Sentiment Analysis and Visualization for Switzerland based on social media

Kirtan Padh, Luis Medina, Tina Fang

EPFL - Switzerland

Abstract

Switzerland is always near the top when we see the list of happiest countries in the world. We try to do an analysis of how happy each part of Switzerland is during different times of the year and what is it that makes it happy by doing a sentiment analysis of social media posts.

Overview

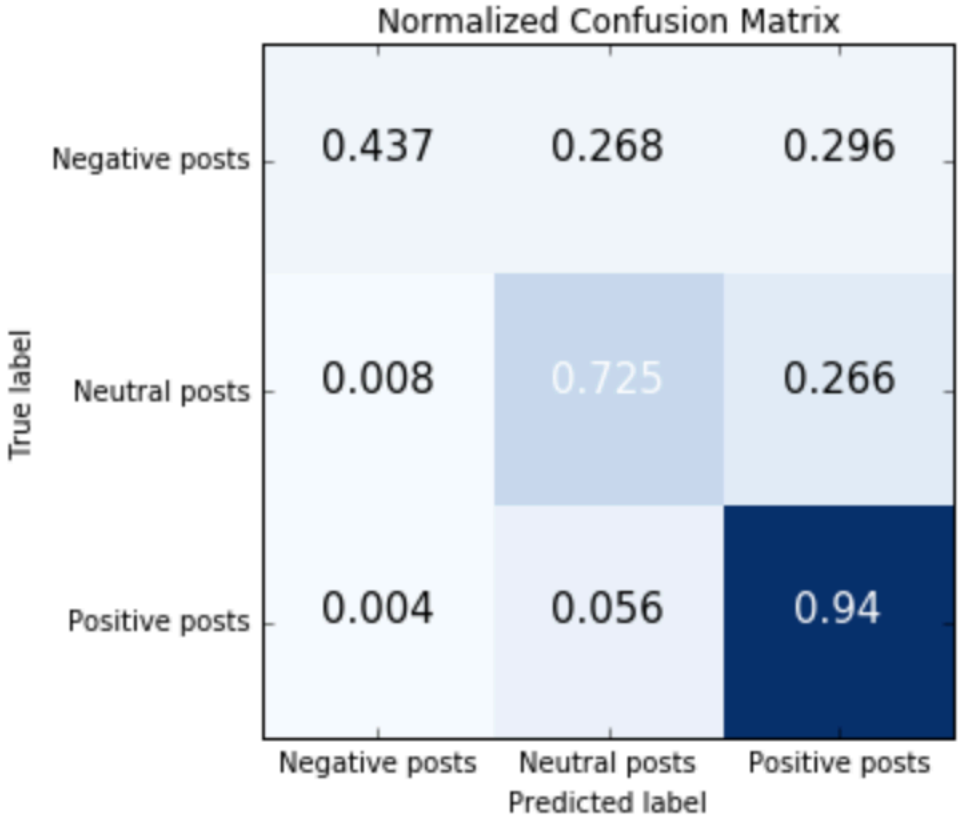
We start by doing some analysis on spark on the cluster and getting the data in the right format to do a sentiment analysis of the Instagram hashtags to get the sentiment for Instagram. We are already given the sentiment for twitter.

We then get the data in the right format and clean it further to do the visualisation, and finally make an interactive dynamic choropleth map of Switzerland which visualizes the average sentiment for each canton. We can filter by time of day, month, gender and language in the dynamic map. All the code and visualisation is available at https://github.com/tbfang/ADA-Project

Machine Learning

FastText is a library implemented by Facebook for word representation learning and text classification. It is based on the idea that linear models with a rank constraint and a fast loss approximation can train on a billion words within ten minutes, while achieving performance on par with the state-of-the-art. Therefore, over the several methods for text classification, FastText has the advantage of being much faster, while achieving high accuracies.

