Implementation of Data Mining from Social Media for Improved Public Health Care

Dr. Mohammed Saeed Jawad
Institute of Computer and Network Engineering
technische universität braunschweig
Hans-Sommer-Straße 66
38106 Braunschweig, Germany
dr.m.s.jawad@gmail.com

Afaf Salem
School of Postgraduate Studies
Department of ICT
Management and Sciences University
Shah Alam, Selengor, Malaysia
Mokhtar.2711@yahoo.com

Professor Wael Adi
Institute of Computer and Network Engineering
technische universität braunschweig
Hans-Sommer-Straße 66
38106 Braunschweig, Germany
w.adi@tu-bs.de

Mohamed Doiher
Faculty of Information and Communication Technology
Universiti Tehnika Malaysia Melaka
Melaka, Malaysia
Jeruslem_2008@yahoo.com

***Abstract*—To improve public health care outcomes with reduced cost, this research proposed a framework which focuses on the positive and negative symptoms of illnesses and the side effects of treatments. However, previous studies have been limited as they neither identified influential users nor discussed how to model forms of relationships that affect network dynamics and determine the accurate ranking of certain end user's feedbacks. In this research, a two-step analysis framework is proposed as the system. In the first level, the system utilized exploratory analysis and clustered users and their useful feedbacks through self-organizing maps (SOM). In the second level, the system developed three lists of negative and positive feedbacks and treatment symptoms caused by implanting the SOM that considered accurate ranking by calculating the frequency of each term of interests. The feasibility of the proposed solution is confirmed as performance evaluations of the system in terms of computational costs. The results showed that these solutions are reasonable computational costs relative to memory and processor usage.**

Keywords—Data mining; social media; medical data; end user feedbacks; positive terms; negative terms; symptoms

I. Research Motivations and Background

Data mining from social media recently gained the attention of many important businesses and industries. Data mining is empowered by the recent advances in big data analysis as well as the network modeling of social media forums and websites, which are integrated to achieve knowledge discovery solutions and to extract useful information from various fields [1]. The health care industry is one of the most important fields that can be significantly enhanced by modern data mining techniques that allow the discovery of certain trend patterns as a product of the social media feedbacks generated dynamically from the experiences and opinions of end users [2]. These techniques can be applied to drug feedbacks in social media which can help manufacturers continuously enhance their products at reduced-costs.

Successful data mining from social media can result in numerous benefits for business owners and manufacturers [3]; however, a number of challenges should be addressed first to reach the acceptable level and wide deployment of this new technology [4]. In addition, classical data mining techniques and algorithms should be empowered with an intelligent pattern recognition tool to predict and visualize the common trends of the data in general. Moreover, it should weigh certain descriptions or feedbacks based on their frequency and eliminate some neutral words as a filtration technique in the preprocessing stage. Different online forums contain feedbacks that should be network-modelled for a more convenient analysis. All these challenges are the main inspirations that motivated this research. The main contributions of this work can be considered as developing a feasible text data-mining solution that can partition different users with certain ID by modeling their existence in different web forums. Furthermore, the accurate clustering of negative and positive feedbacks and the visualization of the overall positive or negative feedback trends in a reasonable computational cost solution is another benefit from this research.

The basic concepts are explained in this section as a background of the research field and the proposed solutions. First, the processes of data collection and mining are considered challenging tasks given a large number of networks studied, and this requires a complex representation of the social network structure. The complexity of such structure is derived from network density and levels of the parents and nodes clustering social media contents. Network clustering involves complex, big, and parallel data processing to cover the analysis of each networks' nodes representing certain user communities. A part of the network usually as small as the sample size is used for data collection.

Traditionally, structuring and modelling social networks can extract useful information as topic trends and the trends of opinions and the linguistics properties play an effective role in this study. At the beginning, certain word filtration techniques are used to remove unwanted words, such as stop and stemming words [5].

