How is Switzerland Feeling? Estimation and Visualization of Canton-Level Moods using Hybrid Fine-Grained Emotion Analysis Methodologies.

By Gökçen Nurlu, Meryem M'hamdi, Riyadh Alnasser
Project supervised by Dr. Michele Catasta, Distributed Information Systems Laboratory
Ecole polytechnique fédérale de Lausanne (EPFL), Switzerland

Rationale

The increasing availability of geolocated data in social media opened new ventures in the exploration and understanding of spatio-temporal movements patterns and possibilities to explain such irregularities in the mobility data. Twitter offers both individual and crowd-based geotagged data at different levels of precision.

The main objective of this project is to understand what triggers mobility in Switzerland by estimating how emotions of users change from one canton to another based on spatio-temporal data. We have followed fine-grained emotion recognition methodology and developped a platform for visualizing happiness in the swiss cantons to provide users and the general public with an interactive interface for mood pattern exploration using multiple time, space and user dimensions.

Methodology

We made use of dataset of 12M geolocated tweets collected in the area of Switzerland. After removing duplicates, dealing with nulls, parsing and other data treatment, we assign each tweet to its corresponding canton based geolocation information. We then chunk the dataset based on the language detected. We have followed a hybrid methodology which combines rule based and learning based to strengthen traditional lexicon spotting.

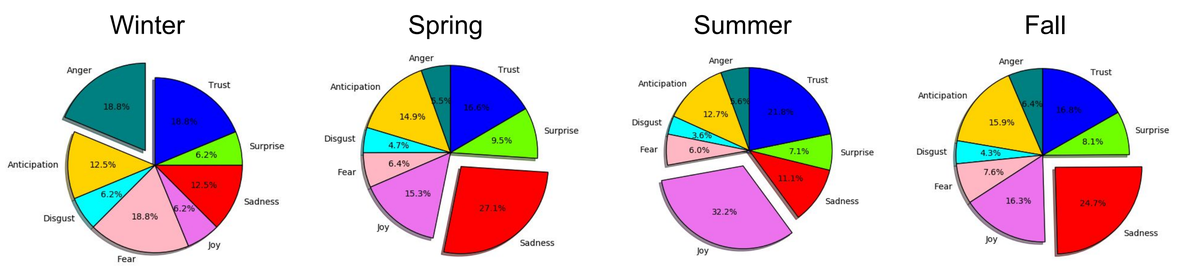

Fig1. Different emotions in different seasons of the year of 2012 in Canton du Jura. People are being more joyful during the summer!

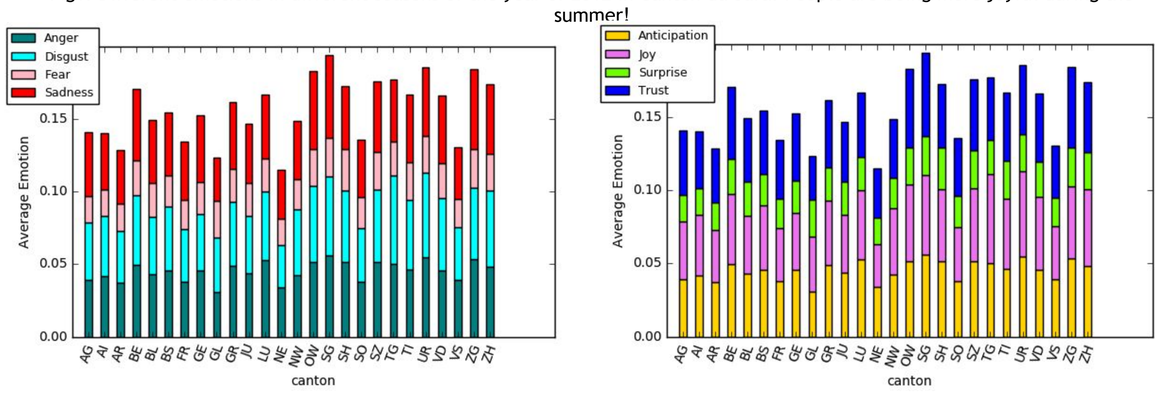

Fig2. The average of different emotions in different cantons in Switzerland for the year of 2015.

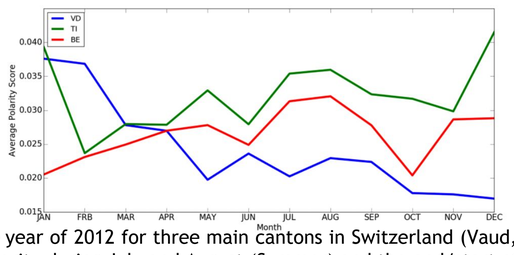

Fig3. Monthly Polarity Score in the year of 2012 for three main cantons in Switzerland (Vaud, Bern, and Ticino). Interestingly we observe a slight increase in the polarity during July and August (Summer) and the end/start of the year (Holidays).

Approach

We provide an interactive interface to visualize difference between variants of described ER methodology:

a) Simple Approach without syntactic analysis and lemmatization
b) Rule Based Lexicon Approach
c) Lexicon Approach extended with Word Embedding
d) Semi-Supervised Approach

We analyze both polarity and multi-category scores per canton at different levels of time granularity which can be played with in the interface and provide a way to visualize tweet density and number of users.

