**Referendum in Catalonia – explorative data analysis on social media presence**

Ville Saarinen, Emmi Lahtisalo, Neli Noykova

**Summary**

In this project we explore the information, incorporated in two data sets extracted from Twitter on 08.10.2017 and 10.10.2017.

1. **Introduction**

Nowadays the conflict between globalism as ideology for creating stable, prosperous, and peaceful world order, and on the other hand the efforts of small nations to preserve their national identity by even creating new independent states within the existing states, is rapidly increasing. According Jonatan Haidt (<https://www.humansandnature.org/the-ethics-of-globalism-nationalism-and-patriotism> ) the year 2016 could become a global turning point for this conflict. Under these conditions the answer of the question about patriotism as positive phenomena is also not so clear because it can be viewed as neglecting “citizen of world (existing country)” identity. Similar phenomena happen also inside the countries. One of the most recent such events was the Catalan Independence Referendum on 1st of October 2017. The Referendum was pronounced as illegal from Spanish Government. The deep understanding of ongoing processes in Spain can help political and sociological scientists to find the reasons for these events, and also to refer the phenomena to similar processes happening in other countries.

The new insights concerning such events could help also politicians from different parties, and also all people who are interested in the current political situation.

Here we are focused on two data sets from 8.10 during the march against the independence of Catalonia, and on 10.10.2017 during and after Puigdemont’s conclusive speech. In our opinion these two events are among the most important ones after the Catalan Independence Referendum on 1st of October 2017.

Our goal is to provide explorative analysis of these data sets in order to find new information, incorporated in the data. We have tried to stay as close as possible to the original data and find and explore the information, involved there.

1. **The data**
   1. **Planning the experiment**

We have tried to obtain data from two of the most popular social networks – Twitter and Facebook.

Due to some technical problems with Facebook Gaphical API we decided to proceed with data sets obtained from Twitter using Twitter API. All obtained/streamed data are in .json format. **The script for obtaining/streaming Twitter data is provided here.**

Before to obtain the used data sets we obtained almost online one sample data set on 10.10.2017 just before the Puigdemont’s conclusive speech. This data involved 53199 posts. We carefully checked the obtained information and after that decided which tags to use for extracting the working data sets.

We were focused on people’s reaction only inside the Spain and Catalonia. One possible way to select them was based on country code. Even under these restrictions the posts were in several languages. After analyzing these data we determined the set of hash tags for the next two requests, as well as the time interval, which we needed for our work.

Since this sample data set was quite small (434 K), we have used it also to check if our scripts work correctly. Only after that we have applied them to the investigated data.

For further extracting we chose the following hashtags:

['coordinates', 'created\_at', 'favorite\_count', 'favorited', 'hashtags',

'id', 'id\_str', 'infl\_score', 'is\_retweet', 'lang', 'parent\_tweet\_id',

'place\_coordinates', 'place\_country\_code', 'place\_name', 'place\_type',

'possibly\_sensitive', 'retweet\_count', 'retweeted', 'source', 'text',

'user\_created\_at', 'user\_defined\_location', 'user\_followers\_count',

'user\_geo\_enabled', 'user\_id', 'user\_name', 'user\_screen\_name',

'withheld\_in\_countries']

* 1. **The data used**

The first data set was collected from 08.10.2017 . It includes 569 980 posts (data points) from time interval 14:26:29 – 21:33:35 and takes 409 520 K memory after the provided preprocessing.

The second data set was collected from 10.10.2017. We were lucky to get bigger data set of 1 499 997 posts from time interval 17:54:57- 23:54:24, which takes 1 081 125 K memory.

One of the possible explanations about the big difference between the number of posts in both data sets is that the march on 08.10 has happened during all the day and people probably have commented it during the whole day, not only for the 7 hours interval in the first data set. On the other hand it is logical to assume that people were mostly interested to comment the Catalonian independence just before Puidgemont’s conclusive speech.

Due to huge size of the second data set, it was first divided to 3 parts. After the preprocessing these 3 parts were merged back to one data file.