The concept of the self-organizing map (SOM) is utilized in particular research fields; in the simplest terms, SOM is a predefined wordlist used to correlate with the large data from the social networks under the tests to extract positive and negative words [6]. Importantly, certain algorithms are used to determine the frequency of certain positive or negative words, so that weighing the word (in another meaning its effectiveness among other words) can be determined. Finally, simple statistical tools are used to identify a positive or negative trend and the most common words describing the symptoms of drug use.

Similar solutions in the literature on disease surveillance for the case of Influenza-related community who share flu-posting online is utilized through the technique *text and structural data mining of web and social media* (WSM) [7] [8]. In the critical analysis of the SOM and WSM techniques in the literature [9], [10], it can be concluded that SOM techniques have more advantages in terms of the capability of investigating the positive and negative feedbacks of treatments. This advantage can be achieved by mapping large dimensional information onto a low dimensional space.

II. Methodologies of the Proposed System

A two-step framework is proposed as an investigatory analysis to evaluate the correlations between user posts and positive/negative words under a drug name. The correlation is obtained by using SOM. Using a network-based approach, the system enabled users and their posts to find the possible partition using complete linkages. The two processes involved with inter-social dynamic maps for reviewing SOM results are described below:

- The correlation between user and judgment.
- The partition between users and their posts.

Regarded as an unsupervised technique, Self-Organizing-Map (SOM) is used to explore the survey dataset based on the artificial neural network. The representation of the SOM data is in multidimensional data such as two or three dimension. Based on the data compression of the vector quantization technique, the SOM process is used to reduce the dimensionality of sectors. The information is stored as a topological relationship within the training sets in a network. Therefore, large data sets are visualized with high dimensionality using SOM. The competitive learning approach of SOM has one neuron unconnected to the input and output layers for each training phase. Although the connection between the neurons is absent, communication exists between each phase through a neighborhood function. The proposed SOM approach is used to summarize and visualize the profiles of individual patients. This visualization process helps determine domain experts. The two perspectives involved in the process of accurately obtaining results are computational and scientific perspectives [11].

By using SOM in the computational perspective, feasibility is examined by extracting useful information from questionnaires. In the scientific perspective, the different types of patient diseases such as type-I diabetes are collected to understand the responses in the diabetes survey and suggest about domain experts to clinicians. By contrast, clinicians are required to take a survey about their patients. The mean, skewness, variance, and frequency are the traditional descriptive statistical methods, but it provides simplified conclusions. Thus, data are analyzed based on statistical machine learning tools with black box by clinicians. The SOM algorithm is used for mining correlations and clustering similar responses within the surveys. If the dimension is higher for clustered responses, then the data is visualized in a two-dimensional grid to reduce data complexity. Complexity is reduced by revealing more meaningful relationships and by understanding the dependencies among the survey responses.

Previously, SOM is used to visually explore data areas such as health, lifestyle, nutrition, financial, gene expression, marine safety, and linguistics. Recently, SOM is utilized to explore questionnaire-based loneliness survey data. The present research also focuses on improving data interpretation by revealing possible associations between the tendency of item nonresponse and the background variables of participants. The flaw conclusion is obtained by item nonresponse which is related to the background variables of respondents, such as age and gender nonresponses. Considering the undetected non-causative relationships between independent and dependent variables, the nonresponse factors affected patient satisfaction.

In the present study, item nonresponse does not refer to participants who fail to return the survey, but to the ones who choose not to respond to all questions. In the proposed approach, the issues involved in this research are included for data analysis. Large surveys have demonstrated that although respondents and non-respondents in patient satisfaction surveys may differ according to several demographic and clinical characteristics, the differences in satisfaction between them tend to be relatively small and non-respondents do not constitute a homogenous group. Many highly sophisticated statistical methods are used as a standard technique to handle the problem of missing responses.

In the existing link method, the idea that missing data are not just a statistical nuisance but also contain valuable information as such is tested simply by including the number of item nonresponses per respondent as an explanatory variable in the models. The expected predictors of patient satisfaction and the SOM are explained in the network-based modeling of social media. The patient satisfaction section is necessary to understand the slightly ambiguous nature of the concept. Patient satisfaction is affected not only by factors related to care receivers and providers but also by factors related to the means in which it is defined and measured. The approach is proposed in both quantitative and qualitative studies for patient satisfaction. The main reason for utilizing the SOM algorithm in data mining and knowledge discovery

techniques is data driven rather than theory driven. However, this reason does not mean that they cannot be applied for confirmatory purposes.

