Sentiment Analysis of Health Tweets

Yu Han

12/27/2020

1. Introduction

2020 was a difficult year for everyone, health then became the most common concern in our lives and frequent topic in the tweets. But different people have different concerns about their health. The goal of this project is to know what type of health problem people are concerned about more and to check how negatively they think about their health using sentiment analysis methods.

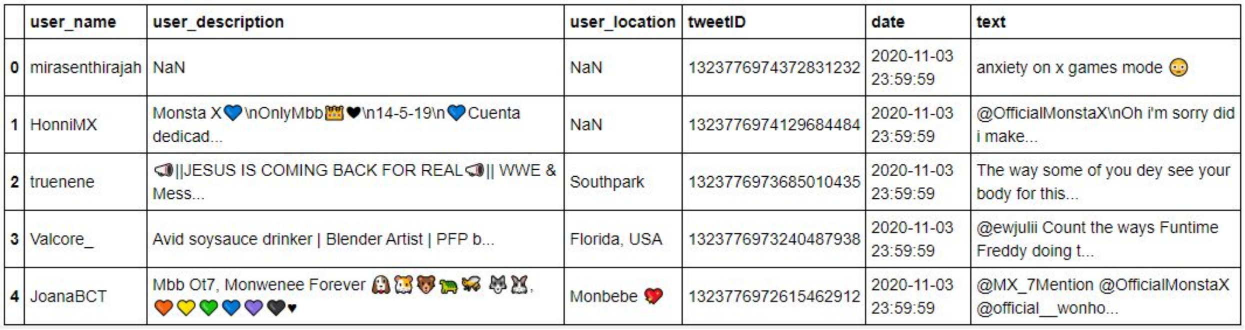
Hopefully, this project could help healthcare organizations identify which health problems people concern more now, then they could take corresponding measures.

2. Data Collection

The tweets data was scraped by using tweepy based on health keywords (link) from Twitter API. The data was collected from 2020-11-03 to 2020-11-10.

I extracted the following information from each tweet (shown in table 2.1):

- tweet.full_text: Text content of the tweet when API is told to pull all contents of tweets that have more than 140 characters
- tweet.text: Text content of tweet (I combined full_text and text to text)
- tweet.created_at: Date tweet was created
- tweet.id_str: Id of tweet
- tweet.user.screen_name: Username of tweet's author
- tweet.user.location: User defined location for the account's profile. Can be a nullable
- tweet.user.description: Text in user bio. Can be nullable



	user_name	user_description	user_location	tweetID	date	text
0	mirasenthirajah	NaN	NaN	1323776974372831232	2020-11-03 23:59:59	anxiety on x games mode 😳
1	HonniMX	Monsta X💙\nOnlyMbb👑🖤\n14-5-19\n💙Cuenta dedicad...	NaN	1323776974129684484	2020-11-03 23:59:59	@OfficialMonstaX\nOh i'm sorry did i make...
2	truenene	📢||JESUS IS COMING BACK FOR REAL📢|| WWE & Mess...	Southpark	1323776973685010435	2020-11-03 23:59:59	The way some of you dey see your body for this...
3	Valcore_	Avid soysauce drinker | Blender Artist | PFP b...	Florida, USA	1323776973240487938	2020-11-03 23:59:59	@ewjulii Count the ways Funtime Freddy doing t...
4	JoanaBCT	Mbb Ot7, Monwenee Forever 🐶🐹🐻🐢🐝 🐺🐰, 🧡💛💚💙💜🖤♥	Monbebe ❤️	1323776972615462912	2020-11-03 23:59:59	@MX_7Mention @OfficialMonstaX @official__wonho...

Table 2.1 Example of the first five rows in scraped tweets data.

