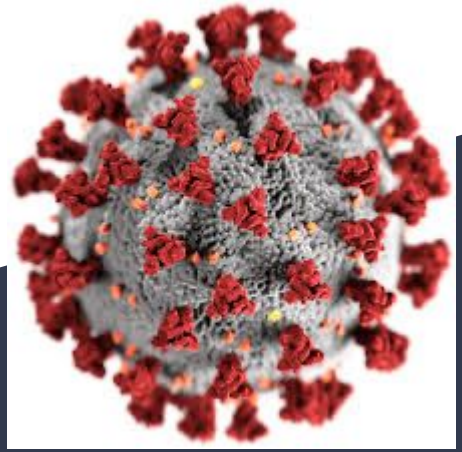
# NLP
# CoronaVirus tweets research

Roman Garayev



# Dataset

COVID-19 tweets dataset. Perform Text Classification on the data. The tweets have been pulled from Twitter and manual tagging has been done then.

The names and usernames have been given codes to avoid any privacy concerns.

Columns:

1) Location

2) Tweet At

3) Original Tweet

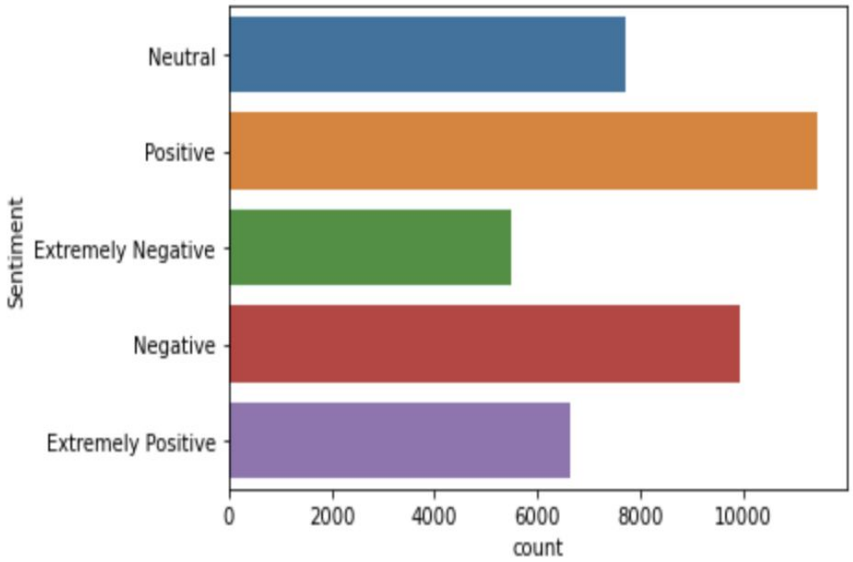
4) Label

| | UserName | ScreenName | Location | TweetAt | OriginalTweet | Sentiment |
|---|---|---|---|---|---|---|
| 0 | 3799 | 48751 | London | 16-03-2020 | @MeNyrbie @Phil_Gahan @Chrisitv https://t.co/i... | Neutral |
| 1 | 3800 | 48752 | UK | 16-03-2020 | advice Talk to your neighbours family to excha... | Positive |
| 2 | 3801 | 48753 | Vagabonds | 16-03-2020 | Coronavirus Australia: Woolworths to give elde... | Positive |
| 3 | 3802 | 48754 | NaN | 16-03-2020 | My food stock is not the only one which is emp... | Positive |
| 4 | 3803 | 48755 | NaN | 16-03-2020 | Me, ready to go at supermarket during the #COV... | Extremely Negative |

# Task

Construct NLP Pipeline, check performance on unigrams, bigrams, threegrams and 1-3grams, apply unsupervised methods in order to reduce dimension

# Target Data - Tweets' Sentiment



# Data Preparation and Tokenizing

Removing:

References

Punctuation

Stop Words (sklearn english stop words and custom)

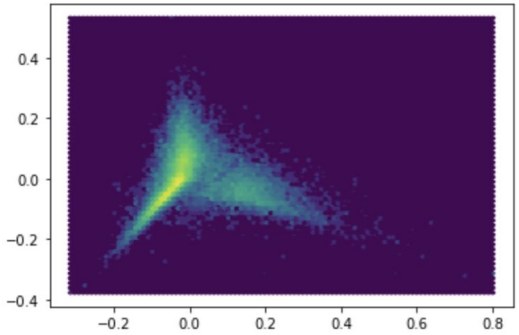
To lower case

Tokenizing through sklearn.Tfidfvectorizer with min_df = 0.004,

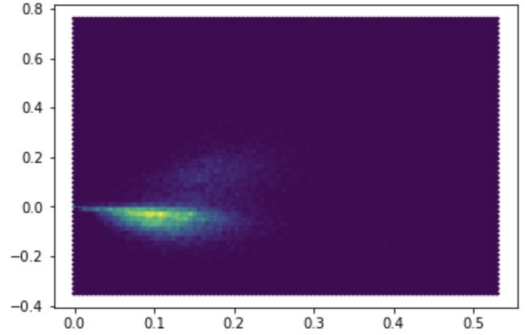
max_df = 0.65 and using casual_tokenize

Dimension Reduction methods allow us to interpret multidimensional data into 2 or 3 dimensions where we can visualize it. Also they allow to avoid "dimension curse". These methods use Singular Matrix Decomposition. The Decomposition spreads matrix into 3 other matrix (2 ortodiagonal and 1 symmetrical). The eigenvectors and eigenvalues of this matrix are used to construct new representation. As we can see on plots, even 4000+ dimension data can be represented as well.

PCA




Truncated SVD



# LDA. SVD Solver. Comparing Accuracy on different datasets.

According to result, unigrams showed the best performance among all.

Metric: Accuracy

| Data | Result. LDA |
| --- | --- |
| unigrams | 0.58 |
| bigrams | 0.36 |
| threegrams | 0.3 |

# Logistic Regression and Random Forest

Among supervised algorithms were used, Logistic Regression with Elastic NET regularization showed the best performance. Unfortunately, because of lack of the computational resources and huge dataset size, I can not use complicated model like Gradient Boosting

Logistic Regression on unigrams: 0.6

Random Forest on unigrams: 0.52