system first determines whether the fact table contains relevant data, and does aggregations only for data that are not contained in any of the fact tables. The storage system is optimized for fast I/O, and the result can be returned as fast as possible.

A ROLAP solution is implemented to support this aggregation and cube construction. Aggregation of the OLAP cube is a combination of on-the-fly aggregation and off-line aggregation of some data for better performance and accuracy. As described in the case study, computation of a cube for a large amount of data and dimensions is extremely expensive. In this algorithm, two variables (cost, and benefit) will be dynamically calculated during the cube construction, and these two variables will determine which aggregation needs to be conducted off-line and which aggregation needs to be conducted on the fly. Cost is determined by the number of dimensions, measures, and hierarchy structures. The benefit compares the amount of data before and after the summarization, and the different level of hierarchies.

The following tasks are performed during the ROLAP construction process.

- Dimension Member Collection building: In this step, the program will iteratively go through each member of each dimension
- Hierarchy relationship building: In this step, dependencies for all the hierarchy members are analyzed, the relationships between dimensions are saved into a key-value mapping table, and this table is used later to determine the order of the execution plan.
- Lookup Bitmap Index building: In this step, a bitmap index is used to organize the dimension data, which will greatly improve the query time.

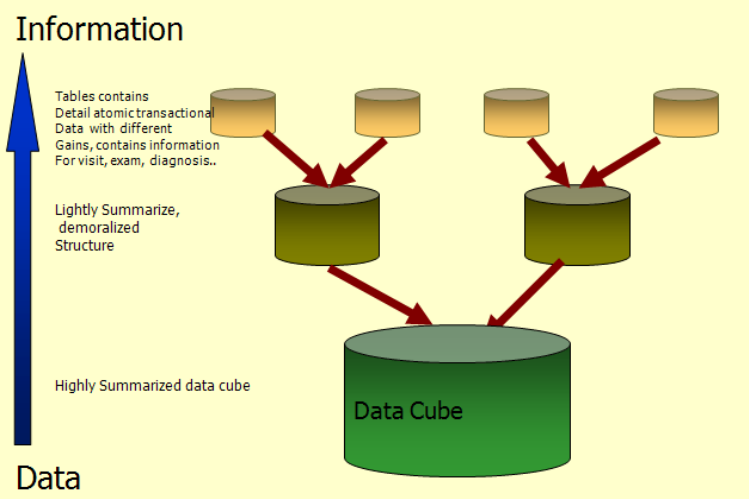



Figure 4-4: Text OLAP aggregation.

### 4.3.4 Ontology Integration

We summarize text corpora based on a given ontology hierarchy, and this framework will seek to generate a data cube for every level of an ontology. Using this pre-aggregated information, we can carry out sophisticated analyses – where a topic hierarchy with roll up/drill down relationships can be performed to enable the aggregation of topic measures at different levels. The following example illustrates how this can be accomplished:

An ADR hierarchy tree is constructed using existing public biomedical resources (Medra[4], GTR [5] or ADReCS[6]). There are normally four levels in the ADR hierarchy tree: the SOC, the HLGT, the HLT and the PT[6], while a more complicated Ontology structure could also be constructed.

[4] http://www.meddra.org/

[5] https://www.ncbi.nlm.nih.gov/gtr/

[6] http://bioinf.xmu.edu.cn/ADReCS/

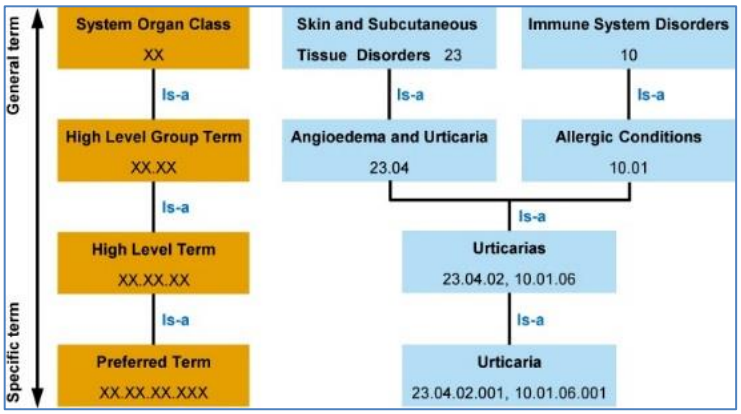



Figure 4-5: Example ADR hierarchy tree from Meddra.

The ADR hierarchy tree will be incorporated as the background knowledge to drive the topic modelling. This model is built on top of hPAM model[49], during the training process, every phase from the ADR ontology can be mapped to certain position within the DAG model structure.

Star schema is used to model the ADR, an aggregation algorithm is executed to efficiently summarize all of the different topic levels that are part of a given ontology. Specifically, we show that OLAP techniques can be used to efficiently obtain aggregations of classes, represented as dimensions (in OLAP terminology) in a set of documents.

As a result, the user will be able to obtain a data cube containing aggregations (COUNT, SUM, AVERAGE, MIN, MAX) on several measurements (e.g. term frequency, number of documents, ADR score) that are related to a set of classes in different hierarchies (Parent ADR, Child ADR, Correlated ADR).

## 4.4 Evaluation

The following example illustrates a case when the user can perform complicated analysis using the MDT.