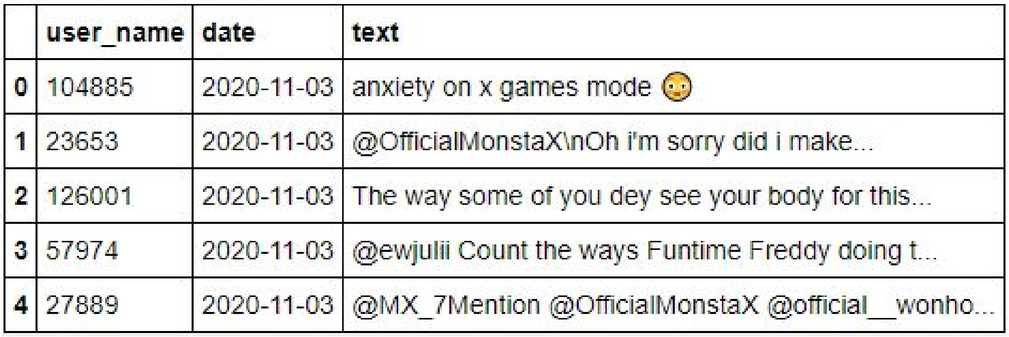
3. Data wrangling

3.1 Regular data cleaning

I first imported all CSV datasets and combined them into one dataframe using Python. Then I checked the missing data and observed nan value on user_description and user_location. I dropped user_description, user_location, and tweetID features because they are not useful for this project.

I then checked the duplicates and found out there is a big amount of duplicates in the data. That's probably because one tweet includes more keywords and can be extracted multiple times. I got 146,046 tweets after dropping the duplicate rows. The final amount of tweets should be enough to do the analysis.

I also changed the datatype of user_name from category to numeric. The dataset then looks like in table 3.1.1.

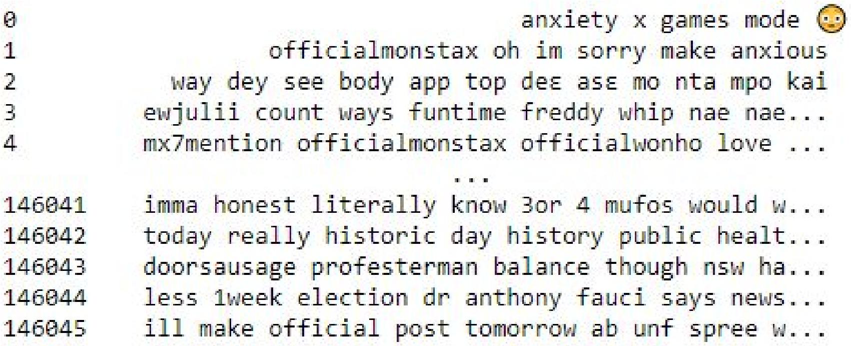


	user_name	date	text
0	104885	2020-11-03	anxiety on x games mode 😳
1	23653	2020-11-03	@OfficialMonstaX\nOh i'm sorry did i make...
2	126001	2020-11-03	The way some of you dey see your body for this...
3	57974	2020-11-03	@ewjulii Count the ways Funtime Freddy doing t...
4	27889	2020-11-03	@MX_7Mention @OfficialMonstaX @official__wonho...

Table 3.1.1 Example of the first five rows in scraped tweets data after regular cleaning.

3.2 Text data cleaning

I applied text data cleaning methods to make text all lower case, remove URLs, remove punctuation, remove common nonsensical text (\n), and remove stopwords. The final cleaning text data example in table 3.2.1.



```
0                                    anxiety x games mode 😳
1                     officialmonstax oh im sorry make anxious
2               way dey see body app top deɛ asɛ mo nta mpo kai
3                 ewjulii count ways funtime freddy whip nae nae...
4                 mx7mention officialmonstax officialwonho love ...
                                        ...
146041    imma honest literally know 3or 4 mufos would w...
146042    today really historic day history public healt...
146043    doorsausage profesterman balance though nsw ha...
146044    less 1week election dr anthony fauci says news...
146045    ill make official post tomorrow ab unf spree w...
```

Table 3.2.1 Example of scraped tweets data after text cleaning.

4. Exploratory Data Analysis (EDA)

4.1 Word frequency for all tweets

I first created a big bag of words from all tweets and created the WordCloud for all word frequency with that list (shown in figure 4.1.1). From the WordCloud figure, we can see that people talked about 'sore', 'sick', 'anxiety', 'pain' a lot. But some frequent words have very little meaning and could be added to a stop words list.

