Analyzing Lyme Disease Help-Seeking on Social Media

Ankita Kamath
Department of Computer Science
Golisano College of Computing and Information Sciences
Rochester Institute of Technology
Rochester, NY 14586
ark4695@g.rit.edu

Abstract—Lyme disease is an infectious disease spread through the bite of the ticks. Recently, has been spread widely here in the United States. This disease is relatively new and can cause long-term health issues. With the growth of the internet technologies and social media, chronic diseases such as Lyme disease is openly discussed on various online forums. This project involves extracting information from some of these online forums like 'Seeking a Doctor', 'General Support' and 'Medical Questions. We perform statistical analysis on the results obtained from Semafor, a frame extraction tool [1]. We developed a narrative frame schema to understand the relationship between different posts. We use machine learning algorithm like LSTM (Long Short-Term Memory model) to categorize the posts into different semantic frames based on context. The extracted information will be used to understand the details of Lyme disease, like symptoms, different stages of treatment, various medicines used by the members of the online forums.

Keywords—Lyme disease; narrative extraction; semantic frames; social media; long short term memory model; frame extraction; confusion matrix;

I. INTRODUCTION

With the growth of internet technologies, social media has been tremendously used for sharing information. Chronic diseases are openly discussed on social media. Users can anonymously share thoughts about their health condition on online forums. Eight out of ten internet users have used online forums to get health information [2]. Many users have admitted having an impact from the contents of social media on how they care about themselves. Thus, social media plays an important role in how a user understands his illness. The users tend to refer online forums for information on lifestyle changes, treatment regime and its effects [3]. Due to this impact of social media on the users, even the health researchers have turned their attention towards the online forums. They are eager to extract and study relevant information shared by the patients on such platforms. They aim to use this platform to spread awareness among the users and to provide better facilities to the patients.

Lyme disease is a tick-borne infectious disease caused by the spirochetal bacteria called Borrelia. It is transmitted to the humans through the bite of the deer tick. A blood test can be used for the detection of antibodies for the spirochetes. The common symptoms of Lyme disease include fever, fatigue, headache and a distinct bulls eye rash. The infections may spread to joints, heart and may corrupt the nervous system if it is not diagnosed and treated in time [4]. It is very difficult to diagnose the disease as similar symptoms can be seen in other diseases. There is a lag of approximately five years from the start and the diagnosis of the disease. Approximately 30,000 cases have been reported to Centre for Disease Control and Prevention by the state health department and the number is increasing every year [4]. There are numerous online forums set up to spread and share knowledge about the Lyme disease. The information shared on these forums may be inaccurate, vague, irrelevant, or ambiguous, and thus, the information may have a negative impact on the users. Hence, there is a need to study the information shared on these forums. We aim to study and analyze these posts and perform mining for relevant information using machine learning algorithms.

We have divided this project into different milestones:

• Understanding the data and developing the narrative frame schema.
• Data preparation.
• Implementation of an LSTM (Long short-term memory) model.
• Evaluation of the model.

Development of narrative frame schema involves understanding the data and developing a process model to identify related semantic frames. A semantic frame is a structure of related terms and concepts. Data preparation involves labeling the posts with the associated semantic frames. The next step is implementing an LSTM model which classifies the posts obtained from online forums into different frames. An LSTM model is an advanced form of recurrent neural networks that excels in learning sequential data. The model will be evaluated by precision and recall. The last step is to perform statistical analysis on the extracted frame by running the Semafor, frame extraction tool on the entire dataset of posts.

The paper has been organized as follows: Section II discusses the previous work on the analysis of the information on social media and Lyme disease. Section III focuses on the implementation that we have performed to get the desired results. Section IV describes the results and analysis. Section V and section VI focuses on the conclusion and future work.
II. RELATED WORK

Diomaiuta et al. [5] proposes an automated system which collects and summarizes the patient's medical history reports. The main components of the system are extraction subsystem and summarization subsystem. Since the medical records are found in both structured computer-based and unstructured narrative text formats, the extraction subsystem consists of Natural Language Processing (NLP) module and query module. The NLP module performs language identification, tokenization, lexical analysis and medical keyword recognition to extract the medical keywords from the unstructured narrative texts. The query module submits the query to get the medical keywords from the structured records. The summarization subsystem performs match and project operation to select the relevant medical attributes obtained from the extraction subsystem. It then performs group operation to aggregate the medical details based on type and keyword label. Thus, this system provides a visual summary of the medical details of the patient to the clinical practitioner. It summarizes the medical problems, prescribed medicines and various treatments given to the patient. This system proves to be efficient for the quick analysis of a large number of medical records stored in different sources.