A. Pipeline of the Different Integrated Algorithms Followed for Effective Medical Data Mining

The methodologies followed in this work can be considered as different algorithms integrated together to complete the desirable partitioning of the end user from different surveys or forum websites and the accurate clustering of each positive and negative individual feedbacks and overall visualization trends. This pipeline of algorithm can be understood as tasks of text-filtrations techniques to ensure uniform text for further accurate, useful, and informative data partitioning and clustering algorithms. The ranking of certain negative or positive terms is considered by algorithms to detect the frequency of this specific term of interest. The completed tasks of all these integrated algorithms are shown as implementation results in the coming section as proof-of-concept. Fig. 1 shows the proposed system architecture for improving health care. The sections involved in the system are described below:

- Preprocessing Dataset;
- Term-Frequency-Inverse Document Frequency (TF-IDF);
- Filter Stop Words;
- Performance Analysis.

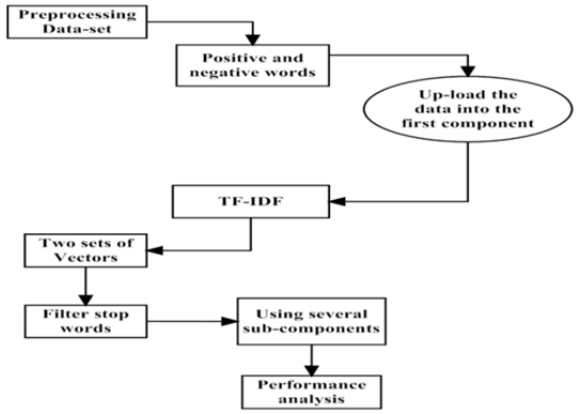

Fig. 1. Data mining architecture from social media.

The unwanted data are removed by the preprocessing filtration stage of the text data set under-study. Then, both words/information are uploaded into the first component. The uploaded data are provided as input to the TF-IDF where the vectors are divided into two sets. The uploaded data are processed in the second component to filter excess noise by using several subcomponents. Thus, the uniform set of variables is measured by the TF-IDF method. Finally, the performance is analyzed using a high dimension of data.

Fig. 2 shows the flow diagram of the proposed network-based modeling of social media for improving health care. From the flow chart, the information of both negative and positive words along with side effects is analyzed using SOM technology. The process is started to choose a dataset. Afterwards, the chosen dataset is preprocessed to remove unwanted information.

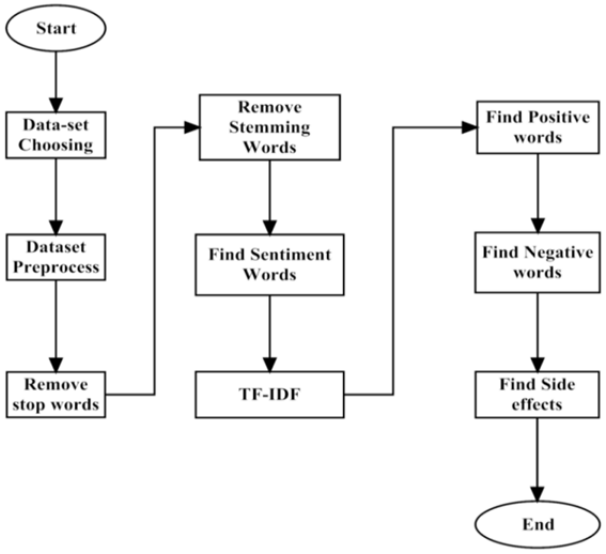

Fig. 2. Flow chart of proposed methodology.

