ADA project - Mining democracy

Ahmad Ajalloeian, Gergely Odor, Ramtin Yazdanian
EPFL, Switzerland

I. Introduction

Twitter is a tremendously popular social network (also notorious if you count Donald Trump's account), based on 140 (recently changed to 280) character pieces of text called "tweets". Twitter has not only been used in a more personal capacity, but also in the direction of political agendas and activism.

The Swiss tweets dataset offers many intriguing questions, especially given the fact that it has geolocation tags. First of all, the locations allow us to understand how Swiss tweeters are distributed throughout the country. Secondly, given Switzerland's relatively rare status as a country with multiple official languages, tweets by Swiss users offer a unique opportunity to gain insight into the relationship between the languages, contents, and locations of tweets. In addition, we are interested in gaining insight into Swiss political activity on Twitter and its distribution across languages and locations. This project will aim to find answers for - or at least gain insights into - the questions listed below.

A. Research questions

1) Does the density of the location of tweets correspond to population densities? Or are the tweets significantly more concentrated in cities?
2) Can we reconstruct the Röstigraben only based on the language of tweets?
3) Are there any spikes in activity and do they correspond to particular events (e.g. referendums)
4) How involved are swiss people in politics on twitter? Which areas are most involved?

B. Literature review

Related to the last question we did a short literature review. We did not find anything on the political activity of the population, however we found articles on the twitter activity of Swiss politicians. It has been reported that in Switzerland, politicians were quite late to adopt twitter; by the end of 2012 only 30% of Swiss politicians had a twitter account [1]. However, this figure has seamed to changed since 2012 since according to a recent news article "For Swiss politicians Twitter profiles are no longer a novelty" [2].

In the final days before the deadline, we found a preprint that had a large overlap with our fourth research question [7]. The only difference between this work and ours is that we worked with a downsampled local dataset, whereas the authors there used the Twitter API to specifically download the political tweets. Also large part of the search for political tweets in [7] were done manually.

II. Methods

The code for the project is available at [3].

A. Acquiring the datasets

We primarily use with the swiss tweets dataset provided by the ADA team (originally by Swisscom). The tweets were queried using a bounding box around Switzerland, which means that not all tweets are actually swiss. We will explain how we filter the ones outside of Switzerland in section II-B2.

We also used several supporting datasets, which were not straightforward to locate and acquire. The geojson file for the boundary of swiss municipalities was downloaded from [4]. The population data was acquired from [5] and the population densities were downloaded from [6]. For detecting political activity we used the data of [7] available at [8].

B. Preprocessing data

1) Reading the data: Parsing the provided dataset was in itself a demanding challenge, so we decided to dedicate a subsection to it. The data was provided in the tab separated values format; columns were separated by \t and rows by \n. However, many of the tweets contained tabs and newlines, these were denoted by \\t and \\n. To make it more complicated the last column (userLocation) sometimes ended with a large number of backslashes which resulted in further ambiguity. Since even after resolving all these exceptions and a few others, we still got rows with varying length, we decided to follow a different approach.

The first three columns of the dataset (id, userid, createdAt (date)) were all not nullable and very robust in terms of syntax. We decided to match a regular expression on

```
\d{4,}\t\d{4,}\t\d{4}\-\d{2}\-\d{2}\s\d{2}:\d{2}:\d{2}
```

which has worked with 100% accuracy. This allowed us to find the beginning of each row, however, separating the cells was still difficult. We decided to split on \t, and merge together the cells [3:-16] as the text cell, since we expected that the last 16 and the first 3 columns had no \t, only possibly 4^{th} (text) column. It turned out that we were wrong, the last (20^{th}) column (userLocation) also had \t, but these exceptions only involved less than 20000 datapoints (out of $\sim$ 20 million), so we neglected these. We will see later that this is a tiny fraction compared to the fraction of datapoints we had to exclude due to other issues with the dataset.

