SwissFeels: Sentiment Map of Swiss Tweets

Brandon Le Sann, Timothée Lottaz, Seth Vanderwilt

EPFL Applied Data Analysis Fall 2016

go.epfl.ch/swissfeels

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

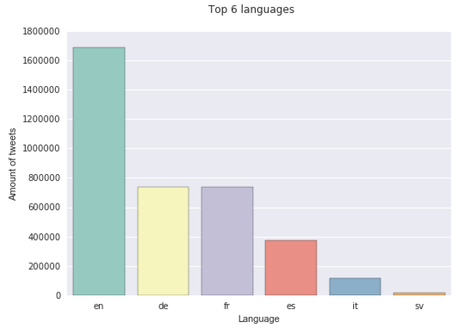
Objective

The goal of our project was to analyze a large dataset of geolocated tweets and construct an interactive sentiment map of Switzerland, similar to that of Happy Maps.
We focused on characterizing the sentiment of the tweets as positive or negative towards a certain entity, i.e. "is this tweet positive or negative about company X?".
The objective was to have an interactive visualization that takes a keyword as input, for example "CFF" (Swiss national railway) and displays the sentiment of each canton on the Swiss map.

Data acquisition

The ADA course staff collected tweets from January to November 2016 that were geolocated in Switzerland. Each tweet was annotated with estimates of its language and sentiment. We filtered the original 50GB dataset to a more manageable collection of approximately 3.7 million tweets.

Figure 1: Number of tweets in each language



Data Format

The following fields were necessary in order to process the tweets:

- `geo_state`, the tweet's source canton
- `sentiment`, the tweet's sentiment, either Positive, Neutral or Negative.

We also decided to keep other interesting fields:

- `author_gender`, which can be MALE, FEMALE, or UNKNOWN
- `lang`, the language of the tweet
- `main`, the raw text of the tweet
- `published`, the date and time the tweet was published

Data Cleaning/Issues

- There was one major issue with the dataset. The **geolocation** of the tweets was not collected prior to July 2016. This made 60% of the data unusable.
- The `geo_state` field was often valid, but we had to filter out some **outliers** that were not Swiss cantons. These represented 0.4% of the data.
- Another minor issue was the **language detection**. Somehow Spanish seems to be spoken a lot more frequently than Italian (a national language)! Looking further into this problem we found that many Italian-language tweets were mislabeled as Spanish.
- Twitter **bots** were a problem that we couldn't satisfactorily address. For example many local radio stations automatically tweet their playlists, which polluted the dataset.
- The **sentiment analysis** algorithm worked poorly on **non-English** tweets.

Result graphs

Figure 2: Can we see the Röstigraben?

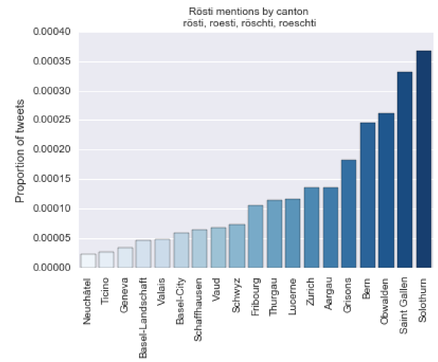


Figure 3: How do the Swiss feel about Hillary and Donald?

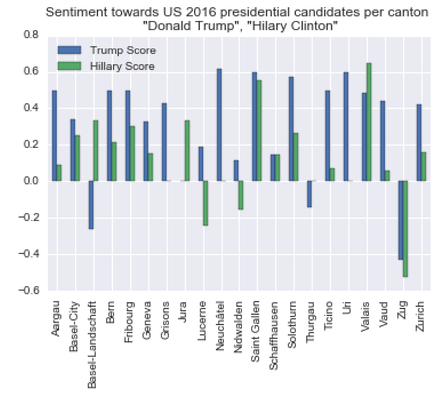
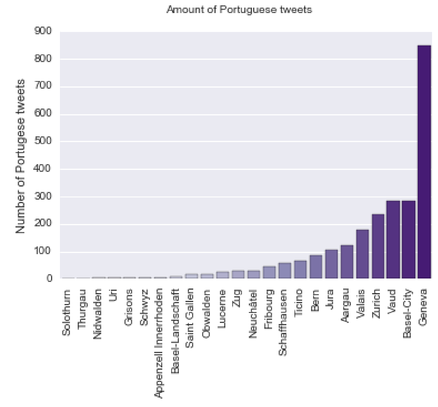


Figure 4: Where is the biggest Portuguese community?



Website

We built an interactive map of Switzerland that displays the mean sentiment of each Swiss canton. Thanks to the search function, it is possible to view the mean of a subset of tweets containing search terms such as "SBB CFF FFS". See Figure 5 for an example of mean sentiment. There's also an option to display a map of the proportion of tweets containing the search terms, as you can see on Figure 6. Some matching tweets are displayed so that the user can verify that his/her query works well.

Figure 5: Per-canton sentiment mean for "Brexit"

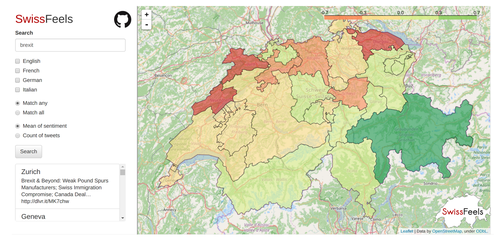
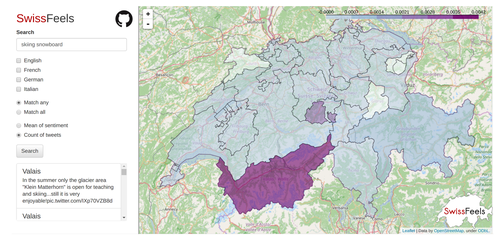


Figure 6: Per-canton mentions for "Skiing" or "Snowboarding"



Conclusion and Future Work

Overall, the SwissFeels project performs quite well. Some queries are polluted by bots or spurious matches, as our current implementation simply searches for string occurrences in the raw text. However, many queries are very clear ("skiing", etc.) and give interesting results. Labeling tweets with entity mentions would provide more reliable search results in the current implementation.