# Twitter sentiment analysis

Nghia Dang
Platini Dacheu

# Task 1
# Metadata

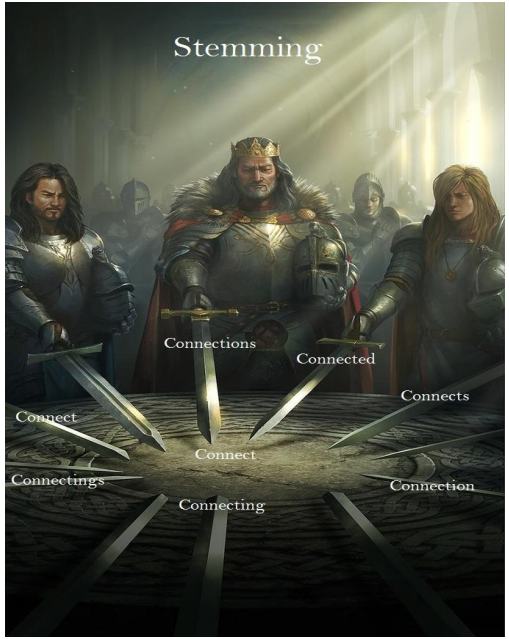
- Type: Supervised learning
- Sub type: Natural language processing
- Task: classify racist or sexist tweets from other tweets.
- Dataset: Tweets gathered from twitters
    - Features: Id, tweet content, label

# Task 1
# Preprocessing

- Stop words removal: if we remove the words that are very commonly used in a given language, we can focus on the important words instead.
    - First ten stop words: ['in', 'yourself', 'becoming', 'never', 'something', 'ten', 'ca', 'they', 'used', 'everyone']

# Task 1
# Preprocessing

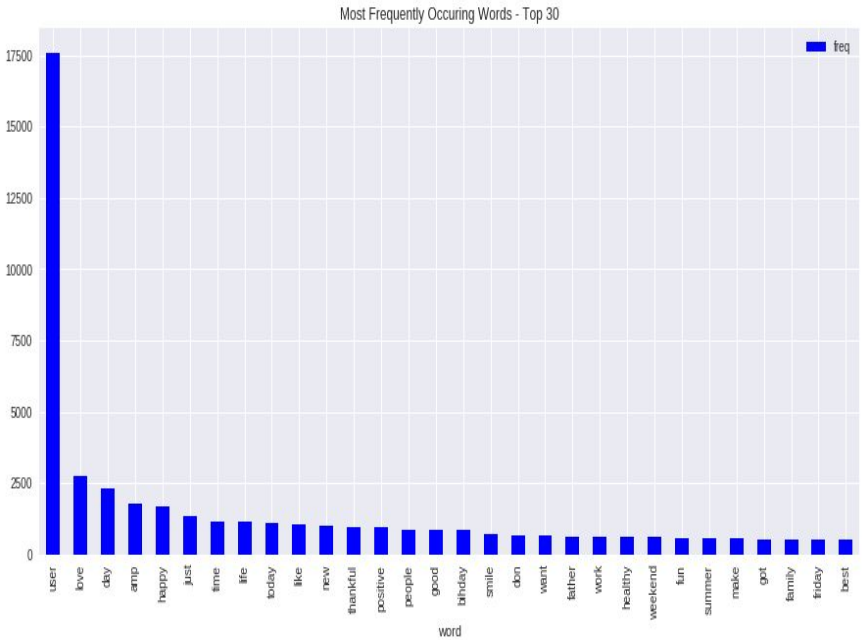
Stemming is a natural language processing technique that lowers inflection in words to their root forms, hence aiding in the preprocessing of text, words, and documents for text normalization.



