Visualizing Health Tweets over Regions and Timestamps

Bonpagna Kann[1], Sihem Amer-Yahia[2], Michael Ortega[3], Jean-Louis Pépin[4], Sébastien Bailly[5]

***Abstract*—Social media has become one of the major data sources for social studies through users' expressions, such as significant moments in their daily life or their feelings and perceptions toward specific discussion topics. In health care, social media is thoroughly used to study people's discourse on ailments and derive insights into the impact of ailments on patients' quality of life. Recently, there has been an increasing interest in applying machine learning algorithms to enhance the prediction of ailments through users' social media data. In this study, nearly 800 million posts were retrieved from Twitter through preprocessing and running the time-aware ailment topic aspect model (T-ATAM) to predict diseases, symptoms, and remedies for two chronic conditions, namely sleep apnea and chronic liver diseases. The study was conducted on English tweets emitted during 2018, most of which were from European countries and the United States. The data were processed using T-ATAM by regions, timestamps, and treatment, namely continuous positive airway pressure (CPAP), to see the differences in the distributions of top diseases along with the top symptoms and remedies in different regions; timestamps; as well as before, during, and after CPAP was introduced. Based on approximately 331,000 tweets related to liver diseases and 1 million tweets on sleep apnea, various visualizations of statistics are displayed, including world maps, word clouds, and histograms. Results of this study indicate that depression and drinking are the leading symptoms of liver diseases; meanwhile, lack of nighttime sleep and overworking are considered the main factors of sleep apnea.**

***Keywords*—Social Media, Topic Models, Visualization, Health Data, Data Analysis.**

I. INTRODUCTION

Social media plays a vital role in providing information about people in their daily life through the large-scale analysis of their posts, representing their activities and perceptions on various topics. Social media has been used in many studies for multiple purposes, including exploring trends and discussing global events such as political perspectives [1]-[3], daily news [4], [5], and natural disasters [6], [7]. Twitter, an interesting example of social media with millions of tweets posted daily, has become an essential source of data for transition detection and prediction [8], [9] in health care [10], [11]. A large body of work relies on tweets in investigating ailments, including cancer [12], [13] and infectious diseases [7], [14].

Even though each message on Twitter provides little value, the combination of all messages can be processed to generate valuable knowledge and derive essential insights. For instance, users of Twitter can express their situation by posting a message such as "I've just caught a cold today" or "I feel my back hurts. What should I do?". This information might not be compelling as it is mainstream for everyone; however, at an aggregate level, combining such posts can provide important statistical information and help monitor the symptoms of ailments in time and space. This practice has been implemented in the United Kingdom [15], [16], and the United States [17], [18]. Studying the symptoms of ailments in tweets can help measure an illness's risk factors and promote health campaigns in different regions and at different periods. In addition, studying treatments of ailments in tweets can help gauge the effectiveness of a given treatment, viz., continuous positive airway pressure (CPAP) leads to a change in discourse in tweets.

The use of social media in comprehending people's perspectives on their ailment conditions and detecting discussion topics and symptoms has been studied in numerous prior research. However, the study and visualization of tweets based on regions and timestamps remain uncovered, particularly in sleep apnea and chronic liver diseases. This paper implements a developed model, the time-aware ailment topic aspect model (T-ATAM) [8], to predict diseases, symptoms, and remedies of sleep apnea and chronic liver diseases by region and time.

II. METHODS

A. Dataset

A subset of tweets emitted in 2018 was extracted using the Twitter API. This study focuses on tweets from English countries, the majority of which are in the United States and European countries. The content includes the timestamp of the tweets along with their geographical attributes. In addition, health dataset information for liver diseases and sleep apnea, including diseases, symptoms, and remedies, were extracted from health care websites with a combination of manual and automatic methods using Beautiful Soup, a Python package for web-scraping and parsing HTML and XML documents in Python. Table I displays the number of diseases, symptoms, and remedies for chronic liver diseases and sleep apnea in the keyword dataset. The data used in this study were obtained from publicly accessible sources, namely Twitter API. The tweets and user data from the database were anonymized to ensure users' privacy.

B. Data Preprocessing

The collected tweets were input into the preprocessing steps shown in Fig. 1. The raw tweets were primarily extracted with

[1,2,3] *CNRS, Univ. Grenoble Alpes, Bâtiment IMAG, 700 Av. Centrale, 38401, Saint-Martin-d'Hères, Grenoble, France (email: [1]bonpagnakann@gmail.com, [2]sihem.Amer-Yahia@univ-grenoble-alpes.fr, [3]michael.ortega@imag.fr)*

[4,5] *INSERM, Univ. Grenoble Alpes, Av. des Maquis du Grésivaudan, 38700, La Tronche, Grenoble, France (email: [4]jpepin@chu-grenoble.fr, [5]sbailly@chu-grenoble.fr)*

[Received: 9 January 2022, Revised: 19 September 2022]

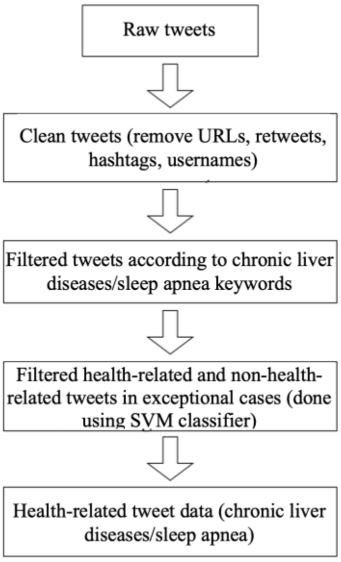


Fig. 1 Data preprocessing diagram.