1) Data Pre-Processing: We tokenize tweets using RegexpTokenizer which splits using regular expression, perform punctuation removal.
2) Syntactic Analysis: We perform part of speech tagging using nltk POS tagger, construct dependency parsing trees using Stanford Dependency Parser and applying handcoded syntactic rules to refine emotional scores based on influence of negation, adjectival, adverbial dependencies. We remove NRE entities and kepp only NAVA (Noun, Adjective, Verb, Adverb) words.
3) More Cleaning: We convert to lowercase and perform lemmatization to normalize terms and remove stopwords.
4) Hierarchical Semantic Analysis extended with Word Embeddings: rather than relying solely on keyword spotting, we try different approaches to distributional semantics (pmi, word2vec) to extend lexicon based on similarity scores with a set of representative words for each emotion category.
5) Word and Sentence Emotional Vector Averaging and Normalization using geometric mean
6) Semi-Supervised Approach on a set of annotated sentences using Domain Adptation.
7) Fine Tuning thresholds and ML parameters.

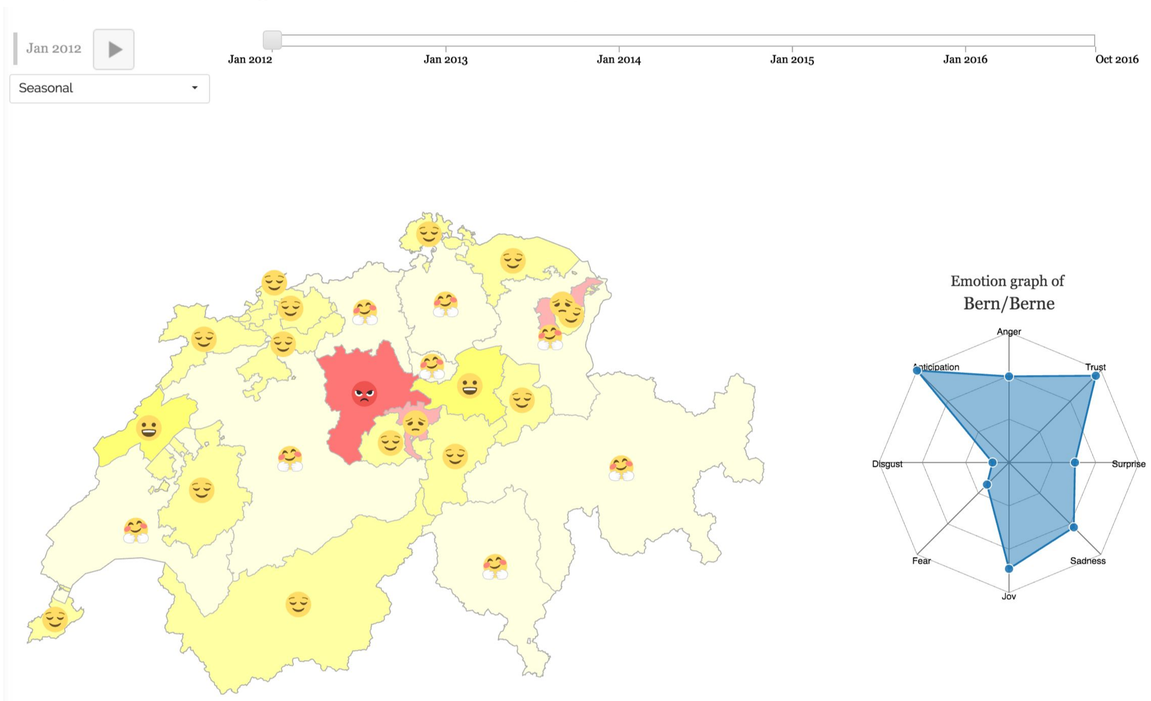

Fig4- screenshot of the visualization tool.

Results

During our analysis , we have interestingly observed that the mood of some cantons changes with the seasons of the year. For example in 2012 Canton du Jura registered different emotions in different seasons, for instance, people are being more joyful during the summer (see Fig1).

Looking at the sentiment of tweets being posted in Switzerland, and how it is changing in different months of the year, we found that during summer and before end of year for some canton the average sentiment of the tweets increases, as opposed to different times of the year (see Fig3).

Fig2 shows the average of different emotions expressed in tweets in different cantons during the year of 2015.

Challenges and Future Work:

- We had to apply a hybrid approach for emotion analysis since we don't have annotated dataset which makes a fully supervised approach leading to a better control of accuracy impossible.
- In future work, it could be interesting to combine a thorough analysis of migration flows with how emotions change. By analyzing the emotion of the hometown vs workplace or destination, we can discover interesting patterns and build a model that predicts happy paths and recommends them to users based on the history of what made them happiest.

More about this project
This project was completed by EPFL Master students as part of Applied Data Analysis Course taught by Dr. Michele Catasta. For further details contact michele.catasta@epfl.ch