Abstract

Sentiment analysis of Twitter activity concerning therapeutic treatments and possible vaccines during the COVID-19 pandemic: analysis of Hydroxychloroquine, Remdesivir, Regeneron, and COVID-19 vaccines.

Keywords: twitter api, sentiment analysis, opinion mining, social networks, python, nltk, textblob, pandas, matplotlib, covid-19 vaccines, covid-19 treatments, naive bayes

Twitter Sentiment Analysis of CVOID-19 Vaccines and Therapeutic Treatments

With the current COVID-19 pandemic there has been a great deal of conversation around possible therapeutic treatments and possible vaccines. The goal of the project is to characterize public sentiment regarding therapeutic treatments and possible vaccines by analyzing Twitter activity using key dates for differentiation.

|  |  |
| --- | --- |
|  | *“It’s estimated that 80% of the world’s data is unstructured, in other words it’s unorganized. Huge volumes of text data is created by social media sites like Twitter.” Schneider (2016)* |

This project does not attempt to answer the cause(s) of the sentiment toward COVID-19 vaccines or therapeutics, it simply seeks to analyze the sentiments of keywords surrounding COVID-19. To accomplish this, I’ve analyzed Twitter activity for three of the primary therapeutic treatments for COVID-19; Hydroxychloroquine Remdesivir, and Regeneron. For comparison, I’ve also pulled Twitter activity for “covid-19 vaccine”.

While this project won’t solve any COVID-19 related issues, my objective is to at least provide insights on what some users on one social media site are saying and feeling.

# Getting Started!!

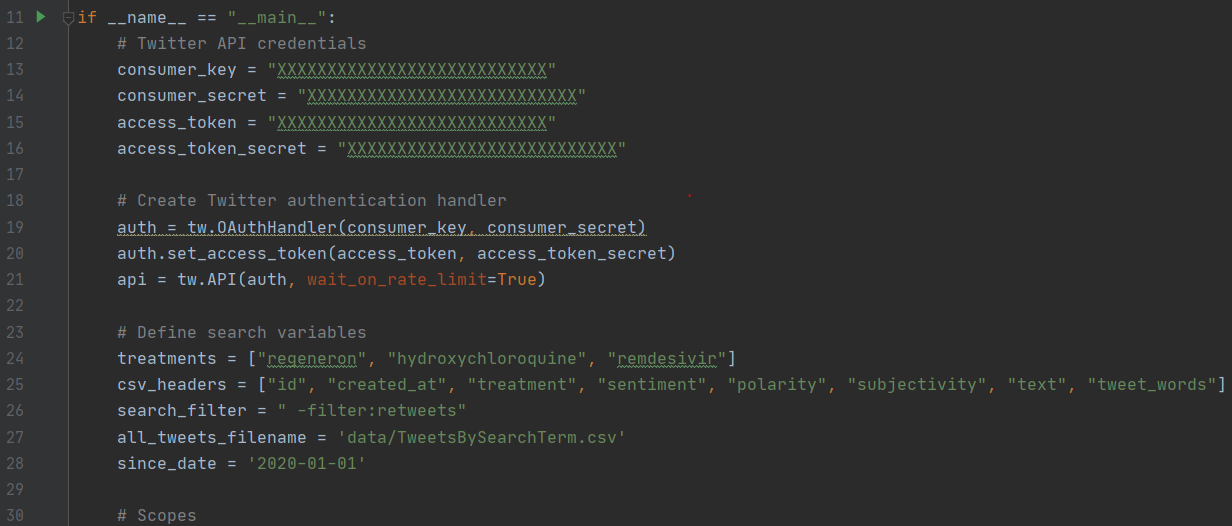
Up front my hunch is that in general there is a more positive sentiment towards vaccines than treatments, simply given the fear factor of the virus and the political environment that exists. President Trump has favored treatments that that serve more as a therapeutic cure, whereas vaccines serve a preventative measure. For this reason, my hunch is that many people like the general idea of not getting the virus at all rather than getting it and having to then deal with the doctor, medication, possible hospitalization, and overall risks.

Let’s see how this theory of mine plays out!

# The Data

The data collected for this project comes solely from Twitter. I collected 600 tweets in total with and even split for two groups of 300. One group represents COVID-19 vaccines, the other group represents COVID-19 treatments - Hydroxychloroquine, Remdesivir, and Regeneron.