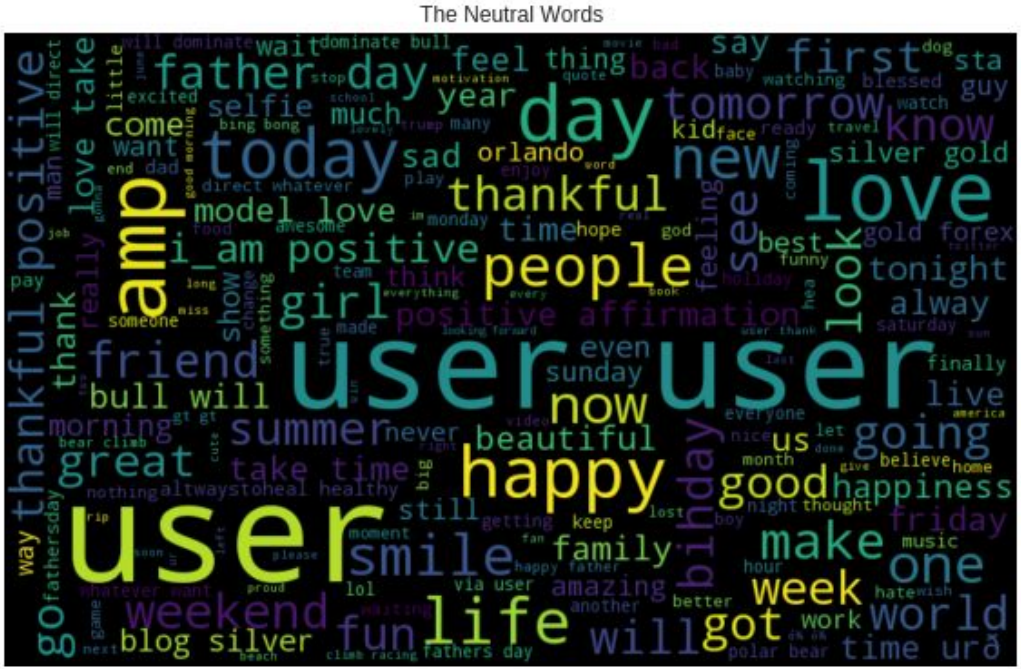
# Task 1
# Pipeline and visualization

- Visualization: Distribution of labels in train and test sets
- Visualization: Variations of length
- Feature engineering: Length of the tweet
- Standard scaling: Normalizing to mean 0 and std 1
    - Neutralize importance of all features
    - Dropping memory from displaying large numbers

