Date Mining of HPV Misinformation Content in Twitter-Sphere: A Network Analytic Approach

Xuantong Mou[1,*,†] Yilin Lan[2,†]

[1] *Southwest Minzu University, Chengdu, China*
[2] *Huaqiao University, Quanzhou, China*
[*]*Corresponding author. Email: 2020240082@mail.chzu.edu.cn*

[†]*Those authors contributed equally.*

ABSTRACT
HPV has attracted much attention from different countries and organizations in recent years, but the vaccination rate of HPV is not high. This paper attempts to develop a "Python program" to track the "Twitter sphere content", so as to explore the impact of social media in the process of HPV transmission and prevention. In fact, "social media" is fixed as the "channel" and tool of script, allowing users to interact and share ideas and content on digital and portal websites, which often becomes the reason for people's "health information" all over the world. This paper mainly takes "Twitter" and other "social media" platforms to track the content of "HPV" vaccination as an example, in order to use Python language to explain the message and communication of "human papillomavirus vaccine" on "Twitter". This paper believes that through the role of social media, we can further enrich the communication channels and produce a variety of methods to make this vaccine popular, and Twitter is one of them. At the same time, there are "four elements" of HPV vaccine error information. This research will give some enlightenment to the development and application of Python.

Keywords: *HPV, Social media, Misinformation content*

1. INTRODUCTION

The “human papillomavirus vaccine (HPV) vaccine defends teenagers and growing adults from the higher risk of the “HPV virus” species that generally cause ninety per cent of the “cervical cancer” and “anal cancers” as well seventy per cent of the “oropharyngeal cancers”. Moreover, some research paper shows that in “US country” one in four people are generally affected by the “human papillomavirus”. Some other research shows that nearly fourteen million people are affected by this virus. This research paper gives a complete preview of this topic to examine the advantages of people knowing regarding the “social media content”, and some applications might be a collision consciousness, understanding, and so on about the “human papillomavirus (HPV)”. It is a common “sexual infection”, and a vaccine is also available to secure the “virus transmission”.

Therefore, some countries organized events to promote the “human papillomavirus vaccine (HPV)”, but “human papillomavirus vaccines (HPV)” uptake remains less in comparison to the availability of the “childhood vaccines” and “adolescent vaccines”. This project has developed a “python program” to track the “Twitter sphere content”. In addition, by utilizing the “social media” platform, various affected countries can spread the news corresponding to the “HPV” virus and the precaution regarding this virus. Therefore, “social media”, fixed for the script as “channels” and tools that permit the users to interact and share ideas, content in the digital as well as the web portals, is frequently becoming the cause of “health information” for the people worldwide. In this research paper, a python program has been created to track the “HPV” vaccination content through “social media” platforms, such as “Twitter”.

The research study aims to explain the “Human Papillomavirus Vaccine” messaging and communications on the “Twitter sphere” by utilizing python language. Further, the aim is to develop the python program for detecting “Twitter sphere content”.

2. LITERATURE REVIEW

This research paper utilizes a “network analytics” strategy to record the “fragmentation” of web behaviors through the “social media” interactions. Therefore, experimental data applied in the research paper has been obtained from the public online tweets on the “Twitter” platform. According to Ortiz, Smith & Coyne-Beasley, in addition, the “text-mining analysis” has been concentrated on the subjects as well as hashtags of online tweets on the “Twitter” platform [1]. Sometimes, societies and people are trying to negatively affect public views about security, and the vaccination value is working in the “online portal” and “social media”. There are various types of methods used to make this vaccine so popular; “Twitter” is one of them. Therefore, it influences choice-making within a few communities. The primary object of this research paper must explore how normal people debate about “data-driven journalism”. Sometimes, social media approach various ways to the people about the “HPV” vaccines. Therefore, while various researchers have examined “social media” utilizing “word embeddings”, python language has been chosen to show the content of the “HPV” vaccines on “social media” platforms such as “Instagram” and “Twitter”. As stated by Ortiz, Smith & Coyne-Beasley, Moreover, social media allows both customers and “health professionals” to share ideas and concepts regarding health problems with many people, whether the data is correct or not that have been shared by the health professionals or customers [1]. There are various “social media platforms” available in the modern world, such as “Facebook”, “Twitter”, “YouTube” that link users.

As stated by Chin et al., therefore, where users can share experiences and opinions about health-related subjects [2]. In addition, “World Health Organization”, various countries suggest “HPV vaccination” happen between nine and fourteen ages, with immature teenage girls’ main target association. However, vaccination for teenage children and growing adults who have still to vaccinate has also been suggested. In some countries, vaccination has been recommended regularly for teenage boys and girls. Sometimes, web portals and “social media” vaccine data have an essential reference for the people to obtain vaccination decisions. Therefore, the influence of the online “HPV vaccine” data on real vaccination ways is valuable. Some studies showed a connection between the expression of the negative “social media HPV-vaccine” data and the consequent negative opinions structure via “HPV vaccines”. The corresponding task has been conducted to achieve a better idea regarding the HPV vaccine data that the researcher has collected from the Twitter social media platform.

According to Featherstone et al., there are various types of stakeholders utilizing “social media” platforms, such as “pharmaceutical companies”, government companies, health care companies, experts, and “news media” to absorb viewers to increase disease knowledge and promote public well-being [3]. But, sometimes, it is not clear what results in this fitness knowledge has on the laypeople. Moreover, “social media”, especially “Twitter”, is an online source that provides various methods to study health-related information and converse regarding fitness and healthcare activities. There has been a current rise in “vaccine” sources that guidance can reach through “social media” platforms. Therefore, many users’ sources distribute incorrect information, talk to opposite medical experts, and worry about “vaccines”. On the other hand, there are various articles available that emphasize the value of health experts to include digital messages.

3. METHODOLOGY

There are some content analyses that have been performed based on approximately 5000 top tweets in “Twitter’s platform” that were mainly concentrated on the “HPV” vaccine. Therefore, Twitter’s were evaluated for the content analysis, classification model, as well as the number of communications. Apart from this, this system has been controlled through the concept of “HPV”. Moreover, the person must gain knowledge of the vaccination by which the virus can be detected. Therefore, Twitter has been defined for this research paper; via “social media”, users can connect with other users [3]. In this research paper, various types of search items have been utilized about the “HPV” and “social media”, including famous “social media” platforms, such as “Twitter”, “Instagram”, and so on. Therefore, the whole search items have been shown in the table below.

Table 1 Search Items

Key Concepts	Search Items
“Social Media”	“Social media” or “social network” or Twitter or “Instagram” or “Twitter” or (online and comment)
“HPV” and “HPV vaccine.”	“HPV” or “human papillomavirus” or “human papilloma virus”

Source: Ortiz, Smith & Coyne-Beasley, 2019 [1]

The above table shows the key concept and search items that the researcher has collected for conducting the corresponding task. In this context, social media includes the construction of communities or channels and supporting partnership and engagement. Therefore, technologies, such as messages and email, permit people to share experiences and opinions with others. As claimed by Agergaard, Smith & Nielsen [4], moreover, “health professionals” most utilize these transmission gadgets as one-way transmission systems, like sending message reminders to the patients about appointments and other health-related information.

In this case, “Twitter API” has been required for retrieving the data from the social media platform, such as “Instagram”, “Twitter”, and so on. Once users have the “Twitter” app for setting up, users can easily access “tweets” in “python language” by importing the required “python libraries” [5]. Moreover, pythons’ libraries are available for accessing the “Twitter API”. In such circumstances, for accessing the “Twitter API”, users require four generally things from the users “Twitter App” page. These types of keys are generally located in the app setting. Therefore, the keys are “consumer key”, “secret consumer key”, “access token key”, and “access token secret key”.

