Using text mining, analyse text messages and online health records

Dr. Savyasachi
[1]Assistant Professor, Department of Information Technology, L. N. Mishra College of Business Management, Muzaffarpur, Bihar

Abstract- Examining Online Health Data Using Text Mining Methods. Online exchange of health information is a different approach. To disseminate information on health-related topics, social network sites like Twitter, Facebook, Reddit, and online support groups for specific conditions are increasingly used. In order to receive counsel, this may need giving personal health information, or it may entail responding to questions from other patients based on their medical histories. This increase in social media usage offers a fresh perspective on how to use user-generated content from social networks to enhance the present landscape of health communication. These internet channels of contact allow health professionals to assist those looking for guidance more quickly. Non-profit organisations and federal agencies can also disperse preventative information in such networks for improved outcomes. Researchers studying health communication might extract information into patient experiences that may be difficult to obtain through conventional surveys by analysing user-generated content on social networks. Getting the signal from the noise is the main challenge in mining social health data. The informal nature of social data, typos, emoticons, tonal changes (such as sarcasm), and ambiguities resulting from polysemous words make it challenging to develop automated methods for extracting insights from such sources.

Keywords— *Analysis, Text Messages, Online Health Records, Text Mining, Health Communication.*

INTRODUCTION

Throughout the past ten years, social media platforms have expanded quickly, resulting in numerous important changes to people's daily lives. Online knowledge sharing and consumption have become widespread thanks to social media sites like Facebook, Twitter, and Instagram, as well as discussion boards like Reddit, Quora, and Stack Overflow. 76% of Americans utilise a social networking service online. 71% of online adults use Facebook, compared to 18% who use Twitter and 3% who use Instagram. One in four US teenagers and 39% of internet Black American kids use Twitter. With more than 100 million daily active users creating more than 140 million tweets since its start in 2006, Twitter has become one of the top ten websites on the Internet (Alexa, 2016). with more than 100 million daily active users creating more than 500 million tweets everyday (Twitter, Inc, 2013). Twitter users, often known as tweeters, can openly express their opinions through 140-character-limit tweets. By replying or utilising a hashtag to join the conversation, tweeters can also participate in real-time dialogues and discussions with other tweeters. Twitter is an asymmetric network; a user can only see the feeds of the users they follow, and they won't see the feeds of their followers until they follow them back. Reddit is a forum on the internet that was founded in 2005 and focuses on issues in news, music, sports, gaming, health, and entertainment. Reddit users, also known as Redditors, can publish text or Web links and receive feedback in the form of comments, upvotes, and downvotes. Reddit utilises an algorithm based on the Newton cooling method and up/down votes to rank posts. Reddit, as opposed to Twitter, allows for lengthier, more in-depth posts that can deliver more information in a better structured fashion. This essay will concentrate on a few subreddits (also known as sub-forums or topic-related discussion groups). Reddit manages the themes using subreddits while Twitter tracks them using hashtags. As the content of social networks gets more in depth, it has changed the way researchers conduct their research and patients seek health information. Everyday social network users create millions of posts. Compared with telephone/volunteer based survey methods to collect data, social network based data collection is an impossible mission for the human mind to analyze the posts manually. Identifying the different aspects of the themes that the posts describe, and evaluating the opinion of each aspect of the posts is computationally challenging. Researchers are developing automated methods to analyze these large, unstructured datasets. In this context machine learning, natural language processing (NLP), and statistical analyses present the possibility of utilizing this massive data for deriving insights from social streams.

ONLINE HEALTH RECORDS

As information is simply facts about anything or someone, health information is simply facts about something or someone's health. It includes the clinical context of the patient, including their medical history, diagnosis, allergies, current therapies, drug side effects, and lifestyle factors that may affect their health (e.g., exercise, smoking, drinking). Personal health records, such as electronic medical records, are typically stored securely and are only accessible to individuals and healthcare professionals. Patients now have a new avenue for information-seeking thanks to the increased popularity of social media platforms and online health forums. By sharing their personal story, and viewing others' posts, patients can have a deeper understanding of what they may face in the future; how to find a good health provider; and where to find support groups. As social media is not as regulated and is available for everyone to post their thoughts, we typically don't know who the posters are in real life or to what extent are they knowledgeable. The following concerns need to be considered:

- The trustworthiness of these posts
- Privacy issues
- Misleading information
- Incomplete data

These health-related data can be used to extract a variety of health-related information, including:

- Public beliefs, perceptions, and attitudes towards products, regulations.
- Latest trends in substance abuse and addiction.
- Factors associated with mental health concerns, such as suicidality, depression, and anxiety.

