

Sentiment analysis of health tweets

Problem and Context:

2020 was a hard year for everyone, thus health became the most common concern in our lives. But different people have different concerns about their health. The goal of this project is to know what type of health people are concerned about more and apply the sentiment analysis to check how negative they think about their health.

Target Clients:

The primary clients this project targets are healthcare organizations. Hopefully, this project could help them identify which health problems people concern more now, then they could take corresponding measures.

Data:

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

I will extract the following information from each tweet:

- tweet.full_text: Text content of tweet when API is told to pull all contents of tweets that have more than 140 characters
- tweet.text: Text content of tweet (I will combine 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 account's profile. Can be nullable
- tweet.user.description: Text in user bio. Can be nullable

Approach:

Multiple steps will be taken to show the information of the data.

1. The scraped tweets data will be exported to multiple CSV files scraped on different days.
2. The multiple CSV files consisting of the data will be imported, merged, and cleaned via Python 3. I will check the missing values, duplicates, then apply text data cleaning to make text all lower case, remove URLs, remove punctuation, and remove stopwords.
3. Datasets will be explored visually via WordClouds and charts.
4. The sentiment analysis method will be applied to know how people are concerned about their health.

Deliverables:

The final draft of the project will be presented via PowerPoint slides. Jupyter Notebooks will be delivered detailing each step is taken and code written for the analysis of the project. A Github repository for the project will be created as well.