Tweet analysis: Do positive people cluster together?

Maciej Dragula, Viktor Crettenand
School of Computer and Communication Sciences, EPFL, Switzerland

***Abstract*—**

I. Introduction

It is said that positive people attract each other. In this project we will investigate this claim using the data set from ref [1]. We will try to understand how positive or negative people cluster together as well as see if a change in the sentiment of a user is correlated to a change of sentiment of his followers and followee. We will look at how the average sentiment of tweets change with the seasons across a year.

A. *The initial plan*

The initial plan for this project was to study the sentiment of people regarding climate change and study how people of the same opinion cluster as well as what proportion of people change their opinion about global warming and what are sociological situations in which people are more likely to change their mind. This however was impossible in practice with the time at our disposal because of the Twitter API's limit on http requests. The problem with collecting tweets on climate change is that most users never Tweeted about climate change. Building a data set with users who Tweeted about the climate, let alone tweeted about it several times is like searching for a needle in haystack especially because of the http request limit. As an emergency backup plan we decided to use the dataset from ref [1] and partially enrich it with the tweet's text.

II. Data set

The first set was to get the access to Twitter API to be able download necessary information for a given Twitter's user. Thus, we have applied for a proper key, and went through a tutorial on a Twitter's website with documentation.

Next, we wanted to download the tweets posted by users gathered in a data set used in [1]. In order to to that, we have used the file describing the network of the users where the connection between two users is specified as a follower-followee relationship. We have encounter a very serious problem which is the fact that many accounts either do not exist anymore or never existed. In a numerous number of cases GET method of most recent tweets made by a particular user, the Twitter API [2] return an information that a given page doesn't exist. The problem is so serious that we easily reached a limit of 3200 requests per hour. This limit is set by Twitter Inc., and we believed it is done because of security reasons, i.e. to prevent downloading information all the time.

We have check that the connection which we established with Twitter's server is correct because we were able to download recent Tweet e.g. published by Donald Trump. Trump's nickname at Twitter is *@realDonaldTrump*, and using the website *https://tweeterid.com* we have found a corresponding Twitter ID. What is more, we finally have found Twitter IDs from the data set given in [1], e.g. 1573741 which corresponds to a user with a nickname *twitterislame*. But unfortunately we were able to find only about a dozen of users which really exist and download their recent tweet which is not enough.

III. Models and Methods

The Sentiment140 [3] is a data set created at Stanford University where each record is a tweet and a sentiment. The sentiment is represented by a binary variable: negative sentiment and positive sentiment.

We have decided to use a Support Vector Machines (SVM) as a learning model, and we trained it on this data set achieving 81% on a test set. Both classes are well balanced so we believe that a in this particular case the accuracy is as well good measure as F1-score.

Next next step would be to predict a sentiment of a tweets downloaded via Twitter API. As we know the connections between the authors of the tweets thanks to the network provide by [1], we wanted to investigate if people who write posts with a positive sentiment cluster together. In order to check this thesis, we have decided to calculate a following ratio:

$$P = \frac{\text{\# recent positive tweets}}{\text{\# all recent tweets}} \tag{1}$$

Then we would like to check, if people who are connected by follower-followee relationship are positive both positive, in other words their P values are similar. We suspect that positive sentiment of tweets posted by one person may influence a sentiment of their follower's tweets.

We believe it might be also interesting to see how the positiveness of tweets of a particular person changes in a time, e.g. in different hours of a day, and how it affects the followers.

IV. Results and Discussion

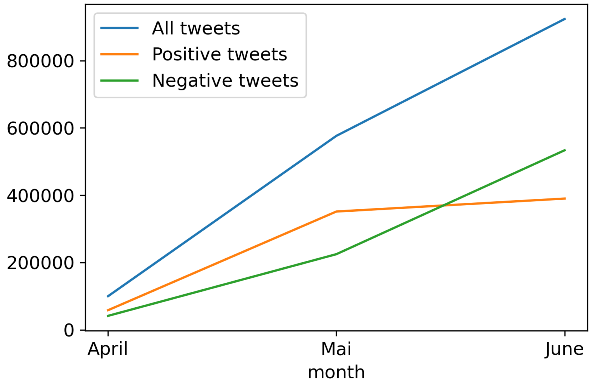

Figure 1. Number of positive and negative tweets in the months of April, Mai and June

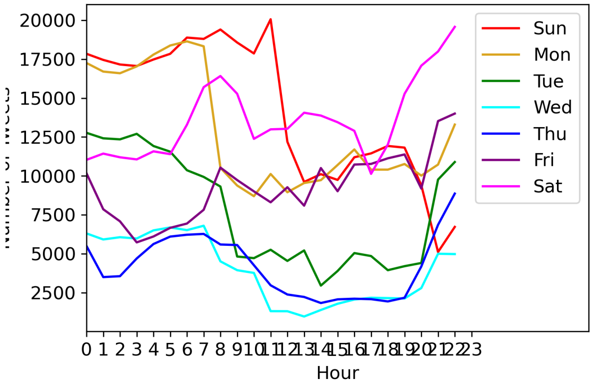

Figure 2. Number of tweets per hour

V. Summary

References

[1] Hai Liang and King-wa Fu. Testing propositions derived from twitter studies: Generalization and replication in computational social science. *PloS one*, 10(8):e0134270, 2015.

[2] Twitter api documentation. https://developer.twitter.com/en. Accessed: 2020-12-18.

[3] Sentiment140 data set. http://help.sentiment140.com. Accessed: 2020-12-18.

Appendix