On the other hand, “social network analysis” is the method of studying social arrangements for the management of “networks and graph theory”. Therefore, a short concept of the “graph theory” has been elaborated on in this research paper. Sometimes, it sinks into “Python code” with “Network X” building and importing social media networks from the existing dataset [6]. According to the network thesis, two basic components have been managed in the overall networking system. These components are “nodes” and “edges”. The term node has been classified as “plug”. Generally, all the nodes have been gathered in a small place and connected with each other. Usually, these nodes are responsible for creating a variety of “network topologies”. Moreover, five “network topologies” have been controlled through the networking system. These are “bus topology”, “ring topology”, “star topology”, “mess topology” and “tree topology”. Usually, these topologies have been configured through the nodes. Generally, all the entities have been demonstrated through the nodes. Moreover, a node can be able to grab its own properties.

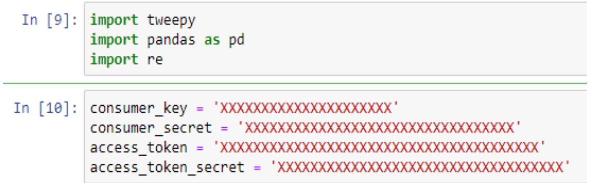
On the other hand, connections between two nodes have been maintained through the edges. Moreover, one node can relate to all the nodes or a single node. Furthermore, an edge has managed the connection strength among two nodes. It can check whether the signal has been maintained in a straight line (smooth) or fluctuates. Moreover, it can perform to solve the noise coming from the signal during the communication [7]. “Real-world networks” and “social networks” have an individual construction that differs from informal mathematical networks. The “phenomenon” requires that existing networks usually have short ways between any linked network members. Therefore, it has been applied to the “existing social networks” or “virtual social network, and for the “physical social networks” like airports and so on. In this case, the network has been constructed from the different datasets, as long a user can show the connection between the nodes. In this research paper, “sentiment analysis” has been utilized to determine whether the writing piece is “positive”, “negative”, or “neutral”. There are different types of networks available, but in this research, “NetworkX” has been developed to analyze various networks [6]. Moreover, the “cascade network” has been implemented in the “network analysis” to investigate the true and false risk messages about HPV vaccines.

Along with that, “Jupyter Notebook” has been utilized to process the methods with respect to the data mining of the “Twitter” dataset from the social media platforms (such as “Twitter”, “Instagram” and others). Further, the dataset regarding this topic for HPV has been processed using data scraping with the necessary “API” details and tokens. In such circumstances, it has been utilized “python programming” to gather public “Instagram” posts with the help of text mining checking. It has gathered data by getting to “Instagram’s public application programming interface, meeting the organization’s terms of administration for public information, and gathering up to 100 new posts each hour [8]. On the off chance that more than 100 posts with a specific hashtag catchphrase were posted each hour, just the latest were recovered. All things considered, all information for this review were gathered preceding the application programming interface conclusion. The information assortment strategy in this review is like that utilized in earlier “Instagram” research.

4. TECHNICAL ANALYSIS

In this research paper, the primary objective of this paper is to characterize the original or anti “HPV vaccine” networks in “Twitter” by utilizing the python language, also explain the misconfiguration of the “HPV vaccine” networks. In this project, the python programming language has been considered for successfully conducting the research paper. “Python language” is very easy and mainly utilized for general purposes. The “Human papillomavirus (HPV) vaccine” is the primary achievement of stopping cancer diseases. Therefore, Twitter has been chosen for collecting the misconfiguration data and utilized in this project. Moreover, different kinds of python codes have been applied in this corresponding research paper. Furthermore, a dataset has been created by using the “twitter” network. This research paper has developed an “ API ” by utilizing the python language corresponding to this research paper.

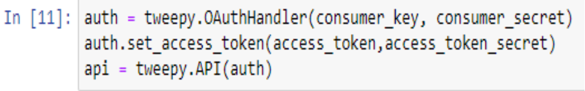
Therefore, various codes and algorithms have been applied in this project to conduct the research paper successfully. Moreover, all the codes have been explained below.



```
In [9]:  import tweepy
         import pandas as pd
         import re

In [10]: consumer_key = 'XXXXXXXXXXXXXXXXXXXXX'
         consumer_secret = 'XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX'
         access_token = 'XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX'
         access_token_secret = 'XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX'
```

Figure 1 Import libraries in Jupyte

The above figure shows the programming codes that have been utilized to find the misconfiguration in the “Twitter” network about the “HPV vaccine”. Here, “import tweepy” has been written to import the “Twitter” data in the data which have been created regarding this research paper. Therefore, “import pandas as pd” has been applied to import the data in python for successfully conducting the research paper. On the other hand, “import re” has been utilized to compile the “regex first” in this project. Therefore, there are four different types created in this project for completely executing the research paper. These keys are “consumer key”, “consumer secret”, “access token”, and “access token secret”. Moreover, all the keys are different job roles in this research paper, and all the keys are different and unique from each other.



```
In [11]: auth = tweepy.OAuthHandler(consumer_key, consumer_secret)
         auth.set_access_token(access_token,access_token_secret)
         api = tweepy.API(auth)
```

Figure 2 Authentication of the user account corresponding to API

The picture shows the python codes that have been utilized in this research paper. Here, “auth” functions have been applied to authenticate the key that has been shown in the above picture. Furthermore, “API” has been created by applying python programming language to this research paper. Moreover, “Twitter. API (auth)” has been written for authenticating the “Twitter” data set in this paper. Therefore, to authenticate the “Twitter API”, “consumer key”, and “consumer secret” has been applied inside python programming codes. Moreover, “auth. set access token” has been used in this code for accessing the consumer credentials, and for this reason “, access token”, “access token secret” have been applied in this python codes for ‘successfully conducting the research paper. Moreover, “Twitter API” has been utilized corresponding to this research paper for collection the twitter information about the “HPV vaccination”.

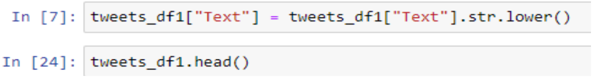
```
In [2]: pip install tweepy
```

Figure 3 Install of tweepy command

In this paper, the “pip install” command has been used for successfully executing the program. Therefore, these codes have been utilized to import “Twitter” data from the “Twitter networks”.

Figure 4 shows some python codes that have been included corresponding to this research paper for successfully conducting the task. Therefore, the “import snscrape module. twitter as sntwitter” code was written inside the program to import the “Twitter” modulus to the dataset. Moreover, “import pandas as pd” has been written to import the data set using the python language successfully. Therefore, some conditions have been inserted in this paper. “Tweets list1 []” has been written for showing the whole “Twitter” users list in this program. Therefore, “if i > 4999” have been included because of the data set range, which means, in this research paper, the data set range has been mentioned as “4999” for showing the users lists. In this code, “for” conditions have been applied to get the items in this paper. Moreover, the username has been chosen in this project as “HPVSideEffects”. “Tweets list 1. Append ([tweet. Data, tweet. Id, tweet. Content, tweet. User. username])” this python code has been written inside this whole program for showing the users list as “tweets id”, “tweet information”, “tweet content”, and “tweet users’ username”. “Tweets. Df 1 = pd. Dataframe (tweets list 1, column = ‘data time, [tweet id], [Text], [username])”, this code has been implemented in this program for incorporating the data regarding the HPV and vaccine information by users id, name, username, and data in the data set which have created for storing the user’s data related to HPV. Therefore, all the codes have different purposes in this project.