During preprocessing, stop and stemming words are removed to handle a large amount of data. Sentiment words are then identified to analyze the response of the disease-affected person. Term-frequency-inverse document frequency (TF-IDF) method is used to discriminate the words as either positive or negative. To determine the side effects of the patient, both positive and negative data are gathered. During patient treatment, possible side effects are found by overlapping all wordlist to the modules. Finally, the process is stopped to analyze the performance of the proposed method.

III. Implementation Results

Data mining provides the essential tools for discovering patterns in data. By using the data mining algorithm, the user can choose a specific file. To choose a file, browse the file selection and enter the file name. A view button is provided to view a particular file. This button shows the data as a preview in the file data box provided as shown in Fig. 3 and 4.

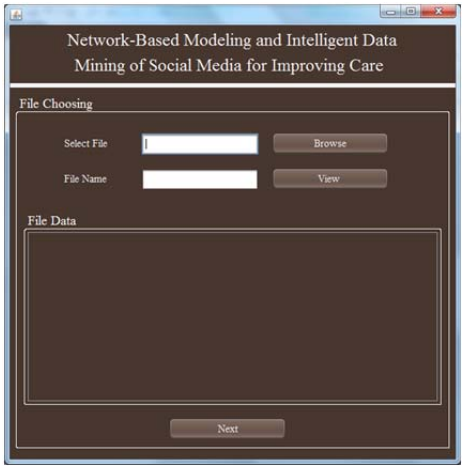

Fig. 3. File choosing window.

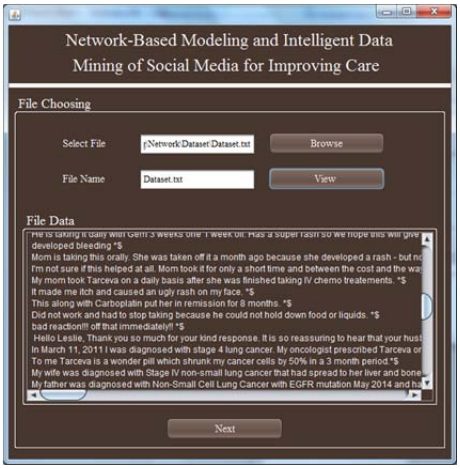

Fig. 4. Dataset file is chosen in the file choosing window.

Unwanted special characters should be removed because of their highly unstructured, noisy nature and informal communication. To remove an unwanted special character, we have to select the file first and enter the file name. The show option will then help the user to preview the special characters removed from the file, as shown in Fig. 5. The file path is obtained to upload this regenerated file. The file is then uploaded by entering the file name. The user can upload the file by clicking the upload button. If the user wants to view the uploaded file, it will be instantly displayed upon upload.

Afterwards, a message for the successfully uploaded file appears, as shown in Fig. 6.

Similarly, the file can be uploaded for each and every user. For example in Fig. 7, 11 users uploaded their files, and the review is shown at the right side of the window. These reviews may contain positive or negative aspects.

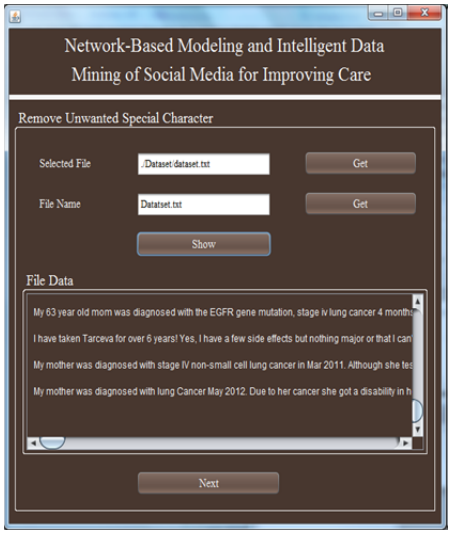

Fig. 5. Removal of unwanted special characters.

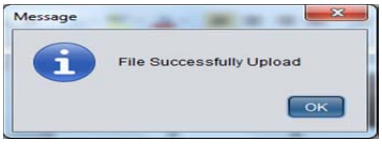

Fig. 6. Message window.