2) Processing geolocation: Out of the $\sim$ 20 million tweets, about $\sim$ 18 million were geolocated. Only about 10 million these geolocations were in Switzerland (the rest came from neighbouring countries in the bounding box). Since our goal was to analyse only swiss tweets, the rest of the datapoints (about 50% of the whole) had to be dropped.

For our later analyses, we had to detect the municipality each tweet came from. We made use of the geojson map which contained the boundaries of each municipality as a polygon, and we used the *shapely* python library to detect which polygon each geolocation is in [**?**]. Since the 10 million tweets came from only less than

15000 unique locations (probably since the geolocations were not the exact locations of the users, rather maybe the location of the transmissing tower each cellphone was connected to), we could compute the municipality information in less than 30 mins.

3) Language detection: Language detection was a crucial step especially for research question 2, but also for detecting political content. For this task, we used the *langdetect* python library, a port of Google's language detection code based on a naive Bayesian filter [**?**]. This is the only task for which we had to use the hadoop cluster. Our simple spark code took 3 hours on ?? nodes.

Unfortunately, by manual inspection we saw that the accuracy of *langdetect* was far from perfect (even though the website claimed 99% accuracy). One of our concerns was that possibly the swiss german dialect could cause problems, so we downloaded a swiss german corpus and tested *langdetect* on the sentences of the text. On the collection of swiss german blogposts we saw an accuracy above 82%, whereas on the "Blick am Abig" newspaper from Zürich our accuracy was above 95% (the majority of tweets were classified as simply german in both cases). After this confirmation, we concluded that the observed errors are probably due to the fact that tweets are short and don't follow the rules of the languages very closely.

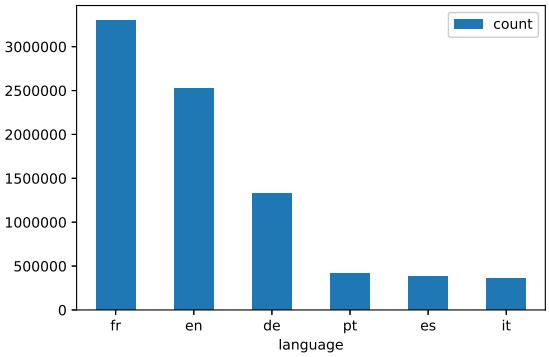

Figure 1. The number of tweets in the top5 languages in the dataset

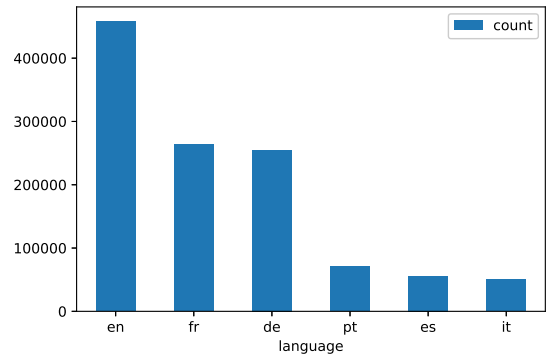

Figure 2. The number of tweets in the top5 languages only for 2016

C. Detecting political content

TODO: Ramtin

III. Results

A. Population density and tweeting location

The distribution of tweets over the municipalities is visualized on Figure 3. On Figure 4 we show a heatmap of twitter activity for only one month (over the whole dataset this was too computation intensive). From these maps show that indeed highly populated areas have higher twitter activity. However, our question was more subtle, we were interested if the

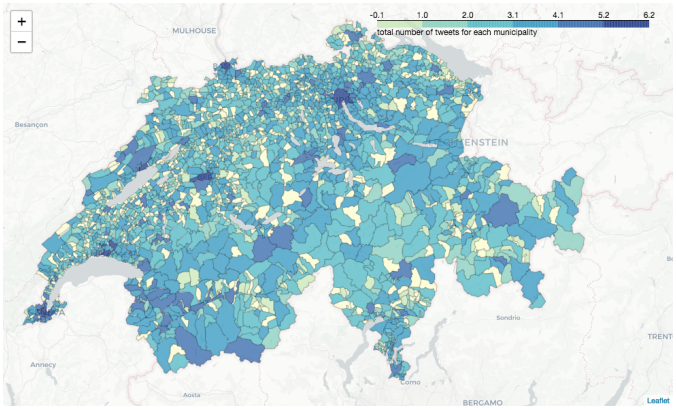

Figure 3. Colors on the map show the logarithm (base 10) of the number of tweets in the whole dataset coming from each municipality

