03_AM_Sentiment_Analysis_Vaccine_Tweets

July 13, 2021

[1]:
```
import pandas as pd
import numpy
```

1 Sentiment analysis

- Preprocessing
  - remove URLs
  - remove Stopwords
  - Lemmatization
- Sentiment Analysis with nltk

[2]:
```
clean_vaccine_tweets = pd.read_csv("../data/interim/cleaned_vaccine_tweets.
↪csv", index_col=0)
```

[3]:
```
clean_vaccine_tweets.head()
```

[3]:
```
                    id                  created_at  \
0  1338158543359250432  2020-12-13 16:27:13+00:00
1  1337840331522453504  2020-12-12 19:22:45+00:00
2  1338544403795881984  2020-12-14 18:00:29+00:00
3  1337735595704115200  2020-12-12 12:26:34+00:00
4  1337850832256176128  2020-12-12 20:04:29+00:00

                                                user  geo  \
0  {'id': 76052772, 'id_str': '76052772', 'name':…  NaN
1  {'id': 1300382181605494800, 'id_str': '1300382…  NaN
2  {'id': 1164717209253552000, 'id_str': '1164717…  NaN
3  {'id': 1316036067754205200, 'id_str': '1316036…  NaN
4  {'id': 1110032180237852700, 'id_str': '1110032…  NaN

                                           full_text  \
0  While the world has been on the wrong side of …
1  @cnnbrk #COVID19 #CovidVaccine #vaccine #Coron…
2  The FDA Authorizes Emergency Use Of The Pfizer…
3  The #FDA finally issues #EUA now comes the pro…
4  There have not been many bright days in 2020 b…
```

```
                                              hashtags             user_id  \
0  ['covid19', 'supplychain', 'logistics', 'vacci…             76052772
1  ['covid19', 'covidvaccine', 'vaccine', 'corona…  1300382181605494800
2  ['pfe', 'pfizer', 'pfizervaccine', 'pfizerbion…  1164717209253552000
3     ['fda', 'eua', 'pfizerbiontech', 'vaccinated']  1316036067754205200
4  ['bidenharris', 'election2020', 'pfizerbiontec…  1110032180237852700

   PfizerBiontech  SputnikV  Sinopharm  Sinovac  Moderna  AstraZeneca  \
0               1         0          0        0        0            0
1               1         0          0        0        0            0
2               1         0          0        0        0            0
3               1         0          0        0        0            0
4               1         0          0        0        0            0

   Covaxin  JandJ user_location coordinates
0        0      0           NaN         NaN
1        0      0           NaN         NaN
2        0      0           NaN         NaN
3        0      0           NaN         NaN
4        0      0           NaN         NaN
```

---

[4]:
```
import re
import nltk
from nltk.corpus import stopwords
from nltk.stem.porter import PorterStemmer
from nltk.sentiment import SentimentIntensityAnalyzer
```

2 NLP

2.1 Preprocessing

[5]:
```
clean_vaccine_tweets["corpus"] = ""
```

[6]:
```
nltk.download('stopwords')
nltk.download('wordnet')
nltk.download('vader_lexicon')
```

```
[nltk_data] Downloading package stopwords to /Users/ayman/nltk_data…
[nltk_data]   Package stopwords is already up-to-date!
[nltk_data] Downloading package wordnet to /Users/ayman/nltk_data…
[nltk_data]   Package wordnet is already up-to-date!
[nltk_data] Downloading package vader_lexicon to
[nltk_data]     /Users/ayman/nltk_data…
[nltk_data]   Package vader_lexicon is already up-to-date!
```

[6]: True

Set lowercase, remove punctuation, remove url: - “no”,“nor”,“not” removed from stopwords because they may be relevant for sentiment - add “vaccine” to stopwords because it has little information value

[7]:
```
def clean_dataset(dataset):
    for i in range(0, len(dataset)):
        #Tokenize and set words to lowercase
        review = dataset["full_text"][i]
        review = ' '.join(re.sub("(@[A-Za-z0-9]+)|([^0-9A-Za-z \t])|(\w+:\/\/
↪\S+)"," ",review).split())
        review = re.sub("[^a-zA-Z]", " ", review)
        review = review.lower()
        review = review.split()

        #stopwords:
        all_stopwords = [word for word in stopwords.words("english") if word␣
↪not in ["no","nor","not"]]
        all_stopwords.
↪extend(["a","b","c","d","e","f","g","h","i","j","k","l","m","n","o","p","q","r","s","t","u"
↪","&amp;","nhs","vaccine","covidvaccine"])
        #lemmatization:
        lemma = nltk.wordnet.WordNetLemmatizer()
        review = " ".join([lemma.lemmatize(word) for word in review if word not␣
↪in set(all_stopwords)])
        dataset["corpus"][i] = review
```

[8]:
```
clean_dataset(clean_vaccine_tweets)
```

```
<ipython-input-7-424768f53dcf>:16: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  dataset["corpus"][i] = review
```