TABLE I
NUMBER OF DISEASES, SYMPTOMS, AND REMEDIES IN THE KEYWORD DATASET

Ailment Type	Diseases	Symptoms	Remedies
Chronic liver diseases	61	169	220
Sleep apnea	15	175	117

the attributes of places, timestamps, and tweet texts and fed into another step to remove URLs, retweets, hashtags, and usernames. After the clean tweets were obtained, they were filtered by the chronic liver diseases and sleep apnea keywords, including diseases, symptoms, and treatments, to ensure that the tweets were specifically related to these two diseases. In addition, a list of stop words was also used to filter unnecessary words (viz., subjects, prepositions, articles, or general words) from the tweet data. The filtered tweet file was then joined with the clean health and non-health-related file and converted into a LIBSVM format. With the Octave support, classifying tweets based on health and non-health-related tweets was conducted using MATLAB. The final output of the preprocessing was the predicted label file, which was used with the tweet file to generate the final health file for the input to the T-ATAM model for both liver diseases and sleep apnea.

C. Data Processing on T-ATAM

1) Review of Time-Aware Ailment Topic Aspect Model (T-ATAM): In this study, the previously proposed time-aware model, T-ATAM [8], was implemented to predict health-related tweets, using time as a variable drawn from a multinominal distribution from the corpus. To begin, the data of mapping tweets to documents with different timestamps and geographical attributes were presented. According to the dataset, a set of posts $\mathcal{P} = \{\rho_1, \rho_2, ..., \rho_n\}$ is considered, where a post is the unit update of user activity on social media, such as Twitter. In a post, there are several attributes, including place identity, timestamp, and post content. Moreover, $\mathcal{G} = \{\mathfrak{g}_1, \mathfrak{g}_2, ..., \mathfrak{g}_n\}$ is considered as the geographical attributes for each post.

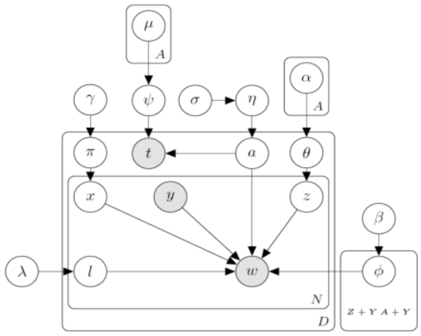


Fig. 2 Time-aware ailment topic aspect model (T-ATAM).

$\mathcal{P}_g$ is regarded as the group of posts originating from region $\mathfrak{g} \in \mathcal{G}$. Additionally, variable $\mathcal{T} = \{t_1, t_2, ..., t_n\}$ represents the temporal attributes to divide the posts to run T-ATAM and obtain the result by month. $\mathcal{P}_g^t$ refers to the set of posts in $\mathcal{P}$ originating from the region $\mathfrak{g}$ during period t. Finally, $\mathcal{D} = \{\mathcal{D}_g^{t_1}, \mathcal{D}_g^{t_2}, ..., \mathcal{D}_g^{t_n}\}$ is used to define all sets of the documents representing the aggregated tweets from region $\mathfrak{g}$ for different timestamps t_i. The description of the terms is described below.

$\mathcal{P}$	:	Posts
$\mathcal{G}$	:	Regions
$\mathcal{T}$	:	Timestamp
$\mathcal{P}_g^t$	:	Posts in region $\mathfrak{g}$ during timestamp t
$\mathcal{D}_g^t$	:	Documents mapping the posts $\rho \in \mathcal{P}_g^t$ to document.

T-ATAM is a time-aware model where the timestamp *t* of each post is treated as a random variable [4]. Before the tweet document was generated, those filtered health-related words associated to a unique disease, including liver cancer, central sleep apnea, or cirrhosis were generated word by word in the document. The word can also be the symptoms or remedies of the disease; back pain and surgery, for example, belong to the symptoms and remedies for liver cancer. During the process, two random variables ℓ and *x* were chosen to identify whether the word was a general-purpose word or background word. Then, the word could be drawn from the vocabulary distribution common to the whole corpus or generated from the distribution topic z. When the word was health-related (case ℓ = 1; x = 1), another random variable y allowed to choose whether this word was aspect-neutral (case y = 0), a symptom (case y = 1), or a treatment (case y = 2). Finally, words were inferred from the diseases associated with the document. However, it is noticeable that in T-ATAM, a timestamp is generated for each document depending on the ailment associated with the considered tweet obtained from the database. As timestamp is an observed random variable, generating one time-distribution per ailment $\{\psi, a \in \mathrm{A}\}$ to generate a timestamp for a given tweet is necessary. Subsequently, the timestamp according to a multinomial distribution with parameter ψ_a, which was obtained from the Dirichlet distribution defined by the vector μ specifically for each ailment, was generated depending on the ailment a connected to the document. The process of the model is illustrated in Fig. 2.

