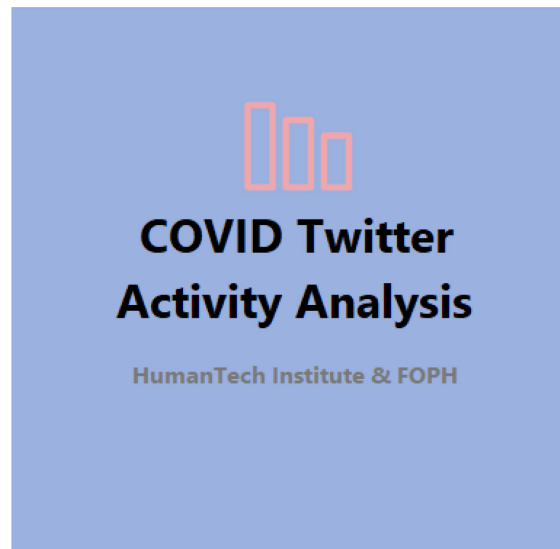




Report – Data Analysis

Version 1.0 - October 2020

HumanTech
Technology for
Human Wellbeing Institute



Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Swiss Confederation

Federal Department of Home Affairs FDHA
Federal Office of Public Health FOPH

1 INTRODUCTION

This document reports statistics and analyses of the data collected by the developed platform. These analyses present different views of the collected data and the information that can be extracted from this data. The report also details the results of classification algorithms that have been trained and evaluated on the data collected by the platform.

In this report, in order to perform the analyses, and generate the statistics and figures, we took into accounts the tweets that were collected by the platform between **Sunday, July 26th** and **Thursday, September 24th**.

Note that the database has been cleaned once on August 31st, after more than one month of data collection, because the system was getting slow when displaying the data on the dashboard. At that time, the total quantity of tweets collected and stored in the database was around 2 million. During this cleaning operation all tweets written by non-Swiss accounts, which are not relevant for our study, were deleted from the database.

This document is organised as follows. In chapter 2, we present the global statistics about the collected tweets. In chapter 3, we present information about the tweet labelling process and resulting accuracy of the algorithms performing the classification of the <Relevant> and <Non relevant> tweets. In chapter 4, we present statistics and different analyses of the tweets classified as <Relevant> by the classifiers. We finally provide a small conclusion in chapter 5.

2 GLOBAL STATISTICS ON TWEETS COLLECTED

Overall, more than 2.8 million of tweets have been collected by the platform for the considered period. Below you can find the global statistics of the number of tweets collected by the platform at 2 different dates.

The Figure 1 corresponds to the system on the 31 of August, before the system was cleaned to remove tweets not localized as swiss tweets. Note that the classification algorithms were not activated at that time, hence the value “0” for the relevant tweets.

The Figure 2 corresponds to the system on the 24th of September, when we started producing this analysis document. We can see that about 1 additional million of tweets were collected during these 3.5 weeks since the 31th of August and about 16’000 of those were identified as produced by individuals living in Switzerland. From the total of 73’734 swiss tweets, the system classified 12’355 tweets as “Relevant”.

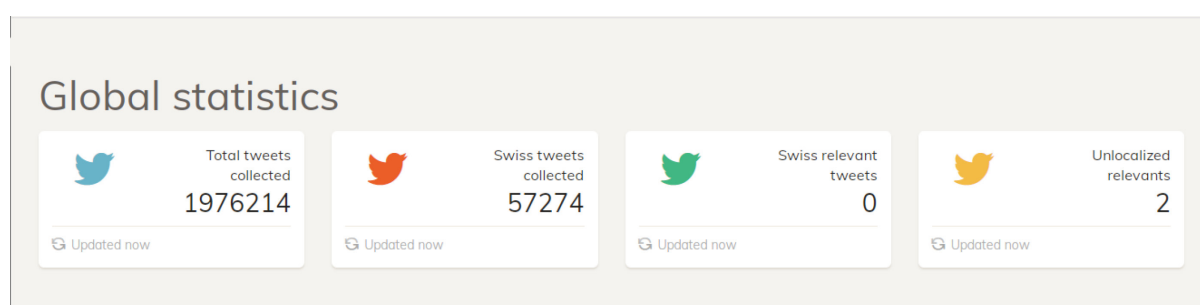


Figure 1: Global statistics before database cleaning (August 31st)

