AGH
Data exploration

Piotr Rząsa, Michał Kawałek

Fake news analysis based on government election in Poland

Analiza fake news na podstawie tweetów o wyborach parlamentarnych w Polsce

First stages

1. Downloading all tweets/re-tweets with hashtags connected to Polish government election in 2018 and authors of those tweets.
2. First analysis did not bring any results - data was to chaotic and give no valuable conclusions.
3. We focused on few particular cities in which we tried to get tweets of main candidates. Unfortunately, they did not used Twitter that much and almost did not used hashtags.
   The amount of data collected this way was much lower than expected. Also because this was second turn of election.

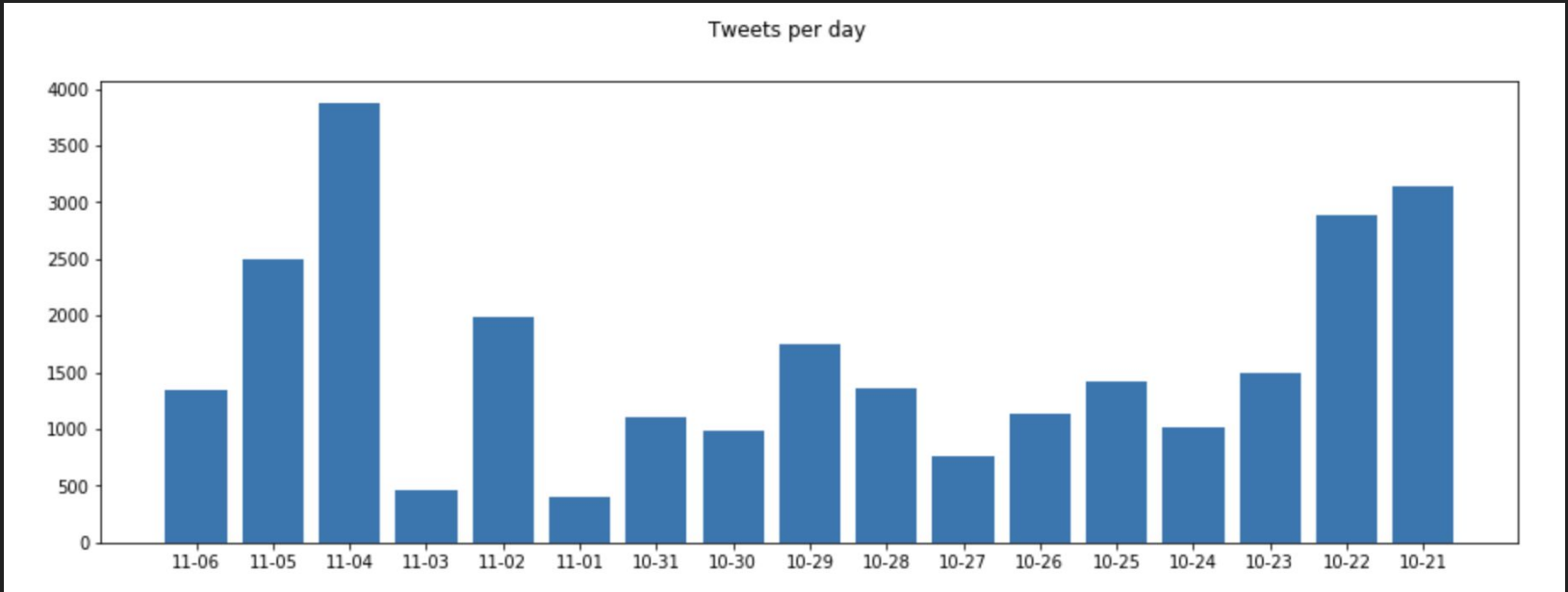
Collected data

Examples of hashtags:

- #wybory
- #Wybory2018
- #NowyPrezydentWiększeMożliwości
- #WspólnieTworzymyPrzyszłośćKrakowa
- #wyborysamorzadowe2018
- #WyborySamorzadowe

- 8882 users
- 31943 retweets
- 4402 tweets

Tweets per day



Fake accounts

We have chosen an algorithm shown below for detecting fake accounts on Twitter:

- For each user we counted metrics, which allowed to classify it as fake account or not.
- The initial value of metric for each user was 1 (we assume it is not fake account).
- Next, for some criteria we multiply this metric by appropriately selected values.
- At the very end, if the value of the metric for a given user was below 0.5, we would consider the account to be fake.

