User-driven topic modeling of patient opinion data

A minor thesis submitted in partial fulfilment of the requirements for the degree of
Masters of Computer Science

Bin Lu
School of Computer Science and Information Technology
Science, Engineering, and Technology Portfolio,
Royal Melbourne Institute of Technology
Melbourne, Victoria, Australia

November 4, 2014

Declaration

This thesis contains work that has not been submitted previously, in whole or in part, for any other academic award and is solely my original research, except where acknowledged.

This work has been carried out since March 2014, under the supervision of Dr Jenny Zhang, Dr Amanda Kimpton, Dr Daryl D'Souza.

Bin Lu
School of Computer Science and Information Technology
Royal Melbourne Institute of Technology
November 4, 2014

Acknowledgements

I would like to take this opportunity to convey my special thanks to my supervisors Dr Jenny Zhang, Dr Amanda Kimpton and Dr Daryl D'Souza for their incredible support during the past year. I was so lucky to have three best supervisors from school which gives me opportunity to extend my knowledge in different directions, and my work progress is constantly reviewed from 3 different perspectives.

Contents

List of Figures

List of Tables

Abstract

Topic models are increasingly being deployed as a powerful technique to discover useful structure or topics in otherwise unstructured, large text collections. Such collections are widely available in the form of web-based forums for consumers to express feedback (opinions, reviews, general commentary) about products and services, thereby allowing providers to improve products and services. Such improvements are dependent on effective decision-making from the data available in these forums. In turn, such vast quantities of (typically content-specific) data call for techniques to accurately classify and summarise forum posts. Topic models have emerged as a useful way of automatically finding useful structure in such large, collections of free-form text data, to aid providers and domain experts in arriving at effective decisions (about products and services). While many studies have addressed this problem, none, to our knowledge, has applied topic modeling to healthcare. In this thesis we develop a topic model for a web-based forum for patient feedback (patientopinion.org.au) about general healthcare issues. Our model provides improvements over the baseline model generated by a popular topic modeling tool (Mallett), via user tag data available in the posts within patientopinion.org.au.

Chapter 1

Introduction

Publicly available web-based forums for consumer feedback about products and services provide valuable information for decision making for both consumers and service providers (of products and services) alike. With the availability of forums such as blogs and social networks, it has never been easier to freely, critically and anonymously express feedback in the form of opinions and general comments about products and services. Analysing such feedback poses challenges, in part because of the vast quantities of user posts, but also because the responses from general consumers is typically expressed in free form text. Both providers and consumers use such forums in a variety of ways. Providers use the data to effect improvements in products and services. Consumers rely on other consumers' feedback to decide on acquisition of products and services, or to validate their own opinions be it about complaints or the comments about improved services. A provider might wish to identify a service (among many of its services) that is causing customers to withdraw their patronage. A consumer might wish to locate a provider outlet with the friendliest customer service. These decisions cannot be made without manually scanning large volumes of postings in the forum. If forums are regarded as collections of large, free form text documents (a post by a consumer may be regarded as a single document) then effective decision making from the feedback in these collections is impossible without systematically analysing, summarizing and subsequently organising the documents to qualitatively enhance the usability of the available data.

And many previous studies focus on analysing product reviews. The focus of this study is to investigate a model that on suite reviews and more specifically, reviews related to health systems. Previous studies have shown that effective regulation of healthcare professions is increasingly recognized as pivotal to improvements in healthcare quality (Bismark and Studdert [2013]). Moreover the effectiveness of regulatory authorities is affected by insufficient resources and inadequate information received (Bismark et al. [2013]).

The Australian Health Practitioner Regulation Agency (AHPRA) is the official Australian Government authority responsible for monitoring the performance and conduct of health practitioners across the 14 health professions. Since 2010 laws in all Australian states and territories require health practitioners to report all “notifiable conduct” that is brought to their attention to AHPRA. The law targets all registered health practitioners in Australia which includes doctors, nurses, dentists and practitioners from 11 allied health professions. The obligation to report notifiable conduct includes an obligation on behalf of employers, education providers and health practitioners. A report can be made if there is a reasonable belief that the behaviour or performance is notifiable. Apart from the role of AHPRA to facilitate the regulation of 14 healthcare professions, there are other forums to address poor performing healthcare practitioners. These include publicly available stories, feedbacks and opinions from patient which can address similar issues faced by AHPRA.

In this thesis we focus on topic modeling as a way to summarise documents in such web-based forums or collections. Specifically, we apply topic modeling to a web-based forum that deals with patient opinion (www.patientopinion.org.au). The following quote summaries the objectives of the forum[1]:

> Patient Opinion was founded in the UK in 2005 and since then has grown to be the UK’s leading independent non-profit feedback platform for health services. Patient Opinion Australia (POA) was established in 2012 and, similar to its UK counterpart, is registered as an independent not-for-profit charitable institution. Patient Opinion is about honest and meaningful conversations between patients and health services. We believe that your story can help make health services better.

A *topic* is a set of highly probable words or terms discovered from the collection of documents that are used to categorise a subset of the documents in the collection. Topic modeling automatically discovers a set of identifying topics that allow for effective organisation of the collection of documents.

Topic modeling has proven a very powerful tool of analysing large quantity of data, on the other hand publicly available healthcare opinions, reviews have been overlooked in previous topic modeling studies.

[1] www.patientopinion.org.au/info/about

In this project we use Latent Dirichlet Allocation (LDA) Blei et al. [2003] as a base topic model, one that has been used in many other studies as a base model and has proven to be effective. The Multi-Domain Prior Knowledge LDA (MDK-LDA) topic model, proposed by Chen (Chen et al. [2013]), extends LDA by introducing a new latent variable s in LDA to model *s-sets*, which is a domain expert pre-defined category or group. Each document is an admixture of latent topics while each topic is a probability distribution over s-sets. Another approach is Aspect-Based Summarization (Garcia-Moya.L and Berlanga-Llavori.R [2013]). It is usually composed of three main tasks: aspect identification, sentiment classification, and aspect rating. This model is used to analyse product review and designed to effectively retrieve features and sentiment for products.