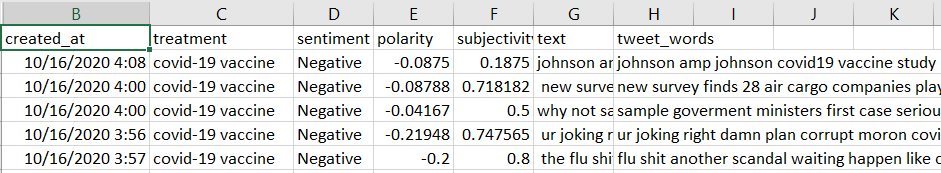
I used Tweepy and the Twitter API to gather the tweets need for this project, and the “csv” client to store the results in a CSV format file.



Because I need to evaluate the sentiments behind COVID-19 vaccines as well as COVID-19 treatments, it made sense to make a list of the search terms, loop through them to get the appropriate tweets for each term, pre-process and clean each tweet, and then write them to a csv file as they are retrieved.

There are other challenges for COVID-19 vaccine. Like, how do I search for them when most people don’t know the vaccine names, therefore they don’t use them in tweets? The answer was to search for a generic “covid-19 vaccine” term and use that to represent tweets about COVID-19 vaccines.

It worked! As you can see the file contains a few extra fields as a result of the pre-processing performed prior to insertion. I describe these steps, methods and reasons below in the Cleaning the Data section.



# Cleaning the Data

I need the real vaccine and treatment content in each tweet, but Twitter tweets are messy in their raw unstructured form. I’ll need to do a good bit of cleaning in order to make sense of them and present them in an NLP model. The average tweets might contain URLs, Twitter username handles, swear words, and number of special characters of combinations of all of the above. Cleaning the data will strip the content of the tweets down to their relative words.

Data Cleaning operation to be performed:

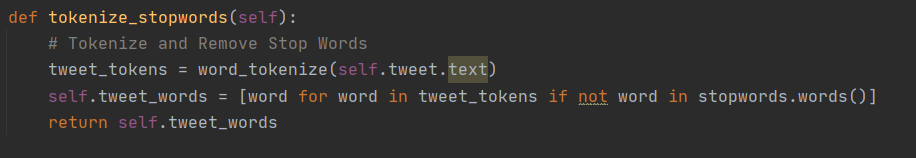
* Remove URLs
* Remove handles
* Remove special characters
* Removed single character words
* Convert all characters to lower case
* Replace ”covid 19” to “covid19” to make things easier



We’re almost at a point where we can look at the data more closely to understand the sentiments around COVID-19 vaccines and therapeutic treatments. Now I can process the words in each tweet and generate the sentiments for each while I’m at it.

Other pre-processing:

* Tokenization
* Removing Stopwords



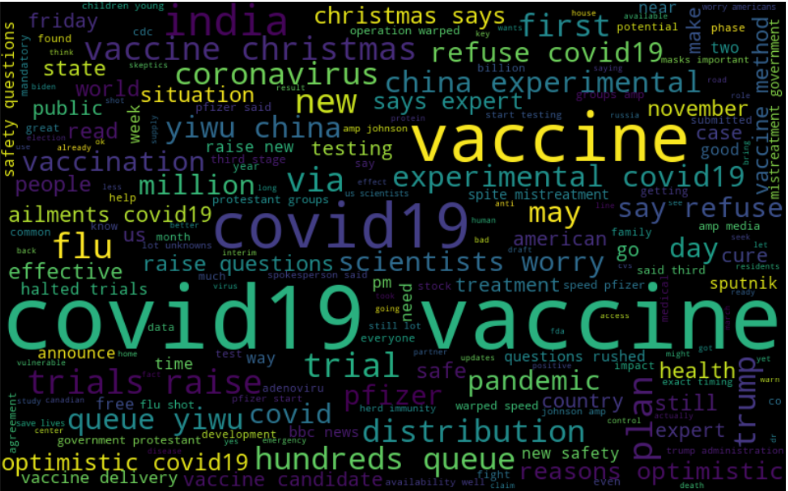
# Exploring the Data

Finally, I can start looking at the data in a bit more detail. To do this I used word clouds, word frequency, and histograms. In producing the word analysis, I ran across a number of words that I felt were not adding value and obscuring the results a bit. They didn’t seem to be vaccine or therapeutic related, so I omitted then from the word clouds and other EDA methods.