RELATED WORK

People are encouraged to publish their ideas and personal information on social media. A single post might not provide insightful health information, but millions of postings on the same subject show a numeric shift that results in a qualitative shift. According to numerous research, compiling millions of posts can reveal information about public health. Some significant public health surveillance examples include influenza detection (Aramaki et al., 2011) and infectious disease outbreaks (Choi et al., 2016). In (Brownstein et al., 2009) it is estimated 37%-52% of Americans seek health-related information on the Internet. Usually, inaccurate or irrelevant information is also available to the public, and it is crucial to identify which information is correct, especially when it comes to health-related information. Inaccurate information could potentially have a negative impact on our well-being. Nowadays, people prefer to use the Internet as a priority option to seek advice regarding their illnesses, drug use, and self-treatment. Chung studied the accuracy of online information regarding the safety of infants during sleep (Chung et al., 2012). The American Academy of Pediatrics has published recommendations for reducing the risk of sudden infant death syndrome (SIDS), suffocation, strangulation, entrapment, and other accidental sleep-related infant deaths. However, these recommendations are given as guidelines by health professionals containing medical jargon that cannot be easily understood by the general public without a related background in medicine. Therefore, people probably enter the keywords related to infant sleep safety into a search engine and may follow the suggestions listed in the search results. Chuang et al. (Chung et al., 2012) analyzed 1300 websites on infant safety sleep (13 keywords and first 100 websites for each). The overall proportion of accurate information is 43.5%, inaccurate information is 28.2%, and 28.4% irrelevant information. They also found that different data sources have huge differences. Government websites (.gov or .state) and organizational websites (.org) achieved the highest level of accuracy: 80.9% and 72.5%, respectively. On the contrary, blogs and personal web- sites had very low accuracy score: 25.7% and 30.3% respectively. Another finding was that different keywords brought different outcomes from as high as 82% accuracy to as low as 18% accuracy. This study shows that the Internet does provide an opportunity for people to seek health-related information, and patients need skills to identify what information is most accurate. On the other hand, the study also shows the quality of keywords is important for health-related information. We can see on line health information still has a long way to go to improve the accuracy of health information. It requires collaboration between health professionals, researchers, and Internet users. Besides static websites, social network sites also contribute a massive amount of health information. The study by Chou et al. (Chou et al., 2009) shows that in 2007, about 69% of American adults had Internet access. Among Internet users, 5% of them joined in an online support group. As health providers mainly focus on clinical outcomes, patients' mental issues usually aren't given enough attention. Through affliction of emotional and physical pain, patients may develop depression and suicidal thoughts. There are several studies on suicide prevention (Kavuluru et al., 2016; Luxton et al., 2012), many identifying the most important posts during the conversion which led to a sentiment shift. By analyzing millions of posts, we may find some patterns which can help health providers and patients' families help patients through difficult times. Monitoring sentiment change during a conversation and manual analysis is time- consuming and unrealistic. With the power of NLP and machine learning, sentiment analysis (also known as opinion mining) is used to classify the polarity of a given context automatically. More details on sentiment analysis will be discussed in section 2.1. By classifying the sentiment of posts during online communication, we can use topic modeling/statistical analysis to summarize and categorize what types of information patients will need for support, what types of support are most helpful (sharing personal stories, general support, information support, etc.), and how to attract patients and keep them in the conversation. The benefits of online health support forums such as Cancer Survivors Network (CSN), Lungevity, and Patients Like Me are immense. Although the numbers of users are far below the numbers of more general social networks such as Facebook and Twitter, online health support forums offer patients the chance to interact with others who have been diagnosed with the same diseases such as lung or breast cancer. As online health support forum members are patients, health providers, patients' families and friends, the posts made by these users are more accurate than blogs and personal websites; when some posts are recognized as inaccurate, other users will quickly move to correct them. Online health-related social media offers an abundance of information for patients, health providers, and researchers. Wicks et al. (Wicks et al., 2010) show that over 70% of Patients Like Me users think the site is "moderately" or "very helpful"; over 50% of patients found the site helpful for understanding the side effects of their treatments; and 42% of patients agreed that site had helped them to find another patient who can help them understand a specific treatment for their symptoms. This shows an opportunity that online health information and communication can provide a critical mass of useful information for different parties.