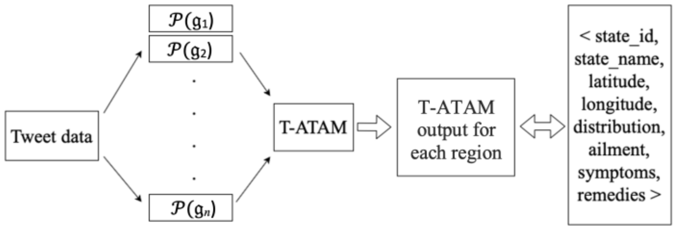


Fig. 3 Data processing on T-ATAM by regions.

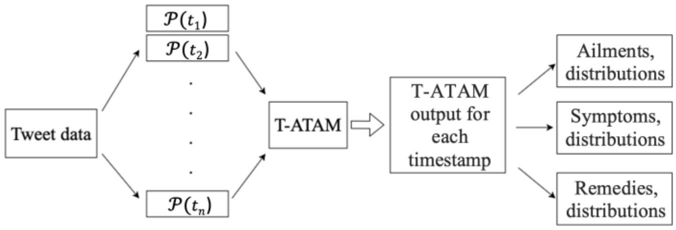


Fig. 4 Data processing on T-ATAM by timestamps.

2) Data Processing on T-ATAM by Regions: This part discusses the prediction of diseases, symptoms, and remedies in each region of the tweet dataset. First, the tweet posts ρ were grouped and divided by their region g before providing each data into T-ATAM to generate the distribution vectors of the diseases in each region. In this case, the top ailment in each region with the highest distribution was selected along with other information, including symptoms and remedies. The process is shown in Fig. 3.

3) Data Processing on T-ATAM by Timestamps: In this case, the changes in diseases, symptoms, and remedies in different periods are observed. The periods from January to December 2018 were set. The tweet posts ρ were grouped and separated according to their timestamp t. T-ATAM was run to generate diseases, symptoms, and remedies for each month. The summary of the process is displayed in Fig. 4.

4) Data Processing on T-ATAM by Treatments: One essential aspect of analyzing health-related tweets is the ability to answer questions related to what happened before and after a given treatment. For instance, the patients' condition after they receive the CPAP treatment. CPAP is a mode of respiratory ventilation used to treat sleep apnea. In this study, the result of the changes in the diseases, symptoms, and remedies was observed from before and after CPAP was introduced. To address those questions, the dataset was split into three main categories for each region: PRE (before CPAP is mentioned in the tweets), DURING (three months after CPAP is first mentioned), LONG-TERM (later than the three months from when CPAP is first mentioned). Then, T-ATAM was applied to obtain sets, as shown in Fig. 5.

D. Data Visualization

After running T-ATAM, the output data were used for visualization in different ways. Fig. 6 illustrates the visualization process of health-related tweets. The output data from the model were used for visualization. After the data had been filtered, the output files were produced with different regions and timestamps. The D3.js was used for various

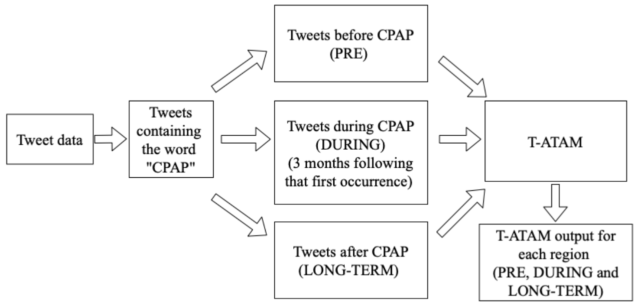


Fig. 5 Data processing on T-ATAM by treatments.

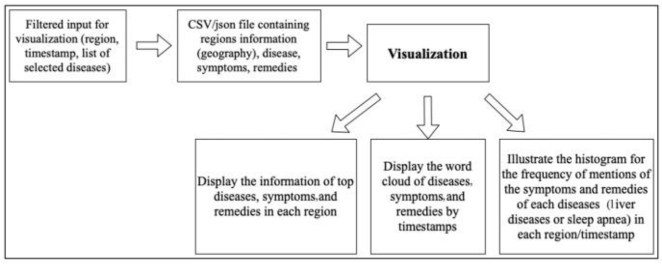


Fig. 6 Visualization process of health-related tweets.

purposes, such as displaying the ailment information in each region on the map, word clouds, and histograms. In order to have proper input data to implement the visualization, a form was created in the front end to ensure that the visualization worked on demand and more interactively. The constraints were divided into spatial, temporal, before/after treatment, and type of diseases requested. The different types of displays are described as follows.

