Data Expedition into the Swiss Twitter Corpus — Workshop Results at SwissText 2018

Ralf Grubenmann[1], **William Fallouh**[1], **Christoforos Nalmpantis**[1], **Mark Cieliebak**[2]
[1]SpinningBytes AG, [2]Zurich University of Applied Sciences
rg@spinningbytes.com, william.fallouh@isen.yncrea.fr,
christofernal@gmail.com, ciel@zhaw.ch

Abstract

Data Expeditions are short, collaborative events focusing on finding interesting patterns and insights in a dataset through interdisciplinary teams. This is a report of the Data Expedition into the Swiss Twitter Corpus expedition hosted by SpinningBytes AG at the SwissText 2018 conference. The aim was to research interesting topics related to Switzerland in the Swiss Twitter Corpus[1]. Two teams with a total of 11 participants were given 140'521 Switzerland-related Tweets with relevant metadata and analyzed topics of their choice during the 4 hour workshop. We explain how the data expedition was organized and discuss some of the results and lessons learned.

1 Introduction

More and more data is available to industry and to researchers, which leads to more and more new avenues of research being available all the time. For this purpose, Data expeditions are a popular tool for educational and research purposes that can quickly produce interesting analyses from a dataset (Radchenko and Sakoyan, 2016; Ciociola and Reggi, 2015; Burov et al., 2016). They allow groups of people to quickly try out new ideas and test hypotheses, leading to new results that might not be found in a more traditional research setting.

Since visitors to the SwissText conference come from a wide spectrum of industrial as well as research backgrounds, we decided to go with this format to give participants the possibility of hands-on experience, giving them an opportunity to exchange ideas with people outside their field and possibly discovering new topics for future research. Since the focus of the SwissText is on Natural Language Processing in Switzerland, giving the participants Tweets from or regarding Switzerland was the natural choice. These Tweets were specifically curated for this workshop, as detailed in Section 2.

2 Swiss Twitter Corpus

The Swiss Twitter Corpus is a collection of over 3 million Tweets related to Switzerland which has been collected since January 2018. Being related to Switzerland, or "Swissness", is defined as either originating in Switzerland, being written by an important Swiss Twitter account or being about one of a number of hand-curated keywords related to a Swiss topic. Additionally, we look at the users profile location being in Switzerland and whether the language of the user is Swiss-German.

For the expedition, a subsample of Tweets was selected by selecting Tweets with Swiss Geocoordinates, Tweets with at least two keywords present and Tweets with a Swissness-Score of at least 3 or more. The Swissness-Score counts how many of the Swissness-Rules apply to a Tweet, for instance a Tweet with two relevant keywords and Geocoordinates in Switzerland would get a Swissness-Score of 3. This results in a sample of 140'521 Tweets that are highly relevant to Switzerland.

Each Corpus entry contains the Tweet text, the name of the user, the date of the Tweet, the Tweet-language[2], the users country-code, the latitude and longitude (if provided), the keywords found, senti-

In: Mark Cieliebak, Don Tuggener and Fernando Benites (eds.): Proceedings of the 3rd Swiss Text Analytics Conference (SwissText 2018), Winterthur, Switzerland, June 2018

[1] https://www.swisstwittercorpus.ch/

[2] as provided by Twitter

ment annotation (based on Deriu et al. (2017)) as well as the Swissness-Score and why the Tweet was included in the set.

3 Data Expedition

The Data Expedition followed the following format. Participants were split into groups of 4-8 people (5 and 6 in our case), with the goal of forming diverse groups as a mix of developers, researchers, designers and storytellers. The participants were then handed the data along with an explanation of the data format and with an introduction to the Data Expedition. The participants then had roughly 3.5 hours to decide on one or more research topics and to analyze and visualize the data and their results. The teams were then able to present their findings to each other. A summary of the findings were presented the following day to the general audience of the conference.

4 Results

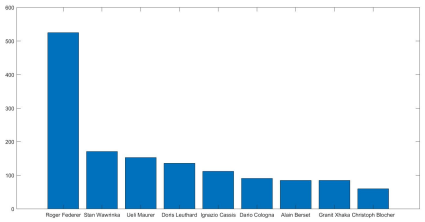
This section details the results and findings of the two teams.

4.1 Team 1

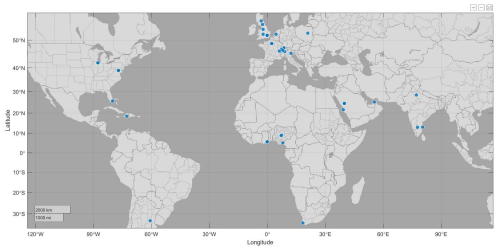
The first team focused on finding interesting patterns related to Swiss celebrities, most of which were included in the keyword set already. They looked into the relative number of Tweets per celebrity to find the most and least popular ones. The most popular celebrity in the dataset was Roger Federer, a famous Swiss Tennis player, and the least popular one was Christoph Blocher, a Swiss politician. The ranking can be seen in Figure 1.

