**Contents**

[1. Introduction 1](#_Toc76304560)

[2. Analysis Section 2](#_Toc76304561)

[2.1 About the data 2](#_Toc76304562)

[2.2 Data cleaning with VIS 2](#_Toc76304563)

[2.3 Data EDA with VIS 4](#_Toc76304564)

[2.4 Sentiment Analysis Methods 6](#_Toc76304565)

[2.4.1 VADER Method 7](#_Toc76304566)

[3. Results 7](#_Toc76304567)

[3.1 VADER Sentiment Analysis 7](#_Toc76304568)

[4. Conclusions 11](#_Toc76304569)

1. Introduction

On 31 December 2019, WHO was informed of cases of pneumonia of unknown cause in Wuhan City, China. A novel coronavirus was identified as the cause by Chinese authorities on 7 January 2020 and was temporarily named “2019-nCoV”.

Coronaviruses are a large family of zoonotic viruses that cause illness ranging from the common cold to severe respiratory diseases. Zoonotic means these viruses are able to be transmitted from animals to humans. There are several coronaviruses known to be circulating in different animal populations that have not yet infected humans. COVID-19 is the most recent to make the jump to human infection.

Common signs of COVID-19 infection are similar to the common cold and include respiratory symptoms such as dry cough, fever, shortness of breath, and breathing difficulties. In more severe cases, infection can cause pneumonia, severe acute respiratory syndrome, kidney failure, and death.

The COVID-19 infection is spread from one person to others via droplets produced from the respiratory system of infected people, often during coughing or sneezing. According to current data, time from exposure to onset of symptoms is usually between two and 14 days, with an average of five days.

日程表

描述已自动生成

Until now, two other recent coronavirus outbreaks have been experienced. Middle East Respiratory Syndrome (MERS-CoV) of 2012 was found to transmit from dromedary camels to humans. In 2002, Severe Acute Respiratory Syndrome (SARS-CoV) was found to transmit from civet cats to humans.

Although COVID-19 has already shown some similarities to recent coronavirus outbreaks, there are differences and specialists will learn much more as they deal with this one. SARS cases totaled 8098 with a fatality rate of 11 percent as reported in 17 countries, with the majority of cases occurring in southern mainland China and Hong Kong. The fatality rate was highly dependent on the age of the patient with those under 24 least likely to die (one percent) and those over 65 most likely to die (55 percent). No cases have been reported worldwide since 2004.

According to the World Health Organization (WHO), as of 2020, MERS cases total more than 2500, have been reported in 21 countries, and resulted in about 860 deaths. The fatality rate may be much lower as those with mild symptoms are most likely undiagnosed. Only two cases have been confirmed in the United States, both in May of 2014 and both patients had recently traveled to Saudi Arabia. Most cases have occurred in the Arabian Peninsula. It is still unclear how the virus is transmitted from camels to humans. Its spread is uncommon outside of hospitals. Thus, its risk to the global population is currently deemed to be fairly low.

1. Analysis Section
   1. About the data

The dataset includes the tweets about COVID-19 since January 2020. It is scraped from the twitter with its API keys. The raw data contains ﻿41157 tweets with the information of users, time, tweet location, and tweets. The objective is to perform the sentiment analysis using the data and find out the public attitudes towards the sudden disease.

* 1. Data cleaning with VIS

Before cleaning and preparing, the raw dataset looks as below.

文本

描述已自动生成

﻿It contains punctuations, symbols, blank, and the like. But what necessary is just those words. Therefore, use ﻿tokenizer to tokenize the tweets. RegexpTokenizer(r'\w+') can extract words from the text and create a list for them.

文本

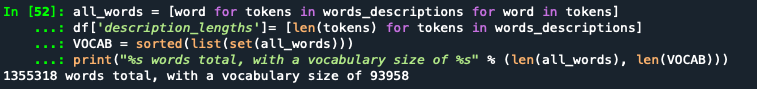
描述已自动生成

﻿Besides tokenization, vectorization can also be deployed to clean and prepare the dataset. First, create an empty list. Then add all the context of tweets to the list, which will be used as the input of Vectorizer. For ﻿CountVectorizer, set input equals to “content”, lowercase=true, and ﻿stop\_words = “English”. Now, ﻿a sparse matrix can be created with it. Next, ﻿convert the matrix to a regular array and get the list of ﻿vocabulary with ﻿get\_feature\_names() of the Vectorizer. With one word per column and one tweet per row, a pandas data frame can be got.

电脑萤幕画面

中度可信度描述已自动生成

﻿What’s more, the number of total words and unique words can be calculated with ﻿tokens. There are ﻿1355318 words from these tweets and ﻿93958 of them are unique.



The most common words can also be counted with ﻿Counter. Among the top 50 popular words, excluding common words, like “the”, “to”, etc., the most frequent words include “﻿coronavirus”, “﻿COVID”, “﻿prices”, “﻿supermarket”, “﻿grocery”, and “﻿store”. Therefore, during the pandemic, people were very concerned about food and daily supplies.

