Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Federal Department of Home Affairs FDHA
Federal Food Safety and Veterinary Office FSVO
Research Management

Alert Centre for foodborne diseases in Switzerland: Identification and localization through social media

Jacky Casas, Omar Abou Khaled
Haute Ecole d'Ingénierie et d'Architecture Fribourg, Boulevard de Pérolles 80, 1700 Pérolles

Key words
Epidemic detection, food intoxication, food safety, tweets classification, tweets localization, machine learning

Aim of the study
This project is based on the project 17AC where a web platform analysing tweets in French to detect food intoxication outbreaks in Switzerland was developed. The goal of this project is first to extend the platform to deal with tweets in the German and Italian languages, and second to optimize the machine learning classifiers and the localization of the tweets in the Swiss territory.

Material and methods
The list of French keywords was revised and then translated in German and Italian in order to collect all the tweets containing at least one of these keywords in real time. The platform was updated to support these three languages in different ways: the interface of the dashboard can now display information specific to language and the backend API can deal with multilingual data after a full rewrite of the code and a change of technology, allowing it to handle more data in a faster way. New functionalities were developed, the first is an interface that allows the user to manually label the tweets (a tweet can be either relevant or not relevant to food intoxication). To help increase the quantity of relevant tweets, the user can also craft his own "fake" tweets. The second functionality is an interface that shows information regarding the three active classifiers (one for each language), as well as previous models. The user can retrain the classifier with a simple mouse click and select one of them to use to classify the future tweets. Other pages allow the user to discover insights about news media and other Twitter accounts.

In order to improve the classifiers score, we needed to have a "big-enough" training set. The first strategy was to label by hand the tweets we collected, but this was time-consuming because very few tweets were relevant in the mass of unrelevant tweets. We then manually crafted "fake" tweets and even crowdsourced this process in the three languages. It worked well for French but not that well for German and Italian, so we finally translated each tweet in the two other languages to increase the training set size.

Regarding the geolocalization of tweets and users, improvements have been done. We defined the term "Swiss influencers" as Twitter accounts that should be followed mainly by Swiss citizens. A list of this kind of accounts was created with currently 785 accounts ranging from Swiss media (televisions, radios, newspapers) to Swiss politicians, local sports clubs, political parties and so on. This list is freely available on Github[1]. We then collected each and every follower of these accounts. Based on the location provided by the users it was possible to create a dataset of self-proclaimed Swiss accounts, that we used later as a ground truth for a machine learning algorithm. Instead of trying to precisely localize users in the Swiss territory, we tried to know if they were from Switzerland or not (other country). Different features like the quantity of Swiss influencers they follow, the use of hashtags, the mention of other Swiss accounts or the places cited in the text were used to train a classifier to detect if a user is Swiss or not.

[1] https://github.com/acknowledge/swiss-twitter-accounts

Results and significance
The new platform ran for entire 248 days. The 209 keywords in the 3 languages generated 14'850 tweets in average on a daily basis (66% in French, 17% in Germand and 17% in Italian).

The classifiers that determine if a tweet is relevant to food intoxication or not achieve all more than 90% accuracy with around 700 tweet samples equally balanced between relevant and not relevant (91.2% for French, 90.0% for German and 91.5% for Italian). We tried to label more tweets in French to see the effect of having more training samples. With a bit less than 1000 tweets, the accuracy decreased to 81.9%. The reason is that the more precise we go, the more difficult it is to detect if a tweet is relevant or not, even for a human.

Each tweet is localized with the four basic algorithms (GPS coordinates, places in text, hashtag place, URLs) and each author is localized with the three basic algorithms (location field, URL and timezone). The algorithm that most accurately localizes wins. With this technique we are able to locate 75.09% of the tweets, with more or less precision. Only 2.31% of these tweets are located in Switzerland. The more advanced algorithm to locate the users with their networks interactions (influencers and other Swiss users), hashtags and more achieved 95% accuracy and 95% F1-score (publication in process). This very good result is therefore counted as the fourth algorithm of the platform to localize users.

During almost a year, tweets and Twitter users were collected and processed. The "Swiss influencer" followers operation was done. All this allowed us to gather a dataset of 177'680 Swiss Twitter users. This corresponds to 23% of the estimated total number of existing Swiss Twitter accounts. With the last version of the classifiers, we can report on average 91 relevant tweets per day published in Switzerland. In the list, a lot of them are borderline tweets that we cannot automatically discard, even with a human brain due to a lack of context. In 2018, only 12 outbreaks caused by the consumption of food were reported. None of them were detected by the system, because it concerned only 153 persons in total. At the end of 2019, an outbreak of Norovirus happened in France and then moved to Switzerland. After analysis of the tweets, 132 of them came from France (for more than 1000 cases) and only 3 from Switzerland, which is too few to be able to detect the outbreak in Switzerland.

Publications, posters and presentations
Casas, J., Zufferey, L., Abou Khaled, O., & Mugellini, E. (2018). Early Detection of Food Intoxication in Switzerland using Twitter. FTAL Conference on Industrial Applied Data Science, Proceedings, 11–12.
Casas, J., Mugellini, E., & Abou Khaled, O. (2020). Early Detection of Foodborne Illnesses in Social Media. In T. Ahram, R. Taiar, V. Gremeaux-Bader, & K. Aminian (Eds.), Human Interaction, Emerging Technologies and Future Applications II (pp. 415–420). Springer International Publishing. https://doi.org/10.1007/978-3-030-44267-5_62
The project was presented at the 21st Emerging Risks Exchange Network (EREN) meeting the 10th of April 2019 at the European Food Safety Authority (EFSA) headquarters in Parma, Italy.

Project *1820AC – Alert Centre for foodborne diseases in Switzerland: Identification and localization through social media*

Project duration *24 months (August 2018 – July 2020)*