Examples of criteria and its weigths

Attributes	Weight
the account has at least 30 followers	0.53
the account has been geo-localized	0.85
it has been included in another user's favorites	0.85
it has used a hashtag in at least one tweet	0.96
it has logged into Twitter using an iPhone	0.917
a mention by twitter user	1
it has written at least 50 tweets	0.01
it has been included in another user's list	0.45
(2*number followers) _ (number of friends)	0.5
User have at least one Favorite list	0.17

First results?

Analysis and detection of fake accounts were to help us detect fake news, but unfortunately it was not helpful at all.

Our algorithm detected near 200 potential fake accounts while we did not find anywhere confirmation that these accounts were actually fake. We checked them only by going to them manually and checking what tweets they are posting. Most of them were actually suspicious, but were they fake… ?

best example: https://twitter.com/St_Janecki

statistics of Mr Janecki: https://foller.me/st_janecki

Fake news - attempt 1

The first approach was a semantic analysis of each tweet and a one-to-one comparison to find similar but, for example, conflicting tweets.

Unfortunately, the algorithm did not work as expected - each method of semantic analysis of tweets failed and the found pairs were not semantically similar to each other.

Used methods:

- bag of words
- calculating the distance between words

Fake news - attempt 2

Visualizing all tweets for:

- from what device tweets come from
- whether the account is verified (it turned out that only 26 out of 8882 accounts are verified)
- number of tweets
- whether they retweet any account fake
- age
- date of registration

Unfortunately, the visualization did not help to detect any groups (too little data)

Change of approach ...

It turned out that we downloaded the data only for tweets containing hashtags, not tweets containing keywords related to the election (we would have a lot more of them, which would make the analysis easier).

Unfortunately, the Twitter API does not allow downloading data after keywords more than 7 days back, and it was already long after the election.

Change of approach ...

The approach has therefore changed towards analyzing the groups of authors.

The check was done for:

- number of followers
- number of retweets
- number of answers
- amount of likes
- number of friends

How ?

We calculate for each user the average values from all tweets for all previously mentioned criteria and build a vector from it. Each user who has a value greater than the average value for a given 'attribute' belongs to a given segment.

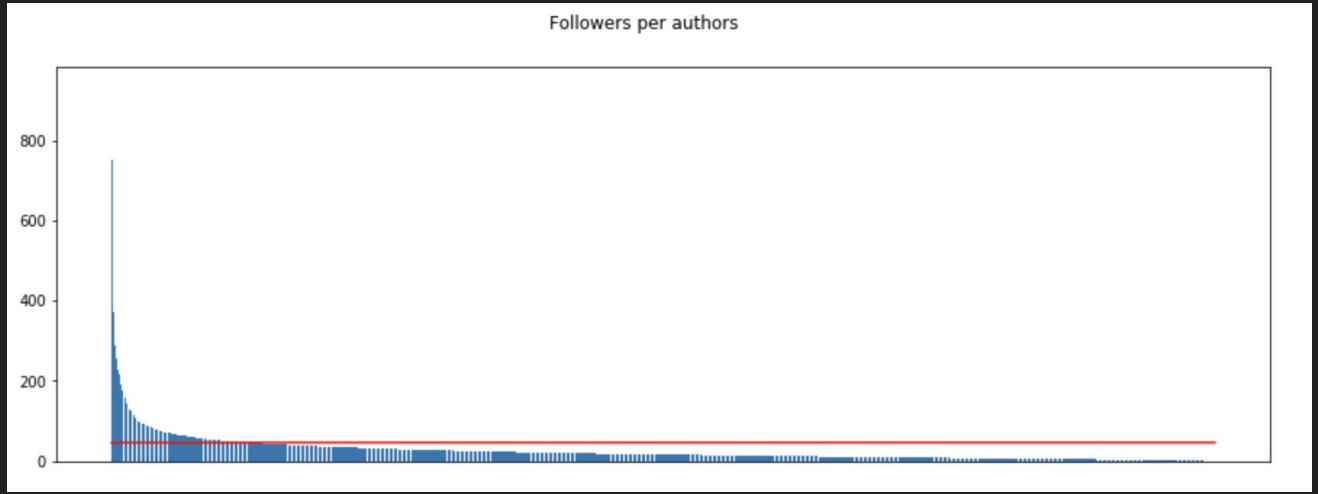
For example:

The user with the number of followers 50 where the average is 25 is in the "users with a large number of followers" segment.

