SENTIMENT ANALYSIS

TWITTER POSTS COVID-19

NATURAL LANGUAGE PROCESSING

Leticia Drasler,

March 2022 - Capstone

OUTLINE

COVID-19

Natural Language Processing

Data

Project

Models

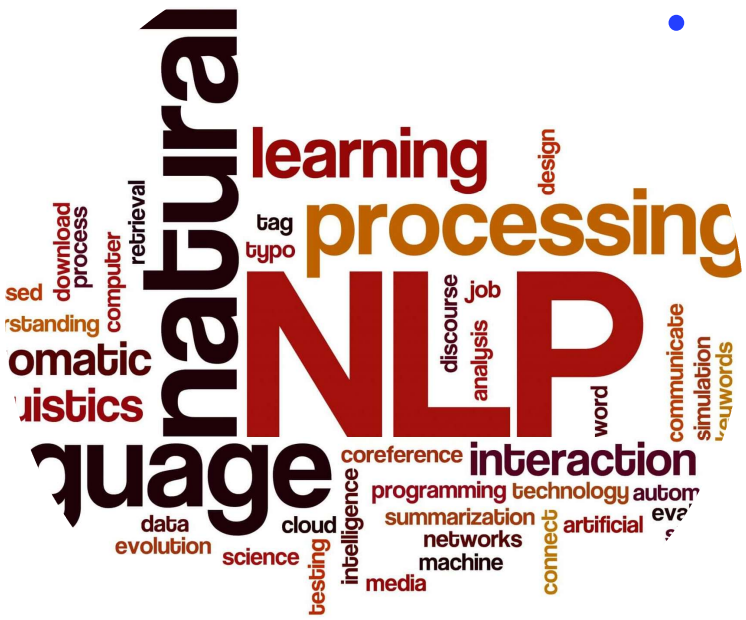
Results

COVID-19

In 2020 the new coronavirus rapidly starts to spread throughout the world, and along with it, fear, and uncertainties. By that scary time, worldwide people started to share their viewpoints on social media even more often than usual. Twitter was one of the most common platforms for people to expose their feelings, opinions, and emotions.

Natural Language Processing

Natural language processing is a machine learning tool that allows computers to interact with humans using human language, understanding manual texting.



SENTIMENT ANALYSIS

Text classification – Twitter posts

Data

- These data of tweet posts were collected through Kaggle, and it is free to anyone
- These data have about 45000 tweets from all around the world.

Free access to the dataset:

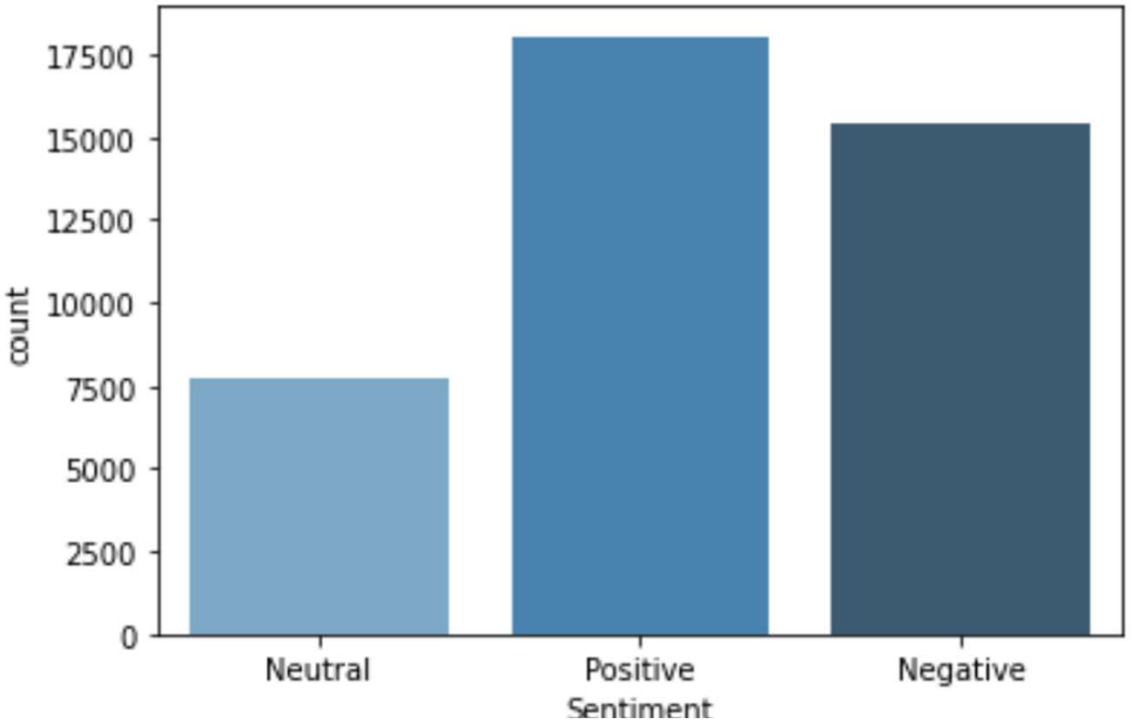
https://www.kaggle.com/ravikumarmn/covid-19-text-classification-using-bert-tpu/data