Due to the unique characteristic of the data from *Patient Opinion*, the existing algorithm has been improved with user input data. LDA has proven to be an effective model and has been used as a based model in many topic modeling studies. We chose LDA as our base model, and incorporated unique features of *Patient Opinion*, specifically the sections of the website that include "What is Good" and "What could be improved". These two sections are completed by users for submission of the story. The template for submission of stories is provided by the website. Generally "What is Good" and "What could be improved" are the main topic or features users want to give feedback about in their stories. Moreover, we take as a starting point that users label their stories accurately.

Our use of user tag data motivates our research question. We seek to answer the following question in this thesis: *Does user-driven input improve topic modeling?*

The project has made the following contribution to the field of topic modeling by using the LDA as the base framework: Introduction of the user specified keywords, and reduction in the number of terms for each topic whilst retaining the quality of the topic. The remaining terms are ranked for each topic by inverse document frequency score which is a common metrics in information retrieval area to measure how much information a word provides, that is, whether the term is common or rare across all document.

Chapter 2

Related Work

Latent Dirichlet Allocation also known as LDA or discrete PCA is a Bayesian graphical model for text document collections represented by bags-of-words (Newman et al. [2009], Blei et al. [2003], Griffiths and Steyvers [2004], Buntine and Jakulin [2004]). The basic idea of LDA is to treat each document in a collection as a vector of word count. Each document is represented as a probability distribution over a number of topics, while each topic is represented as a probability distribution over a number of words. The model allows sets of observations to be explained by unobserved groups that explain why some parts of the data are similar. Generally, only a small number of words have high likelihood in each topic and each document only presents certain number of topics. Following is the equation of collapsed Gibbs sampling:

$$p(z_{id} = t \mid x_{id} = w, Z^{\neg id})\alpha \tag{2.1}$$

$$\frac{N_{wt}^{\neg id} + \beta}{\sum_w N_{wt}^{\neg id} + W\beta} \frac{N_{wt}^{\neg id} + \alpha}{\sum_w N_{wt}^{\neg id} + T\alpha} \tag{2.2}$$

where $z_{id} = t$ assigns topic t with i^{th} word in document d, and word w currently observed indicated by $x_{id} = w$. $Z^{\neg id}$ is the vector of all topic assignments not including the current word. N_{wt} represent integer count arrays, and α is the parameter of the Dirichlet prior on the per-document topic distributions, β is the per-topic word distribution.

2.1 Multi-Domain Prior Knowledge in Topic Modeling

As mentioned before, LDA is a powerful topic modeling framework, however recent studies found that these unsupervised models may not produce topics that conform to the user's existing knowledge(Chen et al. [2013]). Chen et al (Chen et al. [2013]) proposed a novel knowledge-based model, called Multi-Domain Prior Knowledge -LDA (MDK-LDA), which is capable of using prior knowledge from multiple domains to help topic modeling in the new domain. A new latent variable “s” is added to model the s-set and this s-set is defined by domain expert, each document represent admixture of latent topics while each topic is a probability distribution over s-set. MDK-LDA uses s-set to distinguish topics in multiple senses. For example the word light can be represented by two s-set: S1 {light, heavy, weight} and S2 {light, bright, luminance}, if light co-occurs with bright or luminance it will be assigned to S2. The words in each topic are ranked by per-topic word distribution $\pi_t(w) = \sum_{s=1}^{s}(\psi_t(s) \cdot \eta_{t,s}(w))$ where $\psi_t s$ is a per topic distribution over s-sets and $\eta_{t,s}(w)$ is per topic, per s-set distribution over word. To evaluate the discovered topics, two human domain experts are employed, top 20 words from each topic are selected based one previous ranking. Each word in a topic is considered correct if both experts agree, otherwise the word is labelled as incorrect. A natural way to evaluated these rankings is to use Precision @ n (or p@n), where n is the rank position. The MDK-LDA was measured against base model with settings p@n for n = 5, 10, 15 and 20. As a result MDK-LDA outperforms base LDA model in each test, the overall average precision score is 0.67 and 0.88 for LDA and MDK-LDA respectively, a paired t-test also conducted with the result of $p < 0.0001$ which means a statistically significance between LDA and MDK-LDA. The study concluded Multi-Domain Prior Knowledge can be used to improve topic modeling. As most topic modeling studies, MDK-LDA studies data of product reviews. The data is from six domains from Amazon.com. And for this model to work, a prior domain knowledge set (or s-set) need to be generated. The process of generating the s-set involves certain level of human interaction. Our study focuses on Australia healthcare reviews which to the best of our knowledge there hasn’t any topic modeling studies tried to exploring this area. We also proposed a new approach called user-driven topic modeling.

2.2 Intrinsic qualitative evaluation of learned topics

With the strong interest within computational linguistics in techniques for learning topics which capture the latent semantics of a document collection, many models have been proposed for example, LDA and MDK-LDA we mentioned above. Extrinsic evaluation has been used to demonstrate the effectiveness of the learned topic in the application domain but, no attempt has been made to perform intrinsic evaluation of the topics themselves either qualitatively or quantitatively, Newman (Newman et al. [2010]) introduces the novel task of topic coherence evaluation. A range of topic scoring models to the evaluation task are applied, drawing on WordNet, Wikipedia and the Google search engine and compared with human scores for a set of learned topics. The evaluation metrics:

$$Mean - D - Score(t) = meanD(w_i, w_j, ij \in 110, i < j) \tag{2.3}$$

$$Median - D - Score(t) = medianD(w_i, w_j, ij \in 110, i < j) \tag{2.4}$$

Given the topic t based on the component terms ($w_1 w_{10}$ and word-similarity measure $D(w_i, w_j)$, the total number of word-pair over 10 words is 45. The experiment collected 55,000 news articles from English Gigaword, and the collection of 12,000 books was downloaded from the Internet Archive. 200 and 400 topics are learned from topic modeling for news and books respectively. Then a total of 237 topics are select from two collection for user scoring. 9 users scored 237 topics on a 3-point scale where 3="useful" (coherent) and 1 = "useless"(less coherent). The inter-annotator agreement (IAA) score is treated as the "Gold-standard", of all the topic scoring methods tested, Pointwise Mutual Information (PMI) is the most consistent performer, achieving the best or near-best results over both datasets.

$$PMI(w_i, w_j) = log\frac{P(w_i, wj)}{P(w_i)P(w_j)}, \tag{2.5}$$

Due to the consistency of PMI, we will use this method to measure topic coherence of learnt topics from Patient Opinion and also for intrinsic evaluation of user-driven topic modeling.

