Abstract

Online forums provide a wealth of information for users and thus is an important platform to discuss topics in groups. Processing the knowledge and understanding behaviour extracted from these user posts makes it easier to provide further insights to the participants of the forums. Forums for healthcare are getting popular within users seeking treatment methods and discussing the health conditions. We look at a case of Lyme Disease patients from an online group and try to solve one of the current researched topic of finding effective treatment methods for Lyme Disease from the questions and answers discussed by the patients, doctors and experts. Our analysis shows trends in between social and linguistic data over a time window. Our methods can be applied to any contextual information found in online forums and finding latent topics from it.

1 Introduction

The amount of information generated by online forums is immense but it’s not quite the knowledge we need. Seeing the potential of processing knowledge from these forums we suggest a machine learning tool to extract the linguistic features of the text and determine useful information from it. We would present some analysis and visualizations to gain further insights about the information in these forums.

Lyme Disease is a tick-borne disease which spreads through bite of infectious ticks to human. The information related to the treatment of this disease is sparse in comparison to some other widely known diseases. This disease though treatable is fatal if left neglected or misdiagnosed due to similar symptoms seen in other disease. A large number of patients reported that the diagnosis was complicated and had a lag of atleast 5 years in treatment between onset and diagnosis[1]. There are specific online groups initiated for the same reason of spreading some information and discussing the disease. Our study aims to help such groups by extracting knowledge through the online forums in order to make posts containing useful treatment information more accessible to users new to
the disease. Mining for relevant information through user posted data is particularly difficult since the data present online is often inaccurate, incomplete, controversial and misleading [2, 3]. In the sections below, we discuss some related work on the topics of Lyme disease and mining online forums, present the data at hand, explain our implementation plan and finally concluding the study with some visualization on the knowledge obtained.

2 Related Work

The research by Mankoff et al.[1] focuses on comparing the views posted by users about their health problems online and gain knowledge on how these users make use of this information and help others. Our work is an extension of this study which has already surveyed and studied patients diagnosed with Lyme Disease and their participation in online health forums. The information generated online can have a major impact on the patients’ actions[4, 5] and hence studied in order to provide a better tool for the users to access relevant information. A lot of interesting insights were gathered by multiple surveys, interviews, feedback and analysis of health related posts online. From their study they discovered the presence of two prominent models of Lyme Disease present throughout the online resources. First, the dominant model which goes by the belief that Lyme Disease is not present outside of the US-Northeast along with the some known set of symptoms and treatment method. The other being the minority model which had some conflicting ideas about the treatment methods for Lyme Disease but a lot of online content was referred to it. The presence of such conflicting ideas is another issue in the lack of proper treatment methods for this disease.

The study by Feldman et al.[6] is about discovering a text mining methodology to extract Adverse Drug Reactions(ADR) from online medical forums which were initially not discovered during the clinical trials of a new drug. The study not just validates the ADRs found in the text from the forum with the one found much later by the FDA but also aims at predicting any unreported ADRs through machine learning. The approach uses an Unsupervised Relation Extraction framework which includes tools for working with natural language in order to extract information and relation from it. The tool uses a grammar parser to work with English language. This framework has its limitations in producing high accuracy but with a low recall. This is rectified by a suggested additional processing step in the flow of the approach. The overall flow includes cleaning the extracted data, acquiring relation pattern of drug-symptoms using the URE framework while manually removing some irrelevant ones, extracting entities and the relations from the URE tool following with a post processing to find additional relations missed by the URE due to it being in a separate sentences or complex verb. This study is very relevant to the goal of our project to find self-treatments for Lyme disease in online forums. A similar entity(Lyme Disease) and a relation(Disease-treatment) extractions. The paper tries to compare different existing solutions as well, which used principles of co-occurrence of drug-ADR terms[7], supervised learning algorithms like Naïve Bayes and SVM
to differentiate true ADRs to noise[8] and lastly some unsupervised learning studies[9, 10] similar to the one in the paper.