文本

中度可信度描述已自动生成 文本

中度可信度描述已自动生成

* 1. Data EDA with VIS

图表, 条形图

描述已自动生成

The bar chart shows the most common words from the collected tweets. Words like “the”, “for”, “to” don’t have realistic meaning, so more attention should be paid on “coronavirus” and “price”. Among 41157 tweets, “coronavirus” appears 14237 times and “price” appears 7300 times. Therefore, the users cared much about whether the price of things like daily necessities would be greatly affected by the epidemic.

图表, 树状图

描述已自动生成

The above interactive plot shows the common words in another way. The square area occupied by each word represents the frequency of the word. When the mouse is put on the box, the specific number of count will appear. Like what shows in the plot, the word “the” as the most frequent words from 41157 tweets appears 40344 times.

Then, a word cloud of common words is plotted.

报纸上的文字

描述已自动生成

The word cloud shows that users would like to discuss the topic around “supermarket”, “online shopping”, “grocery store”, and “toilet paper” on the twitter, which reflects that people have put great attention on how the daily life would be influenced by the pandemic, such as whether the food or daily necessities would be in short supply.

From the analysis part, a sentiment analysis with VADER method will be performed. So tweets will be classified into three categories: “positive”, “negative”, and “neutral”.

图表, 条形图

描述已自动生成

The plot displays the analysis result that nearly half of tweets held a positive attitude towards COVID-19, while another half held the opposite view. Only nearly 7500 tweets considered it as neutral. More information will be present in the analysis section.

* 1. Sentiment Analysis Methods

Sentiment Analysis also known as Opinion Mining, which is a sub-field of Natural Language Processing (NLP), aims at identifying and extracting opinions through sentiments, attitude, and emotions of the writer within a given text, such as tweets, reviews on shopping website, customer emails, social media messages and comments, surveys, etc.

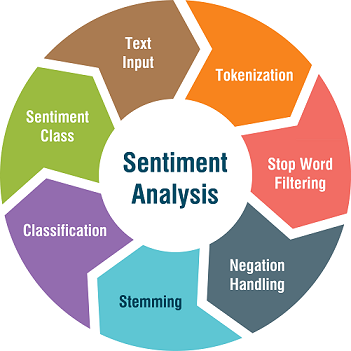
There are primarily 3 ways:

1) Rule-based methods

2) Feature-based methods

3) Embedding-based methods

Rule based are the methods which involve set of manually crafted rules to identify subjectivity, polarity, or the subject of an opinion. These involve techniques like Stemming, tokenization, part-of-speech tagging, parsing, Lexicons (i.e. lists of words and expressions). Generally, two lists of polarized words (e.g. negative words such as bad, worst, ugly, etc and positive words such as good, best, beautiful, etc) are prepared. The number of positive and negative words that appear in a given text are counted. If the number of positive word appearances is greater than the number of negative word appearances, the system returns a positive sentiment, and vice versa. If the numbers are even, the system will return a neutral sentiment. Methods involve AFINN, Bing Liu’s lexicon, MPQA subjectivity lexicon, SentiWordNet, TextBlob & VADER.



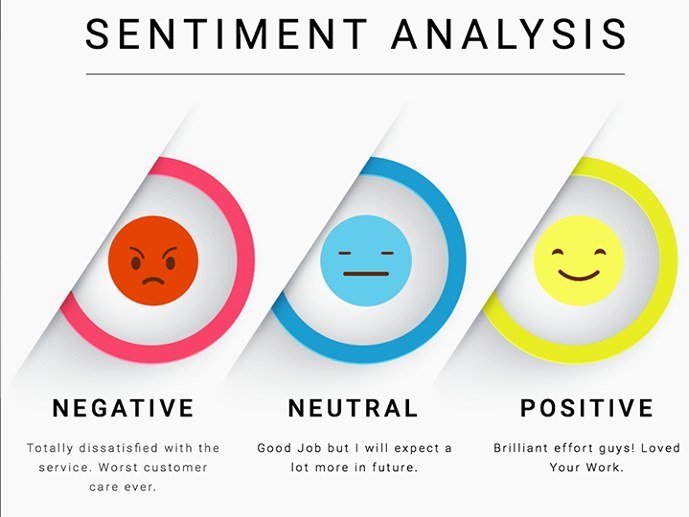
Since rule-based are generally naive in nature as don’t take into account how words are combined in a sequence. Hence, feature-based methods relying on ML techniques like SVM, Decision trees, etc. are used where sentiment analysis is modeled as a classification problem. The first step in a machine learning text classifier is to transform the text extraction or text vectorization, and the classical approach has been bag-of-words with their frequency. More recently, new feature extraction techniques have been applied based on word embeddings (also known as word vectors). This kind of representations makes it possible for words with similar meaning to have a similar representation, which can improve the performance of classifiers

Last is embedding based which involve FastText & Flair

* + 1. VADER Method

VADER (Valence Aware Dictionary for Sentiment Reasoning) is a rule-based method used for text sentiment analysis that is sensitive to both polarity (positive/negative) and intensity (strength) of emotion. It is available in the NLTK package and can be applied directly to unlabeled text data.