Hermann et al. [6] introduces a technique for semantic frame identification using the distributed representation of the words. The author describes semantic frame parsing which breaks down the semantic context of a sentence into frames and analyzes the arguments that fulfill the frame's semantic role. The semantic frame parsing consists of frame identification and argument identification. The frame identification involves extracting all the syntactic dependents of the predicate and mapping them in a vector space and learning the embeddings of each possible frame label. The author describes the embedding process of frame identification by context representation extraction which is tested on two variants i.e., direct dependents and dependency path and learning which helps in speeding the process by learning only the embeddings for the frame labels. The author further describes the argument identification model used in the semantic frame parsing. Thus, the author proposes a technique which outperforms the previous FrameNet style semantic frame parsing and delivers similar results to one of the best parser PropBank SRL.

Yang et al. [7] proposes and implements a semi-automated method for collecting and analyzing health-related information in social media. The method consists of collecting, filtering, analyzing and visualizing the data extracted from the social media. The collection method involves extracting health-related information from various sources like weblogs, microblogs, wiki and other online forums. While extracting this information, the authors found that the weblogs and online forums focused more on disease symptoms and treatments whereas the wiki and encyclopedia provide information about the procedures. The collection method also involves automatic searching and harvesting information. The filtering method involves analyzing the collected information and removing the irrelevant information. Filtering can be automatic or manual and it requires linguistic knowledge and domain knowledge regarding the medical entities. The analysis method involves identifying trends in the information by using classification. The visualization presents the final data in a format which is easily understandable by the human eye.

Akimoto et al. [8] proposes a narrative generation system which is a computer program generating narrative texts. These narrative texts are derived from human and machine creativity. The authors use two concepts for building this system - Structural techniques for generating and transforming narratives and generation control mechanism. The architecture of the system consists of three different phases: Story phase, discourse phase, and expression phase. The story phase is the structure which includes the event as the main element and it is described by the conceptual representation. The Discourse phase is how to form a narrative from a sequence of events. The Expression phase is the surface representation using natural language. The authors explain the different techniques for story and discourse. They explain generation control mechanism using the hierarchical parameters. The authors explain the importance of narrative extraction and how their technique was useful in generating the narratives.

Kitron et al. [9] performs the spatial analysis of Lyme disease in Wisconsin. Human cases of Lyme disease were associated with the tick distribution and vegetation coverage by surveillance measures. Tick Distribution was determined by the Wisconsin Division of Health based on the observations by researchers and the information submitted by the public. The author used normalized difference vegetation index (NDVI) to calculate the degree of greenness, forested areas, and the tick spread during different times of the year. It was calculated using the near-infrared and visible wavelength values. Spearman rank correlation was used to measure the relation between county-level NDVI values, tick distribution, human population density, number of cases by county of exposure and residence, and incidence rates. The spatial autocorrelation showed significant spatial clustering of areas with regards to the tick distribution.

Smith et al. [1] implements a frame extraction tool which extracts the semantic frames from an English text. This tool finds the part of the text which identifies the semantic frames and labels the frame's argument. The input to this tool is a simple English text and output is in the form of JSON file. The process of this automatic analysis consists of preprocessing, target identification, frame identification, argument identification and output. The preprocessing includes lemmatization, tagging the part-of-speech and syntactical parsing using the mJavaParser. The target identification consists of heuristically identifying the frame evoking words and phrases from the input text. Frame identification involves training the log-linear model on FrameNet data with the full text frame annotations. The model produces a probability distribution over identified frames and selects the highest scoring frame. Argument identification involves training the same FrameNet data. The model identifies the group of words that represents the role of each labelled frame instance. The output is in the JSON file consists of text of the input sentences and the semantic frame information predicted by the tool.

III. METHOD

A. Data

The dataset used for analysis of LymeDisease is collected from http://flash.lymenet.org/scripts/ultimatebb.cgi. This
dataset is a collection of various online forums such as Seeking a Doctor, General Support, and Medical Questions, as shown in figure 1.

Fig. 1. Lymenet flash website.

The data extracted from these online forums have been stored on cmhost2.cs.rit.edu in the JSON (JavaScript Object Notation) format. The json structure of the data can be seen in figure 2.