1) Data Visualization on a Map: The output data from T-ATAM by region, including the top ailment with its distribution, symptoms, and remedies, are shown on a map.

2) Word Cloud: The output data by region, timestamp, and before/after treatment can be selected to generate a word cloud.

3) Histogram: The output data based on spatial and temporal granularities can be used to generate the histogram upon user selection, including diseases, symptoms, and remedies.

III. RESULTS

After preprocessing the tweet data, the health-related tweets were obtained with respect to chronic liver diseases and sleep apnea. According to the number of tweets in each step shown in Table II, nearly 800 million tweets were retrieved, with approximately 331,000, and 1 million health-related tweets were collected for liver diseases and sleep apnea, respectively. Additionally, after running T-ATAM and generating the output data, thirteen diseases were generated for liver diseases and eight for sleep apnea, as shown in Table III.

In the sleep apnea section, the study is focused on the use of key terms. It was found that users had different ways of expressing the word "snoring" and "CPAP" in their tweet posts; the number of mentions of these words was not significantly high since users might mention more simple words in daily life. According to the dataset, the word "snoring" appeared in 1084

TABLE II
NUMBER OF TWEETS IN EACH STEP OF PREPROCESSING

Preprocessing Steps	Number of Tweets	
	Liver Diseases	Sleep Apnea
Retrieved tweet data	797,703,079	797,703,079
Clean tweet data	374,709,349	374,709,349
Filtered tweet data by keyword datasets	3,474,225	7,545,953
Health-related tweet data (support-vector machine classifier)	331,479	1,063,445

TABLE III
THE DISEASES OF CHRONIC LIVER DISEASES AND SLEEP APNEA

Sleep Apnea	Chronic Liver Diseases
1. Chronic bronchitis	1. Alcoholic Liver Disease (ALD)
2. COPD	2. Hepatitis B
3. Mixed sleep apnea	3. Gallstones
4. Obstructive sleep apnea	4. Liver cancer 5. Hemochromatosis
5. Positional sleep apnea	6. Primary biliary cholangitis 7. Hepatic encephalopathy
6. Non-Cheyne-Stokes breathing central sleep apnea (CSA)	8. Cirrhosis 9. Hepatocellular carcinoma 10. Metastatic liver cancer
7. CSA	11. Portal hypertension
8. Emphysema	12. Acute alcoholic hepatitis
	13. Nonalcoholic fatty liver

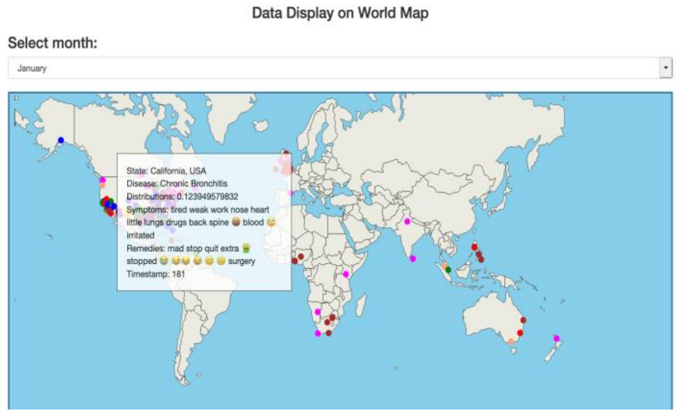
tweets in 395 regions while "CPAP" in 55 tweets in 41 regions. The study on the impact of CPAP on the users' tweets was also conducted in order to observe the changes in the diseases, symptoms, and remedies of sleep apnea in each region. The different terms which have a similar meaning to "snoring" include "snore", "snores", "snored", "snoring", "wheeze", "wheezing", and "wheeze". At the same time, words that are similar to "CPAP" and its related therapy include "mask", "masks", "positive airway pressure (PAP)", "Philips", "compliance", "humidifier", and "therapy".

