Data Mining Techniques in Online Health Communities

Cassandra Mikkelson and Cali Sweitzer
College of Life Sciences
Thomas Jefferson University
Philadelphia, Pennsylvania
e-mail: {cassandra.mikkelson | cali.sweitzer}@students.jefferson.edu

***Abstract*—Online health communities are an untapped domain of unlimited data on patient sentiment towards drugs and medical devices that can provide academia and industry an inside scope of in demand research according to patient responses. These communities are often found on social media platforms, such as Facebook and Reddit, where patients who have similar medical histories connect to share their experiences, advice, and support for each other. This review explores how data mining methods, specifically machine learning and Natural Language Processing (NLP), can be applied to analyze large data sets derived from user-generated responses on social media and health databases. Methods discussed include sentiment analysis, clustering algorithms, and text classification models as effective tools to generate new knowledge on patterns within online health discussions. The paper also highlights potential applications of data mining to improve pharmaceutical research, enhance drug monitoring, and identify adverse events in terms of post-market surveillance for regulatory bodies like the U.S Food and Drug Administration (FDA). Lastly, challenges related to data transformation, cleaning, and privacy concerns are addressed along with proposed augmentations to improve data quality.**

Keywords-data mining; online health communities; Patient Sentiment; Sentiment Analysis; Healthcare Data Transformation.

I. Introduction

In the age of social media, online communities have become a haven where patients facing health challenges can exchange insights and share common experiences. Online health communities are formed on social media platforms including Facebook groups and subreddits on Reddit, where patients and caregivers come together to share their experiences, advice, and support for others within their communities. Patient-driven platforms including Patientslikeme, Health Union, and Healthboards are networks specifically formulated for user camaraderie in the healthcare setting, unlike those support groups naturally formed on other social medias. These platforms and online health communities empower patients to act as healthcare consultants, in the form of reviewing drugs, devices, surgeries, and specific healthcare providers [1]. Consequently, these communities generate an abundance of self-reported data on patient sentiment, which provides valuable insight into patient satisfaction connected to health services.

Data mining combines machine learning, algorithms, statistical analysis, artificial intelligence, and database management systems [2]. Once a database of interest is defined, the data is transformed to complement the model that is created. The model is then tested, evaluated, and interpreted to generate new knowledge through the generation of a custom report. Data mining enables the user to analyze data across different dimensions that recapitulate useful information that can inspire new ideas.

Data mining techniques can serve as potential tools to extract previously hidden information and patterns from online communities. A machine learning approach that most effectively scans, processes, and summarizes social media data would be NLP, which includes text classification (e.g., Support Vector Machines (SVM) and naïve Bayes classifier), sentiment analysis, clustering algorithms (e.g., Self-Organizing Map (SOM)), and supervised learning algorithm (e.g., decision tree).

The results generated by mining of large patient-derived datasets could inform the pharmaceutical industry about in demand medical interventions according to patient needs. Consequently, research outputs could positively impact current pre-clinical and clinical trials to streamline desired research according to patient sentiment and break the barrier between the bedside and benchtop.

This review article highlights various data mining techniques that could be utilized to collect and transform data from online health communities. Section I introduces the concepts of online health communities and data mining. Section II proposes different data mining methods that could help reduce the complexity of data obtained from these communities, including sentiment analysis, SOM, SVM, and the naïve Bayes classifier. Section III explores how data collected from such studies could be applied by the U.S. Food and Drug Administration (FDA) to identify adverse drug reactions and improve efficiency in drug production, such as vaccine development. Section IV addresses challenges related to data transformation and cleaning, proposing augmentations like web scraping and the Levenshtein distance method to address issues associated with data collection from online forums. Finally, Section V concludes the article, emphasizing the underutilization of data generated by online health communities and its potential to positively influence academic and institutional medical research.

II. Proposed Methods and Techniques

As patients increasingly turn to online communities and health platforms for reviews, advice, and solidarity, the development of mining techniques to analyze this user-generated data has become more essential. Alnashwan et al. described three data-driven approaches to elucidate patient sentiment within these online forums: sentiment analysis, content analysis, and topic analysis. Sentiment analysis is a broad field of study with the goal of identifying and characterizing the emotional tone of a body of text [3]. Sentiment analysis is widely used in the context of understanding patient values, attitudes, and preferences towards medical providers, prescriptions and treatments, and adverse effects. It serves as a classification model within supervised machine learning, where predictions are made and validated through associated characteristics. Classic sentiment analysis groups mined posts based on three categories: positive, negative, and neutral. Some studies in current literature also calculated the degree of emotion using a numerical scale, e.g., -5 to +5, based on keywords within the text. Alnashwan et al. hypothesized that classifying medical posts on a binary (positive/negative) or polarity (degree of sentiment) based scale would not be sufficient to encompass the broad and complex nature of online health-related text [3]. As such, the authors suggest a bottom-up categorization approach, in which posts are manually mined for specific sentiment-based keywords and subsequently grouped into seed categories. The multitude of seed categories is then further filtered into six core categories based on the predominating sentiment, examples including treatment inquiry, symptom confusion, and seeking general information. Once categorized, different techniques of data mining under the umbrella of sentiment analysis can be employed. Such techniques include the use of machine learning and lexicon-based text classification systems.