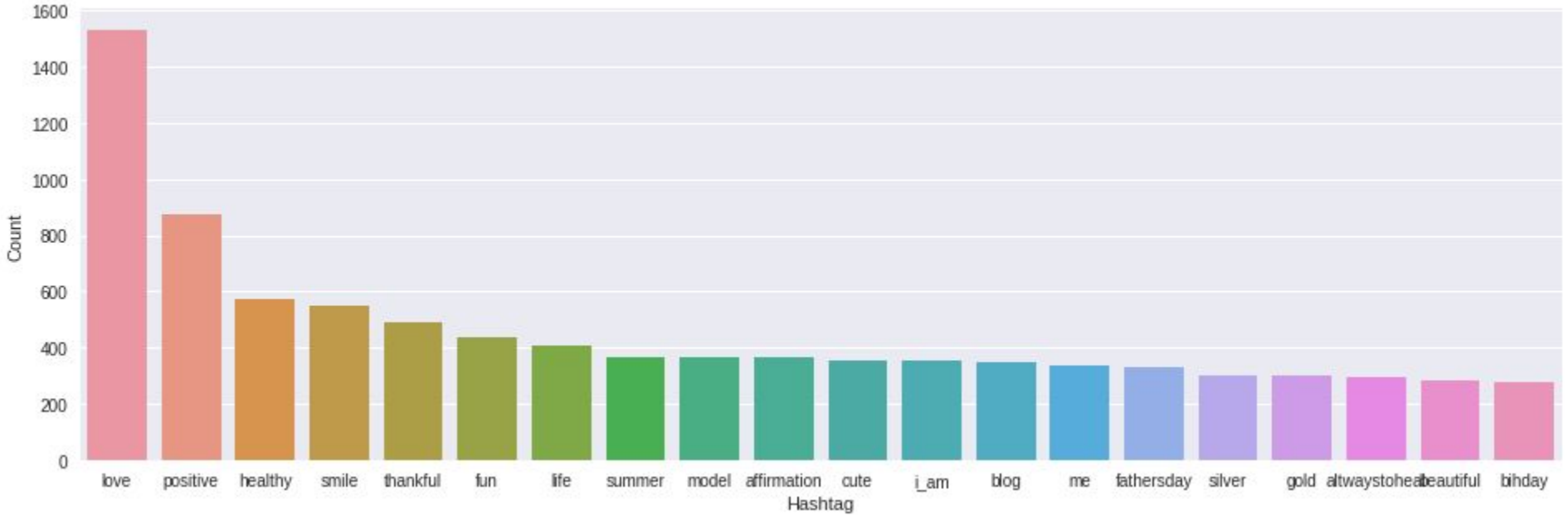
# Task 1
# Count Vectorizer



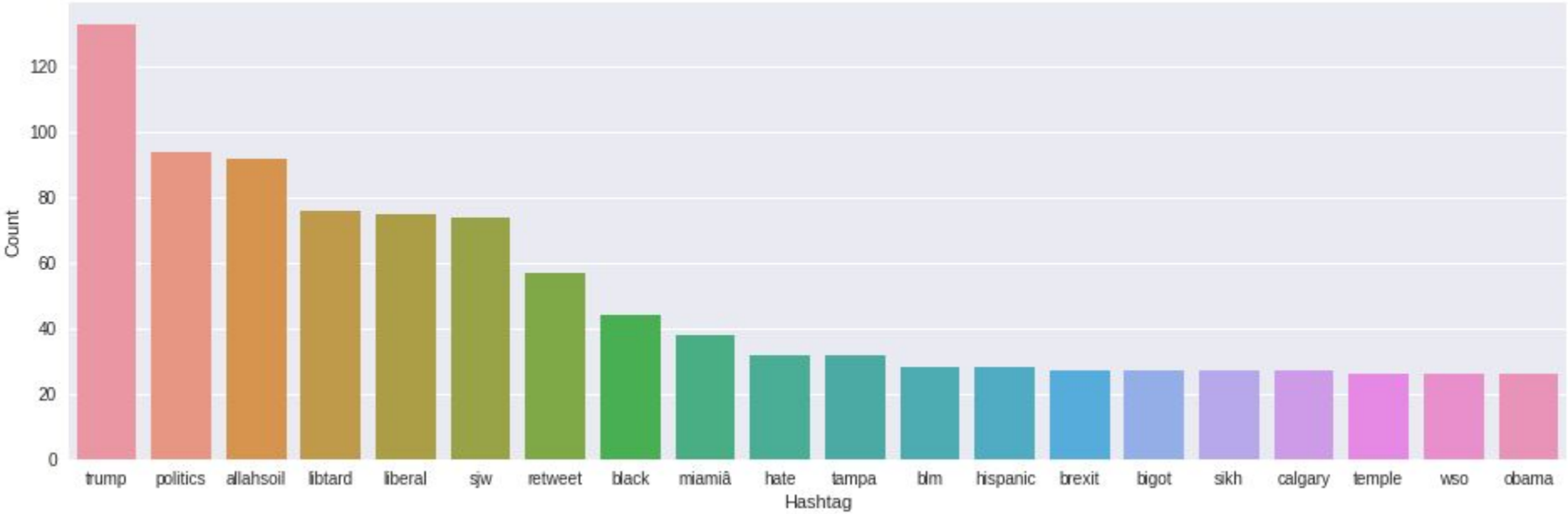
# Task 1
# Word Cloud



# Task 1
# Positive HashTag distribution



# Task 1
# Negative HashTag distribution



# Task 1
# Validation

```
from sklearn.model_selection import train_test_split

x_train, x_valid, y_train, y_valid = train_test_split(x, y, test_size = 0.25, random_state = 42)
```

# Task 1
# Word2Vect Model: Similarity

```
model_w2v.wv.most_similar(positive = "dinner")
```

```
/usr/local/lib/python3.6/dist-packages/gensim/mat
integer` is deprecated. In future, it will be tre
  if np.issubdtype(vec.dtype, np.int):
[('spaghetti', 0.6905485391616821),
 ('#prosecco', 0.6454159021377563),
 ('#wanderlust', 0.6233653426170349),
 ('coaching', 0.5992934703826904),
 ('podium', 0.5991446375846863),
 ('#restaurant', 0.5981520414352417),
 ('#boardgames', 0.597622275352478),
 ('sister!!', 0.5959595441818237),
 ('fluffy', 0.5959100127220154),
 ('#marbs', 0.5941186547279358)]
```