2.3 Best Topic Word Selection for Topic Labelling

Naturally, not all topics are equally coherent, and the higher the topic coherence, the easier the label selection task becomes. Lau (Lau et al. [2010]) proposed a method to select best topic word for topic labelling based on word probability model. The conditional probability used in the study:

$$P(w_i \mid w_j = \frac{P(w_i, w_j)}{P(w_j)} \tag{2.6}$$

where $i \neq j$ and $P(w_i, w_j)$ is the probability of observing both w_i and w_j in the same sliding window, and $P(w_i$ is the overall probability of word w_i in the corpus. Then average of conditional probability for word w_i is calculated by:

$$avg - CP1(w_i) = \frac{1}{9}\sum_j P(w_i \mid w_j) \tag{2.7}$$

for j = 1 1, $j \neq i$, as the study focus on top 10 topic words. The flipped situation also tested, where the most representative word may evoke (rather than be evoked by) other words in the list of ten word, the equation becomes:

$$avg - CP1(w_i) = \frac{1}{9}\sum_j P(w_j \mid w_i) \tag{2.8}$$

The study also tested the pointwise mutual information (PMI) approach we mentioned in previous section, the average of word w_i over top 10 words in the topic is given by:

$$avg - PMI(w_i) = \frac{1}{9}\sum_j PMI(w_j, w_i) \tag{2.9}$$

The testing data used in this study is exactly same as Newman's (Newman et al. [2010]) we mentioned in previous section. The PMI-scores are calculated for 200 topics from NEWS and 400 topics from BOOKS. 60 topics are selected with high PMI-score, and 60 topics with low PMI-score, from both group, resulting in a total of 240 topics for human evaluation. Each topic is scored with 3-point scale by human where score 3 means a most useful or coherent topic. The words in each topic are ranked by average PMI and conditional probabilities CP1 and CP2, top 3 words from each approached are selected as best words candidates. They are evaluated against weighted scoring function:

$$Best - Nscore = \frac{\sum_{i=1}^{N} n(w_{rev_i})}{\sum_{i=1}^{N} n(w_i)} \tag{2.10}$$

where w_{rev_i} is the i^{th} term ranked by the system and w_i is the i^{th} most popular term selected by annotators; rev_i gives the index of the word w_i in the annotator's list; and n(w) is the number of votes given by annotators for word w. As a result, each approach shows the strength in certain topics against base model, the performance isn't consistent across 3 approaches, however, in combination as inputs to a re-ranking model, the result is consistent and shows improvement against baseline. In our study of Patient Opinion Australia, PMI-scores for topics are used to evaluate topic coherence and it is measured by average of number of word-pairs, the reason is we produce topics with variable number of terms. To rank the words in each topic, we use Inverse Document Frequency given by:

$$idf(t, D) = log\frac{N}{\mid \{d \in D : t \in D\} \mid} \tag{2.11}$$

Chapter 3

User-driven Topic Modeling

Although LDA provides a powerful framework for extracting latent topics in text document, sometimes the learned topics extracted are lists of words that do not convey much useful information (Newman et al. [2009]). To solve the problem base LDA model had been extended either by incorporating human judgment into the model-learning framework or creating a computational proxy that simulates human judgments (Chang et al. [2009]) as per the MDK-LDA model (Chen et al. [2013]) we discussed in Chapter 2. However, solutions like MDK-LDA require a certain level of human interaction which limits the size of the training data. Part of the training constraints are defined by domain experts which means the result could contain certain level of subjective factors.

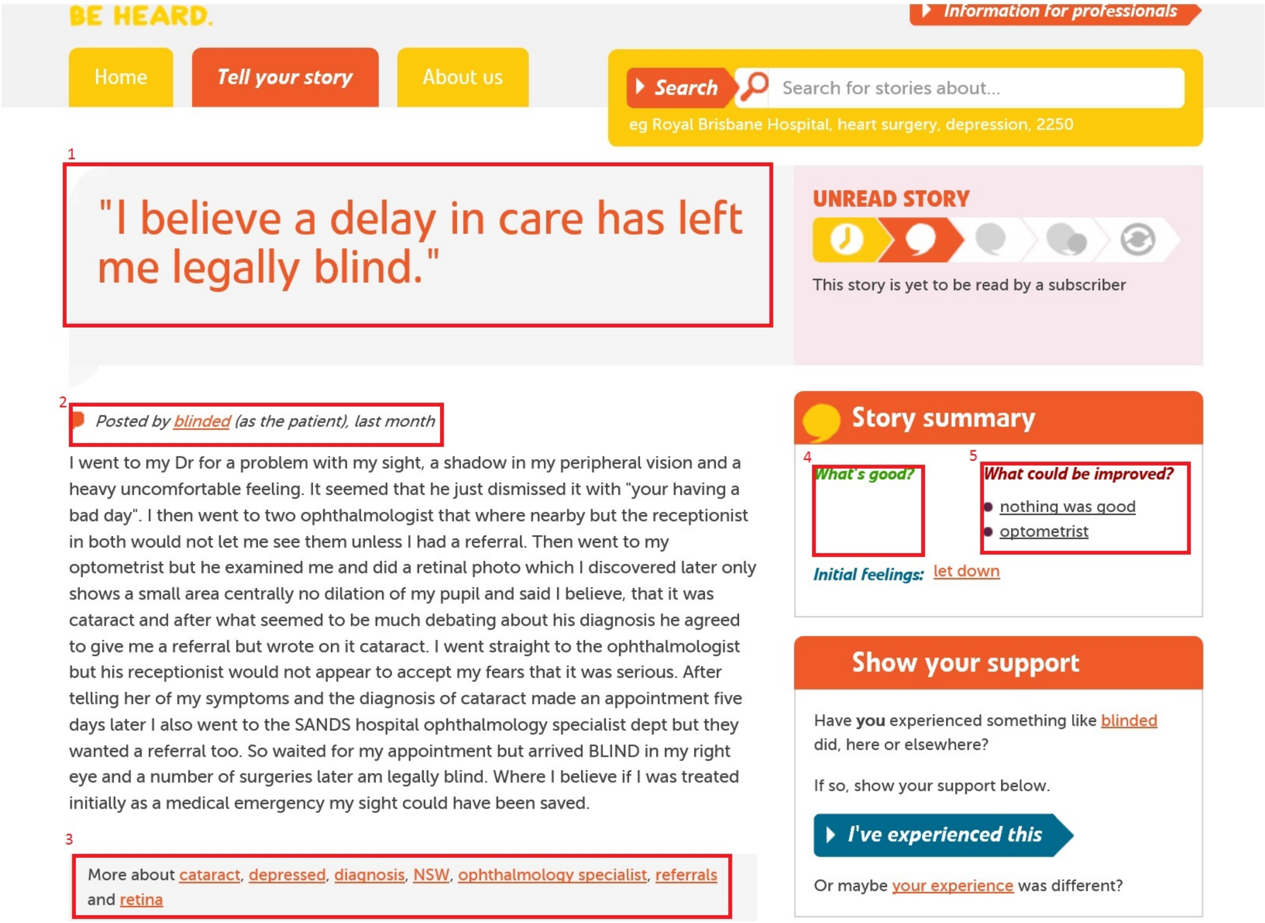
Due to the unique characteristic of the data of Patient Opinion, we have the opportunity to minimize the input from human judge during training phase by employing use user input to simulate human judgment, hence to produce the topic modeling result with less subjective factors.