**Example 1**. A biomedical analyst wants to analyze the ADR pattern and drug usage amongst pregnant women on Twitter. Table 4-1 shows a sample set of topics for the medications and the dimension structure that might be used by researcher. When analyzing the

ADR and drug usage patterns and their link with outcomes amongst pregnant women from Twitter data, the analyst may be interested to include drug categories A and B, with one drug for each category. For each of these criteria, the measurement is the outcome of the pregnancy, and a distribution of the ADR term.

| Topic | PositiveOcc | NegativeOcc | ADR |
|---|---|---|---|
| Category A | 84 | 59 | Sleepy, illness.. |
| Category B | 291 | 451 | Depression, nausea |
| novartis | 15 | 2 | headache |
| dicyclomine | 23 | 1 | bloating |

Table 4-1: Topics measurement.

There are several additional analysis the user can perform: (i) drug category risk, aimed at querying the pregnancy outcome based on drug usage for each drug category; and (ii) drug ADR, the ADR effect score associated with each drug. Depending on the user's requirement, the measures can be aggregated at different levels, either in drug category level, or drug name level. Those different levels are shown in Figure 4-6.

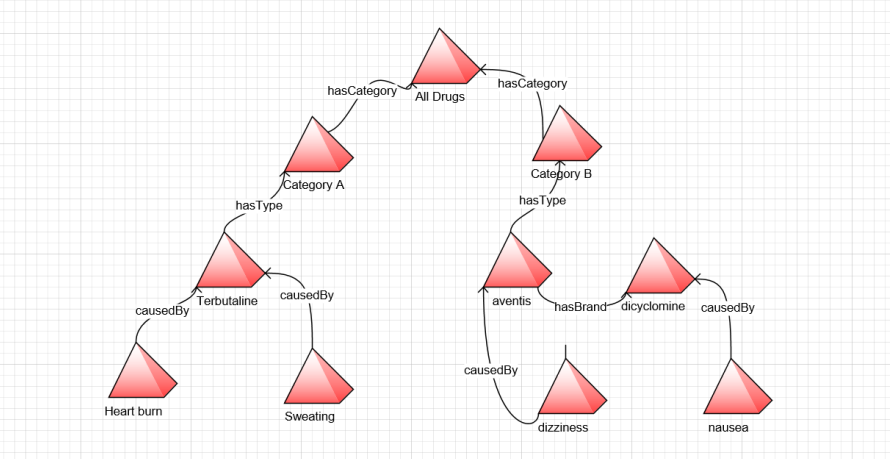



Figure 4-6: Topic hierarchy for the medication subject area; with inter-topic roll-up relationships.

It is also possible to perform some additional use cases. Figure 4-6 is an example of how to deep drive into the text narrative level to compare the topic/word distribution. The cell contains the word distribution for two month pregnancy or two week pregnancy, drug category X and A, and the word distribution for each dimension. By comparing different trimester period, the user can query the difference between ADR terms distribution.

At the level of “Two Month Pregnancy”, it provides an ADR term distribution of all the narratives from this topic, and then the user can drill-down to “DRUG Category A”, which allows the user to query summary of the text documents from sub category of “headaches”. This kind of navigation can help the user to study and compare the drug ADR mentions cross different trimester and drug categories, which could be very useful to study the medication usage and safety issue on pregnant women.

| Time Dimension | DRUG Category | ADR Word Distribution |
|---|---|---|
| Two month Pregnancy | X | headaches 0.035, chest pains 0.58, inflammation 0.254,<br><br>birth defects 0.21, |
| Two weeks pregnancy | A | Feeling sick 0.125, jittery 0.029, |

Table 4-2: Example slicing and dicing.

## 4.5 Summary

In this chapter, an effective information retrieval and management system for health related topic modeling are introduced. This information retrieval and management system is a variation of modern data warehouse/OLAP system which is built on top of star schema. This system includes four major components: topic generation, dimensional modeling and aggregation algorithm. We proposed a solution to model the topic ontology hierarchy into the databased, and semantic aggregation into OLAP.

With this framework, we did some pilot work to store and manage the complex ontology relationship among ADR mentions in a data warehouse system, with different level of aggregations, this kind of ontology relationship has been difficult to represent in a user friendly way for researcher to perform analytics works under current technology. We also evaluated the effectiveness of this framework using different data sets.

For the method we proposed, there have been several main approaches from related publications, our approach extends and improves those approaches,

1. Generating Ontology from topic modeling system- Some recently work has shown that relationships can be built between hierarchical topic modeling and Ontology, in [50], [51] and [52], Ontology was automatically generated from the topic modeling system, however, those approaches reply on using key phases to build relationship between the entities found in the text, and in our approach, we start with existing ADR with Ontology pre-defined.
2. Modeling topic modeling using MTD OLAP - Text data warehouse hierarchies for text mining has been proposed in some studies. In [53][54] a multidimensional model was proposed to integrate text data in a traditional EDW, which supports ontology hierarchies. However, this model is for structured data only and is not able to store topic information. Some other approaches integrate topic modeling with text data warehouse [55]. In this approach, traditional data warehouse and OLAP are extended to incorporate a topic hierarchy, the topic modeling outcome is stored and aggregated. Structure relationships between topics are also modeled in the OLAP cube. In [56], an approach is proposed to model topic hierarchies as DAGs of topics, and a model is defined to manage these topics.