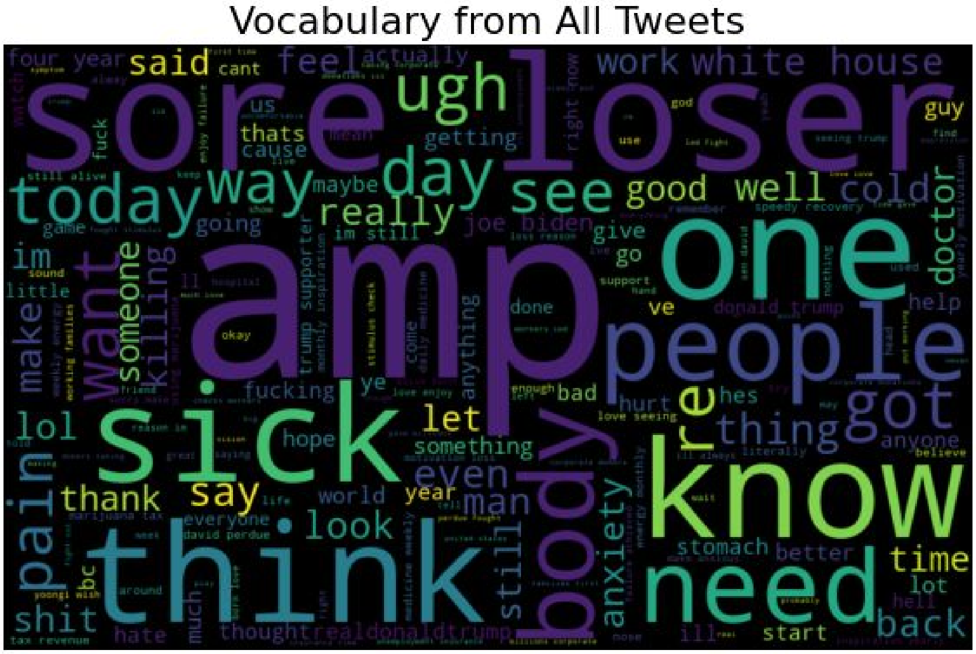


Fig.4.1.1 WordCloud for all word frequency in all tweets.

I also checked the frequency of each word in all tweets, found out that 10% of tweets include Trump, and 38% of tweets about Trump were related to the election. I then dropped those tweets including Trump and the election since I only concern about the health problem in this project.

I filtered the bag of words by removing the top most common words that count > 5000 and that is not in the health keywords list, then plotted the WordCloud again (shown in figure 4.1.2). After removing unmeaning words, we can see that our data looks more sense. And we can see that people are talking about 'sick', 'ill', 'body,' pain', 'anxiety', 'ill', 'hurts' more often.

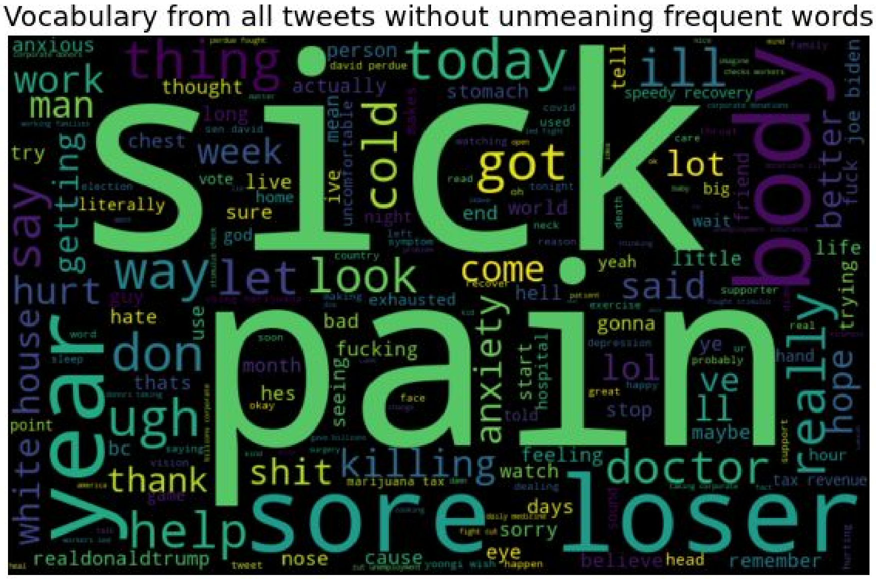


Figure 4.1.2 Vocabulary from all tweets without unmeaning frequent words.