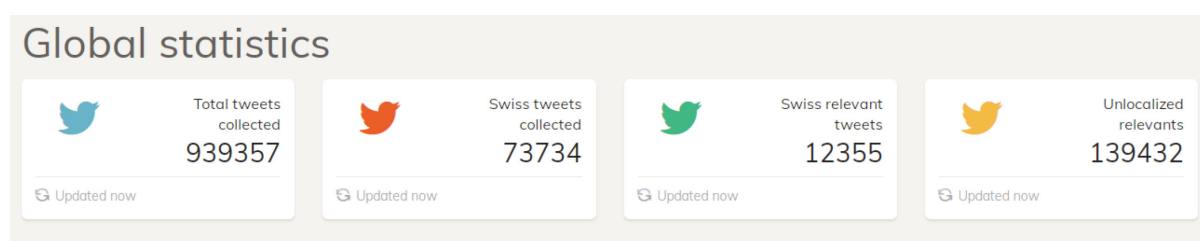


Figure 2: Global statistics on September, 24th

2.1 NUMBER OF TWEETS COLLECTED OVER ALL ACCOUNTS

The graph below shows the number of tweets collected from all Twitter accounts, regardless of their localization. We see that the system collected between 15’000 and 25’000 tweets per day. These tweets contain at least one of the keywords specified previously in the project (see [google sheet](#)).

Note that during a short period, the system collecting the data was down (27-31.08.2020). This explains the sudden drop of tweets visible on the figures below.

Global Statistics

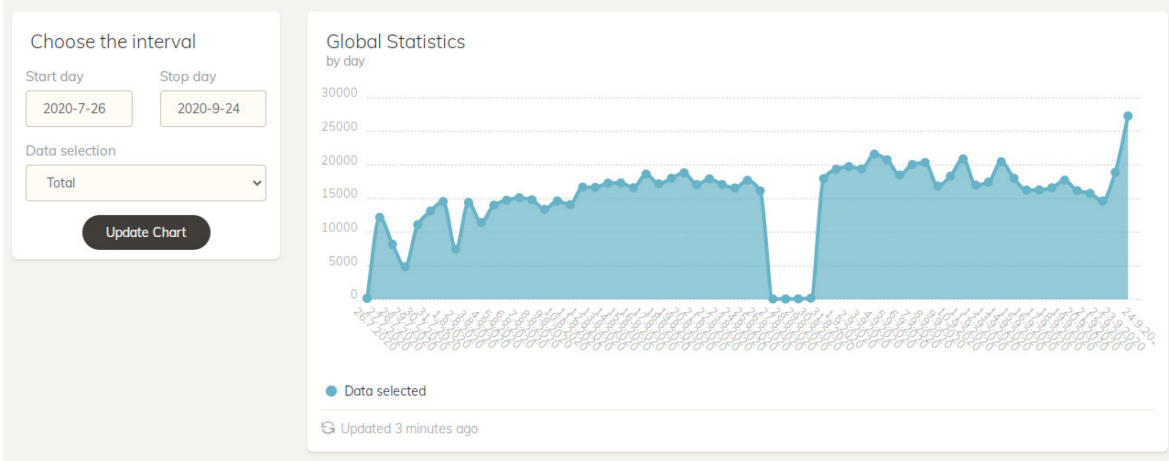


Figure 3: Global statistics per day - All accounts

2.2 NUMBER OF TWEETS COLLECTED FROM SWISS ACCOUNTS

The graph below shows the number of collected tweets from accounts localized in Switzerland by the platform. We see that the system collected between 1000 and 2000 tweets per day.

We can also observe that the number of tweets collected seems to decrease in the last weeks. This would indicate that COVID related topics are being less discussed on Twitter as time goes by. This tendency is not visible over all accounts, which might indicate a tendency only for the Swiss population.