The research by Bahja et al.[11] analyzed is a review and conformance of methods useful in identifying patient feedback text from online forums. The research works on patient data obtained from the National Health Services(NHS) website in UK which is a portal for patients to provide their feedback and ratings of various NHS clinics. The method proposes a sentiment analysis approach to provide a better understanding of the patient’s overall experience of the treatment received, the facilities provided and any limitations if any. The flow paper starts with describing the data at hand which is not exactly a QA forum like the previous study but a text review of a medical service received. They follow the sentiment analysis with topic modelling to further categorize the text into abstract topics. They test the sentiment analysis approach using various classifiers like Naïve Bayes and SVM, which later proves to be efficient in classifying a patient feedback to negative and positive very accurately. This approach may not be very useful in the Lyme disease dataset as the syntax of the text is far more randomized compared to providing only negative and positive feedbacks. Yet, we are more interested to look into topic modelling, the next step in this study which uses Latent Dirichlet Allocation (LDA) to further identify the recurring topics in the feedback. This method uses unigram and bigram modelling to find all such topics in the text and then align the results with that of sentiment analysis. This allows the scores obtained via sentiment analysis to be labelled to the departments identified using topic modelling approach and finds the least and most rated departments in the clinics being reviewed. The study uses single fold cross validation and multifold validation to evaluate the performance of the system. This study in particular can be used to label the dataset in the Lyme disease research.

MacLean et al.[12] tries to identify the different phases of withdrawal, recovery and relapse in opioid drug abuse users in online forums like Forum77 and Medhelp.org. Using Prochashka’s Transtheoretical Model (TTM) they track the changes in behavior of the patients prescribed for opioid using their discussions over time starting from the time they start using the forums till either relapse or recovery. The data is in the form of posts (740,045) and discussions (80,582) from several users (51,152) from various online communities focused on substance abuse. The proposed approach uses TTM develop the taxonomy of the changes in behavior and mapping them to the phases of addictions. Some of these taxonomy is used to manually label the posts which are then used to train and label the entire set using Conditional Random Field(CRF) modeling with accuracy of 67.6%. The challenges faced while doing this is the initial labelling of the phases as rubric which was developed using domain specialists. Another goal of this study is to identify features for classification. The linguistic features and words used by the users were then analyzed using Linguistic Inquiry and Word Count (LIWC) which generated 80 LIWC variables. Additional features are obtained using some more metrics and then used by a CRF model to label all the posts. Finally, the users are classified as weather they succeed(recover) or fail(relapse) the process of addiction. The study is an important one which
can be related the use case of identifying treatment methods in Lyme Disease patients as the structure of complexity of the text being analyzed is similar and the number of features obtained are large.

Another novel study by Jha et al. [13] is a cancer stage prediction from text in online forums. The researchers work with breast cancer forums and try to predict the stage the cancer is in from the patient’s health discourse online. The method works by classifying the four stages in cancer (I - IV) using the text posted by the patients while maintaining ethics in handling the data. They also prove that by analyzing the social network of these patients, their stages in cancer can be predicted. They use three different approaches here: Text based stage prediction to classify the stages using the text, network based stage predictions to use the social media of the patients and a hybrid approach to improve their process. For the text-based approach, they pre-process the posts by removing any patient signature (also any user defined stage information). The SVM classifier in weka works only on unigrams and bigrams but was efficient in classifying the four stages. While the network based approach works on the user’s interactions with others and analyzing the edges which are within 5 posts of each other. A cosine similarity is calculated for all such edges. They seem to have generated a valid classification and prove that online discourse by patients can reveal their stages in breast cancer and can be expanded to other studies as well.

3 Data

Approximately 144000 posts have been gathered over the period of 8 years (2001 - 2009) from http://www.lymenet.org/ containing posts from a number of threads. There are about 5886 users including domain experts and patients who participated in this forum. The structure of the data obtained was a SQL dump. Various pre-processing steps had to be performed in order to use the data for the analysis some of which were:

- Converting to CSV and JSON
- Converting dates to a proper format
- Extracting user list from posts
- Cleaning posts to remove blank lines and HTML style tags
- Using Networkx to create social network graph

4 Methods

We perform a number of exploratory data analysis on the cleaned data but with further modifications as required for the process. We try to obtain as much information from the data as possible during experimentation and hence went
through with the following methodologies. Code for all phases of our research is available at https://github.com/harsh1592/LymeDisease.

4.1 LDA Topic Modelling

We ran LDA (Latent Dirichlet allocation) topic modelling to find the estimated number of topics in the corpus and then to report on the most relevant topics. We ran LDA on a random 80-20 percent split over 20 iterations to train the model. This was done for a range of topics from 5 to 20. In-order to determine the best number of topics we graphed the per-word perplexity against the number of topics and the result in fig 1 shows a change at 10 topics.