One of many examples of machine learning techniques includes the use of Microsoft's Azure Machine Learning software, which serves as a resource for data scientists and machine learning engineers to generate programs such as NLP using artificial intelligence. These tools are developed through machine learning in which the user builds algorithms that allow the computer to continually learn based on predictive models [4]. Such models including NLP where the computer becomes able to interpret and categorize informal text are crucial for data mining of patient sentiment in online health communities.

This broad concept of NLP and sentiment analysis includes the use of lexicon-based text classification systems: content analysis and topic analysis. Content analysis is a research method used to extract meaningful content within a large body or dataset of text by analyzing and grouping relationships of high frequency words, phrases, or themes [4]. In conjunction with content analysis, this NLP technique can also employ topic analysis, which aims to identify overarching topics within a body of text based on a probabilistic model [4].

Simultaneously, Jawad et al. proposed two techniques for text classification that can be used together to identify patterns in patient sentiment from social media in an article for the Proceedings of the 2017 Future Technologies Conference [5]. A SOM categorizes input vectors based on a wordlist into a neural network, hence creating clusters based off words defined as positive and negative. This model would utilize the Term-Frequency-Inverse Document Frequency (TF-IDF) to vectorize text files by assigning text with a numerical statistic that would interpret the frequency of a word in a document relative to the whole document. TF reflects the frequency of a given word and IDF reflects the rarity of a word. The implementation of this model would output results that categorize responses in terms of a positive or negative sentiment, which would be useful to pharmaceutical companies to gain knowledge on patient assessment of their products.

Alternatively, techniques used in data mining from mobile health apps would be translational to mining for online health communities. Fallah et al. compiled a systemic review on the common data mining methods correlated with health apps. They found that the top three successful methods with the highest level of accuracy were cloud-based SVM, decision tree, and naïve Bayesian [2]. After data has been vectorized, the SVM classifies the data on a binary scale and generates a separating hyperplane line that separates the two groups. This method would best be used for a yes/no research question (e.g., is the response to a product positive or negative?). The decision tree mimics a tree, with population classified in branches that construct a tree with roots, internal nodes, and leaf nodes. Nodes reflect choices made in a decision that splits into a branch that represents the outcome of the decision. Data is split into parent nodes and child nodes, to decide the category of the text file. It is commonly used for creating classifications based on a prediction algorithm. Meanwhile, the naïve Bayes classifier vectorizes the text files into multi-dimensional numerical probability values that are used as input for the SOM, SVM, or decision tree for the final classification step. Probability values are based on the probability of a text that contains pre-defined words is equal to the probability of finding these pre-defined words in a category [6].

Through data mining methods, the complexity and breadth of public online data are reduced to expose undiscovered patterns in common patient sentiment and reveal previously unreported ailments. Researchers must design their model according to how their question would categorize their data. For example, SVM is most appropriate for binary classification, while SOM is best for multiple categories and would complement a study to discover the adverse effects of a product from patient reviews.

III. Applications

The applications of data mining in the healthcare setting are robust, allowing for the harnessing of vast amounts of relevant online medical data. One application of data mining in healthcare includes the identification of adverse drug reactions. Some relevant adverse reactions are not apparent until after clinical trial testing and approval by the FDA, as factors which may cause these adverse reactions are often difficult to account for in the clinical trial period. After approval by the FDA, important factors that may cause adverse reactions include long-term use, co-exposures to other drugs, environmental and dietary variances, as well as genetic differences that may not have been probable to account for during clinical trials [7].As such, databases such as the FDA Adverse Event Reporting System (FAERS) give critical insights into relevant drug reactions. As outlined in the most recent FDA White Paper on Data Mining from 2018, several techniques are applied to FAERS safety reports to explicate possible adverse events [8]. Disproportionality methods are used to identify statistically significant associations between medications and events. One such method includes the use of the Proportional Reporting Ratio (PRR) in which the degree of reporting of an adverse event for a particular drug is compared to the same event occurrence amongst all reports of all drugs within the FAERS database [8]. Thus, this robust data serves as a baseline for the occurrence of any event, allowing for associations to be made based on disproportionate reporting. Statistical methods beyond this data mining technique are then employed to further validate causative rather than correlative relationships [8].

Another clinically relevant database used by researchers and data miners alike to extrapolate patient-derived data includes the FDA's Vaccine Adverse Event Reporting System (VAERS). Data mining for patient sentiment in the context of vaccine efficacy and reactogenicity has become increasingly relevant following the COVID-19 pandemic. As described by Dror et al., vaccination compliance relies on a personal risk-benefit perception which can be skewed by misinformation and perceived side effects that may not align with scientific evidence [9]. Data mining and subsequent analysis of such reports proves as an effective tool for minimizing misinformation regarding vaccine reactogenicity, potentially enhancing vaccine uptake [10]. Data mining of VAERS reports provides a powerful dataset for understanding public sentiment related to vaccines, with direct relevancy to vaccine uptake.