The group then focused on analyzing Tweets about Roger Federer further. First, they looked at the language and geographic location of relevant Tweets, noticing that most Tweets originate in Switzerland, but that Roger Federer is also a popular topic world (See Figure 1). A majority of the Tweets was written in English, followed by German and French, with almost no Tweets being in Italian.

To finish their analysis, they looked into common words occurring together with Roger Federer as well as Hashtags related to the topic. A lot of the associations found were to be expected, like "Wawrinka", an important opponent of Federer, though some were surprising to the participants, like "Rotterdam", which they couldn't find an explanation for.



(a) Popularity of Swiss celebrities measured by number of Tweets.



(b) Geographic location of Tweets about Roger Federer.

Figure 1: Visual results of the first team.

4.2 Team 2

The second team decided to create a Twitter-based tourism guide of Switzerland. Specifically, they wanted to recommend top locations for a visitor to Switzerland in the months of February to May. They compared manually curated tourism guides with the Twitter data. To this end, they performed case studies for four different Swiss destinations.

Destination	# of Tweets
Geneva	7536
Lausanne	3379
Sion	1969
Zermatt	909
Verbier	312

Table 1: Popular destinations around lake Geneva sorted by number of Tweets.

Lake Geneva: Geneva was listed as the most popular tourist destination in multiple guides. The task was hampered by the existence of a town called "Lake Geneva" in the United States of America, which had to be filtered out. The team created a ranking of towns around lake Geneva by popularity (number of Tweets), as seen in Table 1.

Brugg: They then analyzed mentions of Brugg, a relatively small town in Switzerland, due to a number of participants coming from the FHNW university situated in Brugg. Due to the small size of Brugg,

only 40 relevant Tweets were found in the dataset and no conclusion could be drawn. Though participants did find an amusing, sexually explicit Tweet that they shared with the other workshop participants.

Lucerne: Lucerne, another popular tourist location, was mentioned in 5115 Tweets, with 227 mentioning the nearby Titlis mountain, a local tourist attraction. The participants didn't find any interesting information regarding this town, though they remarked on the Queen Victoria exhibition taking place there, which was of interest to the British team member.

Lugano: Next, the second team wanted to see if the mountains around the city of Lugano were mentioned in the dataset, since those are purported popular tourist locations. Surprisingly, the mountains were only mentioned a total of 14 times, even though Lugano itself was mentioned 2792 times.

City	% of positive Tweets
Bern	61.2
Luzern	65.9
Basel	71.4
Zrich	72.6
Lugano	86.7
Geneva	89.8
Lausanne	90.5
Zermatt	94.4

Table 2: Percent of positive Tweets (Positive Sentiment larger negative Sentiment) in various Swiss cities.

To round off their analysis, the team members looked at the distribution of sentiment annotations for mentions of Swiss cities (see Table 2). They couldn't find any overwhelmingly negative Swiss cities, but noticed that in general, the Italian and French part of Switzerland is more happy than the German one.

5 Discussion

We organized and executed a data expedition into Swiss Twitter data with a group of 11 people. The participants were very motivated and interested in the topic at hand and discovered several new and surprising insights from the data. Even though the total time available for the analysis was only 3 hours, the teams quickly settled on a topic to study and produced the first results. The workshop itself was praised by several participants and received positive feedback in general, pointing to data expeditions being a useful and easily introduced tool in education and research.

In the future, it might be useful to let participants chose their role in advance, to ease team formation. Producing general statistics about the data in advance and adding scaffolding code for participants to use might help participants finding a suitable topic and speed up development, at the risk of biasing participants towards certain avenues of exploration.

Overall, the expedition was successful and the format will likely be repeated by us in the future.

Acknowledgments

We would like to thank all the participants in the expedition (In no particular order): Stephen, Khalil, Nathan, Jacky, Stefan, Ela, Alma Karalic, Christoph Sess, Michael Sladoje, Alexandru Dimofte, Matthias Sommer.

We would also like to thank the organizers of the SwissText conference for the opportunity to lead this workshop.

References

AV Burov, AV Baranov, and AV Tagaev. 2016. Data expedition as an effective tool of creating a culture of working with open data of the future state and municipal officials. In *Proceedings of the International Conference on Electronic Governance and Open Society: Challenges in Eurasia*. ACM, pages 167–170.

Chiara Ciociola and Luigi Reggi. 2015. A scuola di opencoesione: From open data to civic engagement. *Open Data as Open Educational Resources* page 26.

Jan Deriu, Aurelien Lucchi, Valeria De Luca, Aliaksei Severyn, Simon Müller, Mark Cieliebak, Thomas Hofmann, and Martin Jaggi. 2017. Leveraging Large Amounts of Weakly Supervised Data for Multi-Language Sentiment Classification. In *WWW 2017 - International World Wide Web Conference*. Perth, Australia.

Irina Radchenko and Anna Sakoyan. 2016. On some russian educational projects in open data and data journalism. In *Open Data for Education*, Springer, pages 153–165.