![Figure 1: Per-word perplexity graphed against the number of topics](image)

The 10 topics found are as listed below,

- The D may vitamin use take body A pain day time one system 2
- test would know get one like use 'm think go need post time It see
- get Lyme insure go LLMD know one test like need think year would
- The use also help It cell may one http mani disease cause body
- get take 'm like think know feel help time day would realli 've symptom It
- get know go time year one would like lyme back thing think day
- Lyme disease patient The tick test infect treatment symptom cause treat may Disease
- The Happy know It one use help like said blood think would Birthday post cell
- use day get like good It would go know take one help Thank
- http Lyme lyme read link post help disease test group doctor get state

4.2 Semantic Analyzer of Frame Representation

We performed frame analysis on our dataset using SEMAFOR [14], a python based toolkit to effectively analyze the frames present in a text. The downside of using SEMAFOR was the slow performance on large text blob. Some of the posts in our dataset contains enormously large posts, the distribution of which is shown in Figure 2 on page 6.
We found the majority 97% of the posts had words between 0 - 6000 and see a break in the trend around the 6500 word count. Post/comment containing more than this number was either found to be spam or an information extracted from a web resource. We concluded such information does not provide a treatment information and rarely would talk about personal experience. Hence, as a solution we divide our dataset into posts containing less than 6000 and the rest containing the remainder. On performing the SEMAFOR we obtain the frames which were then merged with the original dataset in order to perform further analysis. The frames to match our obtained dataset are yet to be finalized.

4.3 Social Network Analysis

We perform some time series analysis to find how our dataset change over time. Figure 3 on page 7 shows the number of threads started and the count of posts made over the period of 8 years. For each sliding 6-week window we construct a graph over all the forum users where there is an edge between each pair of users where one of them comments on a thread made the other. Figure 3 on page 7 shows the graph statistics of the social network.

On each of these graphs we measure stats similar to those described in [15]:

1. Diameter
2. **Average shortest path length**: An alternative to diameter.
3. Largest connected component (lcc)
4. Number of connected component (ncc)
5. Average degree
6. Number of triangles on graph
7. **Average clustering coefficient**: for a given node x, the clustering coefficient is the proportion of pairs of x’s friends who themselves are friends.
8. **Average core number**: The k-core of graph is a maximal subgraph in which each vertex has at least degree k. The core number of a vertex is k if it belongs to the k-core but not to the (k + 1)-core. It is as a computationally tractable alternative to the clique number.
9. **Closeness centrality**: for a node u this is defined the reciprocal of the sum of the shortest path distances from u to all other nodes in the graph, times the length of the longest possible path on all graph of the same size.
10. **Core number centralization:** The sum of differences in coreness between each node and the node of maximum coreness in the graph, divided by the sum of differences in coreness between each node and the maximum coreness possible over all graphs of the same size.

### 4.4 Linguistic Analysis

We start off with generating a simple word cloud representation of all the posts. This visualization shows us the most important as well as frequently used words in all the forum data. We keep the time window similar to the one used throughout the analysis. We performed stopwords removal before generating the word cloud. The output in figure 4 on page 8 shows the final visualization. Looking at the first and last shows a lot of variation.

### 4.5 LIWC

Next, we perform LIWC analysis. Since an online forum acts very similar to a network of users, a graph representation of the data is the most suitable for such analysis. We use Networkx library for Python to create these graphs and perform time series analysis on a sliding window. In order to create a post-comment structure we use the first post in each thread as a starting post while the rest as the comments for the initiator post. The data spans from the year 2001 to 2008 and is divided into 65 windows, each sliding by 6 weeks from the...
Figure 4: Word cloud visualization of each year starting from 2001 till 2009

Figure 5: LIWC scores of all posts.
previous. Figure 5 to 6 on page 8 shows the LIWC scores varying over time in the forum. While Figures 15 to 18 are LIWC scores of the comments that users have made over time.

![Figure 6: LIWC scores of all posts.](image)

We made use of the Python multi-threading to improve the performance of getting LIWC scores by utilizing all the CPU cores available and dropping the posts containing more than 6500 words as explained earlier. The improvement in processing was increased by 200%.

### 4.6 Geo-locating posts

A lot of the user posts dealt with location information in order to locate a Doctor or report on some personal cases. This provides an opportunity to map these locations to see visual trends and more importantly to help aid some of the threads in the forum related to finding Lyme Physicians. We have designed a solution which will extract location entities from a text and generate JSON containing the date and locations obtained. Currently, we map these locations obtained through the time window to see the spread of the disease and the interactions of users over various places in the United States.
5 Conclusion and Future Work

From our analysis of the social network, we can conclude it went through three phases of development. Starting from a very small user base in early 2001 the network shows a large number of connected components and a drop in the triangles and cluster coefficient. The large number of connected components
in the early days suggest a small community of people starting to know each other, which then dropped significantly as the group grew in size until the year 2007. The number of user participation was constant there after and hence the rise of the largest connected components again. Similarly, the low triangles and cluster coefficient suggest a similar behaviour due to the members not knowing each other.

