Early Detection of Foodborne Illnesses in Social Media

Jacky Casas(✉), Elena Mugellini, and Omar Abou Khaled

University of Applied Sciences and Arts Western Switzerland, Fribourg, Switzerland
{Jacky.Casas,Elena.Mugellini,Omar.AbouKhaled}@hes-so.ch

Abstract. Alert Center is a platform aiming at detecting outbreaks caused by food toxin infections and food intoxications in Switzerland. It does this by analyzing tweets and sending alerts to the Federal Food Safety and Veterinary Office (FSVO) when a risk is detected. The platform is composed of four main parts: a real-time extractor that targets tweets based on a list of curated keywords, three classifiers (one for each main spoken language) that isolate tweets related to food toxin, a system that locates tweets on the Swiss territory and a web-based dashboard to visualize the results. Combining localization algorithms of tweets and users allows the system to locate 75.09% of the tweets, 2.31% of which were located in Switerzland. In addition, a list of Swiss Twitter accounts corresponding to 15% of the total estimated number of Swiss accounts has been created.

Keywords: Data analysis · Localization · Classification · Risk assessment · Twitter · Food outbreaks detection

1 Introduction

Food toxins infections and food intoxications, in general, can cause outbreaks. If we could prevent these outbreaks, we could avoid health risks for the population, reduce health costs and save a lot of money for companies that have to pay the salaries of their employees in recovery. In Switzerland, the Food Safety and Veterinary Office [1] is responsible for assessing the risks and taking measures to combat the spread of these epidemics. The current system for learning about a risk takes time. Sick citizens usually go to see their doctor after a few days of illness, considering that they are not just waiting for the disease to pass. Cases of poisoning are then reported to the cantonal doctor, who will himself notify the FSVO. The whole process can take from a few days to one or two weeks! The effect of the measures taken at that time is therefore relative. Indeed, sick people will probably already be cured at best, and an epidemic will turn into a pandemic at worst.

Nowadays some people tend to share details about their personal life on the internet and specifically on social networks. If that is the case, it is, therefore, possible that some of them may share the fact that they are sick for some reason. These reasons could be the consumption of poisoned or rotten food, purchased in a store or restaurant for

T. Ahram et al. (Eds.): IHIET 2020, AISC 1152, pp. 415–420, 2020.
https://doi.org/10.1007/978-3-030-44267-5_62

example. If this is the case, it should be possible to retrieve this information by analyzing social networks, to find the information directly at the source.

The Twitter social network was chosen to carry out the experiments. Several reasons have confirmed this choice, including the fact that the data are public, unlike other social networks at the moment, and the fact that the data collection is easy and in real-time.

1.1 Challenges

This project brings different challenges. Indeed, Switzerland is a small country with currently about 8.5 million people. To make the detection even harder, four different languages are used (German, French, Italian, and Romansh), not to mention English. Twitter is also not widely used in the country and most of its use is for political or media reasons. The number of Swiss Twitter users as of January 2019 is approximately 765'000 [2].

1.2 Case Study

To validate our hypotheses and assess the feasibility of such a project, a case study was conducted. An outbreak of gastroenteritis due to water contamination infecting about 1200 people raged in the Le Locle area in July 2015. This represented almost 10% of the local population. Tweets published during this period were collected and analyzed to determine if it was possible to detect the epidemic. The results showed a correlation between the number of tweets mentioning symptoms and the reported cases. 9 tweets were concerned. These pre-results gave the green light to continue the research [3].

2 Methods

The platform is composed of four main components: the tweets extractor, the localization system, the classifier, and the visualization, as seen in Fig. 1. The four parts are explained in detail in the following sections.

2.1 Tweet Extractor

The role of this extractor is to collect in real-time the tweets that are published on Twitter. A list of keywords has been defined by the FSVO. This list is divided into three categories: reasons, impacts, and locations. The use of these words, alone or in combination, can identify intoxications. These expressions contain words from standard language but also from colloquial language to fit regional expressions. There are about 90 words per language, and three languages (German, French and Italian).