Overall, there are some existing approaches to overcome those challenges within our research, each representing some ways to resolve part of the problems, but there has been no work which provides an end to end framework, which is from topic modeling to data warehouse management. To resolve those challenges. Moreover, none of the proposed solutions completely

resolved the challenges to integrate topic hierarchy with the multi-dimensional database ontology. Comparing to those approaches, our approach is a more complete end to end approach, and it is optimized for biomedical text mining.

# 5 LITERATURE REVIEW

## 5.1 ADR on Social Networks

Currently, it has been a thriving area to mine social networks for biomedical purposes in a computerized way. Normally, software programs are used to identify any possible event which may be the indication of an adverse event. One approach is to use narrative report. One of the simple ways is to use lexical based approach, which match each words or tokens in the text document against dictionary [57] [58], and then keyword search using a web search engine and MEDLINE. This approach is somewhat like backward chaining option mining, which has good performance when the target state is rare. Some improvements over this approach are also being studied, such as the lexical match to Unified Medical Language System (UMLS)[59]. This is also based on the keyword match, but the problem with this approach is that the positive value is relatively low.

Some advance NPL methods have achieved better performance. By using these methods, better information and mean representation can be extracted from the document or medical report, which can be as accurate as works from domain expert, and those methods are much more accurate than the lexical method. A number of natural language processing systems have been developed on top of those advanced NPL methods, such as the rule based technique or pattern matching system [60].

Pilot research on detection ADR mentions from social media forum was done by Leaman et al. [10], in this research, forum posts from DailyStrength were crawled and lexicon approach was adopted to extract ADR from posts. This lexicon based approach has been widely used in some other research to identify and extract ADRs from other social media web sites [61] and [62]. Beside lexicon approach, in recent year, some advanced text mining and machine learning techniques have been applied in this area, those approaches include association rule mining and supervised/unsupervised classification. In [63] and [37], association rule mining is applied to associate ADR with other information to extract ADR-related tasks from user-generated health text. In [61] and [63], supervised text classification approaches, especially

Support Vector Machines (SVMs) are applied to classify ADR related social media posts from non ADR posts.

### 5.2 Topic Modeling with Background Knowledge

While topic modeling is a well-studied area in text mining, there have been a lot of approach to incorporate background information to help drive topic models. The basic intuition is to provide domain knowledge as priors to improve the topic quality. Those approaches can be grouped into “seed words” or “Non seed words” approaches. One popular approach is “Seed words” based approach, in which “seed words” are used to guide the topic model. A variation of LDA is proposed in [36], where background information is defined as the topic location of specific words in a corpus, and this background information can be used to guide the topic model. in [36], another LDA variation with background knowledge is also proposed, this LDA can incorporate knowledge represented as First-Order-Logic, which is a set of rules for topic generation. In [5], authors proposed a system to select seed words which are related to topics, then use those seed words to help generating more accurate topics. Some other approaches to incorporate background knowledge have been proposed, those approaches are not “seed word” based. Some approaches have made use of a set of related words to create topics. An approach is proposed in [64] to use tuple of related words to guide the topic model, and the result shows that this approach can help to improve the performance of infrequent words, in [65], authors proposed an approach to expand the tuple related words approach by using a group of words.

### 5.3 Topic Modeling In Text Data Warehouse

It has been a thriving research topic to integrate unstructured data in a multidimensional text data warehouse or OLAP cube in recent years. Some approaches have been proposed to enhance OLAP analyses then use it for unstructured (text) data management [45], [53], [56], [66], [67]. These approaches can be classified into two main areas: OLAP aggregation for on unstructured data, and data warehouse modeling (star schema) to integrate an ontology with a multidimensional hierarchy.

In [30], the authors proposed a framework to combine the keyword search and OLAP technique, this method of querying a multidimensional text database can be done on top of the OLAP cube. During this process, a dynamic dimension approach is proposed, the dynamic dimension is constructed on the fly by extracting frequent and related words from the document, and the keyword based query is also materialized during the run time. In [22], the author conducted pilot research on combining text information with OLAP cube. In this model, text data is aggregated by information retrieval measurements, which makes the summarization possible. On top of text cube approach, In [68], an approach to incorporate topic and cube is proposed, this model is called Topic Cube, this is a very similar approach as our approach. The topic cube extended traditional OLAP cube and text OLAP cube by incorporating topic hierarchy, this approach provides similar analytical power for topic cube as traditional cube. In the topic cube, topic and word distribution are saved in the cell level of the cube, and topic dimensions are defined so that users can analyze unstructured text data from different topics, and compare these with other topics in different level. However, the topic model method in this approach is pLSA model, which cannot be applied in our project because our distribution comes from a variance of LDA with background knowledge. Other similar approaches have been proposed with different aggregation methods. In [39], a conceptual topic cube which supports online OLAP cube type of query of unstructured log data was proposed. This approach also provide nontraditional dimensions other than time and location, those nonstandard dimensions of topics and concepts can facilitate the analysis of unstructured data. In addition, distributed algorithms for learning model parameters was developed. In [48], another contextual text cube model called CXT-Cube which considers several contextual factors during the OLAP analysis was proposed. CXT-Cube comes with several different types of dimensions, each one related to a contextual factor. A study sharing some of the similarities with our research is described in [69]. In this approach, a framework is developed to extract information from Twitter, the extraction result is loaded into a data warehouse. However, those approaches are focused on extracting structured information, such as the relationships between users, instead of the semantics meaning in our

research, since those relationships are structure data, there is no challenge to model them in data warehouse. Of all those approaches, little attention has been paid to explore and extract tweet topics.