In general we see the trends in online Lyme Disease forum which were expected of a small community in the start. The LIWC scores in both the appendix section and the LIWC section tends to show an increase in the health related topics in the community. The increase in positive emotions shows the good health of communication between the members and strong positive belief.

The word cloud representation shows the words being used through the time window and a majority of them related to Lyme Diesase and the medication names. This visualization guides the future work to the discovering of a particular post in a given year. This will help in finding treatment related information and medications used in the forum.

The geo-mapping of the user post was a very interesting visualization to perform on our dataset as it will lead to a very sound base in order to analyze one of the topics board related to finding Doctors. Also, geo-mapping of the post shows a very similar pattern of spread of the post and the reported cases of Lyme in the CDC research.

Our study will help the community to perform similar analysis on such large
dataset in a very short time. Reducing the amount of time taken to analyze a large text and annotating it with LIWC scores at the same time helped us to focus on the problem at hand and carry more experiments. The efficient multi-processing Python tool we built can not only be used in the future study of the same forum but in any similar NLP problem.

References


[10] Xiao Liu and Hsinchun Chen. Azdrugminer: an information extraction system for mining patient-reported adverse drug events in online patient


## A

### Work Progress

<table>
<thead>
<tr>
<th>Week</th>
<th>Work Completed</th>
<th>Hours Taken</th>
</tr>
</thead>
</table>
| Week 1 | 1. Downloaded and installed Anaconda python  
2. Checked out any necessary code.  
3. Constructed github repo. | 14 |
| Week 2 | Background reading, begin coding, data analysis. | 12 |
| Week 3 | Further reading and coding.  
Inspected for signs of self-treatment / rumors of therapies (i.e., keywords) | 13 |
| Week 4 | Further reading and coding. | 10 |
| **Milestone 1** | Basic background reading finished.  
Development platform secured. Methods, tools established and installed. | | 
| Week 5 | Ran time-series-based analysis of LIWC for a variety of time windows | 18 |
| Week 6 | 1. 15 citations in bib file  
2. Wrote one paragraph summary of at least six papers.  
3. Wrote abstract, introduction | 17 |
| Week 7 | 1. Inserted results in paper (in results section).  
2. Inserted data dictionary for sql data  
3. Github repo in report. | 17 |
| Week 8 | Continued to try and get SEMAFOR up | 20 |
| **Milestone 2** | System mostly built and tested, Methods section of paper written. | | 
| Week 9 | 1. Discuss high-wordcount comments (to go w/ wordcount histogram and what they look like. Present a couple of examples  
2. Inserted several of the interesting time series windows into your report.  
3. Documented graph windowing process in methods (look to eNABLE paper for reference). | 25 |
| Week 10 | 1. Generated one tfidf wordcloud per year of data.  
2. Thought about building a classifier that can tell which year a comment was made.  
3. Read 2 papers by Mankoff | 23 |
| Week 11 | 1. Replicated all results on data from 2006 forward only.  
2. Made wordclouds black-on-white. | 24 |
| Week 12 | 1. Created charts for:  
a. ALL forums for all data  
b. ALL forums for last two years  
c. For each forum:  
i. Last two years | 18 |
| **Milestone 3** | Results section written | | 
| Week 13 | First draft of poster and final paper. | 15 |
| Week 14 | Second draft of poster and final paper. | 15 |
| Week 15 | Third draft of poster and final paper. | 18 |
Social Network Analysis (2001-2009)

Below are some more results from gathered from the analysis of the social network considering the time window of 2001-2009

![Figure 10: LIWC scores of all posts for past 8 years](image-url)
LIWC scores (2001-2009)

Below are some of the LIWC scores for an extended time window of 2001-2009

Figure 11: LIWC scores of all posts for past 8 years
Figure 12: LIWC scores of all posts for past 8 years
Figure 13: LIWC scores of all posts for past 8 years
Figure 14: LIWC scores of all posts for past 8 years
Figure 15: LIWC scores of all comments for past 8 years.
Figure 16: LIWC scores of all comments for past 8 years.
Figure 17: LIWC scores of all comments for past 8 years.
Figure 18: LIWC scores of all comments for past 8 years.