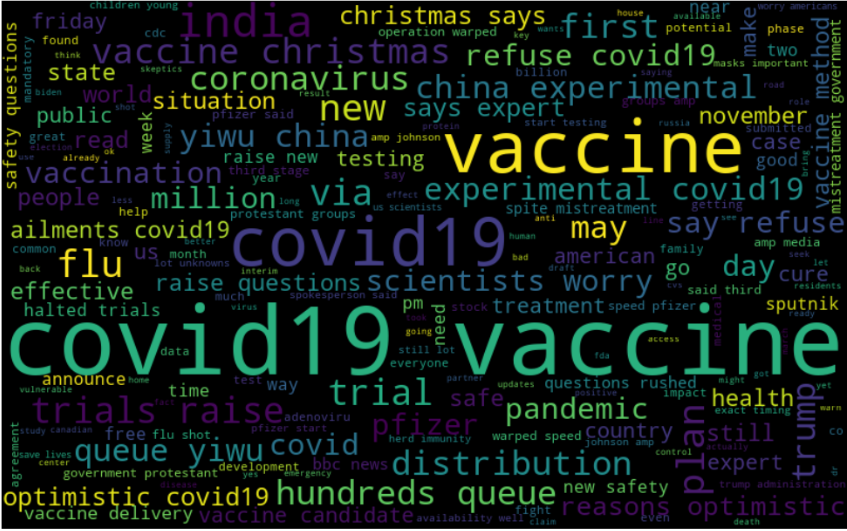


Word Clouds

I found the word clouds to be fascinating seeing all a collection of key words from these COVID-19 vaccine tweets below. A few of the words that help me get a glimpse at sentiment are refuse, experimental, effective, worry, need, and safety.

  
Word cloud for “covid-19 vaccine” tweets

What struck me about the word cloud for therapeutics is the words like experimental, raise questions, safety questions, halted trials, and testing. Overall, they seem to show a more factual questioning of the vaccines without an emotional reaction. But I’ll need to look closer because word clouds do not provide sentiment, just the words.

  
Word cloud for Covid-19 therapeutic tweets (Hydroxychloroquine, Remdesivir, Regeneron)

## Words Frequency

Seeing to vaccine and therapeutic related words was insightful and helpful, but we need to put that part of the story in context by adding word frequency counts. Well, word frequency is showing another part of the picture with regards to possible issues they raise when discussing vaccines or therapeutics. Many of the top words are very predictable as they were part pf the search terms used to retrieve the tweets. I left them in simply to show their frequency.

What’s interesting are some of the words at the bottom of the lists. For instance, “Christmas”. It ranks 8th for vaccines and 24th for therapeutics. I assume that users are interested in when these treatments will be available, my hunch is that they are talking more often about vaccines by Christmas, and not so much with the therapeutics because the therapeutics largely already exist.

|  |  |
| --- | --- |
|  |  |
|  |  |

## Common Words – Histograms

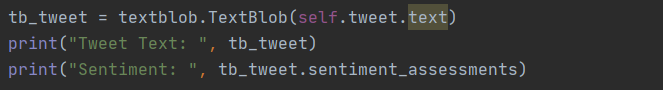
I’ve added histograms for word frequency; however I don’t think gleen much from this look at common words. Perhaps scaling would help but I think the grid view with number is more helpful with common and frequent words.

|  |  |  |  |
| --- | --- | --- | --- |
| |  |  | | --- | --- | |  |  | |  |

# Machine Learning

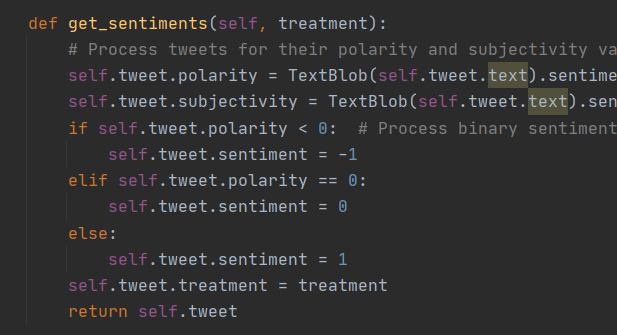
## Unsupervised Rules-Based Machine Learning Approach

Here I use TextBlob to simple model lookup dictionary. In this code I provide TextBlob each tweet’s text and receive sentiment values in return. This method is unsupervised so there is no training, but does provide



The printed results on this particular tweet came out of the API return for search term “covid-19 vaccine” and return .5 for both polarity and subjectivity. So, at least for this one COVID-19 vaccine related tweet the sentiment is “positive”, and the subjectivity tells us that the tweet is regarded as being subjective or and opinion. A score of 0 or less would have indicated that the vaccine tweet was more objective or factual. This method was applied to each of the 600 tweets that I retrieved via Pandas, Tweepy, and the Twitter API. An additional column was inserted that contains a sentiment lexicon to classify our COVID-19 treatment tweets positive, neutral, or negative.