### 5.4 Comparison

In this section, we compare our approach with other similar popular approaches, we will first explain the similarity between our approach and other four main topic modelling approaches which incorporates background knowledge, and then we will discuss the major difference between my approaches and the other approaches from 4 different perspectives.

Classic topic modelling as an unsupervised text mining technique, has been shown to be a flexible, fast, and cost-effective method to perform text mining in the biomedical domain. Because LDA is a well-defined probabilistic model and can be customized, using classic LDA as based model, wide variety of customizations and extensions to the base model have been developed [36] and thereby overcome some of the limitations of topic modelling, which include:

- Uncorrelated topics – Because the Dirichlet topic distribution cannot capture correlations, it is very common to see an overwhelming number of uncorrelated topics from the output topics.
- Bag of words – One assumption that LDA makes is the “bag of words” assumption, in this assumption, there is no order of the words in the document. While this assumption is unrealistic, this will prevent us from discovering some more complicated semantic structure and relationships of the texts, such as, the relationship and hierarchy between topics.
- Unsupervised - sometimes weak supervision is desirable, e.g. in sentiment analysis. One of the drawbacks of an unsupervised learning process such as topic modelling is the interpretation of the topics. Although topics come with hidden information on how those words are grouped together, interpretation of those topics manually is a subjective process and it has a dependency on the background and knowledge of the person who labels those topics.

Because of the limitation of classic LDA, incorporating knowledge with topic models has been a thriving research area, and different approaches have been proposed. Our approach and other approaches share the similarity that they all aim to overcome those three limitations of un-supervised topic modelling by incorporating additional priori knowledge outside the text documents. Those additional knowledge includes document labelling, metadata information of document, relationships between document, and links between documents. All those approaches try to improve the topic modelling process by providing additional guideline on topic modelling process by domain knowledge. The frameworks which were presented in [70] and [71] are the most similar approaches to our approach. Both their frameworks and our framework can be considered as a variation of traditional LDA model, in that their frameworks, a set of rules using must-link and cannot-link words is incorporated into topic models, in the form of seed words, which is very similar to our approach. In [72], they expanded on this work and provided a system to improve the topic model performance interactively by providing feedback. In [73], a semi-supervised topic modelling approach was proposed to integrate semantic information hidden in text articles with reviews from domain expert, topic model is used in this research for clustering. All of those approaches share the same assumption of the multinomial word distribution (bag of words) to model a topic.

Difference between our approach and the other four approaches are compared. These are described as follows

- **LDA Model Extension**: Different LDA model extensions can be loosely categorized into three categories [36]:
    1. Guiding topic modelling process by providing additional document information, such as labels, relationships and links. (LDA+X),
    2. Word distribution variation (φ-side).
    3. Topic distribution variation (θ-side).

Our approach takes advantage of the structured biomedical knowledgebase. Comparing to other approaches, the differences are:

- LDA vs pLSA: Our approach and the approaches taken in [36], [71], [72] are a variance of the LDA model. The approach in [76] is based on the pLSA model, LDA, and PLSA and are both widely used topic modelling methods. The pLSA treats topics as word distributions, uses probabilistic methods and topics are allowed to be non-orthogonal. LDA is similar to pLSA, the difference is that LDA uses Dirichlet priors to draw the document-topic and topic-word distributions. This prevents over-fitting and gives better results. LDA has been proven to work pretty well for short text utterances such as Twitter topic modelling and subsequent classification [77][78], and since Twitter data is the main input data in our research, LDA is preferred in this research.
- Unsupervised vs. Semi-Supervised: As previously described, all of these approaches try to incorporate background information to help drive the topic models. However the foundations of background knowledge are quite different, the approach in [70] is based on using information about the words themselves, which does not take into consideration the structure of the background knowledge. The approach in [72] and [71] provides a system to allow user feedback to drive the must-link and cannot-link tuples. During the modelling process, user input or labelling is required, which is more user intensive than our approach, because our approach is the seed word approach and is fully unsupervised. Our approach has the advantage of utilizing the existing domain knowledge without human annotation or user labelling.
- Uncovering more sophisticated structure/hierarchy in the texts: Another fundamental difference between our approach and other approaches is the ability to produce hierarchical topics. In classic LDA, there are no relationships between of topics. In the real-world document, it is common for document contain hierarchical relationship, e.g. healthy related topics would have sub-categories such as sports, disease and drug. None of the other four approaches can produce topic hierarchy/Ontology, although there has been some work to produce topic ontology [46], this work does not incorporate ontology as background knowledge.

- Domain: Our approach is aimed at topic modelling within the biomedical area, where public reference data has been widely available and well defined in the form of databases, dictionaries and ontologies. Our approach is designed to use these knowledge sources to augment unannotated text. The other four approaches are more generic and not optimized for the biomedical domain.
- Performance/Parallelization: In our approach, we propose a hierarchical distributed architecture including model parallelism on top of Spark to handle a large number of LDA parameters as well as data parallelism to handle massive training corpora. None of the other four approaches has this kind of optimization for big data.
- OLAP/Ontology integration: In our approach, we propose a new approach to combine a traditional multidimensional text database with a probabilistic topic model with background knowledge and incorporate the Ontology with the multidimensional database Hierarchy. This has never been adopted by any of the other four studies.

| Refer-ence | LDA/ PLSA | Back-ground Data | Label -ing | User input | Domain | Ontology | OLAP | Paralleli-zation | Bag of words |
|---|---|---|---|---|---|---|---|---|---|
| [70] | LDA | Seed word | Y | N | general | N | N | N | Y |
| [71] | LDA | Logic tuples | Y | N | general | N | N | N | Y |
| [72] | LDA | Logic tuples & user input | Y | Y | general | N | N | N | Y |
| [73] | PLSA | User input | Y | Y | general | N | N | N | Y |
| Our approach | LDA | Seed word & Ontology | N | N | biomedi-cal | Y | Y | Y | Y |

Table 5-1 Related literature classification and comparison.