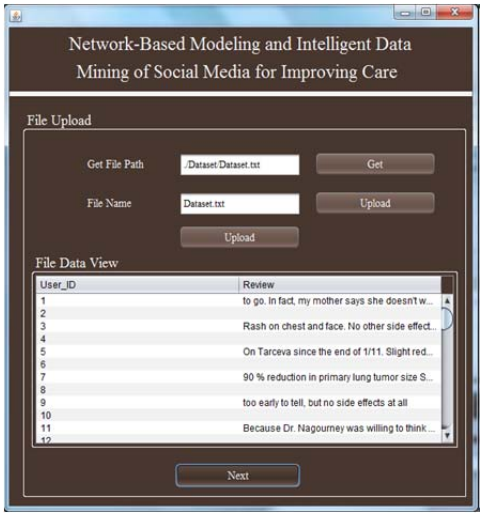

Fig. 7. File upload window.

When data analysis is driven at the word level, commonly occurring words known as stop words should be removed. A user can either create a list of stop words or use a predefined language for specific libraries.

For example, in the context of search engine, if user searches, "what is data mining," then the search engine attempts to find web pages containing the terms "what," "is," "data," and "mining." The search engine finds numerous pages which contains the terms "what" and "is" than pages containing data mining information because the terms "what" and "is" are commonly used in English language. Thus, if a user can neglect these two terms, the search engine could focus only on retrieving pages that contain the keywords: "data" and "mining," which would bring up pages of interest.

Hence, stop words must be removed in the following techniques:

- **Supervised machine learning** – removing stop words from the feature space;
- **Clustering** – removing stop words to generate the clusters;
- **Information retrieval** – preventing stop words from being indexed;
- **Text summarization** − excluding stop words from contributing to the summarization scores and in computing the rough scores.

Examples of useable stop word lists:

1) Determiners such as: the, a, an, and another.
2) Coordinating conjunctions such as: for, an, nor, but, or, yet, and so.
3) Prepositions such as: in, under, toward, and before.

To remove the stop words, we have to view the stop-words list. We then obtain the database values to remove the stop words and to review their removal (see the preview window as shown in Fig. 8). In this study, 15 user IDs are shown with removed stop words content data.

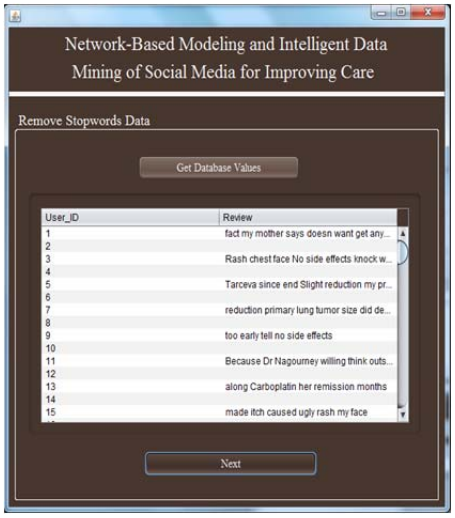

Fig. 8. Removal of stop words data.

Similarly, stemming words can be removed. In information retrieval, stemming is the process of minimizing the words derived from the base word. The stem should not be the same as the base word. For example, words such as "worked," "eating," and "really," the removed stemming words are given as "work," "eat," and "real" (see Fig. 9). The database values can be obtained from the review of users. In the elimination of sentiment words, the user must find the sentiment words in the list. By clicking the find sentiment word option, the word list including "annoyed," "past," "normal," "few," "major," "small," "helpful," "rash," and "side" is shown. The list reflects both positive and negative words presented in Fig. 10. To determine the positive words alone, click the find positive words button. The "positive words found successfully" message will be displayed, as shown in Fig. 11. The sentiment word list also has negative words, which are previewed next to the list box of positive words. However, when the user wants to find negative words alone, it can be successfully derived, as shown in Fig. 12. To obtain negative words, only click the button shown in Fig. 13.

