Sentiment Analysis of HealthCare Big Data: Concept & Analysis

Fuchun Yang, University of Michigan – Ann Arbor

Abstract: We've entered an era in which, in developed countries, people's life expectancy can be as high as 80 years old. As more and more people care about how to increase their own longevity, people naturally are more aware of their own health conditions. With the development of technology, it is much easier to gain information than it was before. This article aims to use sentiment analysis to understand the polarity of posts on the subreddit health. 5000+ posts will be collected via reddit APIs. To do so, we can have a better understanding of what kind of information people care about on an international level.

Keywords: Healthcare, Sentiment analysis, data analysis, Python, R, Web scraping

1. Introduction

There are billions of new information generated every day. We talk about the new products we just bought with strangers online, providing our most authentic reviews, while brands spend a large amount of time cold-calling customers and begging for feedback. How to use the unstructured data in a more efficient way? This is where Natural Language Processing comes in. Natural language processing (NLP) has become one of the most popular topics in the recent few years. It has a wide application, including sentiment analysis, text extraction, text classification, etc. for example, the Siri on iPhone uses NLP. It can break down sentences into parts or other linguistic features. NLP can allow companies to handle a large amount of data and save much time and money. It also can structure the unstructured data sources and allow for further analysis [1-3]. As the data being generated everyday increases, NLP will only become more and more important.

This study mainly uses sentiment analysis, which is a technique in NLP that can be used to determine the positivity, negativity, and neutrality of data. For example, if a company launches a new product and it wants to know customers' reviews, the company can use NLP to collect posts from all social media platforms [1]. In this way, the reviews are authentic, and it saves much money for the company from cold-calling customers to receive feedback about what features customers are happy and unhappy with. Companies can also use sentiment analysis to understand the perception of their brands so that they can make business decisions. Overall, sentiment analysis has broad applications, including market intelligence, politics, entertainment, etc.

In the healthcare field, sentiment analysis can be very useful, too. Big data analytics in healthcare shows that up to 80 percent of healthcare data is unstructured, and therefore goes largely unutilized, since mining and extraction of this data is challenging and resource intensive [3]. This study aims to use web scraping and sentiment analysis in subreddit health. Even though this amount of data is only the tip of the iceberg, it still allows us to have a better understanding of people's attitudes and opinions on health-related topics.

2. Methodology

This project is separated into two parts. The first part is web scraping data from r/health, and the second part is sentiment analysis. In the first part, by using APIs and web scraping, 6000+ data points are collected from r/health and saved into a file [5]. In the second part, it is more about data processing and data analysis. Unnecessary columns, symbols, URLs are removed. Then tokenization, stemming, and text classification are used by using TextBlob on the processed data points for further sentiment analysis. Finally, several plots are created to give more intuitive information. Figure 1 shows the project workflow, and each step will be discussed in detail.

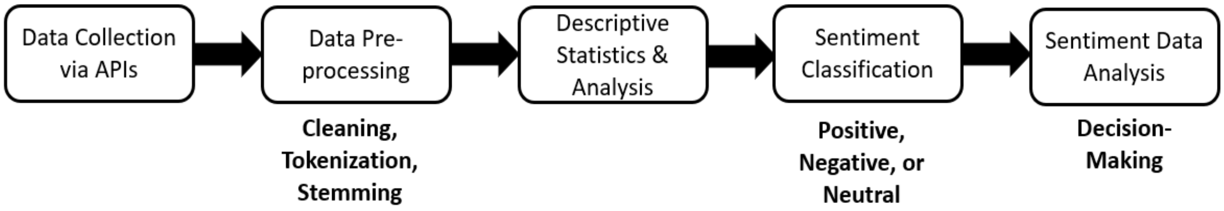
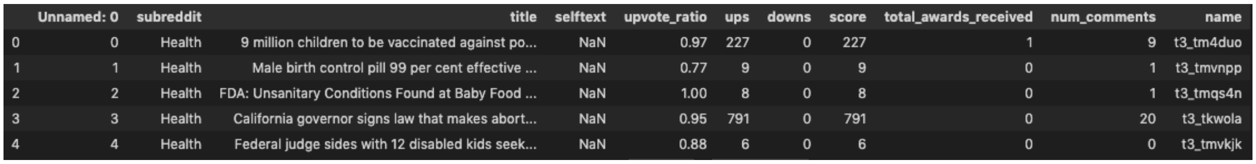


Figure 1. Project Workflow

3. Data Collection

A total of 6,473 best posts in the health subreddit were collected and stored in the csv file. Using the package requests in Python and personal private access tokens and keys for authentication, I was able to access Reddit APIs. The data includes subreddit, title, selftext, upvote_ratio, ups, downs, score, total awards received, number of comments, and name of the user. Figure 2 is a sample of the data.