[9]:
```
clean_vaccine_tweets.head()
```

[9]:

	id	created_at	\
0	1338158543359250432	2020-12-13 16:27:13+00:00	
1	1337840331522453504	2020-12-12 19:22:45+00:00	
2	1338544403795881984	2020-12-14 18:00:29+00:00	
3	1337735595704115200	2020-12-12 12:26:34+00:00	
4	1337850832256176128	2020-12-12 20:04:29+00:00	

	user	geo	\
0	{'id': 76052772, 'id_str': '76052772', 'name':…	NaN	
1	{'id': 1300382181605494800, 'id_str': '1300382…	NaN	
2	{'id': 1164717209253552000, 'id_str': '1164717…	NaN	

```
3  {'id': 1316036067754205200, 'id_str': '1316036…  NaN
4  {'id': 1110032180237852700, 'id_str': '1110032…  NaN

                                          full_text  \
0  While the world has been on the wrong side of …
1  @cnnbrk #COVID19 #CovidVaccine #vaccine #Coron…
2  The FDA Authorizes Emergency Use Of The Pfizer…
3  The #FDA finally issues #EUA now comes the pro…
4  There have not been many bright days in 2020 b…

                                          hashtags              user_id  \
0  ['covid19', 'supplychain', 'logistics', 'vacci…             76052772
1  ['covid19', 'covidvaccine', 'vaccine', 'corona…  1300382181605494800
2  ['pfe', 'pfizer', 'pfizervaccine', 'pfizerbion…  1164717209253552000
3     ['fda', 'eua', 'pfizerbiontech', 'vaccinated']  1316036067754205200
4  ['bidenharris', 'election2020', 'pfizerbiontec…  1110032180237852700

   PfizerBiontech  SputnikV  Sinopharm  Sinovac  Moderna  AstraZeneca  \
0               1         0          0        0        0            0
1               1         0          0        0        0            0
2               1         0          0        0        0            0
3               1         0          0        0        0            0
4               1         0          0        0        0            0

   Covaxin  JandJ user_location coordinates  \
0        0      0           NaN         NaN
1        0      0           NaN         NaN
2        0      0           NaN         NaN
3        0      0           NaN         NaN
4        0      0           NaN         NaN

                                              corpus
0  world wrong side history year hopefully bigges…
1  covid corona pfizerbiontech bbcnews nytimes bb…
2  fda authorizes emergency use pfizer pfe pfizer…
3  fda finally issue eua come problem transportin…
4  not many bright day best bidenharris winning e…
```

2.2 Sentiment Analysis with nltk

[10]:
```
clean_vaccine_tweets["sentiment"] = dict
clean_vaccine_tweets["sentiment_compound"] = 0.0
```

Determine sentiment via NLTK.Sentiment: - Values range from [-1, 1] - -1 is negative, 0 is neutral, 1 is positive

[11]:
```
def sentiment_score(dataset):
    sia = SentimentIntensityAnalyzer()
    for i in range(len(dataset)):
        dataset["sentiment"][i] = sia.polarity_scores(dataset["corpus"][i])
        dataset["sentiment_compound"][i] = dataset["sentiment"][i]["compound"]
```

[12]:
```
sentiment_score(clean_vaccine_tweets)
```

```
<ipython-input-11-9dda2f859171>:4: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  dataset["sentiment"][i] = sia.polarity_scores(dataset["corpus"][i])
<ipython-input-11-9dda2f859171>:5: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  dataset["sentiment_compound"][i] = dataset["sentiment"][i]["compound"]
```

[13]:
```
clean_vaccine_tweets
```

[13]:

	id	created_at	\
0	1338158543359250432	2020-12-13 16:27:13+00:00	
1	1337840331522453504	2020-12-12 19:22:45+00:00	
2	1338544403795881984	2020-12-14 18:00:29+00:00	
3	1337735595704115200	2020-12-12 12:26:34+00:00	
4	1337850832256176128	2020-12-12 20:04:29+00:00	
…	…	…	
118267	1407688257177935872	2021-06-23 13:13:28+00:00	
118268	1407699323035557888	2021-06-23 13:57:27+00:00	
118269	1407699835856330752	2021-06-23 13:59:29+00:00	
118270	1407682599515000832	2021-06-23 12:50:59+00:00	
118271	1407696190578249728	2021-06-23 13:45:00+00:00	