3.1 Description of Data

The data from Patient Opinion of relevance to this study includes:

- The title of the story and the summary of the story by the user.
- The "Author role" and "Time of the post", author role can be the patient, patients relative, carer or doctor.

- The "More about" section completed by the website moderator who inserts relevant tags to the story.
- The most important fields are "What is Good" and "What could be improved", these field are inserted by the user, the fields indicate what user thinks the story is about, and we use these fields to simulate user judgement in topic modeling.


BE HEARD.
Information for professionals
Home
Tell your story
About us
Search
Search for stories about...
eg Royal Brisbane Hospital, heart surgery, depression, 2250
1
2
3
4
5
I've experienced this

"I believe a delay in care has left me legally blind."

UNREAD STORY

This story is yet to be read by a subscriber

Posted by blinded (as the patient), last month

I went to my Dr for a problem with my sight, a shadow in my peripheral vision and a heavy uncomfortable feeling. It seemed that he just dismissed it with "your having a bad day". I then went to two ophthalmologist that where nearby but the receptionist in both would not let me see them unless I had a referral. Then went to my optometrist but he examined me and did a retinal photo which I discovered later only shows a small area centrally no dilation of my pupil and said I believe, that it was cataract and after what seemed to be much debating about his diagnosis he agreed to give me a referral but wrote on it cataract. I went straight to the ophthalmologist but his receptionist would not appear to accept my fears that it was serious. After telling her of my symptoms and the diagnosis of cataract made an appointment five days later I also went to the SANDS hospital ophthalmology specialist dept but they wanted a referral too. So waited for my appointment but arrived BLIND in my right eye and a number of surgeries later am legally blind. Where I believe if I was treated initially as a medical emergency my sight could have been saved.

More about cataract, depressed, diagnosis, NSW, ophthalmology specialist, referrals and retina

Story summary

What's good?

What could be improved?
- nothing was good
- optometrist

Initial feelings: let down

Show your support

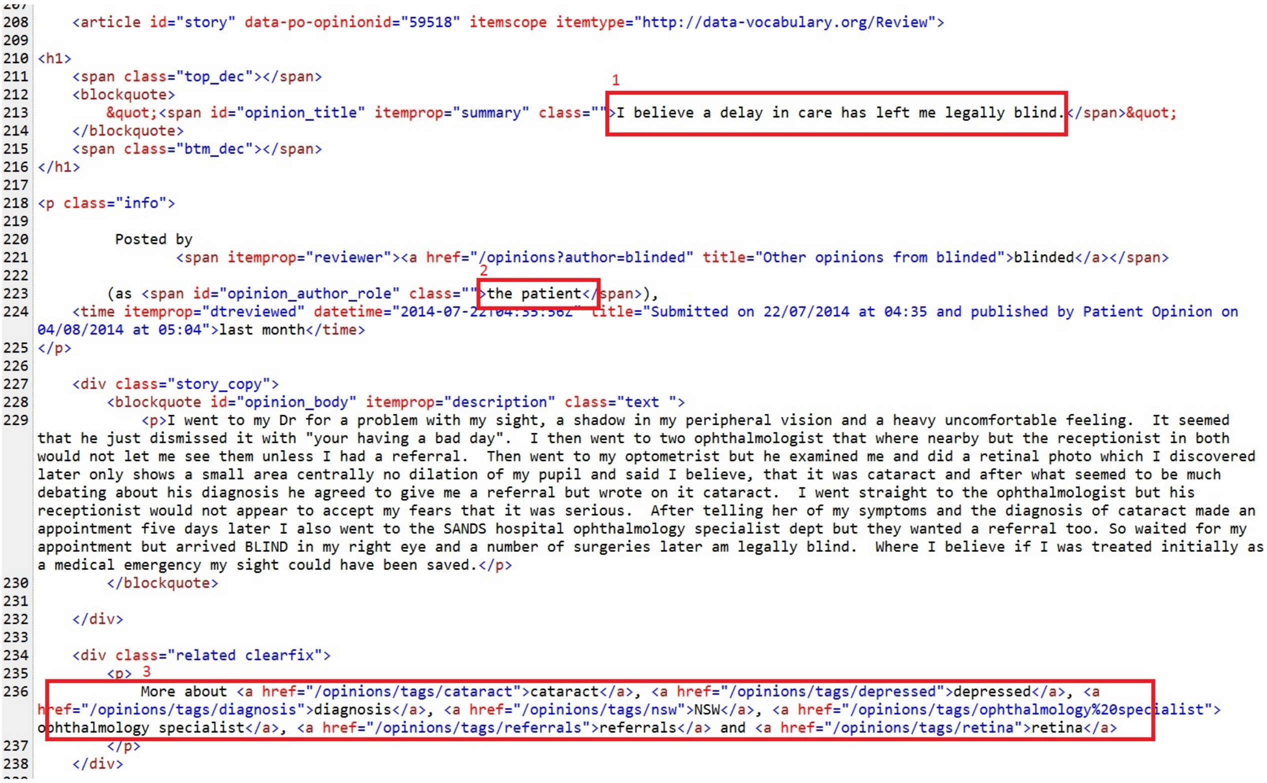
Have **you** experienced something like blinded did, here or elsewhere?