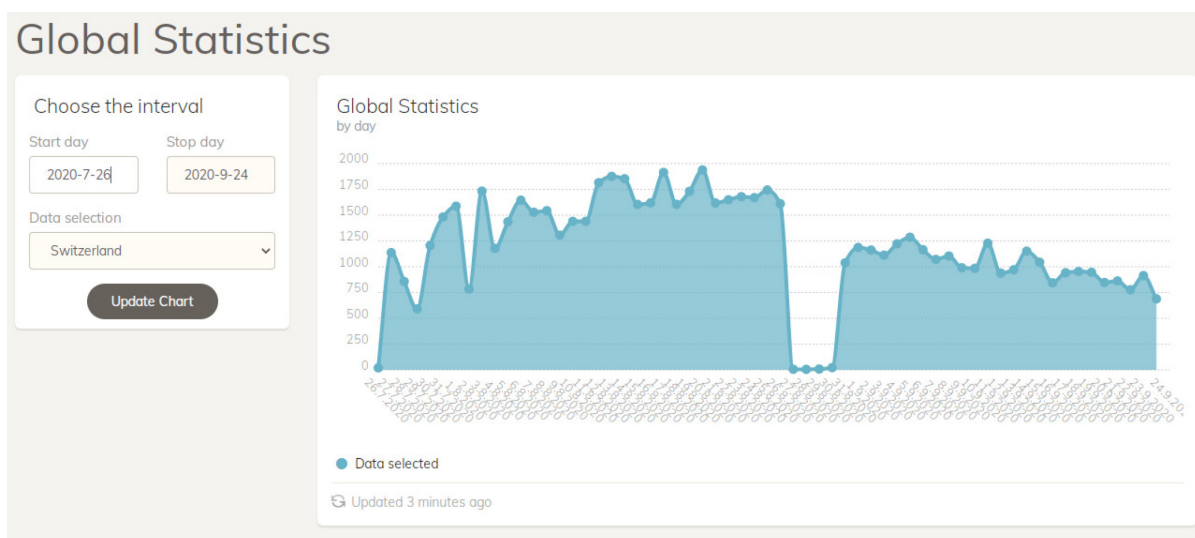


Figure 4: Global statistics per day - Swiss-based accounts

2.3 MOST USED KEYWORDS IN ALL COLLECTED TWEETS

Below, you will find the **10 most used keywords** in all the collected tweets. It shows the average amount of hits per day and the total number of use for each keyword.

Keywords Ranking by period

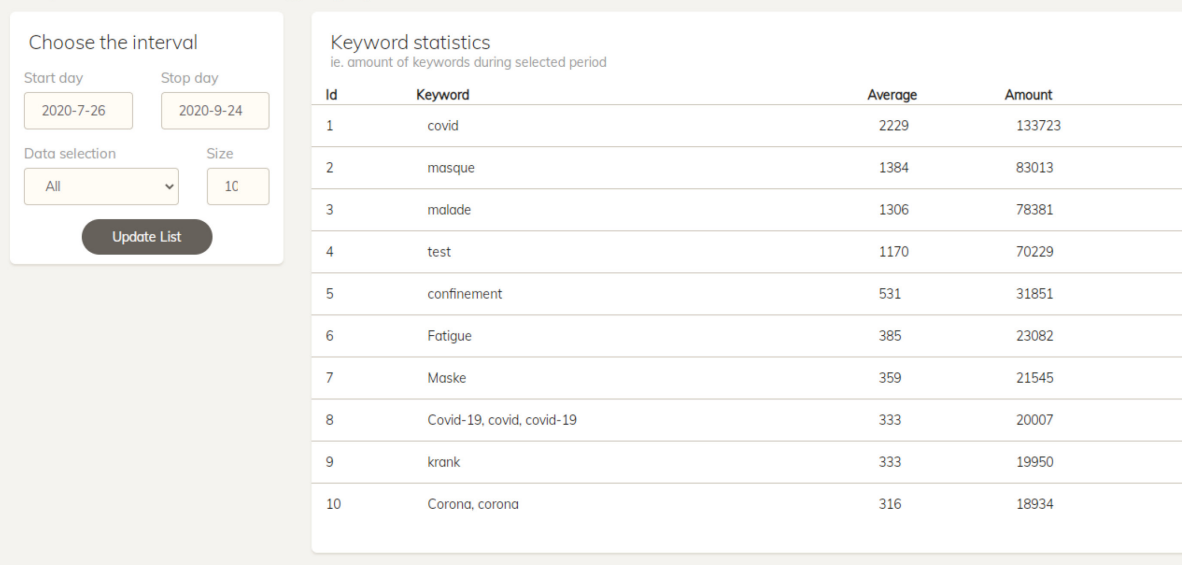


Figure 5. Most used keywords in all collected tweets.

2.4 MOST USED KEYWORDS IN COLLECTED SWISS TWEETS

Below, you will find the **10 most used keywords** in all the collected tweets **localized in Switzerland** by the platform. It shows the average amount of hits per day and the total number of use for each keyword.