Both data sets have been preprocessed in the same way (described in the script Catalonia – Pre-processing.ipynb). For easier further analysis (e.g. removing duplicates) the ‘id\_str’ column values have been casted from integers to string values. For easier access of the data some restructuring was provided so that the nested columns were transformed to separated ones.

The original data contains a lot of unnecessary or repeating information, which was dropped.

We have observed the same phenomena reported earlier about Twitter data – most of the coordinates were not available. Unfortunately there were quite many missing data about the ‘place’, but the available information in this column was more comparing to ‘coordinates’.

Since the information involved in the fields ‘coordinates’ and ‘geo’ was practically the same, we removed the last one.

**2.3. Data privacy**

The both data sets do not directly involve any personal information directly related to users (such as real names, photos, phone numbers, etc.). But they involve some information, which could be indirectly associated with some personal data. Therefore we decided to ensure data anonymization using some hash function for this purpose.

We present only graphical results from our explorative analysis.

Here we use some names of public persons with high activities on Twitter. Since they are public persons, this should be in agreement with the privacy rules.

The public data has anonymized – **cloud - provide link to data sets**.

**Hash function for anonymization** – to be described more detailed when it is used, why and how.

1. **Data Analysis**

**Used techniques:**

In our work we use almost all techniques about the explorative analysis learned during the first two exercise sessions of the course (sentiment analysis: lowercasing, remove punctuation, removing stopwords, apply stemming, calculating word frequency, etc.). We show also some new graphical presentations, needed for the purposes of our investigation.

**3.1 Posts activity of some public persons**

As a first glance we decided to check some of the author names in the posts, which occur very often.

**Mikko Kärnä – by one or other reason in did not found them in user names in both data sets.**

**Posts count in time**

Next we decided to investigate how the posts activity varies in time. We were interested to see the distribution of posts amount in scanned time period.

We observed that in most cases there were more than one posts for every second of DataStamp. Therefore in order to get unique time we first group data by column ‘created\_at’. After that the data is resampled to get the number of posts per hour.

Code:

#PLOT THE NUMBER OF POSTS PER HOUR

import json

import pandas as pd

import numpy as np

data=pd.read\_json('8oct\_pre\_processed.json')

#add one column with 1s,necessary to count number of posts

#in resampled time interval

Number\_posts = []

for index, row in data.iterrows():

Number\_posts.append(1)

data['Number\_posts'] = Number\_posts

#take only two columns from data

data1 = pd.DataFrame(data=data, columns=['created\_at','Number\_posts'])

#group by repeating times

data2 = data1.groupby(['created\_at']).count()

#group the data per hour

per\_hour = data2.resample('1h',how='sum')

#see the number of posts per hour

per\_hour

#-----------------------------------------------

**Results for the data set from 08.10.2017:**

per\_hour

Out[56]:

Number\_posts

created\_at

2017-10-08 14:00:00 46834

2017-10-08 15:00:00 80353

2017-10-08 16:00:00 80466

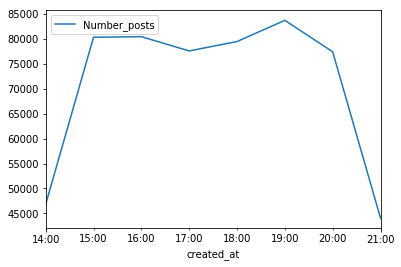
2017-10-08 17:00:00 77606

2017-10-08 18:00:00 79489

2017-10-08 19:00:00 83746

2017-10-08 20:00:00 77428

2017-10-08 21:00:00 44058



Note: Since the data in the beginning and at the end are taken only for about 30 min., but not for 60 min., the corresponding number of posts values are about half less.

**Observation:**

The posts activity during the observed time period does not vary a lot. The most active time when people were posting on Twitter, was around 19:00.

**Results for the data set from 10.10.2017:**

We repeated this analysis for the data from 10.10.2017. Since we have got problems with the merged file, we show here the results from 3 smaller data sets from this day.