If so, show your support below.

Or maybe your experience was different?

Figure 3.1: Patient Opinion Story Sample

Each story also includes some non-mandatory fields which are not in the screenshot but provide very important information for example, the location information as some stories specify a particular hospital or clinic and most stories at least have state information. A quick overview of the data illustrates the potential of using the data from a public forum for monitoring of healthcare professionals and institutions. Patient Opinion Australia established was in 2012, and to date contains 624 stories in total. It's allied site Patient Opinion UK was founded in 2005 and has more than 80,000 stories. As the scope of this project's focus

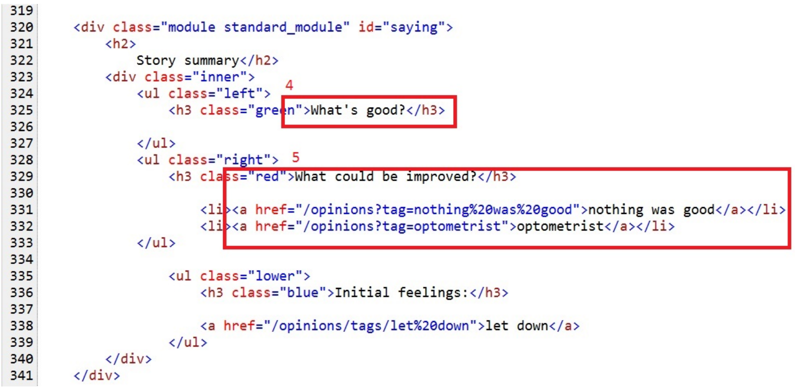


```
    <article id="story" data-po-opinionid="59518" itemscope itemtype="http://data-vocabulary.org/Review">

<h1>
    <span class="top_dec"></span>
    <blockquote>
        &quot;<span id="opinion_title" itemprop="summary" class="">I believe a delay in care has left me legally blind.</span>&quot;
    </blockquote>
    <span class="btm_dec"></span>
</h1>

<p class="info">

         Posted by
                <span itemprop="reviewer"><a href="/opinions?author=blinded" title="Other opinions from blinded">blinded</a></span>

        (as <span id="opinion_author_role" class="">the patient</span>),
    <time itemprop="dtreviewed" datetime="2014-07-22T04:35:56Z" title="Submitted on 22/07/2014 at 04:35 and published by Patient Opinion on
04/08/2014 at 05:04">last month</time>
</p>

    <div class="story_copy">
        <blockquote id="opinion_body" itemprop="description" class="text ">
            <p>I went to my Dr for a problem with my sight, a shadow in my peripheral vision and a heavy uncomfortable feeling.  It seemed
that he just dismissed it with "your having a bad day".  I then went to two ophthalmologist that where nearby but the receptionist in both
would not let me see them unless I had a referral.  Then went to my optometrist but he examined me and did a retinal photo which I discovered
later only shows a small area centrally no dilation of my pupil and said I believe, that it was cataract and after what seemed to be much
debating about his diagnosis he agreed to give me a referral but wrote on it cataract.  I went straight to the ophthalmologist but his
receptionist would not appear to accept my fears that it was serious.  After telling her of my symptoms and the diagnosis of cataract made an
appointment five days later I also went to the SANDS hospital ophthalmology specialist dept but they wanted a referral too. So waited for my
appointment but arrived BLIND in my right eye and a number of surgeries later am legally blind.  Where I believe if I was treated initially as
a medical emergency my sight could have been saved.</p>
        </blockquote>

    </div>

    <div class="related clearfix">
        <p>
            More about <a href="/opinions/tags/cataract">cataract</a>, <a href="/opinions/tags/depressed">depressed</a>, <a
href="/opinions/tags/diagnosis">diagnosis</a>, <a href="/opinions/tags/nsw">NSW</a>, <a href="/opinions/tags/ophthalmology%20specialist">
ophthalmology specialist</a>, <a href="/opinions/tags/referrals">referrals</a> and <a href="/opinions/tags/retina">retina</a>
        </p>
    </div>
```

Figure 3.2: Patient Opinion Story Sample Source 1

is on the Australian healthcare system, the data from the United Kingdom will not be used but it is evident that the healthcare system is attracting feedback from general public. The information collected from those sources has the potential to improve and influence conduct of healthcare professionals and institutions. From Patient Opinion Australia we collected 659 unique terms in user specified field out of 624 stories.

The most frequent terms are: "care" which appears 399 times in 278 stories; "service" which appears 150 times in 141 stories; "staff" which appears 148 times in 116 stories; and "hospital" which appears 141 times in 118 stories. The terms where split into to two groups: "Good" and "Need to Improve", we have 412 user specified terms in "Good" out of 467 stories, while the "Need to Improve" includes 408 terms in 264 stories. An interesting observation is the general order of term frequency in "Good" group matches overall count, while "Need to Improve" group shows some differences. Instead of "service" and "staff", "hospital, doctor, communication" follows after "care" in this group. It suggests stories relate to "service, staff" are more likely get positive feedback compared to "hospital, doctor, communication". However, both groupings illustrate the interest from general public as we do not want to overlook topics which positive. When data was analysed by the topic modeling algorithm all



```
<div class="module standard_module" id="saying">
    <h2>
        Story summary</h2>
    <div class="inner">
        <ul class="left">
            <h3 class="green">What's good?</h3>

        </ul>
        <ul class="right">
            <h3 class="red">What could be improved?</h3>

                <li><a href="/opinions?tag=nothing%20was%20good">nothing was good</a></li>
                <li><a href="/opinions?tag=optometrist">optometrist</a></li>
        </ul>

            <ul class="lower">
                <h3 class="blue">Initial feelings:</h3>

                <a href="/opinions/tags/let%20down">let down</a>
            </ul>
    </div>
</div>
```

Figure 3.3: Patient Opinion Story Sample Source 2

user specified terms were treated as a single group. Table 3.1 displays the term frequency per state of Australia and New South Wales and Queensland are the most active states. Table 3.2 includes the "Good" group term frequency per state. The term "information" has been mentioned 54 times out of 49 stories with 42 stories are from New South Wales. These findings suggest patients in New South Wales are more satisfied in "information" related topics than other states. Table 3.3 includes the result for "Need to Improve" group term frequency per state. The term frequency of "care, doctor, communication, staff" are markedly higher for Queensland than the other states .