Statistics

Authors followers:

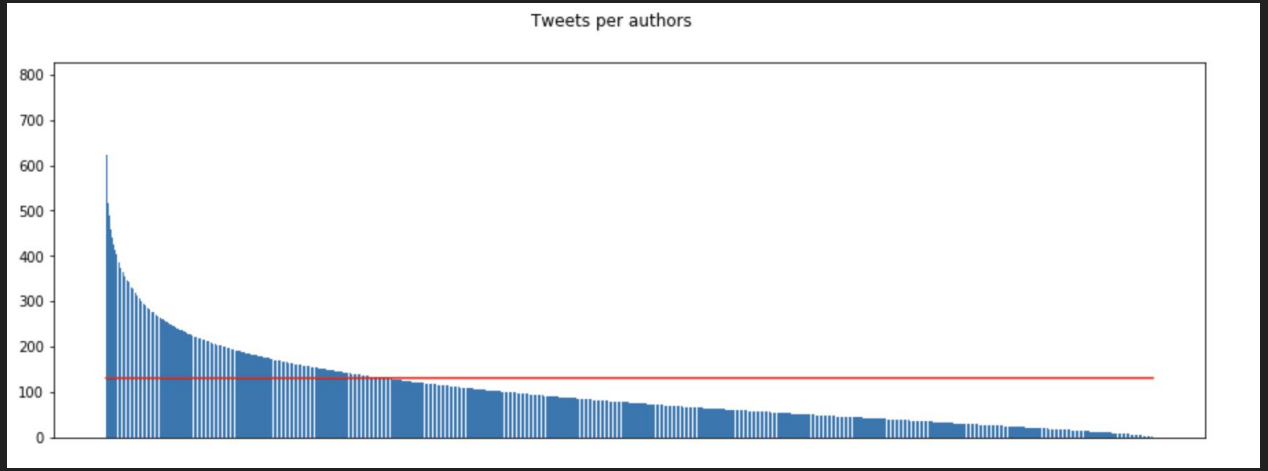
- avg 2184.17
- min 0
- max 875777
- stand dev. 19551.87



Statistics

tweets in general:

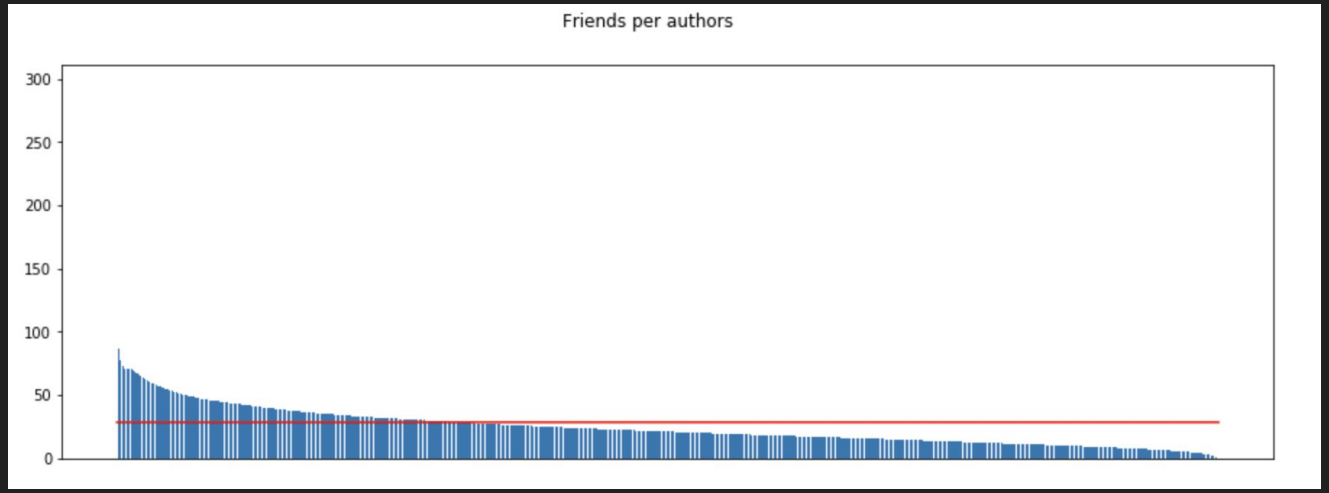
- avg 17173.35
- min 0
- max 620478
- std dev. 32765.08