10oct\_preprocessed1.json:

In [11]: per\_hour

Out[11]:

Number\_posts

created\_at

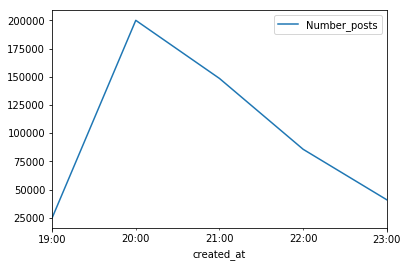
2017-10-10 19:00:00 24868

2017-10-10 20:00:00 199962

2017-10-10 21:00:00 148479

2017-10-10 22:00:00 85724

2017-10-10 23:00:00 40966



10oct\_preprocessed2.json:

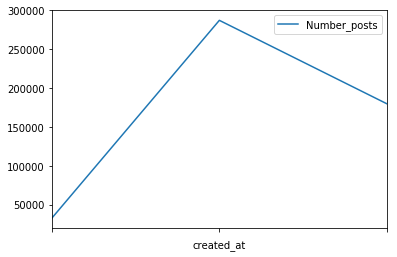
Number\_posts

created\_at

2017-10-10 17:00:00 32782

2017-10-10 18:00:00 287283

2017-10-10 19:00:00 179934



10oct\_preprocessed3.json:

per\_hour

Out[35]:

Number\_posts

created\_at

2017-10-10 14:00:00 61264

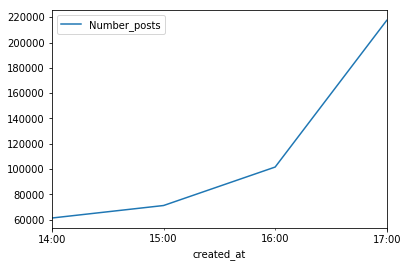
2017-10-10 15:00:00 71140

2017-10-10 16:00:00 101555

2017-10-10 17:00:00 32782

2017-10-10 18:00:00 287283

2017-10-10 19:00:00 179934

2017-10-10 17:00:00 217552

Or when we combine the results from three parts:

2017-10-10 14:00:00 61264

2017-10-10 15:00:00 71140

2017-10-10 16:00:00 101555

2017-10-10 17:00:00  **250 334**

2017-10-10 18:00:00 **287283**

2017-10-10 19:00:00 **204803**

2017-10-10 20:00:00 **199962**

2017-10-10 21:00:00 148479

2017-10-10 22:00:00 85724

2017-10-10 23:00:00 40966

**Observation**

Exact time of Puigdemont’s conclusive speech: 18:00:00.

So, we observe that the most active posts time on Twitter was between 17:00 and 20:00, with maximum number of posts at 18:00, when Puigdemont’s speech began.

Additionally we apply some new classification methods and construct a classifier.

**3.2. Simple exploratory analysis of the data set on 08.10.2017 (link to Catalonia 8oct.ipynb)**

We found the Top50 tweeters with no original tweets. Thus the date suggests some robots, which only retweets original posts most probably automatically. First 15 most retweeters without original posts are retweeted 215 677 tweets, which is quite significant amount.

We have calculated a simple influence score using the following formula:

*Sinfl*=*α*∗|*favorites*|+*β*∗|*retweets*|Sinfl=α∗|favorites|+β∗|retweets|,

where *α* = 1 is a weight given for favorite, and *β = 3* is the weight given for retweet.

We also forum the most active users, both with original and retweeted posts, as well as the most active users who posted only original tweets.

We also show top 20 users with highest influence score.

After finding the most frequent hashtags and removing retweets from this subset we decided which subset of hashtags to use in our further analyses.

**3.3 Heat map – geolocations**

We have restricted our data sets only from Spain.

**3.4 Most retweeted posts**

We are interested to investigate whish posts are most retweeted and if this happen automatically.

The both data sets show presence of several robots.

**3.5 Words frequency**

Since the different posts are in several different languages, we have extracted only the posts in English and Spanish and thus formed two subsets. We took only the rows where the column ‘lang’ == ‘en’ of ‘lang’ == ‘es’. **Link to the code chunks for extracting both datasets** .