Each number we had showed so far is individual case for example, we assumed the issue with "service", "staff", "hospital" are all isolated cases, in reality an issue could be "bad service from staff"; "issues with staff in hospital", topic modeling is a process to discover the connections between terms. Our project intend to employ traditional topic modeling technique to discover the latent topics among documents, and then increase the accuracy of the result with the help of user specified keywords.

Table 3.1: Overall Term Count Over States

Term	TF_{ALL}	DF_{ALL}	DF_{NSW}	DF_{VIC}	DF_{ACT}	DF_{TAS}	DF_{QLD}	DF_{SA}	DF_{NT}	DF_{WA}
care	399	278	61	28	4	3	114	21	2	5
service	150	141	51	7	1	0	51	5	1	5
staff	148	116	32	12	1	0	45	4	0	4
hospital	141	118	18	18	4	3	47	10	1	4
doctor	102	84	16	7	2	1	40	6	1	0

Table 3.2: Group "Good" Term Count Over States

Term	TF_{ALL}	DF_{ALL}	DF_{NSW}	DF_{VIC}	DF_{ACT}	DF_{TAS}	DF_{QLD}	DF_{SA}	DF_{NT}	DF_{WA}
care	250	178	48	13	2	2	73	11	2	6
service	125	119	47	5	1	0	42	4	1	5
staff	124	98	30	1	1	0	35	3	0	3
hospital	80	68	8	9	1	2	31	5	1	2
information	54	49	42	0	0	0	4	1	0	1
doctor	46	40	11	2	1	0	17	4	1	0

Table 3.3: Group "Need to Improve" Term Count Over States

Term	TF_{ALL}	DF_{ALL}	DF_{NSW}	DF_{VIC}	DF_{ACT}	DF_{TAS}	DF_{QLD}	DF_{SA}	DF_{NT}	DF_{WA}
care	149	115	16	15	2	1	47	12	1	3
hospital	61	54	11	8	3	0	15	4	1	1
doctor	56	44	5	5	1	1	23	2	0	0
communication	26	26	2	1	1	0	15	5	0	0
service	25	23	4	2	0	0	9	1	0	0
staff	24	22	2	1	0	0	12	1	0	1

3.2 Preprocessing of Data

To be able to feed the data to Mallet, the original data was normalized. The story body and story identification (ID) are extracted from website. All stories are condensed into one single file with each story formatted to a single line start with the story identification. The leading story identification is used by Mallet for labelling the composition result. Some internal data structures are also defined for post processing. All data is converted to lowercase. Unique user specified keywords are collected. This collection is used to filter out the words in each topic that is generated by LDA at a later stage. A list of related document identification to each word is also collected and indexed for fast lookup for document frequency. Figure 3.4 illustrates the basic data structure: documents that contain term 'education' are '58192, 59160, 58385, 59177', the numbers are the document identification from Patient Opinion Australia, the original identification is used for easier retrieval of stories.

3.3 Using user input to improve topic modeling result

Topics learned from LDA may sometimes not convey much useful information. This may be caused by overfeeding the result set for example, it includes top 19 words for each topic (based on the settings, the total number in each topic can be configured) and some words may not be relevant to the current topic but statistically significant to the topic. Our goal is using user input to reduce the noise while retaining as much information as possible for a topic. The process is given as follows:

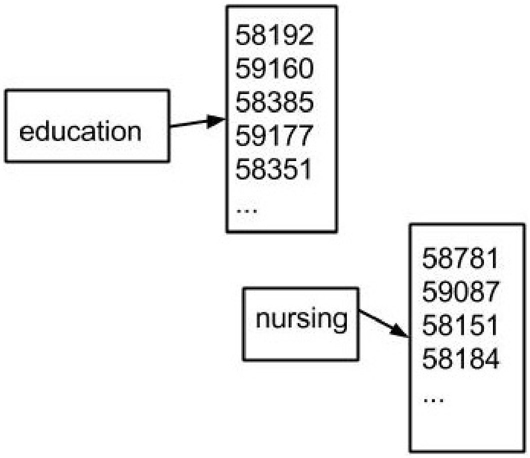


Figure 3.4: Patient Opinion Story Sample

1. Collect unique words from user specified field as S set.
2. Generate a set of topics T with LDA model.
3. Calculate result set R as: for each topic $t \in \{1, ..., T\}, r_n = t_n \cap S$
4. Treat the result T as collection of document, calculate Inverse Document Frequency for each term with $idf(t, D) = log\frac{N}{|\{d \in D: t \in D\}|}$, where N = 100 represent the 100 topics generated with LDA.
5. Re-arrange terms in set R with idf score, so the more significant word appear in front of each topic.

When collect unique user specified keywords the phrase is break down to words, because the result topic generated by LDA is collection of words. Inverse Document Frequency of each word is calculated in the scope of topic, it is the measurement of how important a term is within topics.

Chapter 4

Experiments and Result

We collected all 624 stories from Patient Opinion by August 2014. The count of unique user specified term is 659. One hundred topics are generated using Mallet [1] with setting of optimize interval equals to 20. The number of unique terms in the overall topic set and also in user specified keywords set is 527, compare to 1440 unique terms in the overall topic set. One of the metrics Mallet provide is Topic Composition which is per topic probability distributions over current document ($P(T \mid D)$) which in other word the probability of T given document D. For each document the sum of composition over total topic should always be 1:

$$\sum_{t=1}^{1} 00P(t \mid D) = 1 \tag{4.1}$$

Table 4.1 includes the topic composition. Rows represent documents with document identification (the document ID is the unique number from Patient Opinion Australia, the number can be used to retrieve the original document, not the index of document as by today there are only 624 stories in total) in first column and the remaining columns represent topic probability distributions over current document.