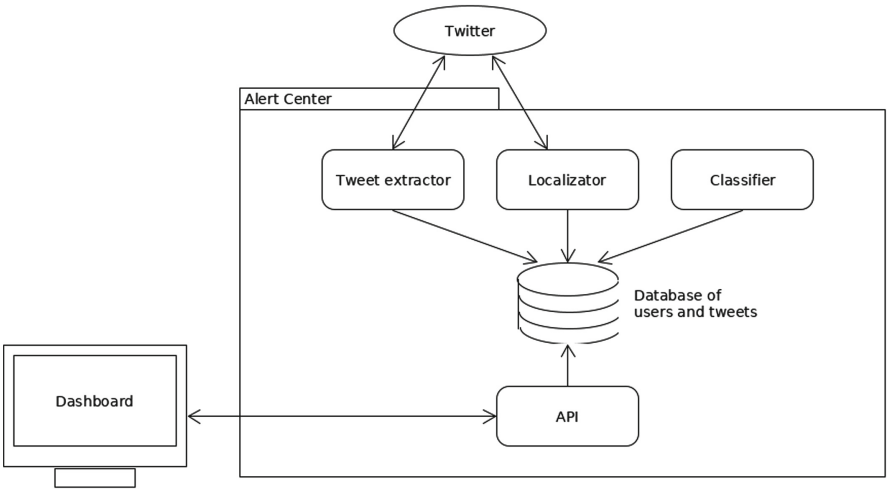

Fig. 1. The architecture of the Alert Center platform

2.2 Localization

Each tweet will be processed by the localizator and the classifier independently. The purpose of the localizator is to determine where a tweet has been published. To do this, several algorithms [4] are applied on the tweet and on the user who published the tweet. Different granularities are possible, from the country to the exact GPS coordinate, via the canton or city precision level. The most consistent and accurate location will then be chosen.

To locate a tweet, the first thing considered is whether the tweet was geolocated when it was published. This is rarely the case, but when it is, it is very useful. If a link is contained in the tweet, the extension will be analyzed, for example, ".ch" corresponds to Switzerland, ".de" to Germany. And then the GeoNames database [5] is used to detect the names of cities, cantons and countries (in different languages) contained in hashtags and tweet plain text [6].

In order to locate users, two pieces of information are taken into account: the "location" and "website" fields that the user indicates in his profile. The techniques of URL country extension and place names are used, as for tweets.

2.3 Classification

The purpose of the classification is to separate tweets that talk about food poisoning from tweets that have nothing to do with the subject. The majority of the tweets collected are unrelated and considered irrelevant. It is, therefore, a binary classification: relevant and not relevant. The classifier used is a simple SVM with a linear kernel whose input features are occurrences of the words used in the tweets weighted using the TF-IDF method. Only the textual content of the tweets are used as features [7].

To keep the performance while limiting the need for training data, it was chosen to have one classifier per language, which gives us 3 classifiers to train (German, French and Italian).

The training data comes from a manual sorting of the collected tweets. Faced with the difficulty of the task, it was decided to add manually created relevant tweets to the dataset. To do this, an online questionnaire asking people to create 10 fake tweets was distributed to some Swiss schools and institutions. Each handmade tweet was then double-checked to ensure relevance.

2.4 Visualization

The visualization was done with web technologies. It is a dashboard designed with HTML/CSS and Vue.js that makes data easy to interpret and interactive with graphics and maps. The data is loaded on the interface through a REST API. Alerts are sent by email to FSVO managers when a threshold of relevant and Swiss tweets is exceeded for a day, allowing them to consult the dashboard to learn more and investigate the potential risk of an epidemic. A weekly email summarizes the events that occurred during the week.

3 Results

In terms of the real-time tweets extractor, the 90 or so keywords per language allow us to retrieve an average of 57,929 tweets per day. The breakdown of languages is as follows: 61.82% for French, 20.2% for Italian and 17.98% for German.