Data

Classes

The dataset was divided into 3 categorical classes, which I turned into numerical.

Neutral: 0
Negative: 1
Positive: 2



The project

For this text classification, I worked with Python programming language inside the jupyter notebook to prepare, preprocess, and analyze these data.

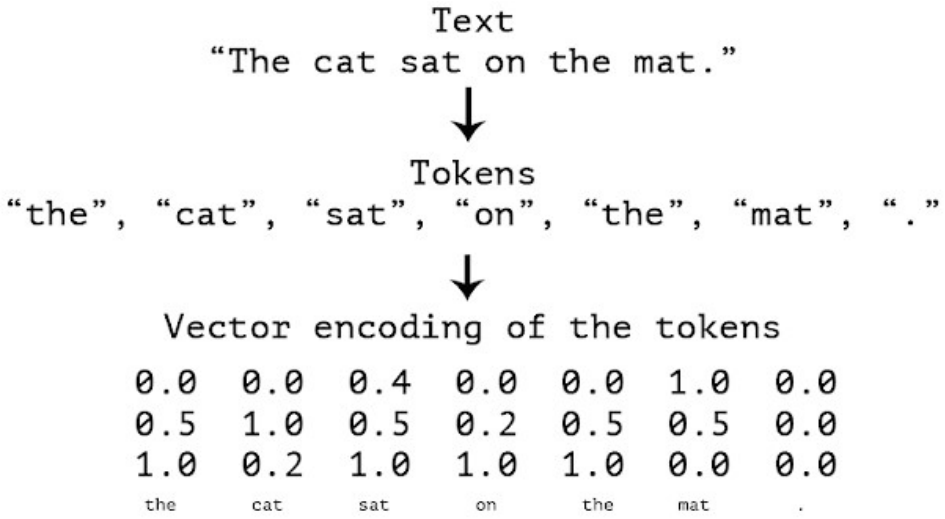
I used many tools such as:

- Pandas,
- NumPy,
- Seaborn,
- Matplotlib,

The project

For the preprocessing steps, I worked with:

- Word tokenizer,
- GloVe, and
- Vectorizer.