WordCloud for Positive Instagram Posts



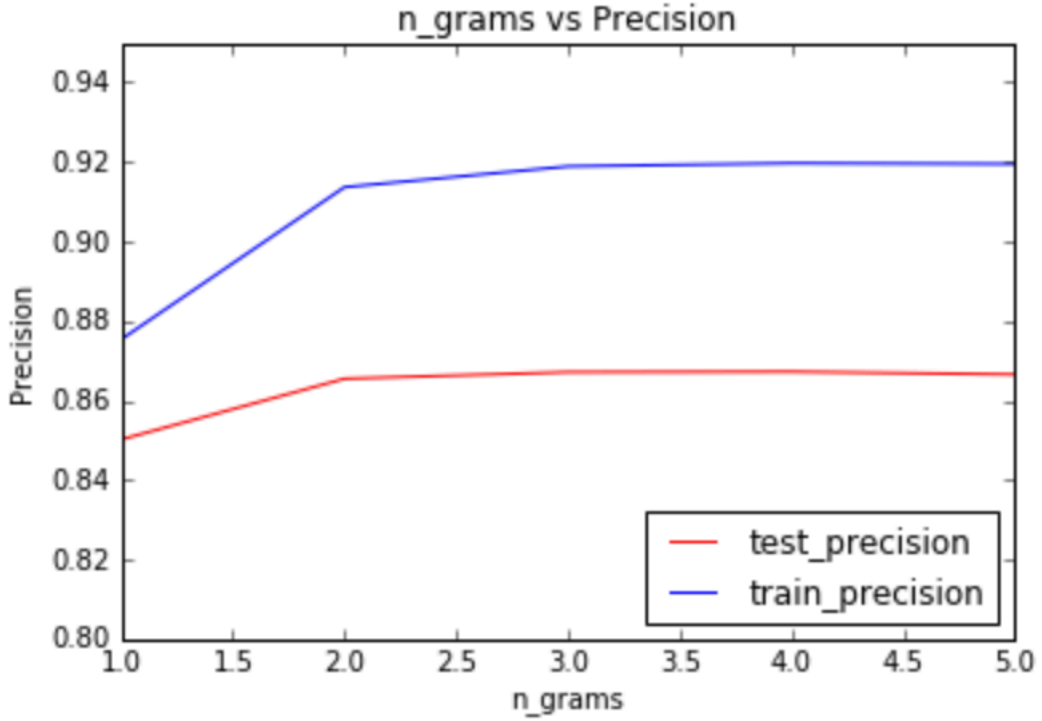
Normalized Confusion Matrix

- Distribution of the sentiments in the training data set:
  - **Positive posts**: 71.98%
  - **Neutral posts**: 23.21%
  - **Negative posts**: 4.81%

The results we are getting in the normalized Confusion Matrix are reasonable considering the distribution of the sentiments in the training data set

Optimal parameters for our model:
for FastText:

- Learning rate = 0.044
- Windows size = 5
- Minimum word count = 3
- Words n-grams = 2