TEXT MINING

Many scientific fields, including statistics, computer science, linguistics, and library science, have contributed to the development of text mining. Text mining approaches deal with unstructured text and concentrate on automated analysis of textual data as a kind of natural language. Despite the lack of a universal definition for text mining, the general method of analysis is accepted. Text mining is also linked to Natural Language Processing (NLP), which is concerned with the study of natural languages. Software options are accessible for analysing social media applications due to the requirement of using automatic techniques for textual data analysis and extracting pertinent information. Text mining tools are used to identify and analyze posts, likes, followers in online social networks to explore people's reactions and behavior.

Moreover, it shows the variation in views and opinions regard different topics. The fundamental process of text mining includes data collection, preprocessing, content analysis, finding and integration.

METHOD FOR TEXT CLASSIFICATION

(i) Traditional Machine Learning- Traditional ML methods include naive Bayes, decision trees, k-nearest neighbors, logistic regression, and support vector machines(SVM). These algorithms are based on feature engineering where a set of discriminative handcrafted features is constructed to improve model performance. Although deep learning has increased in popularity recently, in most Kaggle competitions currently, competitors are still winning by traditional methods. Especially when data is structured, a human can find good feature representations to train ML models. On the contrary, deep learning is adept at finding features from unstructured data such as images, audio, video. Such models can extract the features that humans cannot easily understand, but are still meaningful to a machine. For a dataset with a few hundred to a couple of thousand training samples, traditional machine learning usually outperforms deep learning methods. Since a deep learning structure is more complex than a traditional method, it has larger parameter spaces it needs to search through and learn. A small dataset cannot fully tune parameters that are generally representative for a domain leading to poor generalization. Usually, the challenge in traditional machine learning is identifying an appropriate model and features while in deep learning it is to search for appropriate architectures.

(ii) Deep Neural Networks- In the sixties, a single layer neural network was introduced and was called a perceptron. The structure was one input layer, one hidden layer, and one output layer. During that time, the machines were simple and could not handle complex operations, until the 1980s, when the multilayer perceptron (MLP) was created by Rumelhart, Williams, and Hinton. In the figure. we can see that all the layers are fully connected, where each node in one layer is connected to all nodes in neighboring layers. This brings an issue when the network gets larger: the number of parameters increases drastically. For instance, with 1000 cells in the hidden layer connected with 1000 × 1000 elements in the input matrix, we need 109 weight parameters and 1000 bias parameters for a single hidden layer. Additional layers will further increase the parameter set size. This limited the size of each layer and the depth of the network due to computing constraints. In addition, this may also lead to overfitting and the network getting stuck in local maxima.