Figure 5 shows the output after successfully executing the program. In this picture, users’ data related to the HPV and vaccine has been shown by “data and times”, “tweet id”, “text”, and “username”. Therefore, all the user’s data connected to the “HPV” and “vaccination” has been stored successfully after executing the program. The data name has been in the above picture by “tweets df 1”. Therefore, all the “tweets id” are different from each other in this output. In the “text” portion of this data set, all the public’s tweets related to the “HPV” and “vaccination” have been stored successfully in this project.

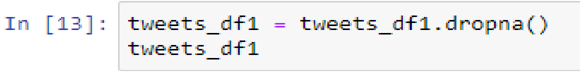


```
In [7]: tweets_df1["Text"] = tweets_df1["Text"].str.lower()

In [24]: tweets_df1.head()
```

Figure 6 Use of str. lower function

The above code has been implemented for changing the text format which has been saved in the data. “Tweets df 1 [“Text”] = tweets df 1 [“text”]. Str. lower ()” has been included for successfully changing the text format.

The output has been shown in Figure 7 after successfully changing the text format in the “text” section of the data set. Here, all the users’ tweets have been changed in lower format. In addition, after successfully changing the format, the previous text has been stored in the data by the “clean tweet” name. In this case, after changing the text format, five attributes have been shown in the data set, such as “data time”, “tweet id”, “text”, “username”, and “clean tweet”.



```
In [13]: tweets_df1 = tweets_df1.dropna()
         tweets_df1
```

Figure 8 Dropping of Null values

The above programming codes have been implemented to drop down the “null value” from the data set after successfully storing the user’s data related to the “HPV” and “vaccination”. For example, “Tweets df 1 = tweets df 1. Dropna ()” has been written in this program for dropping the null values. Here, “tweets df 1” refers to the data set name.

Figure 9 shows the output after successfully executing the drop command in this program. Therefore, all the null values have been removed from the data set. Moreover, only not null values are shown after executing the whole python program.

The python codes in Figure 10 have been inserted to remove the punctuation from the user tweets related to the “HPV” and “vaccination”. Here, the whole program has been written inside the “def” functions. Therefore, after declaring the functions, a code has been written to execute the above functions in this project successfully. Moreover, the output is also shown in the above figures. In the output section, all the “punctuation” functions are removed from the data set. Furthermore, after successfully removing the punctuation functions, the output is saved in the data set by the name of “clean text”.

Figure 11 has shown that the tweets from the “Twitter” data have been split into several parts for decoding the words to evaluate the meaning of every word specifically. Further, the tokenization function has been used for this splitting of words from the tweets. On the other hand, the output of the above program has also been shown in the above picture. In the output section, all the words are shown separately for evaluating the word meaning in this program. Therefore, all the user’s tweets are shown like “worldwide”, “walkouts”, “11032021”, “will”, “demand”, and so on.

Figure 12 shows the programming code for removing the unnecessary gap between the words. The “lemmatization” process has been utilized to execute this task, successfully corresponding to this research paper. Furthermore, “ tokenized_tweet =tokenized_tweet. Apply (lambda sentence: [lemmatizer. Lemmatize (word) for word in sentence]) tokenized_tweet. Head (10)” programming codes have been implemented for successfully conducting the program. In this figure, the outputs of the “lemmatization” have also been shown. Sometimes, users give unnecessary gaps between the words. The above command removes the gaps from the data set. For importing the “lemmatizer” process, the “from nlkt. stem import WordNet Lemmatizer” command has been utilized in this case.

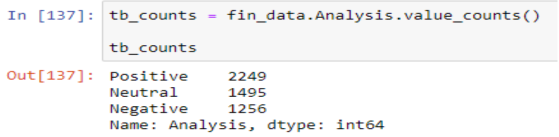
Figure 13 has been shown the “URL” removing codes in the above pictures. Moreover, sometimes, users provide various types of websites links or any other links. Therefore, the above codes have been utilized in this research paper to remove the “URL” in this research paper links. Therefore, after successfully cleaning the links, data has been stored in the “clean_tweet” file. Moreover, the output of the above codes has been shown in the figure. In the output section, all the outputs have been removed by utilizing the above commands.

The python codes in Figure 14 have been implemented in this program for measuring the sentiment and polarities. Sometimes, users tweet positive comments or sometimes negative commands. Therefore, to measure the polarity, the above codes have been utilized in this program. Here, “def” functions have been utilized, and sentiment functions have been applied inside the “def”. The above command finds the sentiment types of tweets and checks the polarities, such as negative comments or positive comments. Further, an output of the above codes has also been shown in the above figures. In the output section, some tweets show negative polarities, and some show positive polarities. For example, the first tweet shows the negative polarities, which means users’ tweets negatively impact others.

Figure 15 shows some graphical outputs of the “sentiment polarities”. Here, the “sns. Displot (tweets df 1 [‘setiment_polarity’])” code has been implemented for showing the graphical representation of the “sentiment polarities” related to the “HPV” and “vaccination”. It has been shown as a seaborn plot for which sns. distplot has been implemented where the sentiments of the tweets are in the negative and positive region. Further, it has been analyzed that there are positive and negative comments in the dataset. Finally, it has been seen that the comments are mostly neutral as the sentiment of the tweets has stayed in the origin region of the graph.

Figure 16 shows the python codes for separating the negative and positive tweets in the data set. Here, “tweets df 1 [‘positive’] = tweets_df1 [tweets_df1. Setiment_polatities > 0]. clean_tweet” codes have been implemented to separate the positive polarities related to the “HPV” and “vaccination” in the data set. Further, “tweets df 1 [‘negative’] = tweets_df1 [tweets_df1. Setiment_polatities < 0]. clean_tweet” codes have been implemented to separate the negative polarities related to the “HPV” and “vaccination”. Therefore, the output picture shows the polarities separately. Here, all the negative and positive tweets have been shown separately in the data set.

Figure 17 shows the python codes for polarity and analysis corresponding to this research paper. Here various types of functions have been implemented to show the polarities separately. Moreover, a “def” module has been included in this program, and an “if-else” condition has also been included in this program. The “find data [‘Polarity’] = fin_data [‘clean_tweet’]. Apply (find_pol)” code has been implemented for finding the polarities. “Find_data [‘analysis’] = fin_data [‘polarity’]. apply (analysis)” code has been applied in this program for showing the polarities in the analysis portion of the data set. The output of the above codes has also been shown in the above picture.

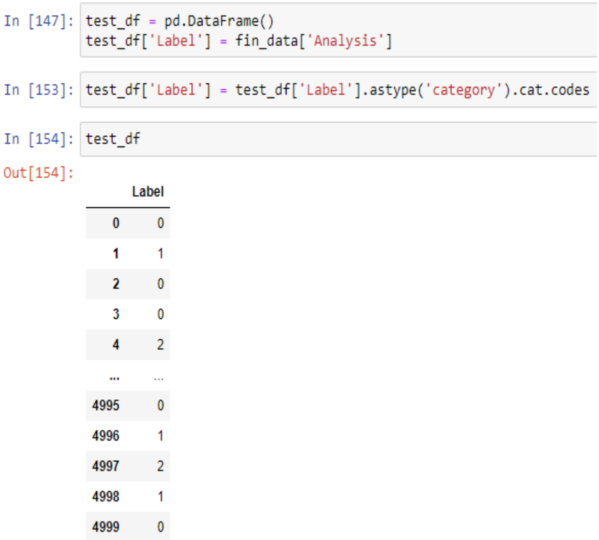


```
In [137]: tb_counts = fin_data.Analysis.value_counts()

          tb_counts

Out[137]: Positive    2249
          Neutral     1495
          Negative    1256
          Name: Analysis, dtype: int64
```

Figure 18 Count of positive, neutral, and negative sentiments

The above programming codes have been written for counting the total number of positive, negative, and neutral tweets of users. In addition, the “tb counts = fin data. Analysis. Value counts ()” codes have been inserted for counting the numbers. The output of the above codes has also been shown in the figure. In the picture, it has been clearly shown the “positive”, “negative”, “neutral”, as well as “name”.