As each topic can be represented as probability distribution of documents and document is formed by terms we can use the sum of composition score to measure topic quality:

$$Score(t) = \sum_{n=1}^{N} Comp(D_{n,t}) \tag{4.2}$$

Where N is set of document that contains at least one term in topic “t”, $Comp(D_{n,t})$ is the

[1] http://mallet.cs.umass.edu/

composition of topic “t” of document “n”. The total composition score for each topic is calculated with the help from term to document identification index built previously (Table 4.2). Clearly, the total composition of documents from original topic words is expected to be greater than its subset which is the same topic but filtered the words those not mentioned by user at all. If the sum of topic composition is small between two groups means the topic we produce has the similar quality as the original one, in other words the documents only contain the terms we omitted (which not mentioned by user at all) do not have significant contribution to the topic, as a result those terms are considered as noise from base model. The average difference of sum of composition between two groups over 100 topics is 1.0829. There are 58 topics which have the difference below this average number and use this number as cut-off boundary, we consider those 58 topics we extracted from base model are good topics as the quality of the topic are relatively higher than other topics. A sample of the topics are selected from the remaining topics. Please refer to Table 4.3.

Table 4.1: Example of Composition

Doc ID	Topic Index	Composition	Topic Index	Composition	...	Topic Index	Composition	Topic Index	Composition
58954	61	0.1171875	91	0.0390625	...	78	0.0234375	72	0.0234375
58832	83	0.076388889	78	0.048611111	...	10	0.048611111	71	0.034722222
58953	83	0.077380952	65	0.053571429	...	29	0.041666667	60	0.029761905
58956	78	0.025	76	0.025	...	65	0.025	62	0.025
58834	42	0.065517241	11	0.065517241	...	58	0.037931034	44	0.037931034
58710	12	0.108208955	18	0.063432836	...	90	0.041044776	71	0.041044776
58952	91	0.044642857	73	0.026785714	...	59	0.026785714	36	0.026785714
58830	94	0.108490566	80	0.051886792	...	99	0.04245283	71	0.04245283
58951	93	0.0625	81	0.052884615	...	0	0.052884615	79	0.043269231
58719	51	0.27238806	27	0.063432836	...	42	0.026119403	22	0.026119403
58716	77	0.041666667	90	0.025	...	73	0.025	62	0.025
58718	83	0.242307692	47	0.05	...	71	0.042307692	11	0.034615385
58839	25	0.070512821	63	0.032051282	...	59	0.032051282	58	0.032051282
58717	43	0.050660793	57	0.046255507	...	92	0.04185022	72	0.04185022
58723	91	0.028301887	68	0.028301887	...	35	0.028301887	99	0.009433962
58965	83	0.097014925	1	0.037313433	...	77	0.02238806	72	0.02238806

Inverse Document Frequency (idf) is used to measure the weight of each term in the scope of whole topic set which means the less the term appears in overall topic set the more important the term is. One thousand one hundred and sixteen out of original 1440 terms appears once in original topic set, so more than half of the term has the highest idf value 2 in the data set. Two hundred and thirty six terms appear twice with idf=1.7 and 55 terms appear 3 times with idf = 1.5. There are also significant variance in the result topics as some topics may remain with the existing order or only have one word shifted, whilst some may look very differently. For example, {program, kate, lifestle, meet, included, learnt, learning, relationship}, only $idf_{kate} = 1.7$, the idf for the rest equals to 2, so the new topic becomes {program, kate, meet, included, learnt, learning, relationship}. Another example {bed, family, care, staff, time,

Table 4.2: Sum of Composition

Topic Index	Sum of Composition of Original Documents	Sum of Composition of Filtered Documents
1	3.920362	3.382284
2	3.742091	3.275150
3	4.396231	3.930039
4	4.275353	3.603821
5	4.569670	4.444751
6	3.153133	2.775077
7	3.368214	2.622281
8	4.541646	4.018145
9	5.399709	5.194911
10	4.498450	3.711763
11	4.255079	4.101976
12	6.755018	6.489172

attending, unit, palliative}, $idf_{care} = 1.3, idf_{staff} = 1.2 and idf_{palliative} = 2$ and the rest has idf equals to 1.7, as a result the new topic should look like {palliative, bed, family, time, attending, unit, care, staff}.

4.1 Topic Coherence Evaluation

Apart from quantitative and qualitative evaluation as above evaluating topic coherence is a component of the larger question of what are good topics, what characteristics of a document collection make it more amenable to topic modeling and how can the potential of topic modeling be harnessed for human consumption (Newman et al. [2010]). The topic coherence is measured with Pointwise Mutual Information (Newman et al. [2011]) (PMI) score:

$$PMI - Score(w) = (\frac{N^2 - N}{2})^{-1} \sum PMI(w_i, w_j), ij \in \{1...N\} \tag{4.3}$$

where

$$PMI(w_i, w_j) = log\frac{P(w_i, wj)}{P(w_i)P(w_j)}, \tag{4.4}$$

Since the number of terms that form each topic isn't normalised we calculate the average of the topic where N is the number of terms in that topic. $(\frac{N^2-N}{2})^{-1}$ gives the number of distinct pairs in N. The measure is symmetric $P(w_i, wj) = P(w_j, wi)$ which means we only measure the difference of topic coherence between original topic and filtered topic.

Table 4.3: Sample topics

Original	Filtered
program healthy kate lifestyle sessions eat meet programme included organised held healthier encouraging learnt foods beneficial learning relationship handle	program kate lifestyle meet included learnt learning relationship
gp local recently records government copy prescription paper multiple tasmania gps referring avail beginning calls surprised cairns super shared	gp local records government prescription paper gps cairns
physio gp mri injury follow shoulder xray week asked hospital discussed full physiotherapist neck stand complaining neurologist princess forte	physio gp mri xray hospital full physiotherapist neck
call waiting phone told back called list unit rang ring explain apparently assumed clerk calling noticed mcewin lyell requested	call waiting phone back list unit clerk calling
father bed family care appears staff time attending difficult dad speak comfort unit incident law visitor awake gosford palliative	bed family care staff time attending unit palliative
time advised team contact causing consultant tumour professional manner independent safe stressful arrival note closed usual considerate empathetic seizures	time team contact consultant professional manner independent empathetic
looked experience er bad partner full approach worry give free skills male chronic terrible running provider building drive welcomed	looked experience er partner full approach skills male building
night stay thing major hospital admission support suggested accommodation fully sydney unable relatives sleep developed tuesday staff added environment	night stay hospital admission accommodation relatives sleep staff environment
waiting wait hours room hour area waited reception number temperature hurt remember panadol minutes geelong sunday impressed time er	waiting wait hours room area reception number time er
child issues jean aboriginal helped understanding knowledge school minds behaviour hay mighty woods anger louise clinician strategies interaction love	child jean aboriginal understanding knowledge school minds behaviour woods strategies
public brisbane system live mater hospital pa run booking advise toowoomba meds qld health expect west lift weekend weak	public brisbane system hospital run booking meds health west lift
health community sarah local support primary medicare rural group libby people murrumbidgee provide part clients topics art groups guest	health community sarah local primary medicare rural group people murrumbidgee clients guest