4.2 Health word frequency for all tweets

Last, I want to know which health words people talk more. I calculated the frequency of each health word in the list from all tweets. Then I plotted the barplot for the frequency (shown in figure 4.2.1). We can see that people talk about 'sick', 'body', 'pain', 'ill', 'cold', 'sore', 'anxiety', 'killing', and 'vaccine' a lot.

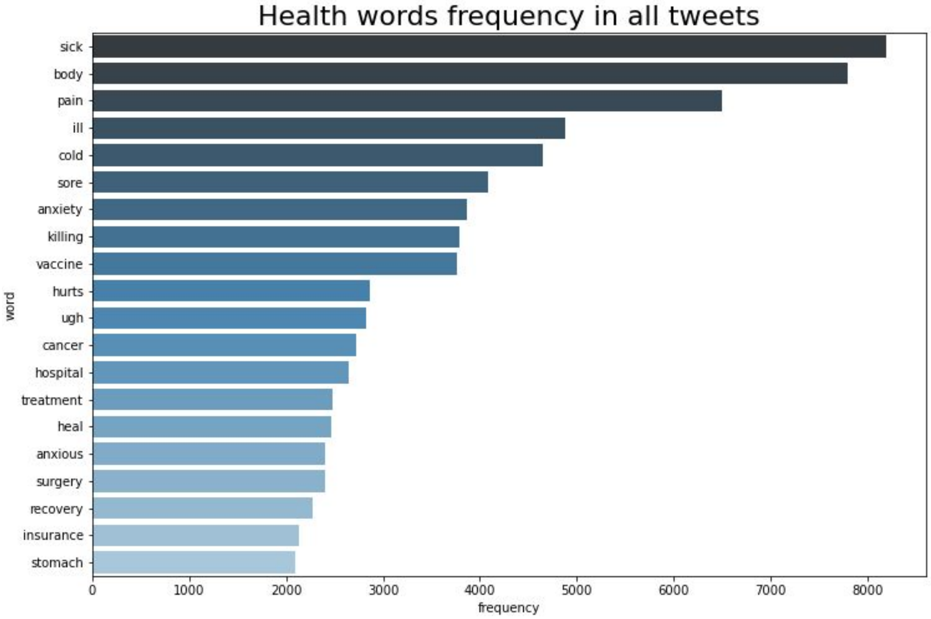


Figure 4.2.1 The frequency of each health word.

5. Sentiment Analysis

5.1 Sentiment labeling with textblob

Here, I want to know how people think when they are talking about health. Is it positive? Negative? Or neutral? I applied the TextBlob function from the textblob library for the sentiment analysis of each tweet. This method will give a polarity score for each tweet. Polarity scores tell us how positive or negative a word is. -1 means the tweet is very negative and +1 means the tweet is very positive. Then I labeled tweets based on polarity score to positive, negative, or neutral. It showed this dataset includes 54,172 positive tweets, 46,499 negative tweets, and 39,711 neutral tweets. The barplot for the sentiment count is shown in figure 5.1.1.

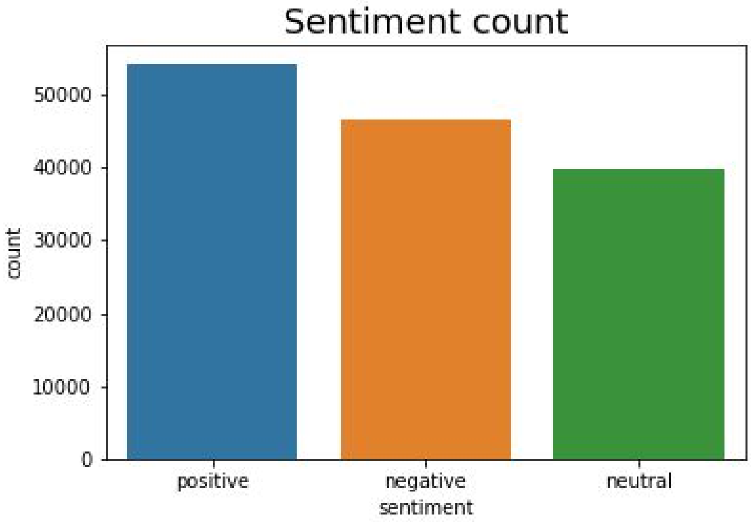


Figure 5.1.1 The tweets count for each sentiment.