The python codes in figure 19 have been inserted for the training of the dataset corresponding to the topic. Therefore, the code “ x = tfidf.fit_transform(tweets_df1['clean_tweet']) ” has been utilized to build the “tfidf” features in the dataset for maintaining the tweets. Along with that, “ train_df = pd. DataFrame.sparse. from_spmatrix (x)”.



```
In [147]: test_df = pd.DataFrame()
          test_df['Label'] = fin_data['Analysis']

In [153]: test_df['Label'] = test_df['Label'].astype('category').cat.codes

In [154]: test_df

Out[154]:
                Label
             0      0
             1      1
             2      0
             3      0
             4      2
           ...    ...
          4995      0
          4996      1
          4997      2
          4998      1
          4999      0
```

Figure 20 Test dataset for the analysis

The above codes have been implemented for testing data regarding this topic. Further, the above code has been utilized in this Jupyter notebook to show the dataset’s extraction regarding this task. Therefore, test_df = pd. DataFrame () has been included in this program for importing the data from the dataset. “test_df [‘Label’] = fin_data[‘Analysis’]” has been implemented for creating the analysis from the dataset to test the dataset. “test_df [‘Label’] = test_df[‘Label’]. astype (‘category’).cat. Codes” has been utilized for creating the category from the dataset to test the dataset as well. The above codes have been inserted in this program for showing the testing data corresponding to this research paper. Along with that, “tfidf” has been used to build the model for the sentiment analysis that has been shown in the figure (fig 18). Moreover, the feature extraction has been processed using the TensorFlow module for sequential models from the training and testing dataset.

5. CONCLUSION

From the above research, it can be concluded that “The Human Papillomavirus” vaccine is so popular that it helps to prevent cancer. Therefore, this vaccine has major advantages for cancer patients. In general, there are various types of use that happen with the HPV vaccine. In this research, all the impacts of social media will be discussed. There are various types of methods used to make this vaccine so popular. Twitter is one of them. In the middle of 2018, this “HPV vaccine” program was populated on social media, especially Twitter. This project shows various kinds of graphical representations of the HPV vaccine advancement. There are various kinds of posts that also take place on the HP vaccine on social media. The anti-vaccine posts also consist in “social media”, especially on Twitter. These elements of HPV vaccine misinformation can be divided into “four broad dimensions”. The first element is misinformation of the theoretical domain. Next are topics on the vaccine debate. The third element is the base of evidence. Therefore, various types of python programming have been applied for successfully conducting the research paper. In this case, different types of fake messages have been shown while executing the python program corresponding to the “HPV vaccination” that also be cleared properly. Further, it can be concluded from the above research that a python program has been developed for tracking the “HPV vaccinations” content in the “social media” platforms, such as “Twitter”. In this study, different types of issues have been found while executing the python program, and that also be clear properly. This innovative idea has an impact on the life of poor people, and this concept can help the developers to analyze the overall process.

REFERENCES

[1] R. R. Ortiz, A. Smith, T. Coyne-Beasley, A systematic literature review to examine the potential for social media to impact HPV vaccine uptake and awareness, knowledge, and attitudes about HPV and HPV vaccination, Human vaccines & immunotherapeutics, 2019, 15(7-8), pp.1465-1475.https://www.tandfonline.com/doi/pdf/10.1080/21645515.2019.1581543

[2] J. Chin, C. L. Chin, S. Panday, A. Ghazanfari, G. Jagadeesan, Z. Wang, R. Caskey, Tracking the Human Papillomavirus Vaccine Risk Misinformation: An Explorative Study to Examine How the Misinformation Has Spread in User-Generated Content, In Proceedings of the International Symposium on Human Factors and Ergonomics in Health Care, 2020, 9(01), pp.312-316.https://journals.sagepub.com/doi/pdf/10.1177/2327857920091069

[3] J. D. Featherstone, J. B. Ruiz, G. A. Barnett, B. J. Millam, Exploring childhood vaccination themes and public opinions on Twitter: A semantic network analysis. Telematics and Informatics, 2020(54), pp.101474.

[4] T. E. Agergaard, M. E. Smith, K. H. Nielsen, Vaccine assemblages on three HPV vaccine-critical Facebook pages in Denmark from 2012 to 2019, Media and Communication, 2020, 8(2), pp.339-352.

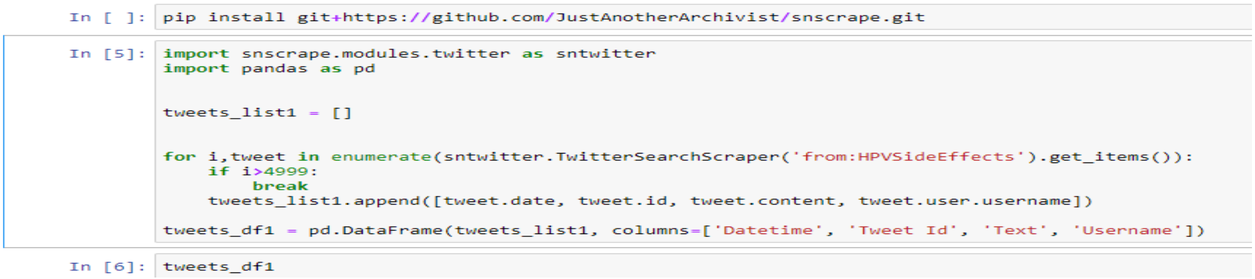
[5] N. Bezuidenhout, Capitalising on social media marketing to raise confidence in COVID-19 public health information and vaccines, Medical Writing, 2021(30), pp.12-15.

[6] B. Sundstrom, E. Aylor, K. B. Cartmell, H. M. Brandt, D. C. Bryant, H. C.Hughes, J. Y. Pierce, Beyond the birds and the bees: a qualitative content analysis of online HPV vaccination communication, Journal of Communication in Healthcare, 2018 11(3), pp.205-214.

[7] L. Singh, S. Bansal, L. Bode, C. Budak, G. Chi, K. Kawintiranon, Y. Wang, A first look at COVID-19 information and misinformation sharing on Twitter, arXiv preprint arXiv:2003, 2020. https://arxiv.org/pdf/2003.13907

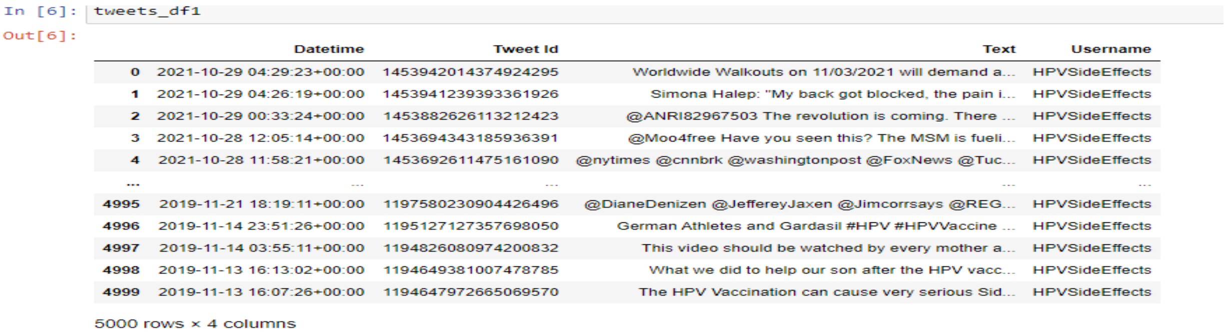
[8] A. Jamison, D. A. Broniatowski, M. C. Smith, K. S. Parikh, A. Malik, M. Dredze, S. C. Quinn, Adapting and extending a typology to identify vaccine misinformation on Twitter, American Journal of Public Health, 2020, 110(S3), pp.S331-S339. https://ajph.aphapublications.org/doi/pdfplus/10.2105/AJPH.2020.305940