I created a function to handle the rules-based TextBlob method to process the sentiments for each COVID-19 treatment related tweet.



### Sentiment Numbers

In taking a look at the lexicon totals, I can now get a more solid understanding of the feelings toward COVID-19 vaccines and COVID-19 therapeutic treatments. While neither category scored higher than 50% positive, therapeutics did score 6% higher in positive sentiments and 2% less in negative sentiments. It would seem that overall Twitter users also are currently more neutral on vaccines and don’t have a strong positive of negative sentiment about the COVID-19 vaccines.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Positive** | **Neutral** | **Negative** |
| **Vaccines** | 102 | 155 | 43 |
| **Therapeutics** | 120 | 144 | 36 |
|  |  |  |  |
|  | **Positive** | **Neutral** | **Negative** |
| **Vaccines** | 34% | 52% | 14% |
| **Therapeutics** | 40% | 48% | 12% |

### Sentiment – Histograms

Below we see that the sentiments for vaccines and therapeutics do not have the same relative distribution with vaccines having a heavy neutral to negative leaning sentiments. While therapeutics is somewhat evenly distributed among all three sentiments, it too leans toward the neutral and negative end. Overall COVID-19 vaccines carry a more significantly more negative sentiment than therapeutic treatments.

|  |  |
| --- | --- |
|  |  |

### Subjectivity - Histograms

While the histograms for sentiment showed a more positive scoring from COVID-19 therapeutics, the subjectivity for both is very clearly into positive territory which tells me that people are speaking in a more factual and objective manner on vaccines and therapeutics.

|  |  |
| --- | --- |
|  |  |

### Scatterplots – Polarity/Subjectivity

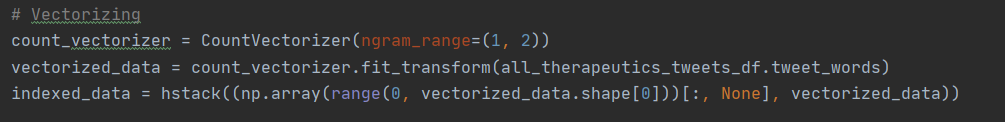
This tells me that as the tweets for both vaccine sand therapeutics become more positive in sentiment, they also become more subjective (an opinion). It also shows that as they begin to have a more negative sentiment, they also tend to be speaking in a more objective of factual manner. This could mean that Twitter users are optimistic about both but grounded in optimism when it comes to whether or not vaccines or treatments are “good”.

|  |  |
| --- | --- |
|  |  |

## Supervised Machine Learning Approaches

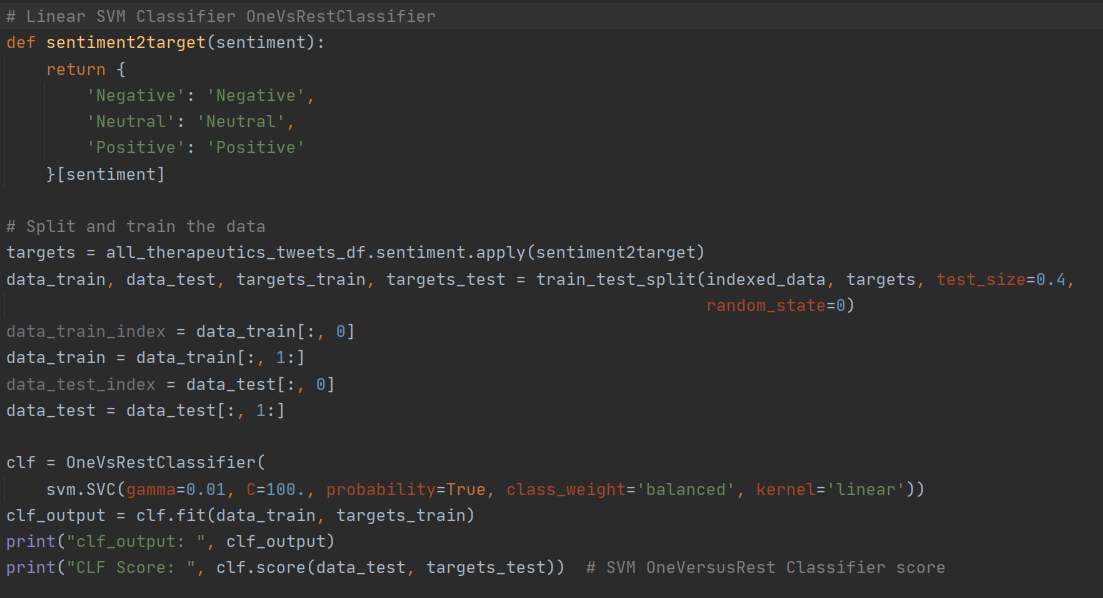
### Vectorizing

Because machine learning algorithms expect data in a mathematical form, I first need to vectorize the tweet data. This means the text data will be converted into vectors for the machine learning classifiers to understand.