	Unnamed: 0	subreddit	title	selftext	upvote_ratio	ups	downs	score	total_awards_received	num_comments	name
0	0	Health	9 million children to be vaccinated against po...	NaN	0.97	227	0	227	1	9	t3_tm4duo
1	1	Health	Male birth control pill 99 per cent effective ...	NaN	0.77	9	0	9	0	1	t3_tmvnpp
2	2	Health	FDA: Unsanitary Conditions Found at Baby Food ...	NaN	1.00	8	0	8	0	1	t3_tmqs4n
3	3	Health	California governor signs law that makes abort...	NaN	0.95	791	0	791	0	20	t3_tkwola
4	4	Health	Federal judge sides with 12 disabled kids seek...	NaN	0.88	6	0	6	0	0	t3_tmvkjk

Figure 2. Sample Dataset and Attributes

4. Data Pre-Processing & Analysis

The data collected from Reddit cannot be directly used for analysis because it has too many symbols and unnecessary details, so this step is to prepare the data. In this step, unnecessary symbols such as "#', "@", "?", etc. are removed from the original data. Then tokenization is applied, which is to separate a sentence into small tokens for sentiment analysis of single words. Stemming is also an important step, as it strips the suffixes (i.e., "ing", "ly", "es") from a word. For example, "play", "playing", "played" are different variations of the word "play." Finally, TextBlob is used to classify the text. TextBlob is a lexicon-based sentiment analyzer. It has predefined rules about words or sentences that helps to calculate the polarity. Sentiments are defined based on semantic relations and the frequency of each word in an input sentence [4]. TextBlob returns two properties, polarity and subjectivity. Polarity lies in the range of [-1,1], where 1 means positive statement and -1 means a negative statement. If the number is closer to 1, it means the posts are more positive; if the number is closer to -1, it means that posts are more negative. Subjective/objective sentences lie in the range of [0,1]. It generally refers to personal opinion, emotion or judgment whereas objectivity refers to factual information. The closer the number to 1, the more subjective opinions posts contain.

5. Results & Discussion

As shown in Table 1, the mean value of polarity is 0.03, and the mean value of subjectivity is 0.3, both two numbers are very low. Figure 3 gives a more complete picture of the polarity and subjectivity in this subreddit. For polarity, it mainly concentrates in the center but has a wide range of polarity. Most of the posts are within the range [-0.1, 0], suggesting that basically the posts can be slightly negative. Extreme positive or negative posts are very few. For subjectivity, it is more right skewed and most of the data concentrates in the left. This suggests that very positive and negative sentiments are very low, and most sentiments are factual information instead of subjective opinions.

Table 1. Descriptive Statistics for Sentiment Attributes

<table>
<tr><th colspan="6">Descriptive Statistics - Polarity</th></tr>
<tr><td>Min.</td><td>1st Qu.</td><td>Median</td><td>Mean</td><td>3rd Qu.</td><td>Max.</td></tr>
<tr><td>-0.60000</td><td>0.00000</td><td>0.00000</td><td>0.03104</td><td>0.08125</td><td>0.75000</td></tr>
<tr><th colspan="6">Descriptive Statistics - Subjectivity</th></tr>
<tr><td>Min.</td><td>1st Qu.</td><td>Median</td><td>Mean</td><td>3rd Qu.</td><td>Max.</td></tr>
<tr><td>0.0000</td><td>0.0000</td><td>0.2000</td><td>0.2648</td><td>0.5000</td><td>1.0000</td></tr>
</table>