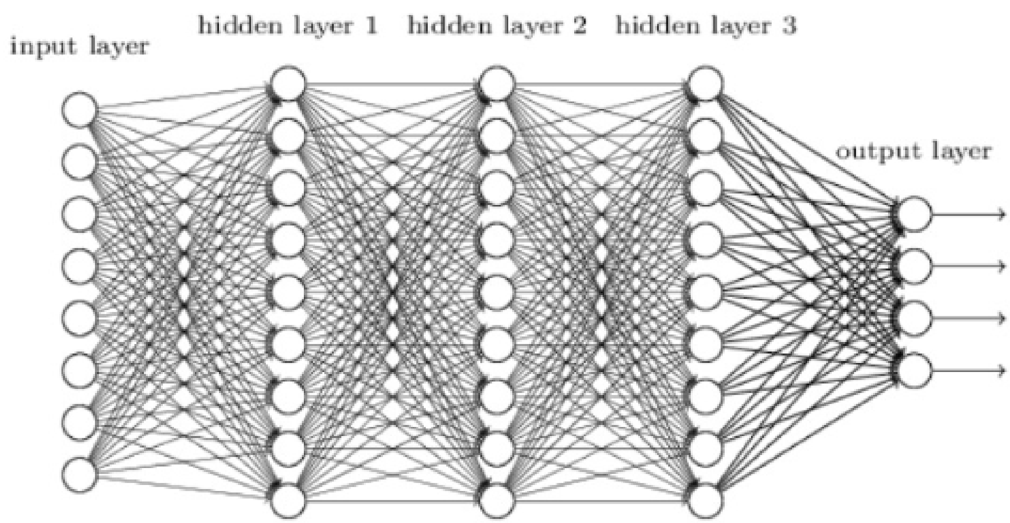

Figure 1- A feed forward deep neural network

(iii) Convolution Neural Network- An example CNN is as shown in Figure 2 for the task of text classification. The difference in this network is that convolution cells only connect with a part of the input cells. It extracts local information from the connected cells in the previous layer. For instance, in the image classification task, the first hidden layer can extract some curves by looking at a part of the pixel matrix, the second hidden layer might know the combination of curves and recognize them as part of the information of an object, and the third layer might know what the object could be and give a probability for each possible class. The purpose of an activation function is making output result from the convolutional layers compressed to a fixed real range so that the range of values that are input to the next layer is controllable; it also introduces non-linearity to extend network abilities to capture more complex functions. The common activation functions include sigmoid, ReLU, Leaky ReLU, Maxout, tanh.

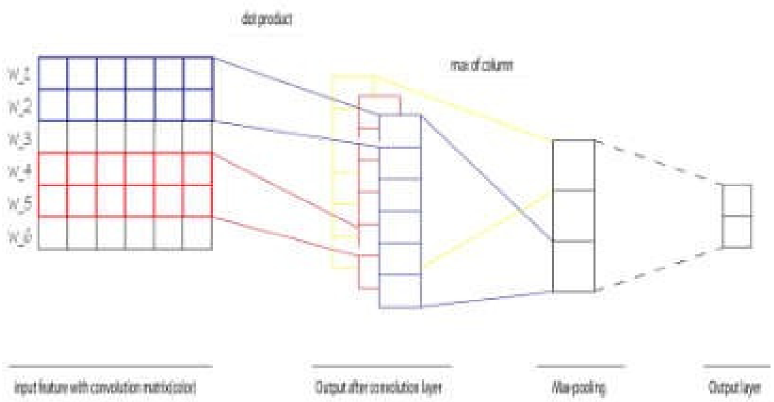

Figure 2- The CNN model with a binary output layer for text classification

TEXT MINING ALGORITHMS

The most algorithms used in analyzing the text in social networking are classification and clustering. Classification is a supervised learning that learn from training process a set of rules. The classification method comprises quantitative approaches to automate NLP to classify each text to a certain category.