	user	geo	\
0	{'id': 76052772, 'id_str': '76052772', 'name':…	NaN	
1	{'id': 1300382181605494800, 'id_str': '1300382…	NaN	
2	{'id': 1164717209253552000, 'id_str': '1164717…	NaN	
3	{'id': 1316036067754205200, 'id_str': '1316036…	NaN	
4	{'id': 1110032180237852700, 'id_str': '1110032…	NaN	
…	…	…	
118267	{'id': 1263779139397382100, 'id_str': '1263779…	NaN	
118268	{'id': 40623001, 'id_str': '40623001', 'name':…	NaN	
118269	{'id': 61611674, 'id_str': '61611674', 'name':…	NaN	
118270	{'id': 126591034, 'id_str': '126591034', 'name…	NaN	
118271	{'id': 231692806, 'id_str': '231692806', 'name…	NaN	

```
                                                  full_text  \
0       While the world has been on the wrong side of …
1       @cnnbrk #COVID19 #CovidVaccine #vaccine #Coron…
2       The FDA Authorizes Emergency Use Of The Pfizer…
3       The #FDA finally issues #EUA now comes the pro…
4       There have not been many bright days in 2020 b…
…                                                     …
118267  #SputnikV Paid #Hyderabad https://t.co/oklatcuWLh
118268  The @WHO said its review of how #Russia produc…
118269  #WHO Finds Production Infringements at #Sputni…
118270  When was the #SputnikV\n\n1. Exploratory Stage…
118271  .@WHO raises concern on cross-contamination, i…

                                                   hashtags  \
0       ['covid19', 'supplychain', 'logistics', 'vacci…
1       ['covid19', 'covidvaccine', 'vaccine', 'corona…
2       ['pfe', 'pfizer', 'pfizervaccine', 'pfizerbion…
3          ['fda', 'eua', 'pfizerbiontech', 'vaccinated']
4       ['bidenharris', 'election2020', 'pfizerbiontec…
…                                                     …
118267                          ['sputnikv', 'hyderabad']
118268              ['russia', 'sputnikv', 'coronavirus']
118269  ['who', 'sputnikv', 'russia', 'covid19', 'coro…
118270                                      ['sputnikv']
118271                                      ['sputnikv']

                    user_id  PfizerBiontech  SputnikV  Sinopharm  Sinovac  \
0                  76052772               1         0          0        0
1       1300382181605494800               1         0          0        0
2       1164717209253552000               1         0          0        0
3       1316036067754205200               1         0          0        0
4       1110032180237852700               1         0          0        0
…                       …              …        …         …      …
118267  1263779139397382100               0         1          0        0
118268             40623001               0         1          0        0
118269             61611674               0         1          0        0
118270            126591034               0         1          0        0
118271            231692806               0         1          0        0

        Moderna  AstraZeneca  Covaxin  JandJ user_location  \
0             0            0        0      0           NaN
1             0            0        0      0           NaN
2             0            0        0      0           NaN
3             0            0        0      0           NaN
4             0            0        0      0           NaN
…           …          …      …    …           …
```

```
118267        0            0        0      0         India
118268        0            0        0      0           NaN
118269        0            0        0      0           NaN
118270        0            0        0      0           NaN
118271        0            0        0      0         India

                    coordinates  \
0                           NaN
1                           NaN
2                           NaN
3                           NaN
4                           NaN
…                             …
118267  [22.3511148, 78.6677428]
118268                      NaN
118269                      NaN
118270                      NaN
118271  [22.3511148, 78.6677428]

                                                    corpus  \
0       world wrong side history year hopefully bigges…
1       covid corona pfizerbiontech bbcnews nytimes bb…
2       fda authorizes emergency use pfizer pfe pfizer…
3       fda finally issue eua come problem transportin…
4       not many bright day best bidenharris winning e…
…                                                     …
118267                            sputnikv paid hyderabad
118268  said review russia produce sputnikv coronaviru…
118269  find production infringement sputnikv manufact…
118270  sputnikv exploratory stage animal eff muppets …
118271  raise concern cross contamination insufficient…

                                                sentiment  sentiment_compound
0       {'neg': 0.125, 'neu': 0.766, 'pos': 0.109, 'co…             -0.1027
1       {'neg': 0.117, 'neu': 0.405, 'pos': 0.477, 'co…              0.8402
2       {'neg': 0.126, 'neu': 0.874, 'pos': 0.0, 'comp…             -0.3818
3       {'neg': 0.137, 'neu': 0.863, 'pos': 0.0, 'comp…             -0.4019
4       {'neg': 0.096, 'neu': 0.6, 'pos': 0.304, 'comp…              0.7347
…                                                     …                   …
118267  {'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound…              0.0000
118268  {'neg': 0.0, 'neu': 0.877, 'pos': 0.123, 'comp…              0.4215
118269  {'neg': 0.256, 'neu': 0.744, 'pos': 0.0, 'comp…             -0.4767
118270  {'neg': 0.06, 'neu': 0.692, 'pos': 0.248, 'com…              0.7184
118271  {'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound…              0.0000

[118272 rows x 20 columns]
```

Sort dataset by sentiment and create a sub-dataset with location only:

[14]:
```
clean_vaccine_tweets = clean_vaccine_tweets.sort_values("sentiment_compound",␣
↪ascending=False, ignore_index=True)
```

[15]:
```
geo_vaccine_tweets = clean_vaccine_tweets[clean_vaccine_tweets["user_location"].
↪isnull() == False].reset_index(drop=True)
```

Export:

[16]:
```
clean_vaccine_tweets.to_csv("../data/processed/vaccine_tweets_with_sentiment.
↪csv")
```

[17]:
```
geo_vaccine_tweets.to_csv("../data/processed/geo_vaccine_tweets_with_sentiment.
↪csv")
```