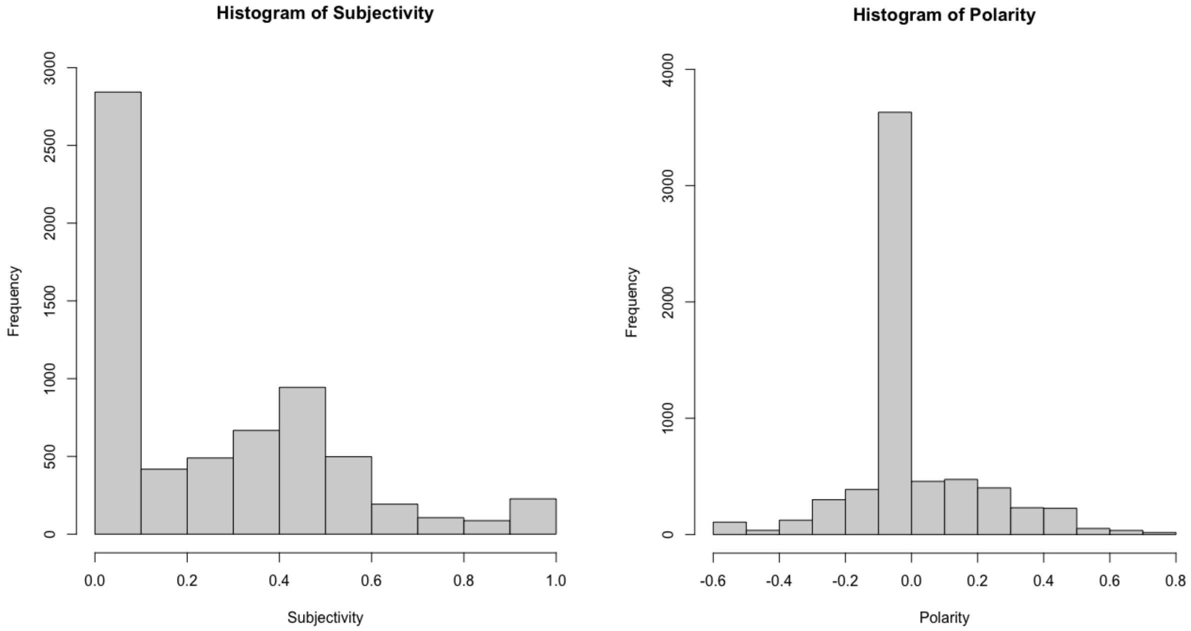


Figure 3. Histogram of Subjectivity and Polarity

Figure 4 presents the number of comments by polarity. As shown by the graph, posts that have a neutral perspective have the most comments. Posts with positive polarity [0,1] have more comments than posts with negative polarity [-1,0]. Both the increase and decrease of polarity are associated with decrease of comments, suggesting that people prefer to leave comments when the sentiments are neutral.

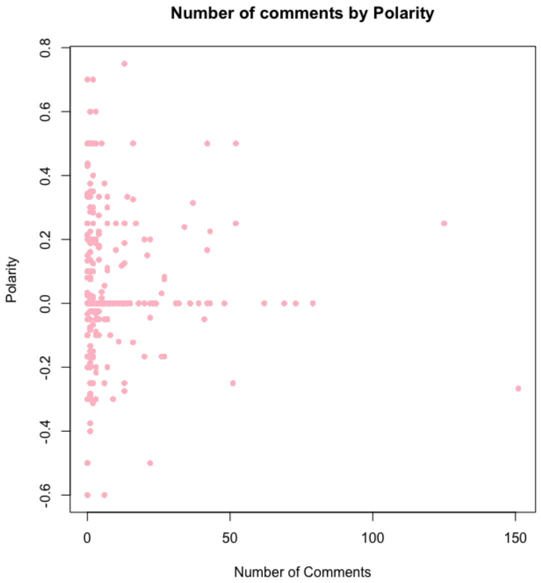


Figure 4. Number of Comments by Polarity

As shown by Figure 5, more than 50% of the data points are neutral, suggesting that people talk about health-related topics in a neutral way. About 27% of people express positive sentiments about health-related topics, mainly about new therapies effectiveness and food beneficial to health. 17% of users express negative sentiments. The topics include possible side effects of food, health risk due to plastics in food, shortage of physicians, etc.

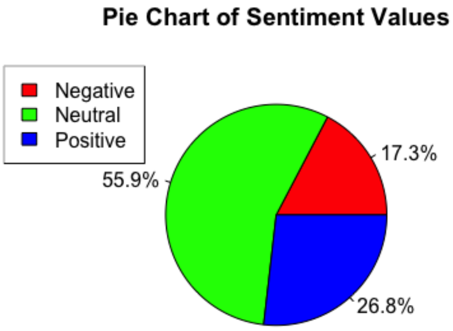


Figure 5. Pie Chart of Sentiment Analysis

Lastly, according to the wordcloud, covid is the most frequently mentioned item. Pandemic, Omicron, and vaccine also stand out. This is reasonable because for the recent two years, we have been living with the pandemic. Risk, brain, and disease also stand out. This shows that people care more about brain-related diseases and risks of health conditions. This confirms what we saw for subjectivity – most of the words in this subreddit are objective.

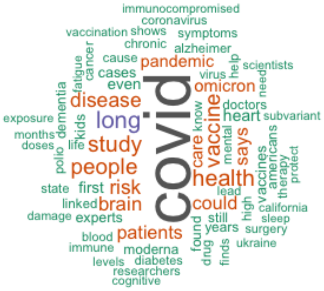


Figure 6. Wordcloud for Sentiments

6. Conclusion

To collect 6,000+ data points from r/health, the APIs and web scraping methods are used. After a series of data processing, TextBlob is used to calculate the polarity and subjectivity of data. Then the data are presented through digestible data visualizations. From the data visualization, several trends are very clear. First, for health-related topics in this subreddit, users post something that is neutral and objective. There is no clear evidence of extreme sentiments and subjectivity. Most of the posts are about discussions of factual information. Meanwhile, as the posts are more neutral, there are more objective factual-based comments. This study is helpful for us to have a sense of the discussion environment in this subreddit. Also, this study helps us to have a better idea of what people generally care about in health-related topics.

7. References

[1] Chowdhary, K.R. (2020). Natural Language Processing. In: Fundamentals of Artificial Intelligence. Springer, New Delhi. https://doi.org/10.1007/978-81-322-3972-7_19

[2] Feldman, R. (2013). Techniques and applications for sentiment analysis. *Communications of the ACM, 56*(4), 82–89. https://doi.org/10.1145/2436256.2436274

[3] Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal, 5*(4), 1093–1113. https://doi.org/10.1016/j.asej.2014.04.011

[4] Barai, M. K. (2021, October 26). *Sentiment analysis with textblob and vader in python*. Analytics Vidhya. Retrieved March 28, 2022, from https://www.analyticsvidhya.com/blog/2021/10/sentiment-analysis-with-textblob-and-vader/

[5] https://www.reddit.com/ Accessed on March 24, 2022