Twitter Sentiment Analysis

It is no longer difficult to understand what people think about a topic by analysing the tweets shared by people. Sentiment analysis is one of the most popular use cases for NLP (Natural Language Processing).

The aim of this assignment is to analyse what Twitter users think about the vaccination against COVID-19.

The assignment is splited into five parts:

1. Data pre-processing and cleanup.
   - Set to lowercase
   - Remove stopwords
   - Remove any url
   - Remove hashtags ('#')
   - Remove ats ('@')
   - Remove punctuation
   - Remove emojis

2. Data Analysis
   - Number of positive/negative/neutral tweets
   - Most common words used in tweets
   - Most common words in positive, negative and neutral tweets
   - Tweets that contain the word 'astrazeneca', 'moderna', 'pfizer' or 'biontech'
   - Number of tweets per month
   - Average number of favorites and retweets
   - Number of negative tweets posted each month

3. Vectorization
   - Bag-of-words
   - Tf-idf
   - word embeddings

4a. Classifying tweets
- SVM
- Random Forests
- KNN

4b. 10-fold Cross Validation
- Precision
- Recall
- F-Measure
- Accuracy

5. Modeling LDA (Latent Dirichlet Allocation)