APPENDIX



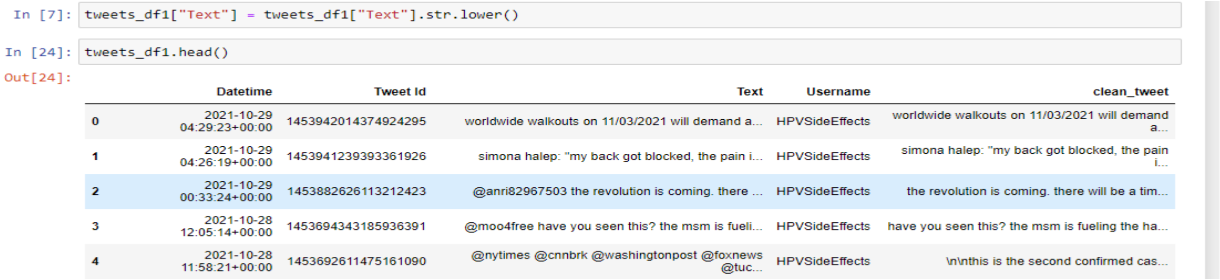
```
In [ ]: pip install git+https://github.com/JustAnotherArchivist/snscrape.git

In [5]: import snscrape.modules.twitter as sntwitter
        import pandas as pd


        tweets_list1 = []


        for i,tweet in enumerate(sntwitter.TwitterSearchScraper('from:HPVSideEffects').get_items()):
            if i>4999:
                break
            tweets_list1.append([tweet.date, tweet.id, tweet.content, tweet.user.username])

        tweets_df1 = pd.DataFrame(tweets_list1, columns=['Datetime', 'Tweet Id', 'Text', 'Username'])

In [6]: tweets_df1
```

Figure 4 Extraction of the account of “HPVSideEffects”



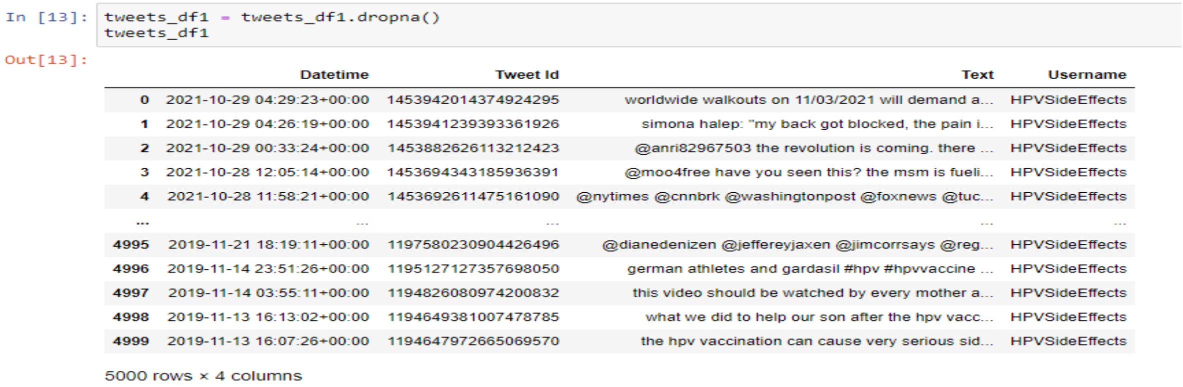
```
In [6]: tweets_df1
Out[6]:
                     Datetime             Tweet Id                                              Text       Username
   0  2021-10-29 04:29:23+00:00  1453942014374924295         Worldwide Walkouts on 11/03/2021 will demand a...  HPVSideEffects
   1  2021-10-29 04:26:19+00:00  1453941239393361926         Simona Halep: "My back got blocked, the pain i...  HPVSideEffects
   2  2021-10-29 00:33:24+00:00  1453882626113212423       @ANRI82967503 The revolution is coming. There ...  HPVSideEffects
   3  2021-10-28 12:05:14+00:00  1453694343185936391       @Moo4free Have you seen this? The MSM is fueli...  HPVSideEffects
   4  2021-10-28 11:58:21+00:00  1453692611475161090  @nytimes @cnnbrk @washingtonpost @FoxNews @Tuc...  HPVSideEffects
 ...                        ...                  ...                                               ...             ...
4995  2019-11-21 18:19:11+00:00  1197580230904426496   @DianeDenizen @JeffereyJaxen @Jimcorrsays @REG...  HPVSideEffects
4996  2019-11-14 23:51:26+00:00  1195127127357698050     German Athletes and Gardasil #HPV #HPVVaccine ...  HPVSideEffects
4997  2019-11-14 03:55:11+00:00  1194826080974200832        This video should be watched by every mother a...  HPVSideEffects
4998  2019-11-13 16:13:02+00:00  1194649381007478785         What we did to help our son after the HPV vacc...  HPVSideEffects
4999  2019-11-13 16:07:26+00:00  1194647972665069570       The HPV Vaccination can cause very serious Sid...  HPVSideEffects

5000 rows × 4 columns
```

Figure 5 Extracted Dataset regarding HPV



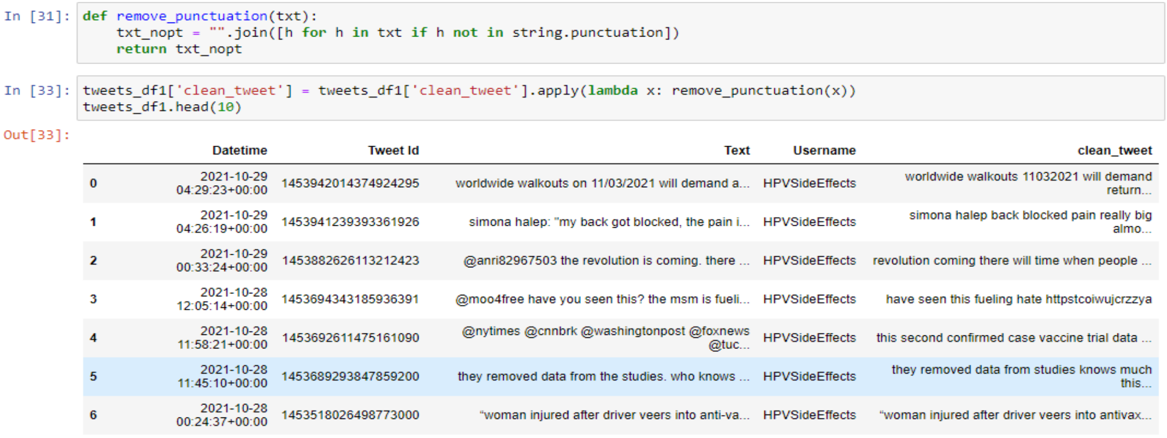
```
In [7]: tweets_df1["Text"] = tweets_df1["Text"].str.lower()

In [24]: tweets_df1.head()
Out[24]:
        Datetime             Tweet Id                                          Text       Username                                  clean_tweet
0     2021-10-29  1453942014374924295  worldwide walkouts on 11/03/2021 will demand a...  HPVSideEffects  worldwide walkouts on 11/03/2021 will demand
  04:29:23+00:00                                                                                                                             a...
1     2021-10-29  1453941239393361926    simona halep: "my back got blocked, the pain i...  HPVSideEffects    simona halep: "my back got blocked, the pain
  04:26:19+00:00                                                                                                                             i...
2     2021-10-29  1453882626113212423  @anri82967503 the revolution is coming. there ...  HPVSideEffects     the revolution is coming. there will be a tim...
  00:33:24+00:00
3     2021-10-28  1453694343185936391  @moo4free have you seen this? the msm is fueli...  HPVSideEffects  have you seen this? the msm is fueling the ha...
  12:05:14+00:00
4     2021-10-28  1453692611475161090   @nytimes @cnnbrk @washingtonpost @foxnews  HPVSideEffects           \n\nthis is the second confirmed cas...
  11:58:21+00:00                                                         @tuc...
```