{ "_id":"$oid":"59c1e7b4cc6b5c0b13590c79"}, "text":"helpplease1, Make a new post asking for a doc in the Bay Area. Or you can private message Robin123 as she might not come back to this thread. To start a new thread/post, you will find the button at the top of the page right under the intro info and right above the profile/search/ec area. It is called Post new topic in sort of 3D font. To send Robin123 a pm, click on the envelop with two people at the top of her post. Good luck"}, "userID":"debilyn"}, "postOrComment":"Comment"}, "dateTime":{"$date":"2012-09-05T18:04:00.000Z"}, "topic":"Doctor in California-LA/Bay Area"

Fig. 2. Extracted posts in JSON format.

B. Data Visualization

In order to understand our dataset better, we generated plots showing the distribution of comments and posts among the users as shown in figure [3][4] using the Matplotlib, a python library for plotting graphs and histogram. This distribution helped us to identify the admin users. Most of the posts from admin users were welcome messages which were not relevant to our analysis. The identified admin users are as follows: " testify", "Lymetoo", "Siciliano", "hopingandpraying", "Keebler", "Tincup", "TF", "Abxnomore", "mbroderick". Figures 5 and 6 shows the distribution of posts and comments per top 10 users excluding the admin users.

C. Approach

We are using GitHub repository [10] to store all the scripts related to LSTM model and bitbucket repository [11] to store the scripts for running semafor over our dataset. The project architecture can be seen in figure 7. In order to analyze the posts based on context, semantic frames from each post were identified. A semantic frame is a collection of related terms or concepts. It is not possible to understand any term without understanding all the related terms. The semantic frames were identified using manual inspection and reference to semantic FrameNet developed by Computer Science Department of UC Berkeley [12]. The details of the identified frames is shown in figure 8. We analyze the semantic frames obtained from Semafor frame extraction tool [1] as shown in 9. Figure 10 shows the mapping of frames detected from Semafor and frames identified by manual inspection of posts.
We have developed a model for categorizing the post based on the identified frames. The frames act as a label for classification. We have used kera sequential model and encoder-decoder sequence to sequence LSTM model for classification.

1) Narrative Frame Extraction: We identified 11 semantic frames based on the topics discussed on online forums. These frames can be used to establish a narrative. We develop a narrative schema based on the identified frames as shown in figure 11. The narratives help us to obtain information about the relationship between different frames.

2) Labelling: The first step of the classification model is to obtain a labelled dataset to train, test and validate the model. The dataset consists of posts and comments which is extracted from the data stored at cmhost2 server in JSON format. The dataset is stored in a .csv file. The dataset of 2000 extracted text was labelled manually as '0' and '1', where '0' represents the post cannot be categorized as a given frame and '1' represents that the post can be categorized as a given frame. The posts can be categorized into multiple frames.

3) Pre-processing: We split the training and testing data in the ratio of 8:2 and store the training and testing data in train.csv and test.csv respectively. The labels are encoded as a one hot vector with using keras multilabelbinalizer library. The posts are tokenized and each unique word is stored as an integer. The sentence is represented by a sequence of integers. The integer representation of the tokens is fed into the embedding layer of both models which converts it to vector representation of words. This vector representation is then fed to the LSTM layer of both sequential model and encoder-decoder model.

4) Classification model: We have used the LSTM model for classification of posts into different frames. LSTM is widely used for text processing and is designed to learn sequential data. It is an advanced form of the recurrent neural network which overcomes the vanishing gradient problem of recurrent neural network. The LSTM consists of three gated layers namely forget gate layer, input gate layer and output gate layer. The forget gate layer uses a sigmoid function which decides what old information can be retained or removed. The input gate layer consists of the sigmoid function to decide what new information can be added to the cell state and tanh function which creates a new vector of candidate values for new cell state. The output gate layer uses the sigmoid function to decide what new information can be added to the cell state and tanh function which creates a new vector of candidate values for new cell state. The output gate layer uses the sigmoid function to decide what the output is going to be. Thus, LSTM model uses memory efficiently by passing only the relevant information to the next cell state, thereby reducing the error and increasing the learning. The LSTM model outputs the sequence of identified frames.
cross-entropy and the optimizer used is RMSProp.

Non-Core Elements: user’s location, user’s location, user’s location.

Name: Medical test frame
Define: Identify the posts where the users talk about insurance issues
Core Elements: user, status of insurance
Non-Core Elements: user’s location, Reason for_not_giving_insurance

Name: Exercise frame
Define: Identify the posts where the users talk about the exercises they are doing to cope with the disease.
Core Elements: user, type of exercise
Non-core elements: user’s location, effects of exercise.