IV. Challenges and Proposed Augmentations

Generating knowledge from a larger data set is generally a challenging and time-consuming task. This is especially so when the data in social media communities contains about two decades of responses, including spelling errors and abbreviations that would make creating a word list an ambitious effort. Transforming and cleaning this unorganized data would take an extended amount of time as these proposed models are best suited for survey responses. To minimize this task, web scraping can be used to extract data from unstructured web browsers into structured data that can be used for analysis. Web scraping is the process of data transformation through computer software, that mimics human behaviors of web exploration to compile data more efficiently than by hand [11]. Meanwhile, there is still the struggle to correct spelling errors and abbreviations, which may be resolved by utilizing Levenshtein distance to identify errors by comparison against a dictionary.

Patient sentiment may vary according to several demographic and clinical characteristics, particularly the social determinants of health—non-medical factors like race, ethnicity, religion, and socioeconomic status—that significantly influence a person's health [12]. As previously mentioned, data mining can be used to extract previously unknown information hidden within a data set. Thus, data mining has the power to segment data according to the social determinants of health and could uncover the diversity in patient sentiment according to factors defined by the social determinants of health. However, extracting such personal data would be challenging and might require social media profile information. To address this, researchers could create tools to categorize responses according to these factors before applying data mining techniques. Web scraping software could also aid in this task.

Studies that result from this work would have to address ethical concerns of informed consent, since consent cannot be obtained from all social media platform users. Even though studies will be compiled of public data from online forums, the collection of data must be aligned with data privacy laws, which protects users from the collection of their names to comply with confidentiality and anonymity. Before the conduction of a study on data mining from online health communities, researchers must ensure compliance with federal and institutional laws and policies along with the privacy policies of the social media platform of interest.

V. Conclusion and Future Work

The goal of this review article is to summarize how data mining can be a useful tool to collect information from online health communities. With the power of data mining, valuable results can be obtained to steer the current pharmaceutical field in the direction of patient-centered research to drive drug and device development toward medical interventions desired by the patient. These methods are currently involved in mining databases, like FAERS and VAERS, for information on adverse drug related events and vaccine uptake. Therefore, the suggested tools are relevant and applicable to be translated to online health communities. Potential techniques were proposed along with the challenges that will be faced and suggested augmentations that will require future work. Further research would include the implementation of these methods and techniques to generate a report from health data obtained from online communities. As a result of these endeavors, upcoming biomedical research would be fueled by patient-centered data.

References

[1] M.-G. Fayn, V. des Garets, and A. Rivière, "Collective empowerment of an online patient community: Conceptualizing process dynamics using a multi-method qualitative approach," *BMC Health Services Research*, vol. 21, pp. 1–19, 2021.

[2] M. Fallah and S. R. N. Kalhori, "Systematic review of data mining applications in patient-centered mobile-based information systems," *Healthcare informatics research*, vol. 23, no. 4, pp. 262–270, 2017.

[3] R. Alnashwan, A. O'Riordan, and H. Sorensen, "Multiple-perspective data-driven analysis of online health communities," in *Healthcare*, MDPI, vol. 11, 2023, p. 2723.

[4] Y. Mao, Q. Liu, and Y. Zhang, "Sentiment analysis methods, applications, and challenges: A systematic literature review," *Journal of King Saud University-Computer and Information Sciences*, p. 102 048, 2024.

[5] M. S. Jawad, W. Adi, A. Salem, and M. Doiher, "Implementation of data mining from social media for improved public health care," *Future Technologies Conference*, pp. 234–240, 2017.

[6] D. Isa, V. Kallimani, and L. H. Lee, “Using the self organizing map for clustering of text documents,” *Expert Systems with Applications*, vol. 36, no. 5, pp. 9584–9591, 2009.

[7] A. Bone and K. Houck, “The benefits of data mining,” *Elife*, vol. 6, e30280, 2017.

[8] U. Food and D. Administration, *Data mining at FDA*, U.S. Food and Drug Administration website, 2018.

[9] A. A. Dror *et al.*, “Vaccine hesitancy: The next challenge in the fight against covid-19,” *European journal of epidemiology*, vol. 35, no. 8, pp. 775–779, 2020.

[10] M. D. Rousculp *et al.*, “Burden and impact of reactogenicity among adults receiving covid-19 vaccines in the United states and Canada: Results from a prospective observational study,” *Vaccines*, vol. 12, no. 1, p. 83, 2024.

[11] M. A. Khder, “Web scraping or web crawling: State of art, techniques, approaches and application.,” *International Journal of Advances in Soft Computing & Its Applications*, vol. 13, no. 3, pp. 144–168, 2020.

[12] M. Marmot and R. Wilkinson, *Social determinants of health*. Oxford Academic, 2005.

Copyright (c) IARIA, 2025.   ISBN: 978-1-68558-247-0