A. *Data Display on the Map*

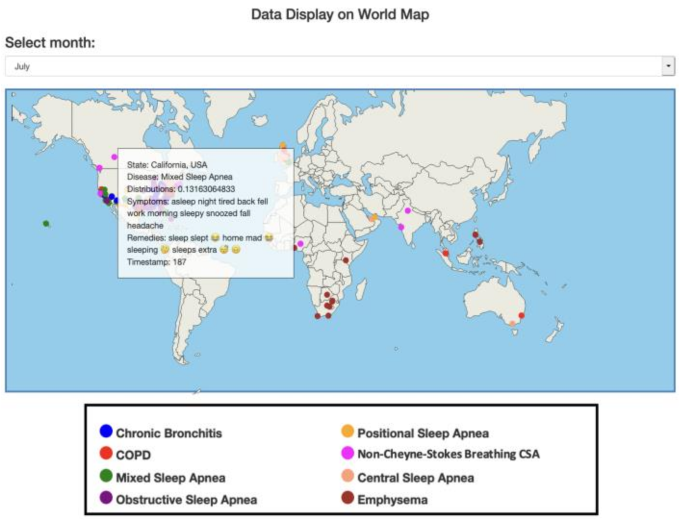
1) Data Display by Months: The output data were monitored for the changes in the diseases of sleep apnea by months. Fig. 7 depicts the changes in sleep apnea diseases in California in January and July. According to the figure, it can be clearly seen that chronic bronchitis was the most popular conversation topic on Twitter in January. Meanwhile, the discussion turned to mixed sleep apnea six months later.

2) Data Display by Seasons: To enhance the observation of the changes in diseases by timestamp, the study of the data based on different seasons, including spring, summer, autumn, and winter, was also conducted. The input data had been separated similarly by months, but the data were grouped by seasons, using the months during the season before inputting into T-ATAM to generate the output.

Additionally, in Fig. 8, for sleep apnea in Nevada, chronic obstructive pulmonary disease (COPD) was mainly tweeted by users in winter. In contrast, central sleep apnea was more



(a)



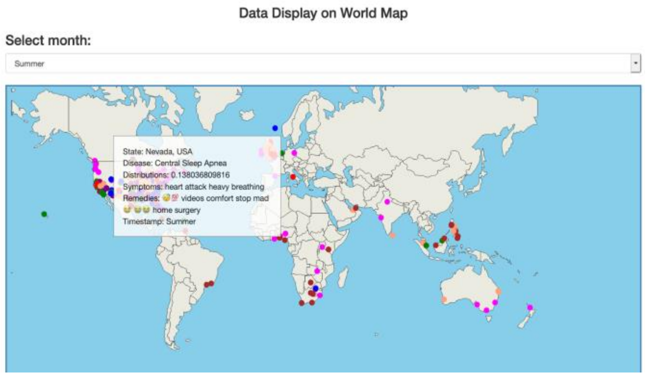
(b)

Fig. 7 Data visualization on a map for sleep apnea by months, in (a) January and (b) July.

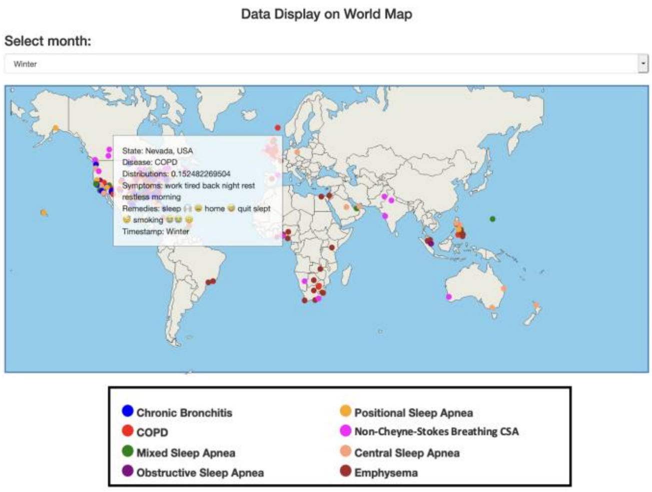
commonly mentioned during the summer with more symptoms of heart attack and breathing problems. Furthermore, for liver diseases, it was also found that more users tweeted about Hepatitis B during winter in California. In comparison, more users talked about alcoholic liver disease (ALD) during the summer of 2018. Due to the reason that most people were talking about drinking alcohol in winter, the result in the summer can be the consequences of the symptoms mentioned in the previous season.

B. *Data Display by Word Clouds and Histograms*

1) Data Display by Months: Running T-ATAM by months is very important as it can provide significant information regarding the changes of diseases of both chronic liver diseases and sleep apnea in different months. For example, in Fig. 9, sleep apnea, obstructive sleep apnea, mixed sleep apnea, and emphysema were the most frequently discussed diseases on Twitter in January. Meanwhile, positional sleep apnea and central sleep apnea showed a significant increase six months later. On top of that, tiredness, lack of night sleep, and heart attack remained the top symptoms of sleep apnea in January and July, with snoring was more common in January. Surgery,



(a)



(b)

Fig. 8 Data visualization on a map for sleep apnea by seasons, in (a) summer and (b) winter.

(a)

(b)

Fig. 9 Word cloud for sleep apnea by months, in (a) January and (b) July.