5.2 Word frequency of tweets for each sentiment

I then plotted WordCloud for each sentiment. The most common words in negative tweets (shown in figure 5.2.1) are 'sick', 'cold', 'sore', 'anxious', 'ill' and some bad words, etc…

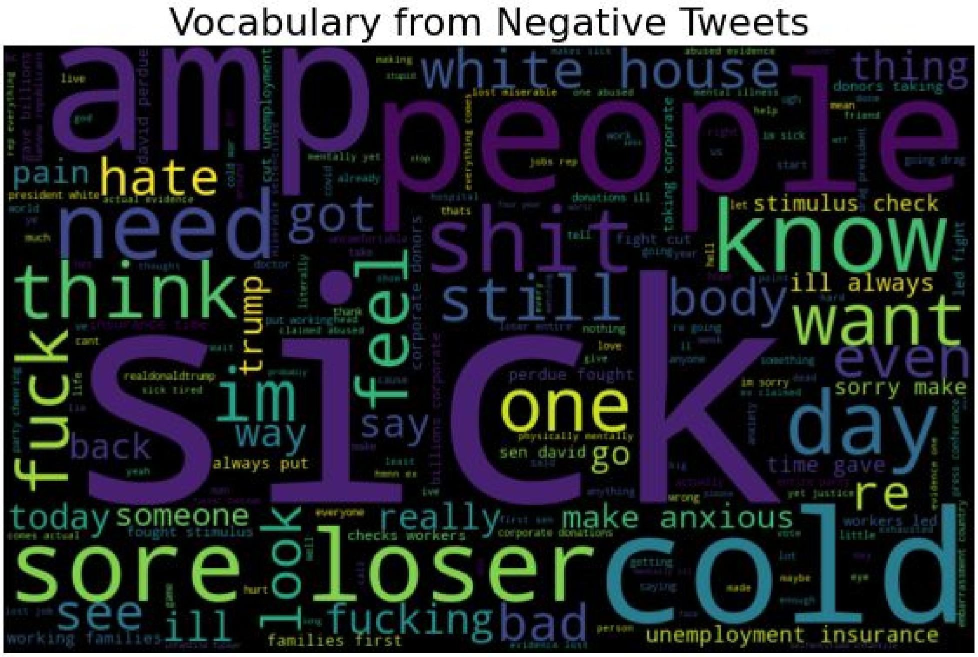


Figure 5.2.1 WordCloud for negative tweets.

Most common words in positive tweets (figure 5.2.2) are 'good', 'well', 'think', 'lol', and 'doctor', etc…

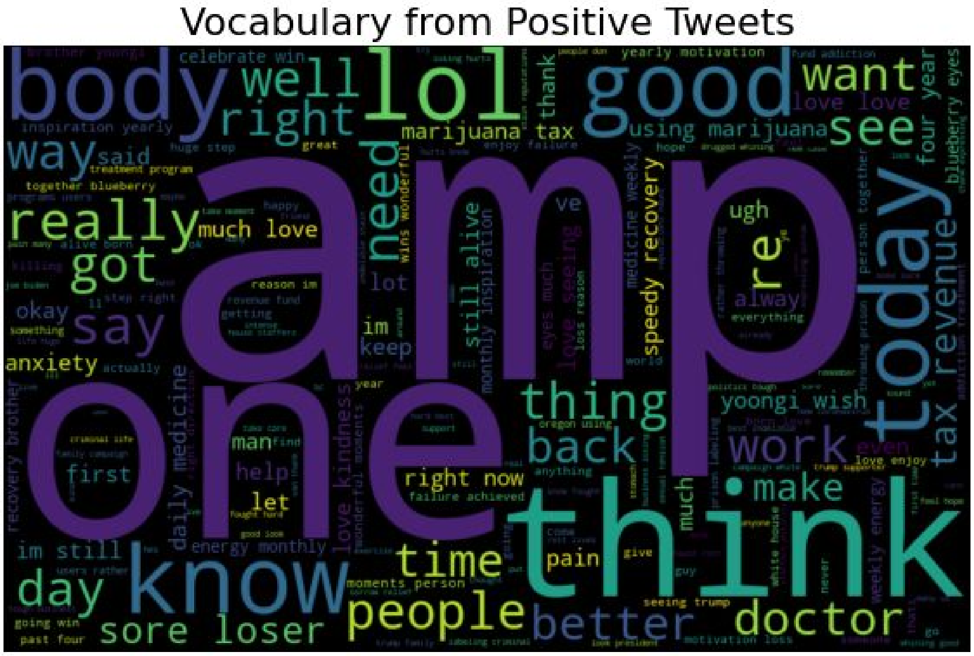


Figure 5.2.2 WordCloud for positive tweets.