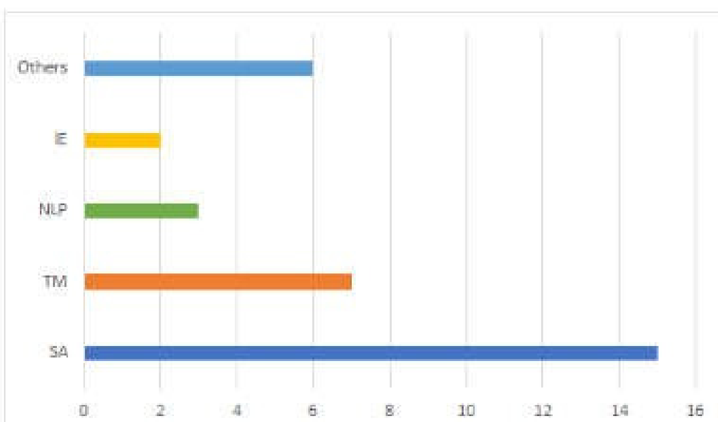

Figure 3- Most applied text mining techniques in online networking

The most common algorithms are K-Nearest Neighbour (KNN), Decision Trees (DT), Support Vector Machine (SVN), and Artificial neural networks (ANN). Clustering is an unsupervised algorithm that grouped the text in clusters. Different clustering techniques include different strategies that can be categorized in three types, naming, partitional, hierarchical, and semantic-based clustering. The studies in this paper cover a variety of text analysis algorithms. Most of the articles focus on classification or clustering. We noted that the number of articles in the dataset that employed clustering and classification algorithms increased recently. Clustering and classification are the most data mining techniques that extensively studied in the context of text.

CONCLUSION

Text mining applications have altogether influenced the inquire about in social systems investigation. Through investigating the inquire about in online social systems, 32 inquire about thinks about were analyzed to supply significant bits of knowledge on the approaches applied to progress decision-making completely different regions utilizing social media platforms. The survey uncovers finding that reply the inquire about questions, the foremost common text mining strategies were assumption examination and subject modeling. Clustering and classification are the foremost information calculations that broadly considered. As information preprocessing is basic step in content mining that might influence the precision of the investigation, the analysts are suggested to depict these steps in detail. The different nature of content information in social media postures numerous challenges, counting the gigantic volume of the information, commotion and etymology issues.

REFERENCES

[1] L. Sorensen, "User managed trust in social networking - Comparing Facebook, MySpace and Linkedin," 2009 1st International Conference on Wireless Communication, Vehicular Technology, Information Theory and Aerospace & Electronic Systems Technology, 2009, pp. 427-431.

[2] M. Naaman, "Social Multimedia: Highlighting Opportunities for Search and Mining of Multimedia Data in Social Media Applications," Multimedia Tools Appl, vol. 56, no. 1, pp. 9- 34, 2012.

[3] S. Rani and P.Kumar, "A sentiment analysis system for social media using machine learning techniques: Social enablement," Digital Scholarship in the Humanities, vol. 34, no. 4, 2018.

[4] D. Sudarsa, S. K. Pathuri, and L. J. Rao, "Sentiment Analysis for Social Networks Using Machine Learning Techniques," International Journal of Engineering and Technology(UAE), vol. 7, no. 2, pp. 473-476, 2018.

[5] K. Jani, M. Chaudhuri, H. Patel, and M. Shah, "Machine learning in films: an approach towards automation in film censoring," Journal of Data, Information and Management, vol. 2, pp. 55-64, 2019.

[6] N. Naw, "Twitter sentiment analysis using support vector machine and K-NN classifiers," Int. J. Sci. Res. Publ, vol. 8, no. 10, pp. 407–411, 2018.

[7] M. Grčar, D. Cherepnalkoski, I. Mozetič, and N. K. Petra, "Stance and influence of Twitter users regarding the Brexit referendum," Computational Social Networks, vol.4, no. 1, 2017.

[8] G. Piatetsky-Shapiro, "Data mining and knowledge discovery 1996 to 2005: overcoming the hype and moving from "university" to "business" and "analytics"." Data Min Knowl Disc, vol. 15, no. 1, pp. 99–105, 2007..

[9] J Cao, K. Basoglu, H. Sheng, and P. Lowry, "A Systematic Review of Social Networks Research in Information Systems: Building a Foundation for Exciting Future Research," Communications of the Association for Information Systems, vol 36, pp. 227-758, 2015.

[10] H. B. Haq, H. Kayani, S. K. Toor, A. Mansoor, and A. Raheem, "The Impact Of Social Media: A Survey," International Journal of Scientific & Technology Research, vol. 9, pp. 341-348, 2021.

[11] D. M. Boyd and N. B. Ellison, "Social network sites: Definition, history, and scholarship," Journal of Computer- Mediated Communication, vol. 13, no.1, pp. 210-230, 2007.

[12] R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins, "Trawling the Web for emerging cybercommunities," Computer Networks: The International Journal of Computer and Telecommunications Networking, vol. 31, no. 11-16, pp. 1481-1493, 1999.