Name: Ask for Advice frame
Define: Identify the posts where the users are asking for advice from other users on medicines, disease.
Core Elements: user, topic of advice.
Non-core elements: user’s location, advice given.

a) Sequential Model: The sequential model is a stack of neural network layers. The model consists of LSTM layer with 64 neurons which is the core layer for prediction, the dense layer feeds all outputs from the LSTM layer to 256 neurons and each neuron provides one output to the next layer. The next layer is ‘relu’ (Rectified linear unit) which is used to speed up the process of training. The dropout layer is used as a regularization measure in order to avoid overfitting. The sigmoid layer is used to obtain the output probabilities in (0,1) range. The loss function and optimizer used for learning are binary_cross-entropy and RMSProp respectively.

b) Encoder-Decoder sequence to sequence model: The Encoder-decoder model is generally used for language translation. The Encoder model as shown in figure 12 consists of LSTM cell which reads in the vector representation of the words and predicts the output. We discard the output and store only the encoder states. These states are then fed into the decoder model which runs for one time-step. The states from the encoder model and labels helps the decoder model to predict the classes. The loss function used for learning is binary_cross-entropy and the optimizer used is RMSProp with a learning rate of 0.0001. The predicted outputs are in probabilities, we select only the classes with the probabilities more than the threshold value.

c) Evaluation: The precision, recall and F1 score are used to evaluate the classification model. The true positives, true negatives, false positives and false negatives are used for the calculation of precision, recall and F1-score.

\[
Precision = \frac{true\_positive}{true\_positive + false\_positive}
\]

\[
Recall = \frac{true\_positive}{true\_positive + false\_negative}
\]

\[
F1\_score = \frac{2 \times precision \times recall}{precision + recall}
\]

The model predicts the probabilities for each frame given an input sequence. We select the frames with probabilities greater than the threshold value.
IV. RESULTS

The precision and recall obtained from Encoder-decoder model is very low as compared to Sequential model as shown in figures 13 and 14. The Encoder-decoder model is a complex model which is used for language translation. The expected output of the encoder-decoder model is a sequence but in our case it is a set of classes. We experimented with this model to see if the model can be used for classification. We may improve the results of encoder-decoder model by increasing the dataset size, balancing the dataset to include equal number of data points for each class.

<table>
<thead>
<tr>
<th>Frames</th>
<th>Precision</th>
<th>Recall</th>
<th>F1-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Seek</td>
<td>17.08</td>
<td>23.45</td>
<td>19.76</td>
</tr>
<tr>
<td>Medical_condition</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>Medical_test</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>Medication</td>
<td>17.33</td>
<td>40.56</td>
<td>24.28</td>
</tr>
<tr>
<td>Insurance</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>Diet</td>
<td>23.46</td>
<td>45.87</td>
<td>31.04</td>
</tr>
<tr>
<td>Exercise</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>Ask_for_advice</td>
<td>35.07</td>
<td>21.65</td>
<td>26.77</td>
</tr>
<tr>
<td>Other</td>
<td>10.96</td>
<td>19.42</td>
<td>14.01</td>
</tr>
</tbody>
</table>

Fig. 14. Results of Encoder-decoder model.

online forums with length less than 1000. The results can be seen in figure 15.

The results seem to be biased towards certain frames like medical_test and seek etc since we have more samples corresponding to those frames in the dataset.

Fig. 15. Posts distributed with respect to frames.

The results may change everytime we run the model, as the model is trained randomly on any samples. There may be cases when the model may not be trained for a particular label due to the size of the dataset.

V. CONCLUSION

From the results and analysis, we can conclude that the semantic frames that we have identified from the extracted posts, obtained from online forums, helped us to categorize the posts into different context based frames. The histogram plots of various frames helped us to determine important topics discussed.

VI. FUTURE WORK

The Frames may be efficiently predicted by increasing the dataset. So, the next step would be to update the dataset and make it more balanced. The relationship between different identified frames can be identified to establish narratives. The Frame Extraction tools like Semafor and Sling can be updated to incorporate our task of efficient frame detection.
ACKNOWLEDGMENT

I would like to thank Dr. Christopher Homan for providing me with an opportunity to work on this project and guiding me through this project, Dr. Minseok Kwon for his valuable feedback on the progress of each milestone. I would like to thank Computer Science Department of Rochester Institute of Technology for preparing me throughout my graduation by providing relevant coursework and resources.

REFERENCES