Figure 7 Output of using str. lower function



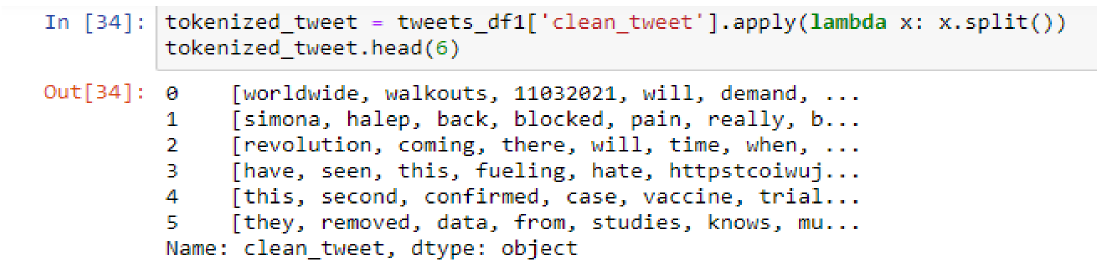
```
In [13]: tweets_df1 = tweets_df1.dropna()
         tweets_df1
Out[13]:
                     Datetime             Tweet Id                                              Text       Username
   0  2021-10-29 04:29:23+00:00  1453942014374924295         worldwide walkouts on 11/03/2021 will demand a...  HPVSideEffects
   1  2021-10-29 04:26:19+00:00  1453941239393361926         simona halep: "my back got blocked, the pain i...  HPVSideEffects
   2  2021-10-29 00:33:24+00:00  1453882626113212423       @anri82967503 the revolution is coming. there ...  HPVSideEffects
   3  2021-10-28 12:05:14+00:00  1453694343185936391       @moo4free have you seen this? the msm is fueli...  HPVSideEffects
   4  2021-10-28 11:58:21+00:00  1453692611475161090  @nytimes @cnnbrk @washingtonpost @foxnews @tuc...  HPVSideEffects
 ...                        ...                  ...                                               ...             ...
4995  2019-11-21 18:19:11+00:00  1197580230904426496   @dianedenizen @jeffereyjaxen @jimcorrsays @reg...  HPVSideEffects
4996  2019-11-14 23:51:26+00:00  1195127127357698050     german athletes and gardasil #hpv #hpvvaccine ...  HPVSideEffects
4997  2019-11-14 03:55:11+00:00  1194826080974200832        this video should be watched by every mother a...  HPVSideEffects
4998  2019-11-13 16:13:02+00:00  1194649381007478785         what we did to help our son after the hpv vacc...  HPVSideEffects
4999  2019-11-13 16:07:26+00:00  1194647972665069570       the hpv vaccination can cause very serious sid...  HPVSideEffects

5000 rows × 4 columns
```

Figure 9 Outputs of dropping of Null values



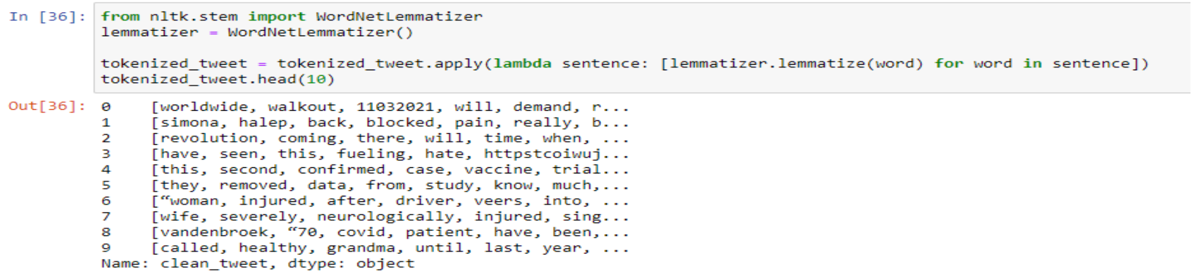
```
In [31]: def remove_punctuation(txt):
             txt_nopt = "".join([h for h in txt if h not in string.punctuation])
             return txt_nopt

In [33]: tweets_df1['clean_tweet'] = tweets_df1['clean_tweet'].apply(lambda x: remove_punctuation(x))
         tweets_df1.head(10)
Out[33]:
        Datetime             Tweet Id                                          Text       Username                                  clean_tweet
0     2021-10-29  1453942014374924295  worldwide walkouts on 11/03/2021 will demand a...  HPVSideEffects   worldwide walkouts 11032021 will demand
  04:29:23+00:00                                                                                                                     return...
1     2021-10-29  1453941239393361926    simona halep: "my back got blocked, the pain i...  HPVSideEffects    simona halep back blocked pain really big
  04:26:19+00:00                                                                                                                      almo...
2     2021-10-29  1453882626113212423  @anri82967503 the revolution is coming. there ...  HPVSideEffects  revolution coming there will time when people ...
  00:33:24+00:00
3     2021-10-28  1453694343185936391  @moo4free have you seen this? the msm is fueli...  HPVSideEffects     have seen this fueling hate httpstcoiwujcrzzya
  12:05:14+00:00
4     2021-10-28  1453692611475161090   @nytimes @cnnbrk @washingtonpost @foxnews  HPVSideEffects   this second confirmed case vaccine trial data ...
  11:58:21+00:00                                                         @tuc...
5     2021-10-28  1453689293847859200  they removed data from the studies. who knows ...  HPVSideEffects  they removed data from studies knows much
  11:45:10+00:00                                                                                                                      this...
6     2021-10-28  1453518026498773000   “woman injured after driver veers into anti-va...  HPVSideEffects   “woman injured after driver veers into antivax...
  00:24:37+00:00
```

Figure 10 Removing of punctuation from the tweets



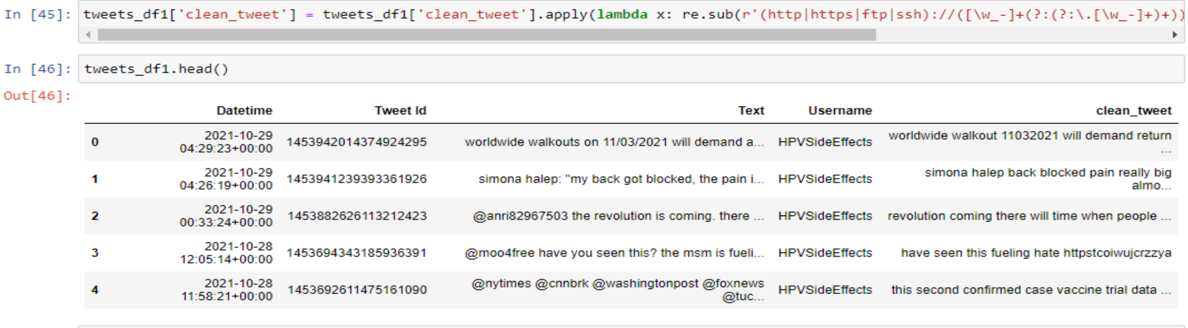
```
In [34]: tokenized_tweet = tweets_df1['clean_tweet'].apply(lambda x: x.split())
         tokenized_tweet.head(6)

Out[34]: 0    [worldwide, walkouts, 11032021, will, demand, ...
         1    [simona, halep, back, blocked, pain, really, b...
         2    [revolution, coming, there, will, time, when, ...
         3    [have, seen, this, fueling, hate, httpstcoiwuj...
         4    [this, second, confirmed, case, vaccine, trial...
         5    [they, removed, data, from, studies, knows, mu...
         Name: clean_tweet, dtype: object
```