**The first 100 most frequent words in English:**

**The first 100 most frequent words in Spanish:**

**3.6 Sentiment analysis – Emmi.**

**Link to the script**

* 1. **Classification results – binary classificator ?**

Manual analysis of small posts subsets.

For classification purposes we first have analyzed some of English language posts manually. The goal was to determine which are pro- and which – against the independence of Catalonia.

We have analyzed two very small subsets – original, not retweeted, including only one hashtag = 'recuperemelseny', written in English, involving more than 15 words, and not referring to URL.

We have got 37 such posts. Two of them (BBC and Reuters) repeated the same text, but it was still their original post because they took it from the same source. This could not influence heavily the classification results because such repetition happens rare.

We found that the agencies, who repeat the same text as their own, are: BBCNews, Reuters, AP, Guardian, Nytimes, BILD.

We have provided the same manual classification analysis for other data set – similar as the first one, only hashtag = 'Barcelona'. We have got 15 posts.

Although we have tried to classify these posts manually, for significantly big part of posts we have got difficulties to provide our decision. Except the posts, which are not relevant, there were quite many posts which refer to Bible. Others refer to some important historical information, which local people would easily understand, but which is unknown for most foreigners. Similar problems appear when people refer to local food, which have some symbolic meaning, and Spanish or Catalonian people easily will understand, but not foreigners.

Thus we have observed that even if we understand the language in the text, in many cases we cannot decide to which group to classify this post.

This surely will cause problems for building and training any kind of classificator, which has to solve this task.

Since we found some supporting information on the net, we next provide sentiment analysis of the Spanish language subset from 08/10.10.17?

We have used the results, published on <http://www.sepln.org/workshops/tass/2015/tass2015.php>

and reported during the TASS 2015 conference about Twitter sentiment analysis. Since these results concern only Spanish language, we have applied them only for the Spanish subsets.

*We also have used pseudocode for naïve Bayes classificator, published on* [*https://web.stanford.edu/~jurafsky/slp3/6.pdf*](https://web.stanford.edu/~jurafsky/slp3/6.pdf) *.*

Description of everything which we have done about this classifier. Also link to the code.

**3.8 Linear model**

1. **Conclusions**

The provided explorative data analyses from two Twitter data sets showed interesting phenomena, which otherwise would be difficult to connect and observe. Thus we found that some public persons like Finnish politician from Lapland Mikko Kärnä, and also user with name Julian Assange, have been very active in Twitter in these two crucial dates.

1. **Open Problems**

We observed high complexity in the posts text, because of which even manual classification is very difficult. Therefore is almost impossible to suggest automatic rules, based on which the decision about pro- or against Catalonian independence could be made.

1. **Future work**

We are still in the beginning of classifier design.

In the future we will classify pro- and against- posts regarding their geographical location.

Here we have used only small amount of the data, about half of the obtained columns. It would be very interesting to investigate the followers of some posts, which form some network.

Inside the posts there are URLs and/or images, which also incorporate some information.

We were interested particularly in these two data sets – from 08.10.2017 and from 10.10.2017, because we have been expected different people with quite opposite opinions to be active in both events. In this initial stage we still did not provided such investigations.

In would be interesting to analyze the posts, which are automatically retweeted from users, which do not have their own tweets, and for which we can assume that they are kind of robots.

Presentation:

* Description on GitHub: scripts, this description
* Presentation slides
* Blog post

Deliverable: Blog post + GitHub.

**Literature**

Romain Thalineau, Monitoring the French Presidential Election on Twitter with Python, 2017.

Mikael Brunata. Scraping, extracting and mapping geodata from Twotter.

Twitter Sentiment Analysis, TASS 2015, September 15th 2015, Alicante, Spain, <http://www.sepln.org/workshops/tass/2015/tass2015.php>

Daniel Jurafsky and James H. Martin. Speech and Language Processing, Chapter 6: Naive Bayes and Sentiment Classification, 2017. <https://web.stanford.edu/~jurafsky/slp3/6.pdf>

Laurent Luce’s Blog. Python, Twitter statistics and the 2012 French presidential election.