The project

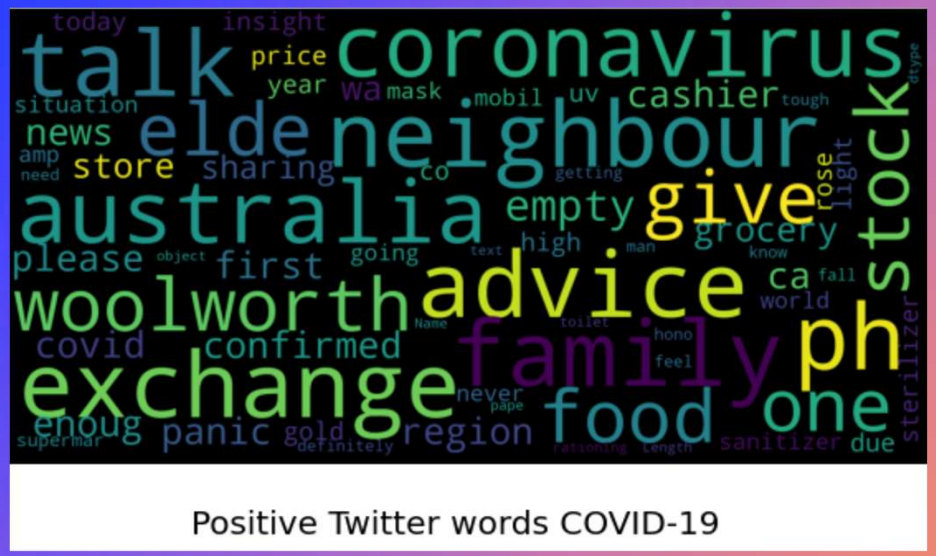
I applied 3 baseline models:

- Random Forest classifier
- Support Vector Machine
- Logistic Regression

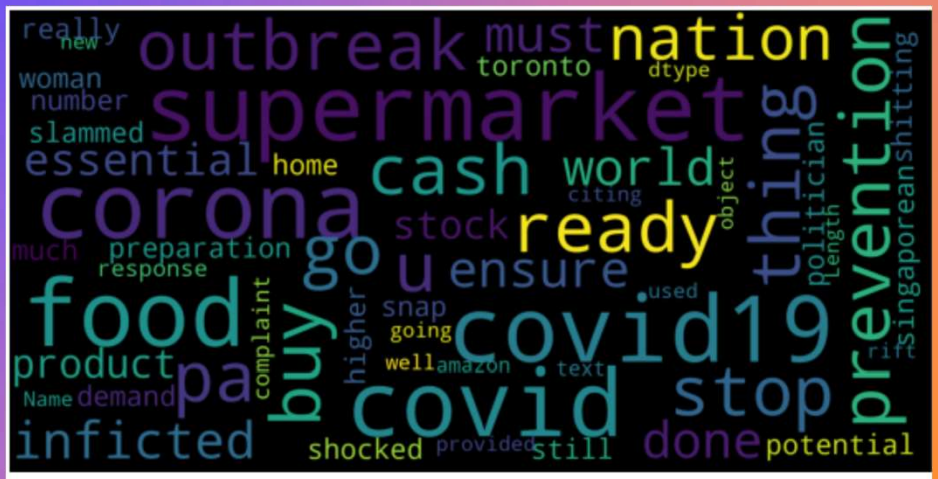
The project

For my final model I built a neural network with embedding, using:

- Keras,
- TensorFlow,



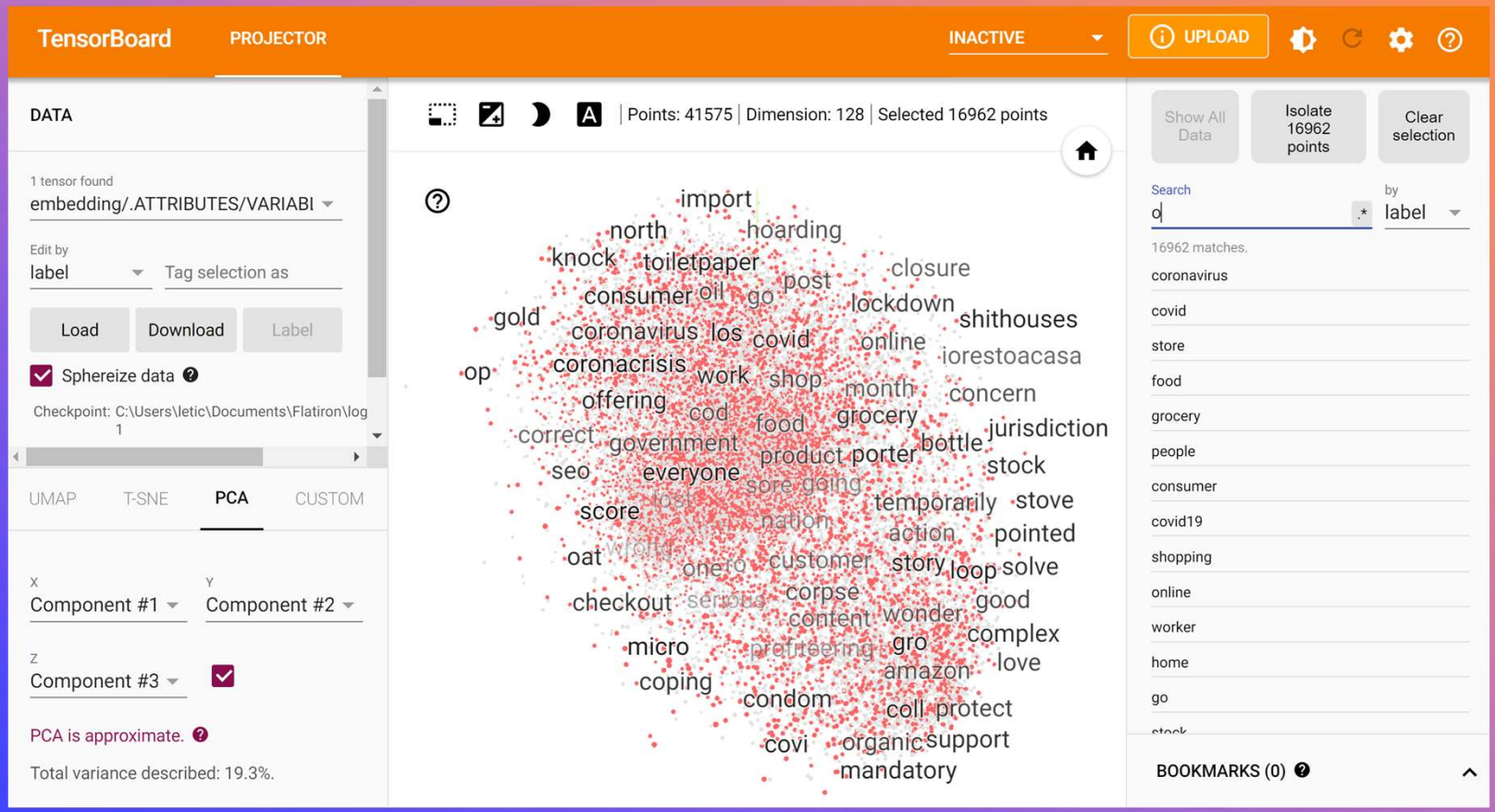
Positive Twitter words COVID-19



Negative Twitter words COVID-19

In these word clouds, we can see some of the most used words on positive and negative tweets

TENSOR BOARD



The models

Baseline model results:

- Random Forest Classifier
  test accuracy of 60%
- Support Vector Machine
  test accuracy of 63%
- Logistic Regression
  test accuracy of 61%

The models

Final results:

Neural Network with Embeddings

We achieved a test accuracy of 82% for our final model.

INTERPRETATIONS AND ANALYSIS

Leticia Drasler

LinkedIn: https://www.linkedin.com/in/leticiadrasler/

GitHub: https://github.com/lddrasler

Contact: leticia.drasler@gmail.com

THANK YOU

Leticia Drasler