Figure 11 Tokenization of the tweets data



```
In [36]: from nltk.stem import WordNetLemmatizer
         lemmatizer = WordNetLemmatizer()

         tokenized_tweet = tokenized_tweet.apply(lambda sentence: [lemmatizer.lemmatize(word) for word in sentence])
         tokenized_tweet.head(10)

Out[36]: 0    [worldwide, walkout, 11032021, will, demand, r...
         1    [simona, halep, back, blocked, pain, really, b...
         2    [revolution, coming, there, will, time, when, ...
         3    [have, seen, this, fueling, hate, httpstcoiwuj...
         4    [this, second, confirmed, case, vaccine, trial...
         5    [they, removed, data, from, study, know, much,...
         6    [“woman, injured, after, driver, veers, into, ...
         7    [wife, severely, neurologically, injured, sing...
         8    [vandenbroek, “70, covid, patient, have, been,...
         9    [called, healthy, grandma, until, last, year, ...
         Name: clean_tweet, dtype: object
```

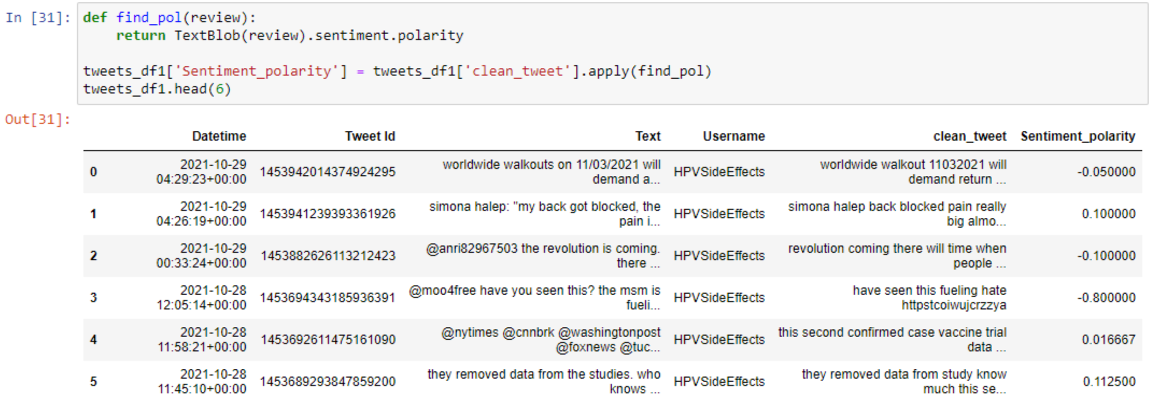
Figure 12 Lemmatization of the tweets



```
In [45]: tweets_df1['clean_tweet'] = tweets_df1['clean_tweet'].apply(lambda x: re.sub(r'(http|https|ftp|ssh)://([\w_-]+(?:(?:\.[\w_-]+)+))

In [46]: tweets_df1.head()

Out[46]:
```

	Datetime	Tweet Id	Text	Username	clean_tweet
0	2021-10-29 04:29:23+00:00	1453942014374924295	worldwide walkouts on 11/03/2021 will demand a...	HPVSideEffects	worldwide walkout 11032021 will demand return ...
1	2021-10-29 04:26:19+00:00	1453941239393361926	simona halep: "my back got blocked, the pain i...	HPVSideEffects	simona halep back blocked pain really big almo...
2	2021-10-29 00:33:24+00:00	1453882626113212423	@anri82967503 the revolution is coming. there ...	HPVSideEffects	revolution coming there will time when people ...
3	2021-10-28 12:05:14+00:00	1453694343185936391	@moo4free have you seen this? the msm is fueli...	HPVSideEffects	have seen this fueling hate httpstcoiwujcrzzya
4	2021-10-28 11:58:21+00:00	1453692611475161090	@nytimes @cnnbrk @washingtonpost @foxnews @tuc...	HPVSideEffects	this second confirmed case vaccine trial data ...

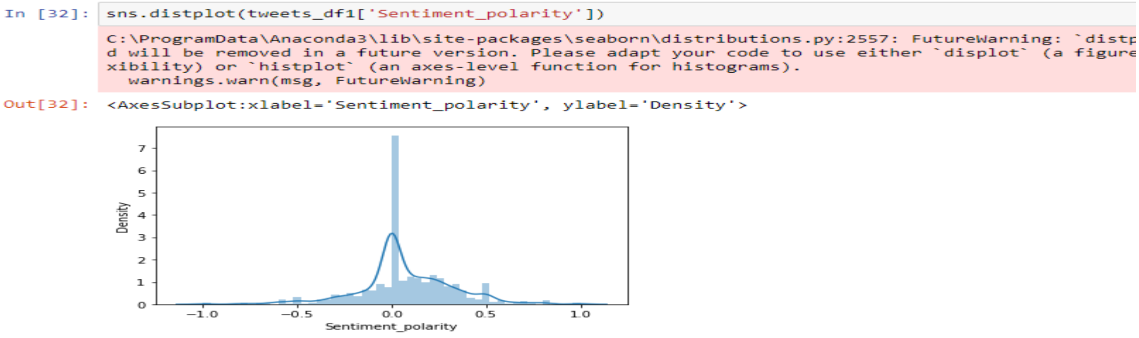
Figure 13 Removing URL from the tweets



```
In [31]: def find_pol(review):
             return TextBlob(review).sentiment.polarity

         tweets_df1['Sentiment_polarity'] = tweets_df1['clean_tweet'].apply(find_pol)
         tweets_df1.head(6)

Out[31]:
```

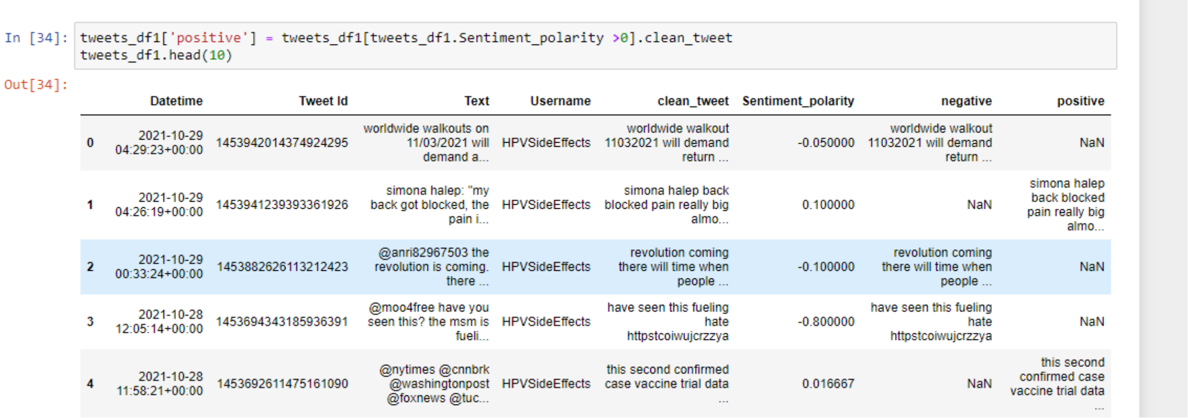
	Datetime	Tweet Id	Text	Username	clean_tweet	Sentiment_polarity
0	2021-10-29 04:29:23+00:00	1453942014374924295	worldwide walkouts on 11/03/2021 will demand a...	HPVSideEffects	worldwide walkout 11032021 will demand return ...	-0.050000
1	2021-10-29 04:26:19+00:00	1453941239393361926	simona halep: "my back got blocked, the pain i...	HPVSideEffects	simona halep back blocked pain really big almo...	0.100000
2	2021-10-29 00:33:24+00:00	1453882626113212423	@anri82967503 the revolution is coming. there ...	HPVSideEffects	revolution coming there will time when people ...	-0.100000
3	2021-10-28 12:05:14+00:00	1453694343185936391	@moo4free have you seen this? the msm is fueli...	HPVSideEffects	have seen this fueling hate httpstcoiwujcrzzya	-0.800000
4	2021-10-28 11:58:21+00:00	1453692611475161090	@nytimes @cnnbrk @washingtonpost @foxnews @tuc...	HPVSideEffects	this second confirmed case vaccine trial data ...	0.016667
5	2021-10-28 11:45:10+00:00	1453689293847859200	they removed data from the studies. who knows ...	HPVSideEffects	they removed data from study know much this se...	0.112500