Table 5-1 summarizes a comparison between other four main approaches and ours in terms of supporting features. Overall, compared to the other approaches, our approach is an

approach with the advantages of better performance, uses existing domain knowledge without annotation or user input, ontology awareness, and OLAP integration. This will be a fully automated unsupervised learning process without the cost of human annotation.

# 6 CONCLUSION AND FUTURE WORK

This chapter concludes the dissertation by summarizing the contributions of the work and discusses some potential future directions.

## 6.1 Conclusion

In this work, an advanced topic modeling framework has been designed to analyze complex longitudinal health information from social media with minimal human annotation. Our proposed biased topic modeling is an approach that incorporates background knowledge in different format and relationships. This biased topic modeling method can help researcher discover knowledge in a timely fashion on top of a large scale data. In this study, we also present a novel approach to classify drugs based on biased topic modeling. This timeline based approach for topic modeling to extract drug usage patter can aid in performing aggregating of latest evidence and preserving of semantics knowledge. This dissertation also explores how to manage information effectively for later retrieval. Nonetheless, some challenges have been identified and then addressed in this dissertation.

By incorporating background knowledge in topic model, we can effectively summarize the semantic information text, which is proven to be useful for various BioNLP tasks. We also propose a scoring system to score the topic modeling output. During the modeling process, all information unit, such as diseases, symptoms, treatments, is connected by the guided modeling. A unified framework is proposed to learn the model from existing timeline, and run the prediction for the unknown timeline. The effectiveness of this prediction framework has been verified with different variance, and the performance has been proven to outperform existing methods. The experimental results show that the background knowledge can significantly improve the prediction accuracy.

At last, the proposed framework is used to build a topic based data warehouse for effectively discover patterns and ontology for ADR from UGC data. The experimental evaluation shows that information retrieval can be done effectively using this warehouse tool.

As the amount of data from social media and publicly available medical libraries increase exponentially, our method provides a low cost approach to discover new ADRs and also provides early warning for drug safety. Using this framework on a dataset of twitter timelines of women that announced their pregnancy, we sought to distinguish what kind of medications they were taking during pregnancy, this framework can be used to obtain a “computable” model that “summarized” their discourse in social media, which makes it possible to perform a control case study on FDA pregnancy category. The intuition for this study is that not all topics are equally important, women taking the more dangerous medications would probably have mentions of adverse effects in greater volume or mention more serious effects. The final results mostly support this intuition.

### 6.2 Future Works

In this section, we discuss future works in several categories.

**Medical Name Variation Library:** Future work can include be a disease and drug name variance library. Our result can be significantly improved from a standard disease, drug variance library, this library can greatly improve the effeteness of the mining. We observed that the in-thread or experienced co-occurrences are more important for ADR discovery. We can thus assign the variance of the medication terms with a different weight or rank for topic modeling.

**Deep Learning and Knowledge Transfer for Text Summarization:** Deep learning has recently shown much promise for NLP and text mining applications. With deep learning, instead of assigned weight to tokens and topics, we can automatically derive better feature representations, which can lead more robust result. Knowledge transfer from one learned training model to another model would be very useful, which means that the knowledge we gained from different reference libraries can be transferred to other domains.

**Context Information in Topic Modeling:** Some future work can be done to optimize topic modeling training. Quality of data is one of the major factors which affect the performance: not all terms from ADR vocabularies are really ADR, sometime they are recommendations, common feelings, symptoms, or somebody else’s experience. We have tried to resolve this

quality issue in the topic modeling building phase by introducing a bag of words of symptoms so the ADR topics can be separated from the symptom topic. Another solution can be introduced is the context information so that we can better differentiate personal experience from recommendation and other person's experience.

**Topic Modeling Enhancement and Visualization:** Some third party libraries such as ADRMine and DLATK Can be used as framework to integrate existing module and visualization, ADRMine can provide words clustering based pre-trained vectors, which can be used to improve the biased topic modeling performance. Visualization can help to discover the topics and the underlining timeline structures, we can use this kind of visualization to discover the corpus structure from different drug categories.

**Integrating Multiple Sources:** In addition to Twitter, which is a general social network, we would like to extend the proposed work to other social media, especially those focused on health related topics, such as DailyStrength. Those social network data contains more health related information. Those data repositories do not contains data size as large as twitter, but most of the information is related to health. It will be interesting to combine DailyStrength data with twitter data, this way we can integrate information from various repositories and consolidate the opinions.

## REFERENCES


[1] B. Chee, R. Berlin, and B. Schatz, “Predicting adverse drug events from personal health messages,” *AMIA Annu Symp Proc*, pp. 217–226, 2011.