Results

Number of Instagram posts in the test data set: 6,845,983

Precision of the created model: 0.866590

Visualization Overview

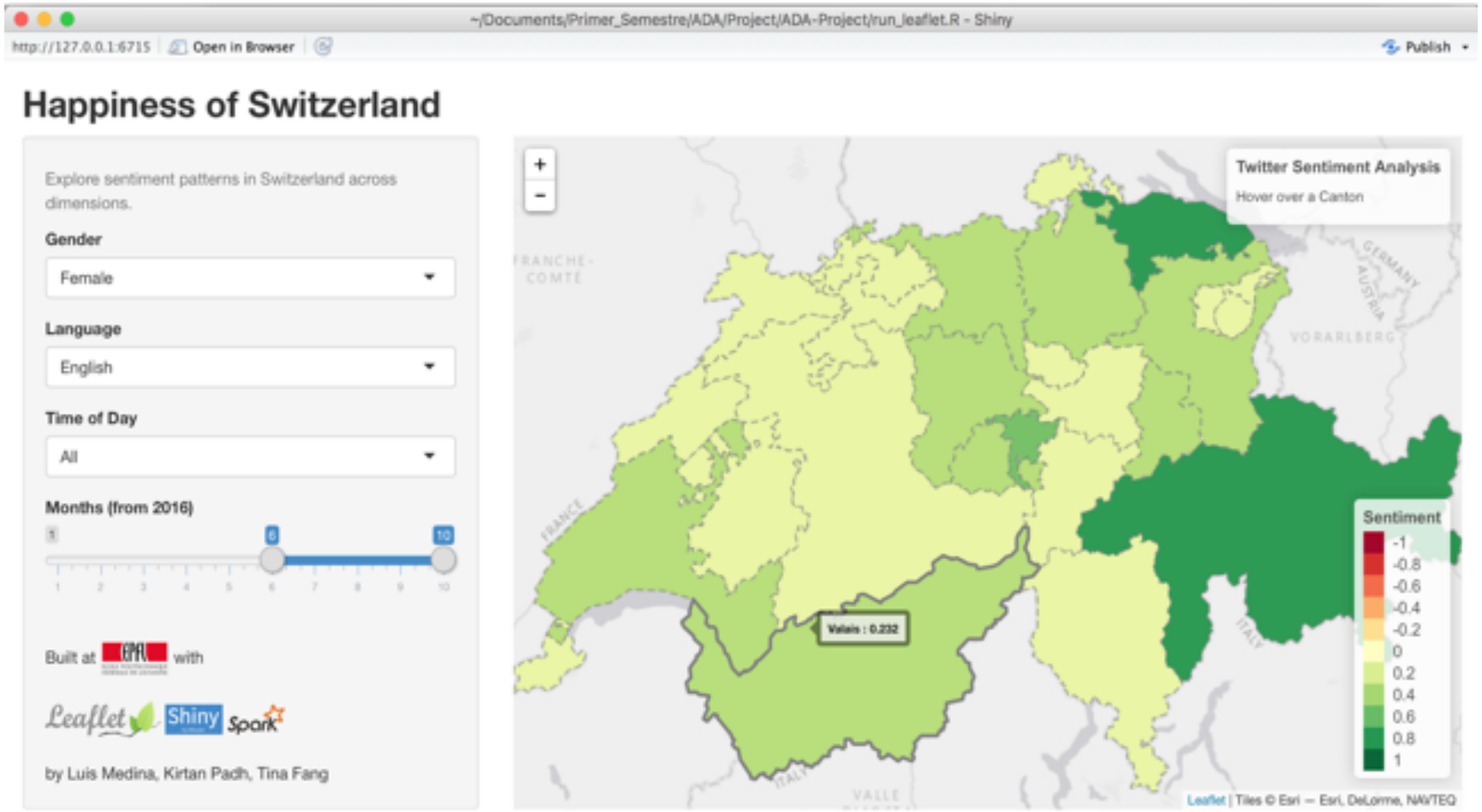
There are 10.5 million tweets from 2016 in Switzerland alone. Canton, gender, language, time of day, and month are 5 dimensions to filter the data by. Out of these dimensions, we obtain 3150 different possibilities for filters. In order to show that to users, we must create an interactive visualization with dynamic filtering.

To build interactive maps, we explored Python libraries such as Folium, Vega, Plot.ly, Bokeh. However, Folium and Vega for LeafletJS and Vincent, respectively, are limited in functionality. For example, hovering highlighting is not possible with Folium. Plot.ly and Bokeh have limited customization for geolocation of the map. We also explored RMaps, which creates interactive choropleths. However, the user interface is less appealing than modern libraries, such as Leaflet for R.

Visualization Methodology

We have used Leaflet and Shiny to build our interactive map in web application. Leaflet is used to build the map visualization with highlighting. We obtained an open sourced Swiss cantons TopoJSON file, and converted it to kml file, which is used to build the base of the Leaflet. Then, we added one single layer to our Leaflet map, representing the average sentiment in the value. The input to our map was an 26-row R DataFrame with canton and average sentiment as columns.

In order to dynamically filter based on user input, we used Shiny to build our web app. We do not want to manually create the DataFrames for each combination, because there will be well over 360 combinations. Hence, we filter and aggregate the full dataframe using parameters from user input and pass that to Leaflet to build the map.



The final choropleth map explores the average sentiment of tweets in Switzerland across the 5 dimensions.

Interactive choropleth web app with dynamic filtering