VADER sentimental analysis relies on a dictionary that maps lexical features to emotion intensities known as sentiment scores. The sentiment score of a text can be obtained by summing up the intensity of each word in the text. For example, words like ‘love’, ‘enjoy’, ‘happy’, ‘like’ all convey a positive sentiment. Also VADER is intelligent enough to understand the basic context of these words, such as “did not love” as a negative statement. It also understands the emphasis of capitalization and punctuation, such as “ENJOY”.



For the advantages of using VADER, it makes a lot of things easier. For example, it does not require any training data, it can understand the sentiment of a text containing emoticons, slangs, conjunctions, capital words, punctuations and much more very well, and it works excellent on social media text. What’s more, VADER can work with multiple domains.

1. Results
   1. VADER Sentiment Analysis

First, import vaderSentiment and call SentimentIntensityAnalyzer object. Next, create a function, which uses polarity scores for knowing the polarity of each text. Then, take some tweet examples to test the function.

文本

描述已自动生成

Here neg (negative), neu (neutral), and pos (positive) represent the proportion of text falling under each category and this proportion will always sum up to 1 (e.g., 0.656 + 0.344 in first tweet).

Compound score is what reflects the overall score. It is the sum of all lexicon ratings which is normalized between -1 (most extreme negative) and +1 (most extreme positive). As per the scoring document ([link](https://github.com/cjhutto/vaderSentiment#about-the-scoring)) it is called as ‘normalized, weighted composite score.’

The typical thresholds standardized for classifying sentences as positive, negative, neutral are:

1) positive sentiment: compound score >= 0.05

2) neutral sentiment: (compound score > -0.05) and (compound score < 0.05)

3) negative sentiment: compound score <= -0.05

However, they can be adjusted for highly positive, positive, neutral, negative and highly negative classes as per the requirements. Besides, VADER generates scores based on emojis, slangs, emoticons, punctuations, etc.

After checking the score function, apply it to the data frame prepared before and adding the result to it as a new column called “scores”.

文本

描述已自动生成

Extract the compound scores out as a new column called “compound”, which will be taken as the final score to classify the sentiment. With everything prepared, the sentiment function can be generated using the standard introduced above and apply it to the compound scores. The classification result will be stored as a column called “sentiment” in the data frame.

图表, 条形图

描述已自动生成

The bar chart shows that nearly half of twitter users took a positive attitude towards COVID-19, while the other half took a negative one. To make the classification clearer, a funnel chart can be plotted as below.

图表, 漏斗图

描述已自动生成

It displays that 44.4% users’ tweets related with the pandemic are positive, while 37.4% of them are negative. Only 18.2% twitter users held a neutral standpoint.

To be more specific, a word cloud about positive tweets is plotted.

文本

描述已自动生成

From the positive word cloud, “grocery store”, “toilet paper”, “food” are very eye-catching. They may reflect that some residents were confident about the supply of daily necessities or food during the epidemic. Words like “oil price”, “demand”, “market” are also striking in the plot, so people may also take positive views on these things believing that the market would recover from the blow of the pandemic.

文本

描述已自动生成

From the negative word cloud, “grocery store”, “supermarket”, “panic” are also very eye-catching. They may reflect that for negative tweets users were still concerned with the supply of daily necessities or food during the epidemic. Compared with the positive word cloud, words like “oil price” and “market” do not show up in the negative plot, so people didn’t take negative views on these things.

文本

描述已自动生成

From the neutral word cloud, “grocery store” and “online shopping” are very striking. They may reflect that some residents did not exclude online shopping to supplement the necessities at home. Besides, “hand sanitizer” becomes obvious in this plot, which means that people paid great attention to the prevention of virus infection

1. Conclusions

For the topic about COVID-19, users of twitter cared more about how the daily life would be influenced by the pandemic, such as whether the food or daily necessities would be in short supply and whether the “oil price” would increase.

44.4% users’ tweets are positive, which can be concluded that some residents were confident about the supply of daily necessities and food during the epidemic and took positive views on the price believing that the market would recover from the epidemic.

37.4% of users’ tweets are negative and words like “grocery store”, “supermarket”, and “panic” are also very common among these tweets. Therefore, some people were still concerned with the situation that the supermarket may run out of products such as “toilet paper”, which also explains why many people went to supermarkets to buy a lot of daily necessities in the early days of the outbreak.

Only 18.2% twitter users held a neutral attitudes towards COVID-19.

图片包含 小, 桌子, 盒子, 年轻

描述已自动生成

Therefore, the best way for the government to appease people’s anxiety during the period of stay-at-home order is to ensure the sufficient supply of necessities so that the people can feel at ease.