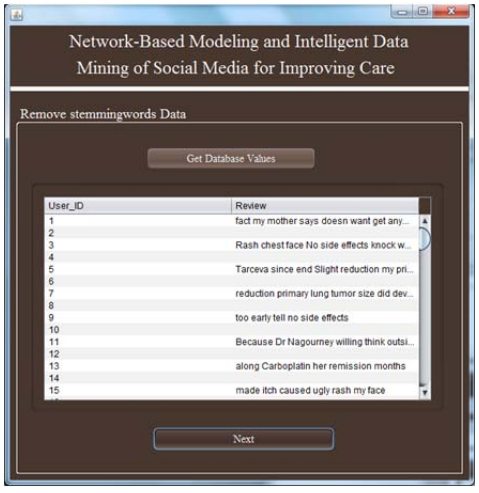

Fig. 9. Removal of stemming words.

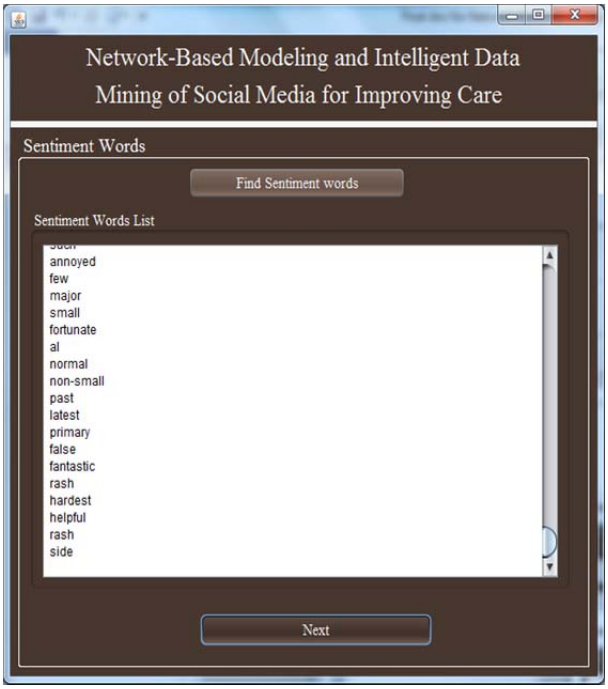

Fig. 10. Finding sentiment words.

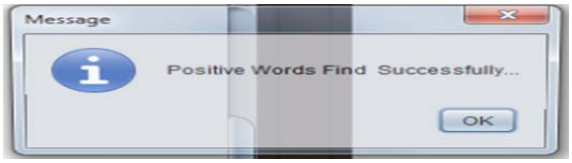

Fig. 11. Successful message.

The successful message for the search for negative words is presented in Fig. 14. The negative words list consists of words like "annoyed," "major," "rash," "hardest," and "false."

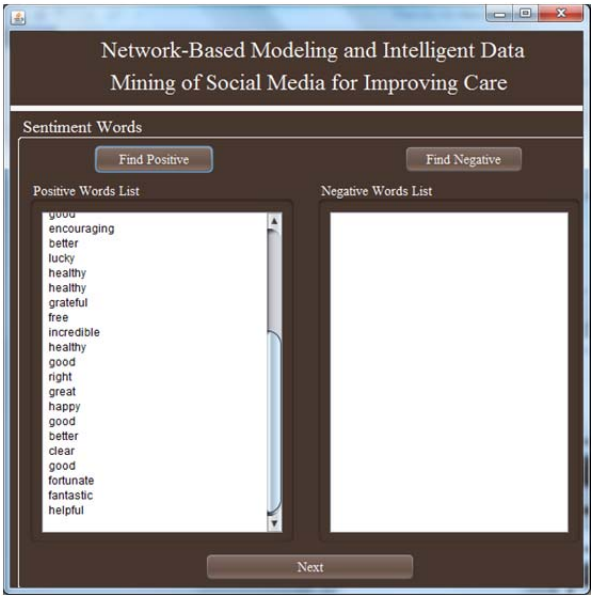

Fig. 12. List of positive and negative words.