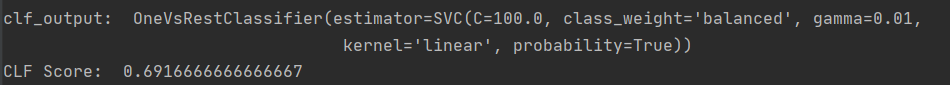


## Linear Support Vector Machine (SVM) Classifier

With SVM we seek to find a hyperplane that divides the data optimally in for each class. For my sentiment analysis I have three class (positive, neutral, and negative), so a different method would be useful. To accomplish this task, I used a OneVsRestClassifier which allows me to find the probability distribution of all three classes (Langkilde, n.d.). As the OncevsRest name implies, each one of the classifiers takes a look at the data point and determines the probability that it belongs to its class.

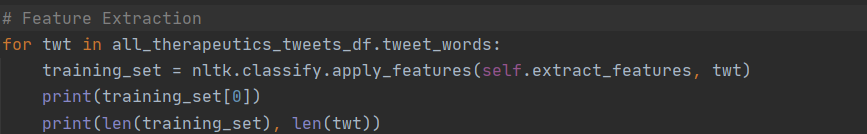


The results output provides us information on the classifier object and the CLF score. The classifier scored a .692 which is a fairly low performance score. With some tweaking this performance could most likely be improved.



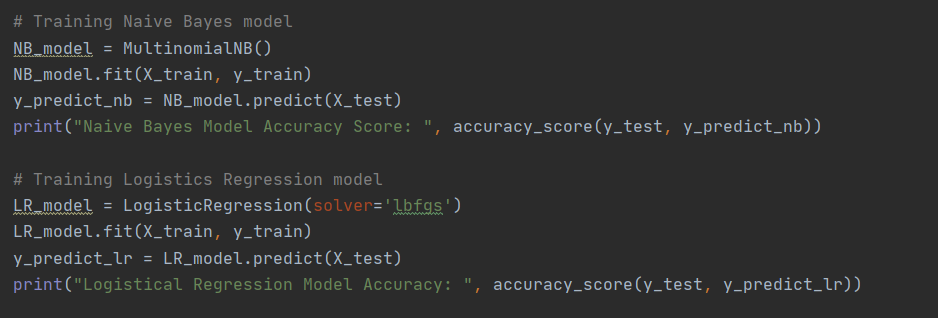
## Feature Extraction

With cleaned data, I can now extract word features from the tweets. Each tweet will get a key for every unique word in the tweet.



## Naïve Bayes & Logical Regression Classification

With feature extraction complete I can now proceed with classification using the naïve bayes classifier which assumes there is no relationship between features. This supervised learning approach will help me make predictions on classified training data.



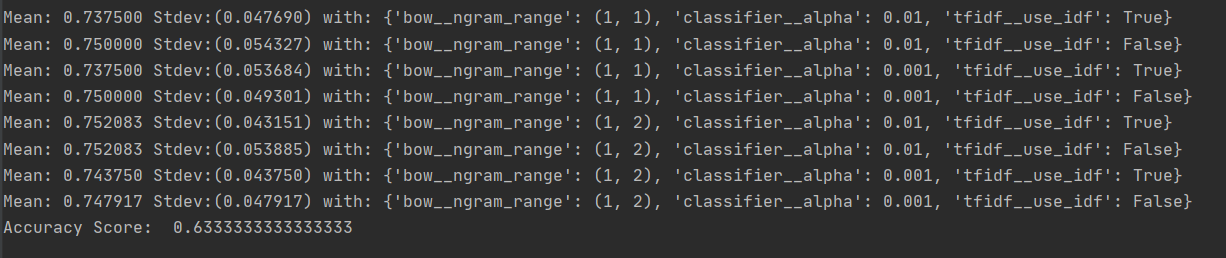
Both models performed exactly the same with scores at .00833. Terrible scores!!