Most common words in neutral tweets (figure 5.2.3) are 'pain', 'sore', 'body', 'anxiety', 'killing', and 'loser', etc... But they are all in the same frequency level.

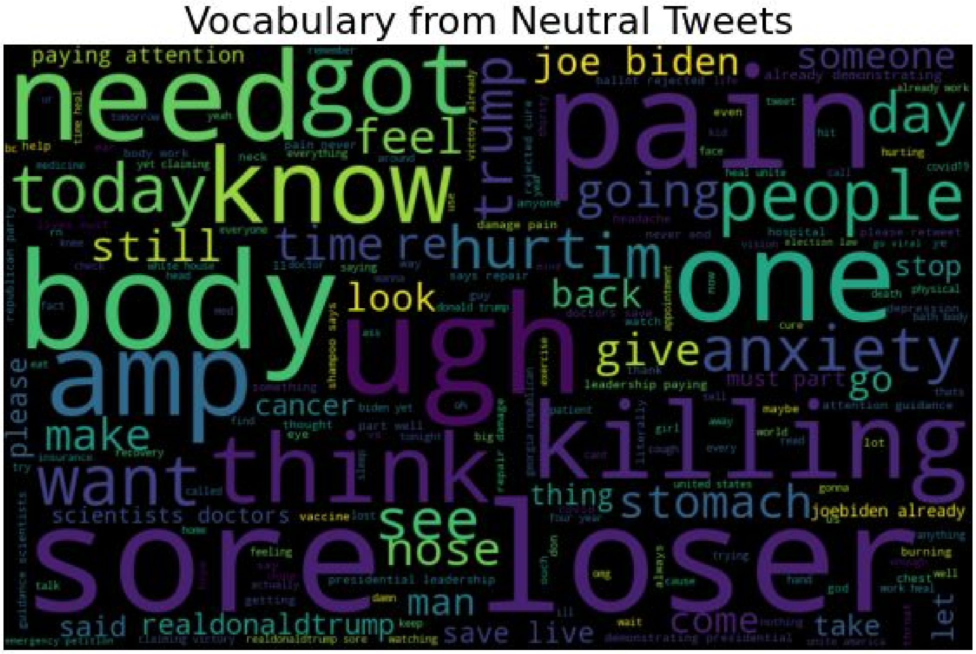


Figure 5.2.3 WordCloud for neutral tweets.

5.3 Tweets' length and word counts in each sentiment

I calculated the average of tweets' length and word counts for each sentiment. Positive tweets showed longer length and more words, negative tweets showed less and neutral tweets showed least (shown in figure 5.3.1 and 5.3.2).

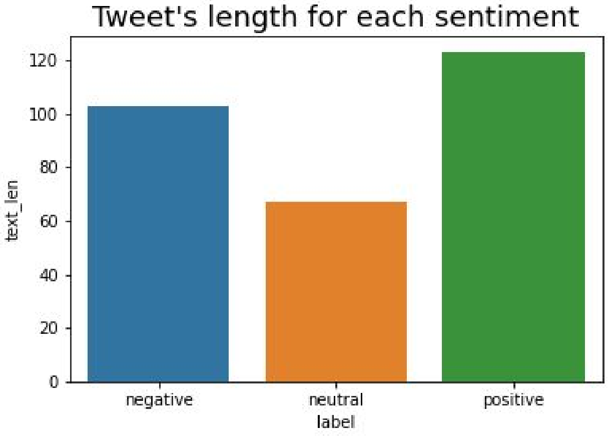


Figure 5.3.1 Tweets' length for each sentiment.