[2] J. Bian, U. Topaloglu, and F. Yu, “Towards large-scale Twitter mining for drug-related adverse events,” *Proc. 2012 Int. Work. Smart Heal. Wellbeing*, pp. 25–32, 2012.

[3] R. Ginn *et al.*, “Mining Twitter for Adverse Drug Reaction Mentions : A Corpus and Classification Benchmark,” *Proc. Fourth Work. Build. Eval. Resour. Heal. Biomed. Text Process.*, no. 1, 2014.

[4] R. Sullivan, A. Sarker, K. O’Connor, A. Goodin, M. Karlsrud, and G. Gonzalez, “Finding Potentially Unsafe Nutritional Supplements from User Reviews with Topic Modeling,” *Pac. Symp. Biocomput.*, vol. 21, pp. 528–39, 2016.

[5] J. Jagarlamudi and H. Daum, “Incorporating Lexical Priors into Topic Models,” In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 204–213, 2009.

[6] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent Dirichlet Allocation,” *J. Mach. Learn. Res.*, vol. 3, no. 4–5, pp. 993–1022, 2012.

[7] L. Wei and A. McCallum, “Pachinko allocation: DAG-structured mixture models of topic correlations,” *ICML ’06 Proc. 23rd Int. Conf. Mach. Learn.*, no. April, pp. 577–584, 2006.

[8] D. Mimno, L. Wei, and A. McCallum, “Mixtures of hierarchical topics with Pachinko allocation,” *ICML ’07 Proc. 24th Int. Conf. Mach. Learn.*, pp. 633–640, 2007.

[9] D. Blei, L. Carin, and D. Dunson, “Probabilistic topic models,” *IEEE Signal Process. Mag.*, vol. 27, no. 6, pp. 55–65, 2010.

[10] R. Leaman, L. Wojtulewicz, R. Sullivan, A. Skariah, J. Yang, and G. Gonzalez, “Towards Internet-Age Pharmacovigilance : Extracting Adverse Drug Reactions from User Posts to Health-Related Social Networks,” *Comput. Linguist.*, no. July, pp. 117–125, 2010.

[11] R. Leaman, J. Yang, R. Sullivan, L. Wojtulewicz, and G. Gonzalez, “Mining Social Network Postings and Biomedical Literature for Early Detection of Adverse Drug Events " *PSB conference session : Personal Genomics* The submitted paper, 2009.

[12] A. Nikfarjam, A. Sarker, K. O ’connor, R. Ginn, and G. Gonzalez, “Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features,” *Journal of the American Medical Informatics Association* 22, no. 3, 2014

[13] S. Liu, W. Ma, R. Moore, V. Ganesan, and S. Nelson, “RxNorm: Prescription for Electronic Drug Information Exchange.” *IT professional*, no. 5, 17-23, 2005

[14] “business week.” [Online]. Available: http://www.businesswire.com/news/home/20121120005872/en/Twenty-percent-online-adults-discuss-health-information

[15] A. Sarker *et al.*, “Utilizing social media data for pharmacovigilance: A review,” *J. Biomed.*

*Inform.*, vol. 54, pp. 202–212, 2015.

[16] X. Chu *et al.*, “KATARA: A Data Cleaning System Powered by Knowledge Bases and Crowdsourcing.” In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*, pp. 1247-1261. ACM, 2015.

[17] F. Geerts, G. Mecca, P. Papotti, and D. Santoro, “The LLUNATIC Data-Cleaning Framework.” In *Proceedings of the VLDB Endowment 6,* no. 9, 2013.

[18] “SIDER.” [Online]. Available: http://sideeffects.embl.de/.

[19] Y. He *et al.*, “OAE: The Ontology of Adverse Events.,” *J. Biomed. Semantics*, vol. 5, p. 29, 2014.

[20] H. M. Wallach, I. Murray, R. Salakhutdinov, and D. Mimno, “Evaluation Methods for Topic Models,” *Proc. 26th Annu. Int. Conf. Mach. Learn.*, no. 4, pp. 1105–1112, 2009.

[21] T. L. Griffiths and M. Steyvers, “Finding scientific topics.,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 101 Suppl, pp. 5228–35, 2004.

[22] J. Chang, J. Boyd-Graber, C. Wang, S. Gerrish, and D. M. Blei, “Reading Tea Leaves: How Humans Interpret Topic Models,” A*dvances in neural information processing systems*, pp. 288-296, 2009.

[23] H. Chan and L. Akoglu, “External Evaluation of Topic Models: A Graph Mining Approach,” *Data Min. (ICDM), 2013 IEEE 13th Int. Conf.*, no. c, pp. 973–978, 2013.

[24] W. Buntine and A. Jakulin, “Applying Discrete PCA in Data Analysis,” *Proc. Twent. Conf. Uncertain. Artif. Intell.*, pp. 59–66, 2004.

[25] M. Rosen-Zvi, T. Griffiths, M. Steyvers, and P. Smyth, “The author-topic model for authors and documents,” *Proc. 20th Conf. Uncertain. Artif. Intell.*, pp. 487–494, 2004.

[26] W. Buntine, “Estimating likelihoods for topic models,” *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 5828 LNAI, pp. 51–64, 2009.

[27] I. Murray and R. Salakhutdinov, “Evaluating probabilities under high-dimensional latent variable models,” *Adv. Neural Inf. Process. Syst.*, pp. 1137–1144, 2009.

[28] J. Foulds and P. Smyth, “Robust Evaluation of Topic Models,” *NIPS Work. Top. Model.*, pp. 1–4, 2013.