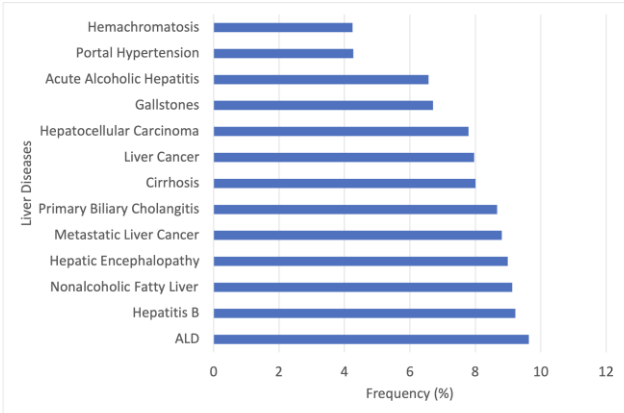


Fig. 10 Histogram of diseases for chronic liver diseases in spring.

sleeping at home, the right diet, and quitting smoking stay the top remedies suggested in social media, while CPAP was mentioned more often in January than in July.

2) Data Display by Seasons: Apart from visualizing the results via word cloud, the results are also displayed via histogram to provide more specific quantitative output data. Based on Fig. 10, ALD, hepatitis B, and nonalcoholic fatty liver were the most common diseases involved in the conversation on social media, with more than 9% in spring; meanwhile, nonalcoholic fatty liver replaced ALD in winter.

According to Fig. 11, drinking was the most mentioned symptom on Twitter related to chronic liver diseases, followed by fever and depression. These top three symptoms of chronic liver diseases had the combination of almost 40% of the total percentage of the symptoms. Then, sleep apnea, obstructive sleep apnea, mixed sleep apnea, and positional sleep apnea were the most popular discussion in social media during the spring. Meanwhile, there was a significant increase in the number of mentions of chronic bronchitis in winter.

Besides heart attack, lack of night sleep, and tiredness, which were the most common symptoms in both seasons, it is also noticeable that more people stated about snoring or breathing problems with problems in their throat in winter. Meanwhile, as can be inferred from Fig. 12, more tweets mentioned depression and falling in summer.

3) Data Display Before/After CPAP: There are several changes in diseases with their symptoms and remedies in each period. In Fig. 13, Fig. 14, and Fig. 15, there is a change in the disease in Tacoma, from chronic bronchitis in PRE and DURING with the symptoms of sleepiness, drugs, as well as nose and throat problems to emphysema in LONG-TERM with the symptoms including fatigue, heart attack, and morning sleepiness. At the same time, no symptom related to breathing problems such as throat or nose problems. Hence, it is interesting to see the changes in the distributions of diseases in each region after CPAP was mentioned in the tweet to see the effectiveness of the treatment and the development of the diseases after one treatment was introduced.

IV. DISCUSSION AND LIMITATIONS

Social media can be used for symptom diagnosis as users describe their conditions in their own words without much hesitation. However, it is also essential to take into consideration that the information on social media is from five different perspectives and limited backgrounds. Hence, it cannot be ensured that the information shared on social media is truthful as it is unverified, short, and unstructured.

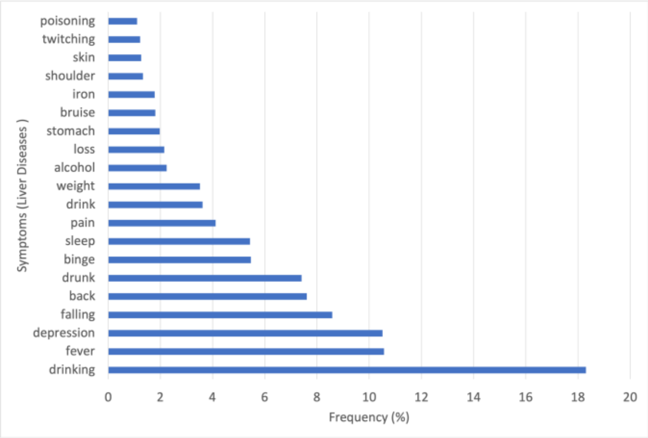


Fig. 11 Histogram of symptoms for chronic liver diseases in winter.

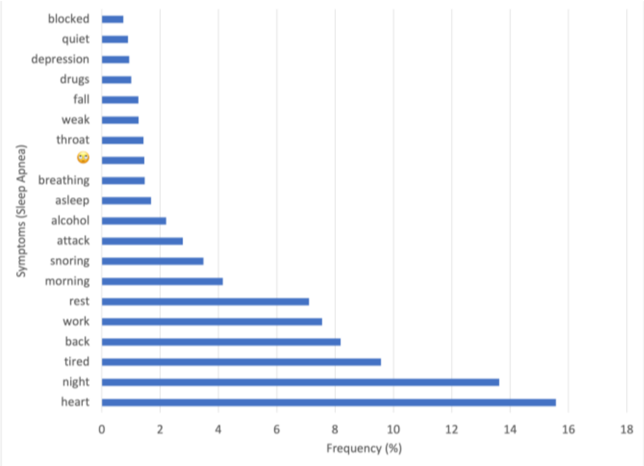


Fig. 12 Histogram of symptoms for sleep apnea in winter.