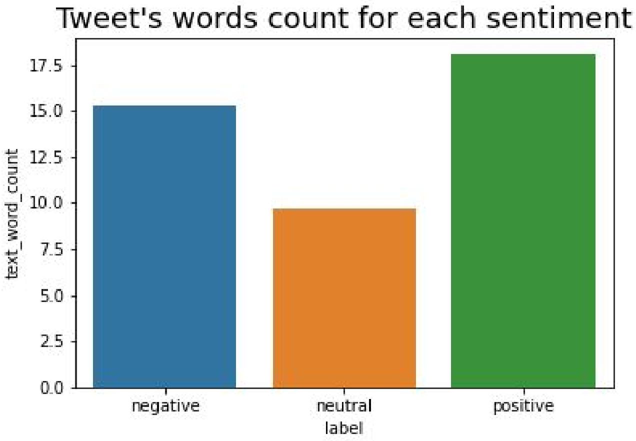


Figure 5.3.2 Tweets' words count for each sentiment.

5.4 Frequency of health words in each sentiment category

I plotted the most frequent health words in each sentiment category (shown in figure 5.4.1). We can observe that the frequency of health words in negative tweets is much higher than in positive and neutral tweets. And the most frequent health words in the negative tweets that are not in positive and neutral tweets are 'sick', 'ill', 'cold', 'anxious', 'exhausted', 'miserable', 'uncomfortable'.

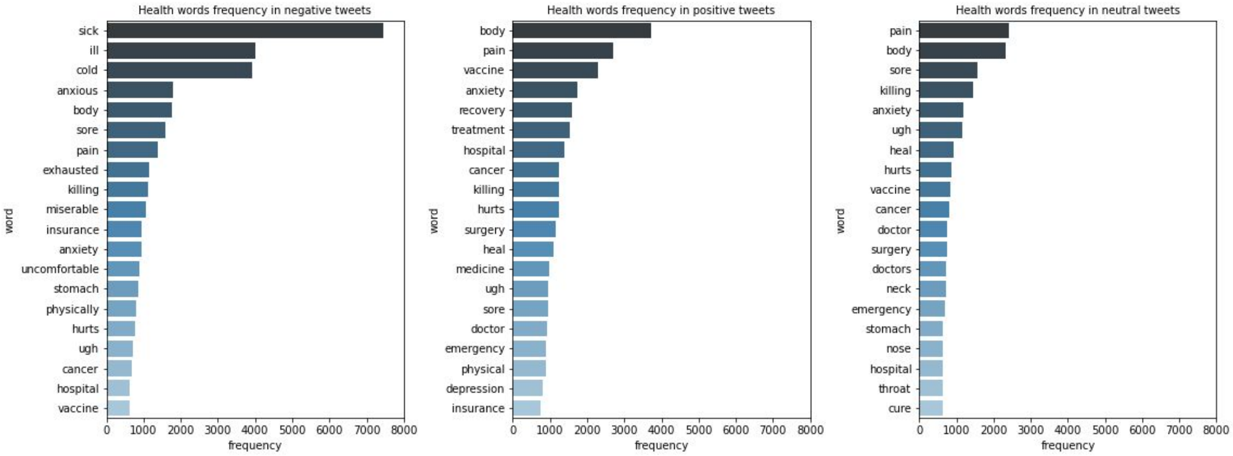


Figure 5.4.1 Frequency of health words of tweets in each sentiment.

Conclusion

In summary, we observe that people talk about 'sick', 'body', 'pain', 'ill', 'cold','sore', 'anxiety', 'killing', and 'vaccine' a lot. Most health tweets are positive. The most common words in negative tweets are 'sick', 'cold', 'sore', 'anxious', 'ill' and some bad words, etc… Most common words in positive tweets are 'good', 'well', 'think', 'lol', and 'doctor', etc… Most common words in neutral tweets are 'pain', 'sore', 'body', 'anxiety', 'killing', and 'loser', etc... But they are all in the same frequency level. In positive tweets, people tend to write more. People tend to write less in negative tweets and least in neutral tweets. The frequency of health words in negative tweets much higher than in positive and neutral tweets. And the most frequent health words in the negative tweets that are not in positive and neutral tweets are 'sick', 'ill', 'cold', 'anxious', 'exhausted', 'miserable', 'uncomfortable'.

Future work

In the future, I would like to collect more data in different periods to identify if my observations are consistent.