[29] I. Hulpu, C. Hayes, and D. Greene, “Unsupervised Graph-based Topic Labelling using DBpedia.” *sixth ACM international conference on Web search and data mining,* pp. 465-474. ACM, 2013.

[30] J. H. Lau, K. Grieser, D. Newman, and T. Baldwin, “Automatic Labelling of Topic Models,” *Proc. 49th Annu. Meet. Assoc. Comput. Linguist.*, pp. 1536–1545, 2011.

[31] Q. Mei, X. Shen, and C. Zhai, “Automatic Labeling of Multinomial Topic Models,” *KDD ’07 Proc. 13th ACM SIGKDD Int. Conf. Knowl. Discov. data Min.* , pp. 490–499, 2007.

[32] A. Karami, A. Gangopadhyay, and B. Zhou, “FLATM : A Fuzzy Logic Approach Topic Model for Medical Documents,” *In Fuzzy Information Processing Society (NAFIPS) held jointly with 2015 5th World Conference on Soft Computing (WConSC), 2015 Annual Conference of the North American*, pp. 1–6, 2015.

[33] C. Shivade and P. Raghavan, “A review of approaches to identifying patient phenotype cohorts using electronic health records,” *Journal of the American Medical Informatics Association.*, no. 2,pp. 221-230. 2013.

[34] C. C. Musat *et al.*, “Improving Topic Evaluation Using Conceptual Knowledge,” *Ijcai*, pp. 1866–1871, 2011.

[35] P. Xie, D. Yang, and E. P. Xing, “Incorporating Word Correlation Knowledge into Topic Modeling,” *Naacl-Hlt*, pp. 725–734, 2015.

[36] D. M. Andrzejewski, “Incorporating Domain Knowledge in Latent Topic Models,” *phd thesis*. University of Wisconsin--Madison, 2010.

[37] M. Yang, M. Kiang, and W. Shang, “Filtering big data from social media - Building an early warning system for adverse drug reactions,” *J. Biomed. Inform.*, vol. 54, pp. 230–240, 2015.

[38] R. Harpaz *et al.*, “Combing signals from spontaneous reports and electronic health records for detection of adverse drug reactions,” *J Am Med Inf. Assoc*, vol. 20, pp. 413–419, 2013.

[39] C. Cortes, “Support-Vector Networks,” *Mach. Learn.*, vol. 20, pp. 273–297, 1995.

[40] “FDA Catergory.” [Online]. Available: https://www.drugs.com/pregnancy-categories.html.

[41] D. M. Blei and J. D. Mcauliffe, “Supervised Topic Models.” *Advances in neural information processing systems,* pp. 121-128. 2008.

[42] D. Inkpen and A. H. Razavi, “Topic Classification using Latent Dirichlet Allocation at Multiple Levels,” *Int. J. Comput. Linguistics Appl,* vol. 5, no. 1, pp. 43–55, 2014.

[43] “Review Author ( s ): Florian Krobb Review by : Florian Krobb Source : The Modern Language Review , Vol . 105 , No . 2 ( April 2010 ), pp . 591-593 Published by : Modern Humanities Research Association Stable URL : http://www.jstor.org/stable/25698773 Acce,” vol. 105, no. 2, pp. 591–593, 2016.

[44] A. Gelbukh, “Document comparison with a weighted topic hierarchy,” *Proc. 10th Int. Work. Database Expert Syst. Appl.*, no. Cic, pp. 566–570, 1999.

[45] E. Gallinucci, M. Golfarelli, and S. Rizzi, “Advanced topic modeling for social business intelligence,” *Inf. Syst.*, vol. 53, pp. 87–106, 2015.

[46] A. Aras, S. Paratap, and M. Bedekar, “Ontological Tree Generation for Enhanced Information Retrieval,” *Int. J. Artif. Intell. Appl.*, vol. 4, no. 4, pp. 41–51, 2013.

[47] G. L. Birbeck, “Revising and refining the epilepsy classification system: Priorities from a developing world perspective,” *Epilepsia*, vol. 53, no. SUPPL. 2, pp. 18–21, 2012.

[48] L. Oukid, B. C. Lyon, B. C. Lyon, O. Boussaid, and B. C. Lyon, “CXT-Cube : Contextual Text Cube Model and Aggregation Operator for Text OLAP,” I*n Proceedings of the sixteenth international workshop on Data warehousing and OLAP*, vol. 27, pp. 27–32.2013

[49] J. Han, “Title: Improving the utility of topic models: an uncut gem does not sparkle,” *PhD Thesis*, The Department of Computing and Information System, The University of, Melbourne, Melbourne, Australia, 2013.

[50] D. Movshovitz-attias and W. W. Cohen, “KB-LDA : Jointly Learning a Knowledge Base of Hierarchy , Relations , and Facts,” *Acl*, pp. 1449–1459, 2015.

[51] M.-A. Rizoiu, J. Velcin, W. Wong, and W. Liu, “Topic Extraction for Ontology Learning,” pp. 38–61, 2011.

[52] Zhu, Xiaofeng, Diego Klabjan, and Patrick Bless. "Unsupervised Terminological Ontology Learning based on Hierarchical Topic Modeling." *arXiv preprint arXiv:1708.09025,* 2017.