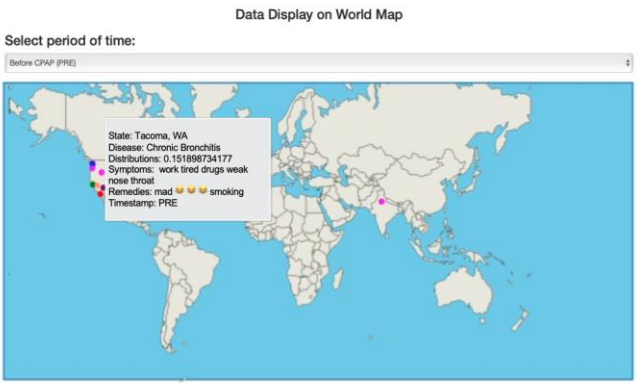


Fig. 13 Data displays before CPAP.

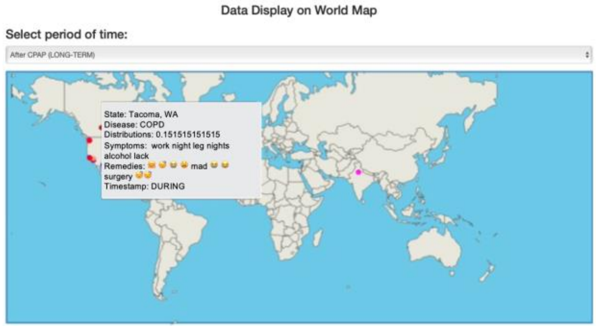


Fig. 14 Data displays during CPAP.

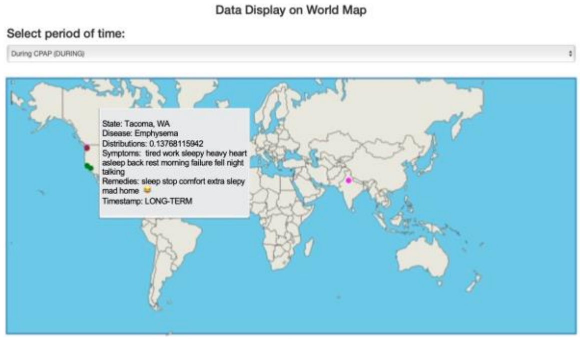


Fig. 15 Data displays after CPAP.

Although geographic coverage is limited due to low internet penetration in some regions, social media also allows users to share and collect information conveniently. According to the collection of tweets, it can be clearly seen that more people have shared information on sleep apnea than on liver diseases. The previous study has mentioned that people who are in the primary steps of liver disease do not feel any difficulties; hence, they might not notice any symptoms or signs of the disease [19]. In addition, sleep apnea is a more common disease and directly affects the daily life of people. For example, when people feel sleepy or exhausted, they may post on social media to merely share their conditions or seek someone to discuss their problems with. Moreover, although drinking is a well-known cause of liver diseases, depression is a more discussed issue, as shown in Fig 12. It can be inferred that depression can be the hidden root of other problems, including drinking alcohol, fever, and weight loss, which contribute to liver diseases. Furthermore, the study showed that the lack of night sleep was actively discussed on social media as one of the main symptoms of sleep apnea leading to other complications, including heart attack and morning sleepiness. According to a previous study with student subjects, college students had the shortest sleep duration and even irregular sleep patterns compared to other adults [20].

V. CONCLUSION

By using the T-ATAM to predict diseases, symptoms, and remedies from tweet posts, helpful information on liver diseases and sleep apnea can be obtained by region and timestamp so that changes can be observed in particular areas as well as a period. In addition, it is also implemented to examine the effect of some treatments and the development of diseases, where data can be filtered and analyzed according to the period before/after the use of that treatment. As a result, this information can be used to understand better the perceptions and experiences of people with diseases and treatments. Even though there are limitations, the outcome of this study can be an advantageous starting point for further studies to better understand and develop technology in the health care sector.

CONFLICT OF INTEREST

The authors declare that this research was conducted and written with no conflict of interest.

AUTHOR CONTRIBUTION

Conceptualization, Bonpagna Kann and Sihem Amer-Yahia; methodology, Bonpagna Kann and Sihem Amer-Yahia; software, Bonpagna Kann and Michael Ortega; validation, Bonpagna Kann, Sihem Amer-Yahia, and Jean-Louis Pépin; formal analysis, Bonpagna Kann; investigation, Bonpagna Kann and Sihem Amer-Yahia; resources, Bonpagna Kann and Sihem Amer-Yahia, Jean-Louis Pépin, and Sébastien Bailly; data curation, Bonpagna Kann; writing—original draft preparation, Bonpagna Kann; writing—review and editing, Bonpagna Kann and Sihem Amer-Yahia; visualization, Bonpagna Kann; supervision, Sihem Amer-Yahia, Jean-Louis Pépin, and Sébastien Bailly; project administration, Jean-Louis Pépin, and Sébastien Bailly; funding acquisition, Jean-Louis Pépin, and Sébastien Bailly.