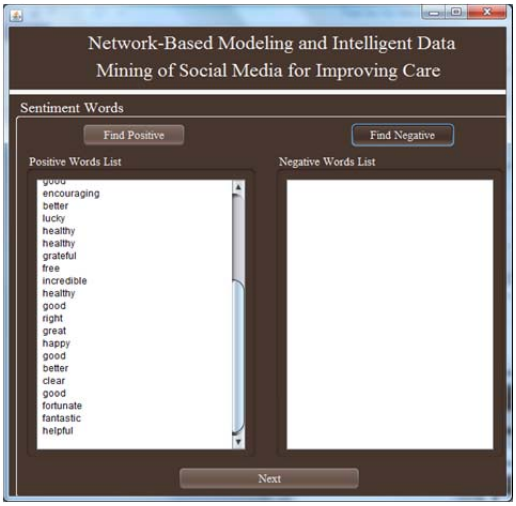

Fig. 13. Finding the list of negative words.

Fig. 14. Successful message for finding negative words.

User can then search for side effect words, which is a list of words that define the disease and its symptoms. To find the side effect words, open the side effect window and then click, "find side effect words." This window shows words such as "painful," "sore throat," "cough," "pain," and "stomach pain," as shown in Fig. 15. The user can analyze the data in positive and negative reviews. Meanwhile, to know the count of the positive and negative words reviewed, proceed to the chart representation, as shown in Fig. 16. The bar chart displays the positive and negative words as a review.

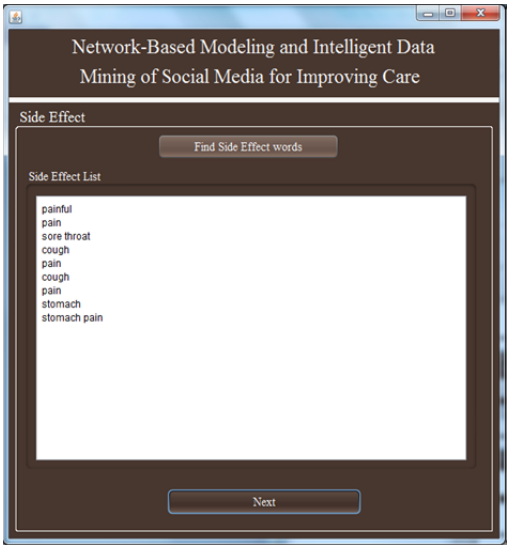

Fig. 15. List of side effect words.

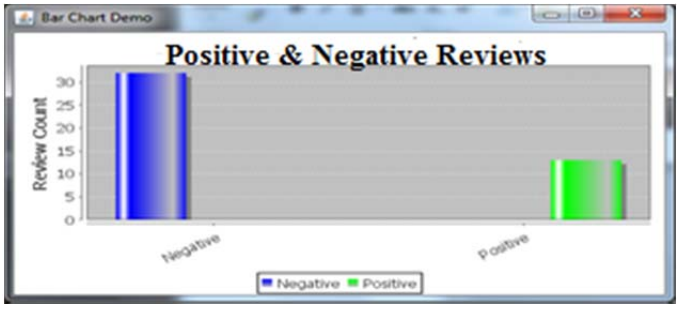

Fig. 16. Review count of users.

IV. Performance Evaluation Analysis

The execution time, processing time, and delay are calculated as follows:

A. *Execution Time*

Execution time is the time spent by the system performing a given task, including the time spent for the run time services on its behalf. Another definition is the *time* for a running program, in contrast to the program lifecycle phases such as link *time*, compile *time*, and load *time*, which is provided by

$$E = \text{CPI}\times\text{I}\times 1/\text{CR}, \quad (1)$$

where E is the execution time, CPI is the cycle per instruction, I are the instructions, and CR is the cycle rate.

B. *Processing Time*

Processing time is the time between the commencement and completion of a process. Another definition for processing time is the total *time* that a *processing* unit used takes for *processing* instructions or steps of a program or an operating system, which is given by

$$P = \frac{1}{TR}, \quad (2)$$

where P is the processing time, and TR (Throughput rate) are the tasks completed/time.