# Model Evaluation

To evaluate the models, performance and accuracy a pipeline is created for grid search parameters and workflow on the model. Grid search will make sure all combinations are tested. In this case the model received an accuracy score of .633 which again is quick low. More tweaking on this model is need for higher accuracy.

The results also show the means, standard deviation, bag-of-words ngram range, classifier alpha and TF-IDF numbers.

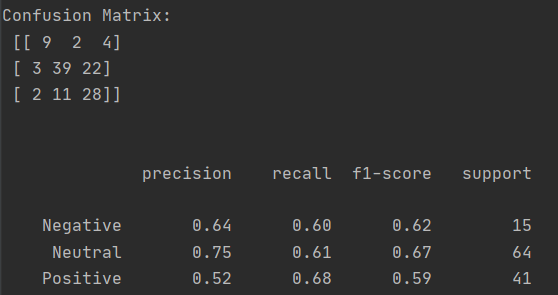


## Confusion Matrix

In the 3x confusion matrix below, the (9, 39, 28) main diagonal provides the correct predictions. Meaning that these are the cases where the model predictions equal the actual values.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | **Predicted** | | |
|  |  | **Pos** | **Ntl** | **Neg** |
| **Actual** | **Pos** | 9 | 2 | 4 |
| **Ntrl** | 3 | 39 | 22 |
| **Neg** | 2 | 11 | 28 |

The classification report shows the figures for precision, recall, an the f1 score for each of the three sentiments. If we look at the numbers for precision, we can see that the mode isn’t very accurate, especially for the positive sentiment tweets for COVID-19 treatments which come sit at just over 50%, the equivalent to flipping a coin. The recall numbers are all in the .60 range indicating a poor ability to find positive instances. The f1 scores really lets us know that this model is pretty lousy over all with .59-.67 scores.



# Conclusion

Overall, I discovered that there is mostly neutral sentiment from tweets around COVID-19 vaccines and therapeutics with mostly subjective content for both. There was significantly more positive tweets around therapeutics and slightly more negative vaccine tweets.

None of the machine learning models proved to be particularly accurate with the SVM classifier scored much higher than Naive Bayes and logical regression classifiers.

# Coding and other valuable resources

I wish I could acknowledge each piece of code I used for inspiration but there are far too many to list. So, I’ve listed a few of the more useful resources I used below.

|  |  |
| --- | --- |
| **Topic** | **Sources** |
| Naive Bayes | https://towardsdatascience.com/creating-the-twitter-sentiment-analysis-program-in-python-with-naive-bayes-classification-672e5589a7ed |
| SVM | https://www.kaggle.com/langkilde/linear-svm-classification-of-sentiment-in-tweets |
| Feature Extraction | https://medium.com/@jerryfadugba/sentiment-analysis-classification-using-2-different-ml-methods-1ba01021dae7 |
| Naïve Bayes | https://medium.com/better-programming/twitter-sentiment-analysis-using-naive-bayes-and-n-gram-5df42ae4bfc6 |
| Tweepy | <https://www.earthdatascience.org/courses/use-data-open-source-python/intro-to-apis/analyze-tweet-sentiment-in-python/> |
| EDA | https://www.kaggle.com/tanulsingh077/twitter-sentiment-extaction-analysis-eda-and-model |
| Training model | https://www.kaggle.com/amar09/nltk-feature-extraction-and-sentiment-analysis |
| Confusion matrix | https://towardsdatascience.com/the-real-world-as-seen-on-twitter-sentiment-analysis-part-two-3ed2670f927d |
| Classification Report | https://medium.com/@kohlishivam5522/understanding-a-classification-report-for-your-machine-learning-model-88815e2ce397 |
| Sentiment Analysis | https://www.analyticsvidhya.com/blog/2018/07/hands-on-sentiment-analysis-dataset-python/ |

References

Langkilde, D. (n.d.). *Linear SVM classification of sentiment in tweets.* Kaggle. Retrieved October 16, 2020, from https://www.kaggle.com/langkilde/linear-svm-classification-of-sentiment-in-tweets

Pozzi, F.A., Fersini, E., Messina, E., Lui, B., (2017). *Sentiment Analysis in Social Networks.* Cambridge. MA: Elsevier

Schneider, C. (2016, May 25). The biggest data challenges that you might not even know you have. *IBM.* Retrieved from https://www.ibm.com/blogs/watson/2016/05/biggest-data-challenges-might-not-even-know/