[53] E. Malinowski and E. Zimnyi, “Hierarchies in a multidimensional model: From conceptual modeling to logical representation,” *Data Knowl. Eng.*, vol. 59, no. 2, pp. 348–377, 2006.

[54] L. García-Moya, S. Kudama, M. J. Aramburu, and R. Berlanga, “Storing and analysing voice of the market data in the corporate data warehouse,” *Inf. Syst. Front.*, vol. 15, no. 3, pp. 331–349, 2013.

[55] C. Zhang, X. Wang, and Z. Peng, “Extracting Dimensions for OLAP on Multidimensional,” *Web Information Systems and Mining*, pp. 272–281, 2011.

[56] U. Dayal, C. Gupta, M. Castellanos, S. Wang, and M. Garcia-Solaco, “Of cubes, DAGs and hierarchical correlations: A novel conceptual model for analyzing social media data,” *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 7532 LNCS, pp. 30–49, 2012.

[57] D. a Giuse and a Mickish, “Increasing the availability of the computerized patient record.,” *Proc. AMIA Annu. Fall Symp.*, pp. 633–7, 1996.

[58] D. M. Rind, J. Yeh, and C. Safran, “Using an electronic medical record to perform clinical research on mitral valve prolapse and panic/anxiety disorder,” *Proc. Annu. Symp. Comput. Appl. Med. Care*, p. 961, 1995.

[59] M. Conway, N. Collier, and S. Doan, “Using Hedges to Enhance a Disease Outbreak Report Text Mining System,” *Proc. Work. BioNLP*, no. June, pp. 142–143, 2009.

[60] M. Fiszman, W. W. Chapman, D. Aronsky, R. S. Evans, and P. J. Haug, “Automatic detection of acute bacterial pneumonia from chest X-ray reports,” *J. Am. Med. Inform. Assoc.*, vol. 7, no. 6, pp. 593–604, 2000.

[61] A. Benton *et al.*, “Identifying potential adverse effects using the web: A new approach to medical hypothesis generation,” *J. Biomed. Inform.*, vol. 44, no. 6, pp. 989–996, 2011.

[62] A. Yates and N. Goharian, “ADRTrace: Detecting expected and unexpected adverse drug reactions from user reviews on social media sites,” *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 7814 LNCS, pp. 816–819, 2013.

[63] A. Nikfarjam and G. H. Gonzalez, “Pattern Mining for Extraction of mentions of Adverse Drug Reactions from User Comments,” *AMIA Annu. Symp. Proc.*, vol. 2011, no. January, pp. 1019–1026, 2011.

[64] J. Petterson, A. Smola, and T. Caetano, “Word features for latent Dirichlet allocation,” *Adv. Neural Inf. Process. Syst.*, pp. 1–9, 2011.

[65] Z. Chen, A. Mukherjee, B. Liu, M. Hsu, M. Castellanos, and R. Ghosh, “Leveraging multi-domain prior knowledge in topic models,” *IJCAI Int. Jt. Conf. Artif. Intell.*, pp. 2071–2077, 2013.

[66] A. Simitsis, A. Baid, Y. Sismanis, and B. Reinwald, “Multidimensional content eXploration,” *Proc. VLDB Endow.*, vol. 1, no. 1, pp. 660–671, 2008.

[67] S. Lee, N. Kim, and J. Kim, “A multi-dimensional analysis and data cube for unstructured text and social media,” *Proc. - 4th IEEE Int. Conf. Big Data Cloud Comput. BDCloud 2014 with 7th IEEE Int. Conf. Soc. Comput. Networking, Soc. 2014 4th Int. Conf. Sustain. Comput. C*, pp. 761–764, 2015.

[68] D. Zhang, C. Zhai, and J. Han, “Topic Cube : Topic Modeling for OLAP on Multidimensional Text Databases,” *Imagine*, 2009.

[69] N. U. Rehman, S. Mansmann, A. Weiler, and M. H. Scholl, “Building a data warehouse for twitter stream exploration,” *Proc. 2012 IEEE/ACM Int. Conf. Adv. Soc. Networks Anal. Mining, ASONAM 2012*, vol. 1, no. August, pp. 1341–1348, 2012.

[70] D. Andrzejewski and X. Zhu, “Latent Dirichlet Allocation with Topic-in-Set Knowledge,” *NAACL 2009 Work. Semi-supervised Learn. NLP*, no. June, pp. 43–48, 2009.

[71] D. Andrzejewski, X. Zhu, M. Craven, and B. Recht, “A framework for incorporating general domain knowledge into latent dirichlet allocation using first-order logic,” *IJCAI Int. Jt. Conf. Artif. Intell.*, no. Ijcai, pp. 1171–1177, 2011.

[72] Y. Hu, J. Boyd-Graber, B. Satinoff, and A. Smith, “Interactive topic modeling,” *Mach. Learn.*, vol. 95, no. 3, pp. 423–469, 2014.

[73] Y. Lu and C. Zhai, “Opinion integration through semi-supervised topic modeling,” *Proceeding 17th Int. Conf. World Wide Web - WWW ’08*, pp. 121–130, 2008.

[74] D. Alvarez-Melis and M. Saveski, “Topic Modeling in Twitter: Aggregating Tweets by Conversations,” *$Icwsm16*, no. Icwsm, pp. 519–522, 2016.

[75] L. Hong and B. Davison, “Empirical study of topic modeling in twitter,” *Proc. First Work. Soc. …*, pp. 80–88, 2010.