C. *Delay*

Delay specifies the system duration for a bit of data to travel across the network from one endpoint to another, which is provided by

$$D = \frac{N}{R}, \quad (3)$$

where D is the delay, N is the number of bits, and R is the transmission rate.

From the execution time, processing time, and delay, the memory usage is displayed in the pie chart in Fig. 17. This figure contains a summary of the memory usage, in which the execution time consumed 56%, processing time acquired 30%, and the delay time has the remaining 14%.

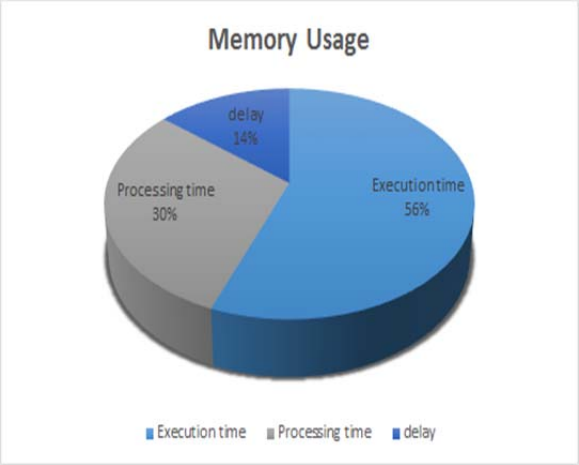

Fig. 17. Memory usage.

V. CONCLUSIONS

The performance of the proposed two-step analysis framework in this research evaluates the positive and negative symptoms of a disease and the side effects of treatment using SOMs. This framework provides the improvement in health care outcomes through an effective mining process. The existing text and structural data mining of the WSM technique provides a cost-effective response. However, some issues such as non-active node are ignored; thus, we chose a two-step analysis framework. The analysis results demonstrated that the memory usage in terms of execution time, processing time, and delay is less than the existing text and structural data mining of the WSM system. Thus, this proposed technique is an enhanced network-based system, which performs better compared with other systems.

REFERENCES

[1] Barbier, G., & Liu, H. (2011). Data mining in social media. *Social network data analytics*, 327-352.

[2] Raghupathi, W. (2016). Data mining in healthcare. *Healthcare Informatics: Improving Efficiency through Technology, Analytics, and Management*, 353-372.

[3] El Azab, A., Mahmood, M. A., & El-Aziz, A. (2017). Effectiveness of Web Usage Mining Techniques in Business Application. In *Web Usage Mining Techniques and Applications Across Industries* (pp. 324-350). IGI Global.

[4] Bello-Orgaz, G., Jung, J. J., & Camacho, D. (2016). Social big data: Recent achievements and new challenges. *Information Fusion*, *28*, 45-59.

[5] Kothapalli, M., Sharifahmadian, E., & Shih, L. (2016). Data Mining of Social Media for Analysis of Product Review. *International Journal of Computer Applications*, *156*(12).

[6] Gharib, T. F., Fouad, M. M., Mashat, A., & Bidawi, I. (2012). Self-organizing map-based document clustering using WordNet ontologies. *IJCSI International Journal of Computer Science Issues*, *9*(1), 1694-0814.

[7] Corley, Courtney D., et al. (2010) "Text and structural data mining of influenza mentions in web and social media." *International journal of environmental research and public health* : 596-615.

[8] Yang, Y. Tony, Michael Horneffer, and Nicole DiLisio. (2013). "Mining social media and web searches for disease detection." *Journal of public health research* 2.1 :17.

[9] Krithika D Renuka and Rosiline B Jeetha. (2017). A Survey and Analysis of Various Health-Related Knowledge Mining Techniques in Social Media. *International Journal of Computer Applications* 158(1):5-10.

[10] Liu, Yuan-Chao, Ming Liu, and Xiao-Long Wang.(2012). "Application of self-organizing maps in text clustering: a review." *Applications of Self-Organizing Maps*. InTech,

[11] Akay, A., Dragomir, A., & Erlandsson, B. E. (2015). A novel data-mining approach leveraging social media to monitor consumer opinion of sitagliptin. *IEEE journal of biomedical and health informatics*, *19*(1), 389-396.