IML Hackaton

Cohen Yoav
Porezky Tal
Merkulov Ilya
Weiss Amit

June 7, 2019

Shortly after we started working on our classifier we reasoned that our main problem is the feature selection. Both choosing them and deciding whether they are binary or not. The choice was often against basic logic or intuition, was made to the appeasing of the test data.

- Our first step was to use the bag of words to maximum results by making use of the inherent methods of the CountVectorizer class such as stop words, both English and Portuguese.
- It is worth noting, that the first step already achieved 70% success rate and from there, we managed to pull a few minor tweaks to raise the success rate by approximately 10 additional percents. Among these 'tweaks':
  - Length of each tweet
  - Tags - We counted the number of appearances of the symbol '@' representing the amount of tags a tweet is containing and the tendency of a user to tag other people in posts.
  - Countries - We examined whether a word from the file countries.txt, a text file containing every country, appear in a tweet.
  - Marks - Various marks and their combinations were examined. Among them: '!' ',' '?' and the regex ...* which detected a continuous use of 2 or more periods.
- Binary - Some features, much to our surprise, proved to be more useful in a binary score function. Describing whether a country was mentioned in a tweet instead of counting the number of countries, raised the success rate by 3%. However, in others e.g tag, no improvement was achieved.
- Correlation - We have also tried adding features describing correlation between length and mark or tag. These proved to be inconsequential.
- We used various classifiers of sklearn and found the best results with MultinomialNB classifier. We have tried to use the nltk library to process the effects of noun adjectives etcetra and emotions core but with no avail!