Each tweet will first be localized with the four algorithms (GPS coordinates, places in the text, hashtag place, URLs). The success rates for these algorithms are as follows:

- 14.51% of the tweets have GPS coordinates
- 42.44% of the tweets contain a place name in the raw text
- 1.29% of the tweets contain a place name in a hashtag
- 4.22% of the tweets contain a URL with a country extension

When we collect a tweet, we also retrieve the account of the user who wrote the tweet as well as the accounts of the users cited in that tweet, if any. By combining the 4 tweets localization algorithms and the two ways to localize users for authors and mentioned users, we can define a localization with a certain degree of accuracy. The algorithm that most accurately localizes wins. Here are the results that allow us to locate 75.09% of tweets:

- 14.51% of the tweets with GPS coordinates
- 18.65% of the tweets with a place name in the raw text
- 1% of the tweets with a hashtag
- 1.31% of the tweets with a URL
- 24.36% of the tweets with the location of their author
- 15.26% of the tweets with the location of a user mentioned in it

So the percentage of unlocalized tweets is 24.91%. It must be understood that location information is only an indication. We cannot be sure, for example, that a user lives close to a person he mentions.

Among all the tweets localized, what is of interest for the current project are the tweets posted in Switzerland. The percentage of tweets located in Switzerland is 1.74% of all the tweets collected, and 2.31% of the tweets localized with our technique.

Let's move on to the classifiers part. To have enough training data, we have crowdsourced the creation of fake tweets. This process is still ongoing and will allow us to collect enough tweets. For the moment we have a few hundred tweets, which is not enough to have significant results yet. We achieve classification results between 75% and 90%.

In order to speed up the process of locating Swiss users, we have developed a system to collect potential Swiss users in advance and preprocess them. To do this, we have defined the notion of a "Swiss influencer" as an account followed by an audience of mostly Swiss people. For example, local sports clubs, politicians or regional media. We have therefore compiled a list of 785 such accounts. This list is available on Github[1] [8]. We then collected one by one all the followers of each influencer, tried to locate them thanks to the "place" field and saved in the database only the Swiss users. So when a new tweet is detected, we won't necessarily need to look for information about its author because it will probably already be in our database. As of now, we were able to collect 120'000 self-proclaimed Swiss accounts, which corresponds to approximately 15% of the total Swiss Twitter accounts. This number is achieved with only 240 of the 785 influencers. The system is still running and we hope to increase this number in the near future.

4 Discussion

In this article, the results are given mainly for the part dealing with localization. But the platform has been collecting data continuously for more than 15 months now and further analysis will be carried out in the future. In the meantime, we will improve classifier performance to detect all relevant tweets.

At the moment, the collection of followers of influencers is still in progress. Indeed, the limitations of the Twitter API make the process a little longer. We aim to create a dataset containing as many Swiss Twitter accounts as possible.

Acknowledgments. We would like to thank the FSVO for the funding of this platform and more particularly Françoise Fridez, Thomas Lüthi and Vincent Dudler, as well as the people who have participated in its design over the years: Laurent Zufferey, Laurent Chassot, Lucas Alborghetti, and Jérôme Vonlanthen.

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

References

1. Federal Food Safety and Veterinary Office (FSVO). https://www.blv.admin.ch
2. Digital Report 2019 for Switzerland. https://datareportal.com/reports/digital-2019-switzerland
3. Casas, J., Zufferey, L., Abou Khaled, O., Mugellini, E.: Early detection of food intoxication in Switzerland using Twitter. In: Proceedings of the FTAL Conference on Industrial Applied Data Science, pp. 11–12 (2018)
4. Zheng, X., Han, J., Sun, A.: A survey of location prediction on Twitter. IEEE Trans. Knowl. Data Eng. **30**(9), 1652–1671 (2018)
5. GeoNames. https://www.geonames.org
6. Schulz, A., Hadjakos, A., Paulheim, H., Nachtwey, J., Muhlhauser, M.: A Multi-indicator approach for geolocalization of tweets. In: Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media (2013)
7. De Silva, L., Riloff, E.: User type classification of tweets with implications for event recognition. In: Proceedings of the Joint Workshop on Social Dynamics and Personal Attributes in Social Media, pp. 98–108 (2014)
8. List of Swiss influencers Twitter accounts. https://github.com/acknowledge/swiss-twitter-accounts