Table 4.4 includes some samples of the PMI scores plus the difference between two PMI included in the last column. As we can see from the result the majority of the topics result in improvement of coherence and the overall average is 0.297. A paired-t test shows the $P = 3.76e^{-22}$ which means the filtered topic is statistically significant from the original one.

Table 4.4: PMI Scores

Topic Index	Original Topic PMI Scores	Filtered Topic PMI Scores	Difference
1	1.864771	2.040466	0.175695
2	1.844573	2.305665	0.461092
3	1.543335	1.928804	0.385469
4	2.145729	2.333656	0.187927
5	1.912059	2.195632	0.283573
6	1.854318	2.341428	0.48711
7	1.748795	2.33346	0.584665
8	1.899982	2.624272	0.72429
9	1.752184	1.80262	0.050436
10	2.152952	2.371954	0.219002
11	1.986177	2.120463	0.134286
12	1.776924	2.030256	0.253332

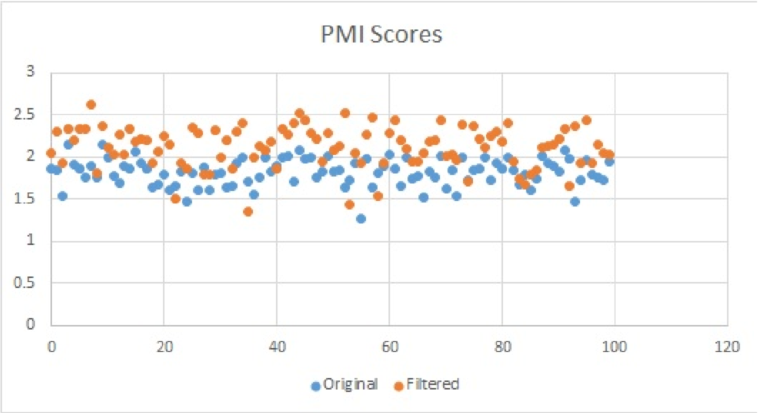


Figure 4.1: PMI-Score

Chapter 5

Conclusion and Future Work

This study shows the possibility of using user input to improve topic coherence in topic modeling of healthcare related blogs. To the best of our knowledge this has not been done before. We proposed a method that use user input words as a filter for the result from LDA model. We successfully reduced the number terms in each topic while still keep the topic meaningful. The terms that have been omitted can be considered as noise which means the documents they associate to do not significantly contribute to the possibility distribution over the topic as this is evaluated by the total composition score. Hence, the overall topic coherence is improved as less noise in the topic. Furthermore, we investigated using Inverse Document Frequency to re-rank the terms in each topic. Unfortunately PMI-score evaluation is symmetric which means the order of each term in topic is not considered. An existing statistical model was not found for the evaluation. Our method of ranking the terms by idf isn't ideal but since the term frequency for each term for in a topic always equals to 1, this approach is reasonable for step one. Future work may utilise this approach and could be expended to count term frequency in the scope of whole collection of documents and idf still from the topic list. If the scope of the factor expended to the whole collection, it is also reasonable to rank the topics not only the terms in each topic, hence the model could make suggestions of which topic is more likely an important one.

Bibliography

M. M. Bismark and D. M. Studdert. Governance of quality of care: a qualitative study of health service boards in victoria, australia. *BMJ quality & safety*, pages bmjqs–2013, 2013.

M. M. Bismark, M. J. Spittal, L. C. Gurrin, M. Ward, and D. M. Studdert. Identification of doctors at risk of recurrent complaints: a national study of healthcare complaints in australia. *BMJ quality & safety*, 22(7):532–540, 2013.

D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.

W. Buntine and A. Jakulin. Applying discrete pca in data analysis. In *Proceedings of the 20th conference on Uncertainty in artificial intelligence*, pages 59–66. AUAI Press, 2004.

J. Chang, S. Gerrish, C. Wang, J. L. Boyd-graber, and D. M. Blei. Reading tea leaves: How humans interpret topic models. In *Advances in neural information processing systems*, pages 288–296, 2009.

Z. Chen, A. Mukherjee, B. Liu, M. Hsu, M. Castellanos, and R. Ghosh. Leveraging multi-domain prior knowledge in topic models. In *Proceedings of the Twenty-Third international joint conference on Artificial Intelligence*, pages 2071–2077. AAAI Press, 2013.

A.-S. Garcia-Moya.L and Berlanga-Llavori.R. Retrieving product features and opinions from customer reviews. *Intelligent Systems*, 28(3):19–27, 2013.

T. L. Griffiths and M. Steyvers. Finding scientific topics. *Proceedings of the National academy of Sciences of the United States of America*, 101(Suppl 1):5228–5235, 2004.

J. H. Lau, D. Newman, S. Karimi, and T. Baldwin. Best topic word selection for topic labelling. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 605–613. Association for Computational Linguistics, 2010.

D. Newman, S. Karimi, L. Cavedon, J. Kay, P. Thomas, and A. Trotman. External evaluation of topic models. In *Australasian Document Computing Symposium (ADCS)*, pages 1–8. School of Information Technologies, University of Sydney, 2009.

D. Newman, J. H. Lau, K. Grieser, and T. Baldwin. Automatic evaluation of topic coherence. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 100–108. Association for Computational Linguistics, 2010.

D. Newman, E. V. Bonilla, and W. Buntine. Improving topic coherence with regularized topic models. In *Advances in Neural Information Processing Systems*, pages 496–504, 2011.