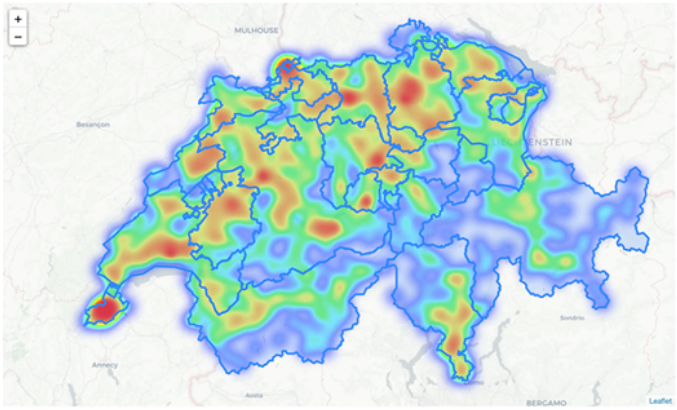

Figure 4. The heatmap of the number of tweets in September 2016.

B. Reconstructing the Röstigraben

Figure 5

C. Political activity

TODO: Ramtin

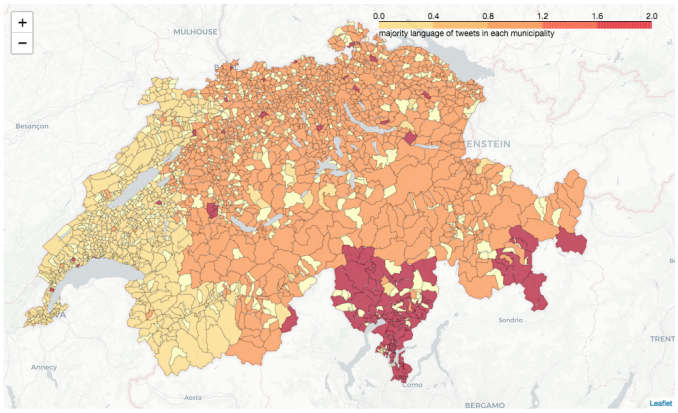

Figure 5. The Röstigraben reconstructed from the language of the tweets. The color encoding is yellow: French; orange: German, red: Italian

IV. Summary

References

[1] A. Rauchfleisch and J. Metag, "The special case of Switzerland: Swiss politicians on Twitter," vol. 18, pp. 2413–2431, Nov. 2016.

[2] A. Fichter, "The Swiss politicians who have mastered Twitter," https://www.swissinfo.ch/eng/directdemocracy/digital-democracy_the-swiss-politicians-who-have-mastered-twitter/42957464.

[3] "Project repository," https://github.com/odorgergo/ada-rag/tree/master/project.

[4] "Switzerland municipalities," https://raw.githubusercontent.com/lvonlanthen/data-map-d3/master/data/ch_municipalities.geojson, accessed: 2017-11-29.

[5] "Swiss population data," https://www.bfs.admin.ch/bfs/en/home/statistics/population.assetdetail.3342066.html, accessed: 2017-11-29.

[6] "The swiss population density data," https://www.atlas.bfs.admin.ch/maps/13/fr/12876_229_228_227/21239.html.

[7] B. Mueller, Christian; Wueest and T. Willi, "Exploring the usefulness of twitter data for political analysis in switzerland - preliminary draft!" *Paper prepared for the Annual Conference of the Swiss Political Science Association*, 2016.

[8] "The swiss political users dataset," https://docs.google.com/spreadsheets/d/1tXUh3ch1fw4Oa7PijB5fusQZFb3GLdGrlJTtIhNw7PY/edit#gid=437242465.