ACKNOWLEDGMENT

This work is supported by the CDP LIFE Project with the cooperation between Laboratoire d'Informatique de Grenoble and CHU Grenoble Hospital.

REFERENCES

[1] S. Dorle and N. Pise, "Political Sentiment Analysis through Social Media," *2018 Second Int. Confer. Comput. Methodol., Commun. (ICCMC)*, 2018, pp. 869–873.

[2] C. Udanor, S. Aneke, and B.O. Ogbuokiri, "Determining Social Media Impact on The Politics of Developing Countries Using Social Network Analytics," *Program*, Vol. 50, No. 4, pp. 481–507, Sep. 2016.

[3] G. Wolfsfeld, E. Segev, and T. Sheafer, "Social Media and the Arab Spring," *Int. J. Press/Politics*, Vol. 18, No. 2, pp. 115–137, Jan. 2013.

[4] N. Hoog and P. Verboon, "Is the News Making Us Unhappy? The Influence of Daily News Exposure on Emotional States," *Brit. J. Psychol.*, Vol. 111, No. 2, pp. 157–173, May 2020.

[5] S. Petrović, M. Osborne, and V. Lavrenko, "Streaming First Story Detection with Application to Twitter: Human Language Technologies," *2010 Annu. Confer. North Amer. Chapter, Assoc. Comput. Linguist.*, 2020, pp. 181–189.

[6] T. Sakaki, M. Okazaki, and Y. Matsuo, "Earthquake Shakes Twitter Users," *Proc. 19th Int. Confer. World Wide Web (WWW '10)*, 2010, pp. 851–860.

[7] S.K.T. Bhavaraju, C. Beyney, and C. Nicholson, "Quantitative Analysis of Social Media Sensitivity to Natural Disasters," *Int. J. Disaster Risk Reduct.*, Vol. 39, pp. 1–10, Oct. 2019.

[8] S. Sidana, *et al.*, "Health Monitoring on Social Media over Time," *Proc. 39th Int. ACM SIGIR Conf. Res., Develop. Inf. Retr.*, 2016, pp. 1–6.

[9] S. Shrivastava and P.S. Nair, "Mood Prediction on Tweets Using Classification Algorithm," *Int. J. Sci., Res. (IJSR)*, Vol. 4, No. 11, pp. 295–299, Nov. 2015.

[10] N.S. Cook, *et al.*, "Patients' Perspectives on COPD: Findings from a Social Media Listening Study," *ERJ Open Res.*, Vol. 5, No. 1, pp. 1–10, Feb. 2019.

[11] S. Gohil, S. Vuik, and A. Darzi, "Sentiment Analysis of Health Care Tweets: Review of the Methods Used," *JMIR Public Health, Surveill.*, Vol. 4, No. 2, pp. 1–10, Apr. 2018.

[12] M.P. Hamm, *et al.*, "Social Media Use Among Patients and Caregivers: A Scoping Review," *BMJ Open*, Vol. 3, No. 5, pp. 1–9, May 2013.

[13] D. Murthy and M. Eldredge, "Who Tweets About Cancer? An Analysis of Cancer-Related Tweets in the USA," *Digit. Health*, Vol. 2, pp. 1–16, Jul. 2016.

[14] Y.T. Yang, M. Horneffer, and N. DiLisio, "Mining Social Media and Web Searches for Disease Detection," *J. Pub. Health Res.*, Vol. 2, No. 1, pp. 17–21, May 2013.

[15] V. Lampos and N. Cristianini, "Tracking the Flu Pandemic by Monitoring the Social Web," *2010 2nd Int. Workshop Cogn. Inf. Process.*, 2010, pp. 411–416.

[16] E.M. Glowacki, A.J. Lazard, and G.B. Wilcox, "E-Cigarette Topics Shared by Medical Professionals: A Comparison of Tweets from the United States and United Kingdom," *Cyberpsychol., Behav., Soc. Netw.*, Vol. 20, No. 2, pp. 133–137, Feb. 2017.

[17] A. Culotta, "Towards Detecting Influenza Epidemics by Analyzing Twitter Messages," *Proc. First Workshop Soc. Media Anal. (SOMA '10)*, 2010, pp. 115–122.

[18] P.M. Polgreen, Y. Chen, D.M. Pennock, and F.D. Nelson, "Using Internet Searches for Influenza Surveillance," *Clin. Infect. Dis.*, Vol. 47, No. 11, pp. 1443–1448, Dec. 2008.

[19] E.A. Tsochatzis, J. Bosch, and A.K. Burroughs, "Liver Cirrhosis," *Lancet*, Vol. 383, No. 9930, pp. 1749–1761, May 2014.

[20] A. Schlarb, A. Friedrich, and M. Claßen, "Sleep Problems in University Students – An Intervention," *Neuropsychiatr. Dis. Treat.*, Vol. 13, pp. 1989–2001, Jul. 2017.