# Task 1
# Classification models

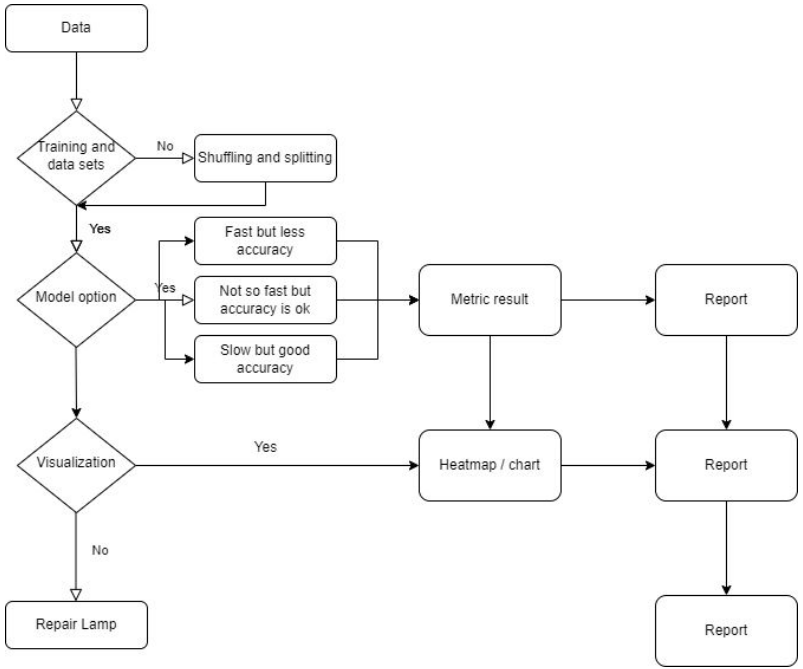
- Random Forest Classifier
- Logistic Regression
- Decision Tree Classifier
- SVC
- XGB

# Task 1
# Critics and proposal

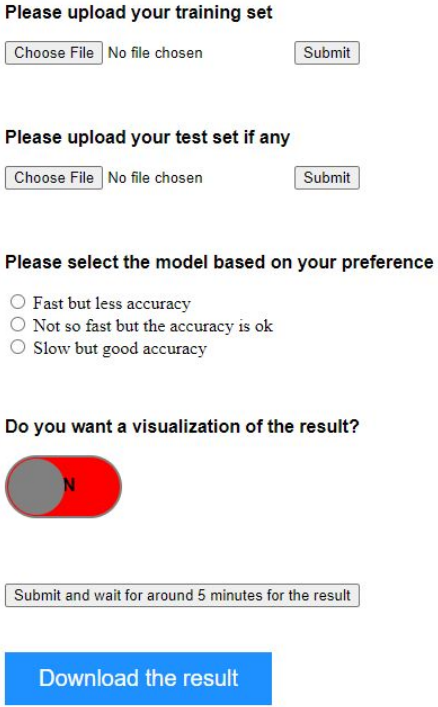
Critic: Using plain train_test_split is unstable for model performance

Proposal: Cross nested-validation

# Task 2
# UI design



# Task 2
# UI mockup



# Task 3
# Challenges

- Compatibility of libraries
- Our roles are sequential, not in parallel.
- Preprocessing is time-consuming: We had to take a fraction of the dataset samples to test. When we applied functions on the entire ones, new values showed up. Thus, we had to debug on the big dataset.

# Task 3
# Lessons learned

- Stratifying sampling the datasets that all unique values exist in the sample dataset, which is convenient for writing functions.
- Visualization helps in analyzing data.
- We should check for the balance of the dataset for appropriate sampling approaches.
- Deepnote is a solution for a teamwork when cells can be run in parallel.

# Reference

https://github.com/sharmaroshan/Twitter-Sentiment-Analysis