Statistics

Users followers

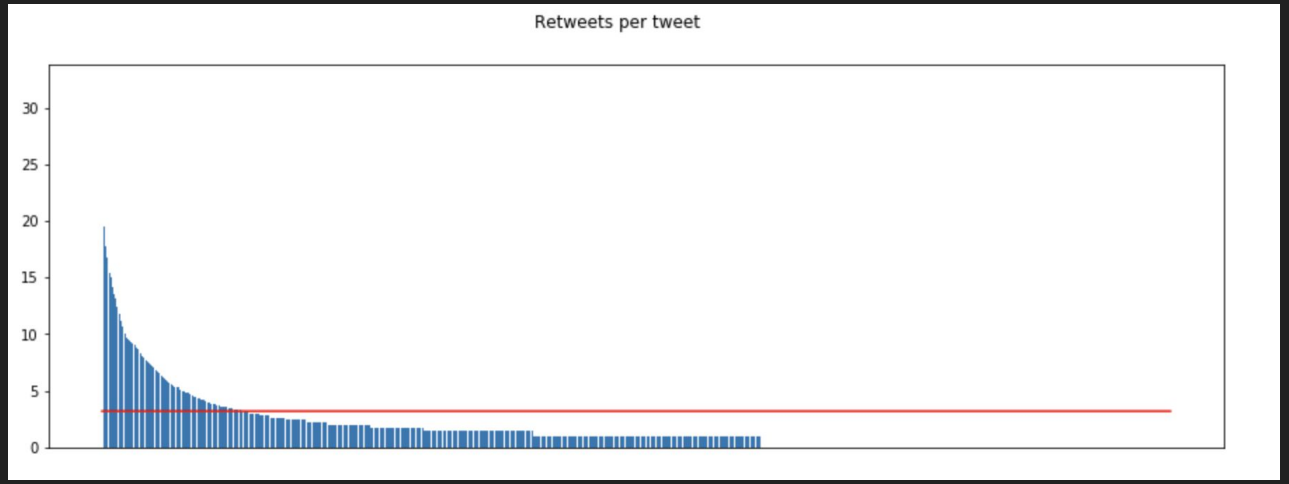
- avg 814.39
- min 0
- max 87814
- std dev. 1418.45



Statistics

number of retweets

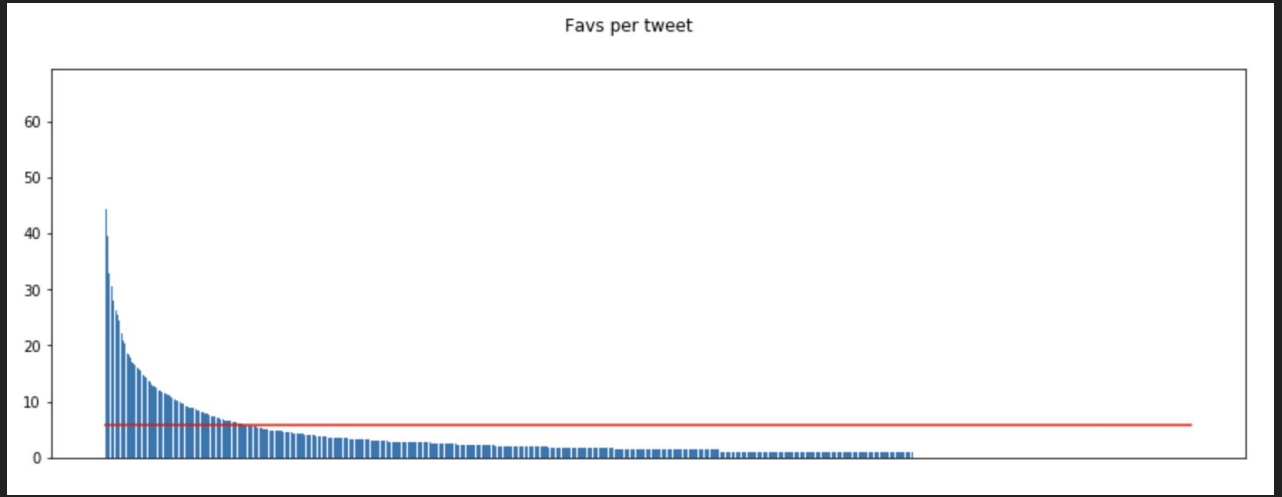
- avg 10.20
- min 0
- max 1038
- std dev. 44.19



Statistics

favs per tweet

- avg 33.62
- min 0
- max 4371
- std dev. 155.01



Groups

- users with a large number of followers - 401
- users with lots of friends - 660
- users with a lot of answers per tweet - 488
- users with a large number of retweets - 185
- users with lots of likes per tweet - 188
- verified users - 26

Conclusions and lessons learned

- irrelevant approach to data collection - lack of a well-thought-out model, too little time, no experience
- Polish texts difficult to analyze when you have do it for the first time
- lack of certainty that something like this could have happened (fake news/accounts)
- easy access to Twitter data and their processing when using Python libraries
- a multitude of options and issues for which analyzes could be carried out