Figure 14 Sentiment polarity of the tweets



```
In [32]: sns.distplot(tweets_df1['Sentiment_polarity'])

         C:\ProgramData\Anaconda3\lib\site-packages\seaborn\distributions.py:2557: FutureWarning: `distp
         d will be removed in a future version. Please adapt your code to use either `displot` (a figure
         xibility) or `histplot` (an axes-level function for histograms).
           warnings.warn(msg, FutureWarning)

Out[32]: <AxesSubplot:xlabel='Sentiment_polarity', ylabel='Density'>
```

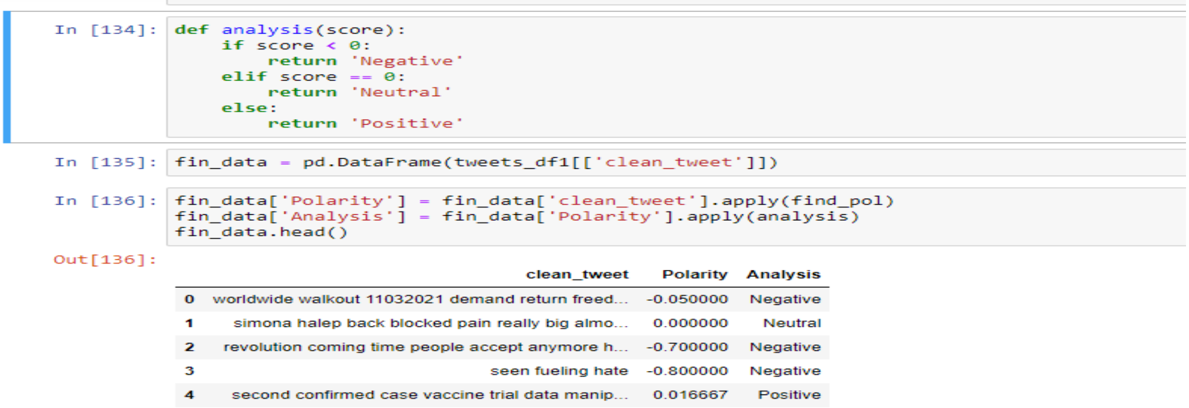
Figure 15 Sentiment graph of the tweets



```
In [34]: tweets_df1['positive'] = tweets_df1[tweets_df1.Sentiment_polarity >0].clean_tweet
         tweets_df1.head(10)

Out[34]:
```

	Datetime	Tweet Id	Text	Username	clean_tweet	Sentiment_polarity	negative	positive
0	2021-10-29 04:29:23+00:00	1453942014374924295	worldwide walkouts on 11/03/2021 will demand a...	HPVSideEffects	worldwide walkout 11032021 will demand return ...	-0.050000	worldwide walkout 11032021 will demand return ...	NaN
1	2021-10-29 04:26:19+00:00	1453941239393361926	simona halep: "my back got blocked, the pain i...	HPVSideEffects	simona halep back blocked pain really big almo...	0.100000	NaN	simona halep back blocked pain really big almo...
2	2021-10-29 00:33:24+00:00	1453882626113212423	@anri82967503 the revolution is coming. there ...	HPVSideEffects	revolution coming there will time when people ...	-0.100000	revolution coming there will time when people ...	NaN
3	2021-10-28 12:05:14+00:00	1453694343185936391	@moo4free have you seen this? the msm is fueli...	HPVSideEffects	have seen this fueling hate httpstcoiwujcrzzya	-0.800000	have seen this fueling hate httpstcoiwujcrzzya	NaN
4	2021-10-28 11:58:21+00:00	1453692611475161090	@nytimes @cnnbrk @washingtonpost @foxnews @tuc...	HPVSideEffects	this second confirmed case vaccine trial data ...	0.016667	NaN	this second confirmed case vaccine trial data ...

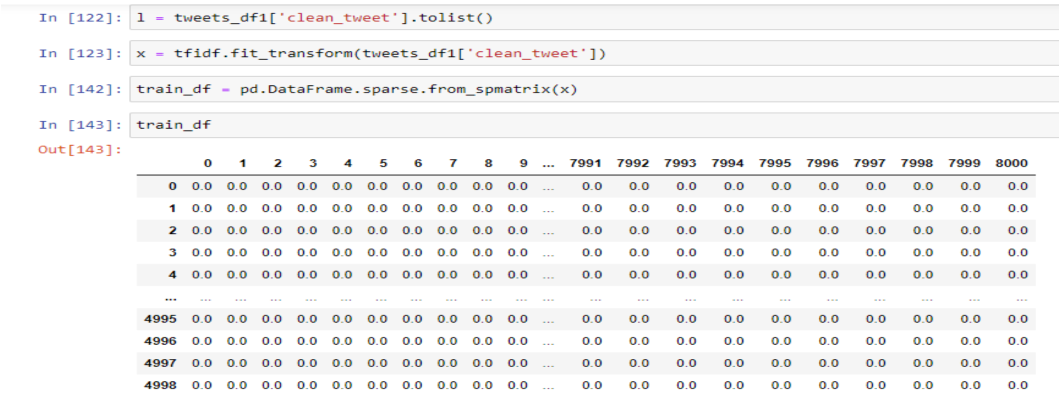
Figure 16 Sentiment of the tweets division in positive and negative columns



```
In [134]: def analysis(score):
              if score < 0:
                  return 'Negative'
              elif score == 0:
                  return 'Neutral'
              else:
                  return 'Positive'

In [135]: fin_data = pd.DataFrame(tweets_df1[['clean_tweet']])

In [136]: fin_data['Polarity'] = fin_data['clean_tweet'].apply(find_pol)
          fin_data['Analysis'] = fin_data['Polarity'].apply(analysis)
          fin_data.head()

Out[136]:
```

	clean_tweet	Polarity	Analysis
0	worldwide walkout 11032021 demand return freed...	-0.050000	Negative
1	simona halep back blocked pain really big almo...	0.000000	Neutral
2	revolution coming time people accept anymore h...	-0.700000	Negative
3	seen fueling hate	-0.800000	Negative
4	second confirmed case vaccine trial data manip...	0.016667	Positive

Figure 17 Analysis of the tweets corresponding to sentiment and stored into analysis



```
In [122]: l = tweets_df1['clean_tweet'].tolist()

In [123]: x = tfidf.fit_transform(tweets_df1['clean_tweet'])

In [142]: train_df = pd.DataFrame.sparse.from_spmatrix(x)

In [143]: train_df

Out[143]:
```

	0	1	2	3	4	5	6	7	8	9	...	7991	7992	7993	7994	7995	7996	7997	7998	7999	8000
0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
4995	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4996	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4997	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4998	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Figure 19 Train dataset for the analysis