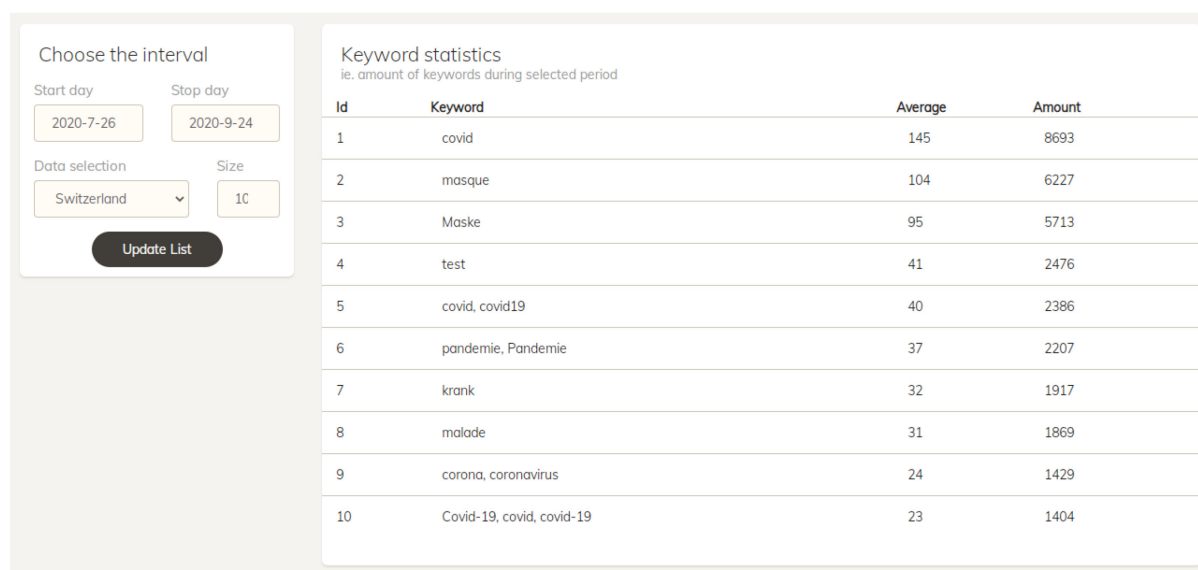


Figure 6. Most used keywords in Swiss collected tweets.

2.5 LANGUAGE REPARTITION

Since the platform collects data in three languages, a pie chart provides an idea of the repartition of these languages among all the collected tweets (not only the ones localized as Swiss). French is usually leading with approximately 60% of the tweets and then Italian and German with each 20%. These numbers are similar to the ones obtained in other projects (collecting data in other contexts, using different keywords) and only reflects the presence and use of Twitter in French, Italian and German-speaking populations worldwide.

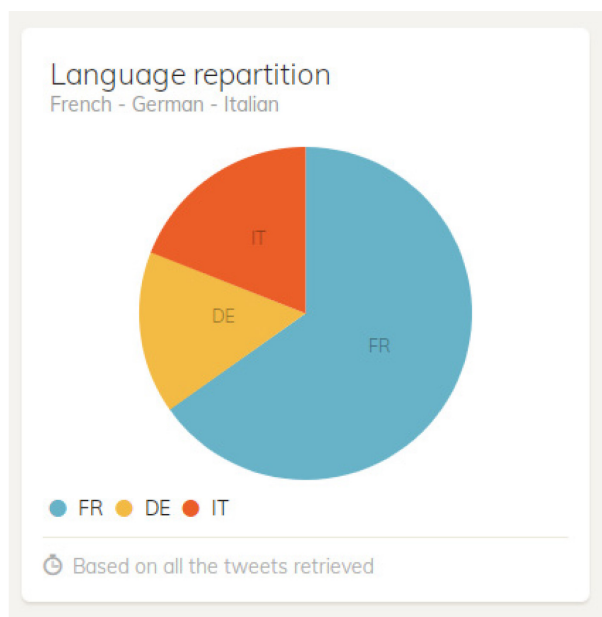


Figure 7. Language repartition of all collected tweets.

3 CRAFTED TWEETS AND LABELLING

3.1 HAND-LABELLING OF COLLECTED TWEETS

Once enough tweets had been collected by the system, we hand-labelled tweets in the 3 languages, based on the criteria defined together. These hand-labelled tweets were then used as a ground-truth to train and evaluate the algorithms that classify 'Relevant' and 'Non Relevant' tweets.

As a reminder, a tweet was manually labelled as 'Relevant' if the content of the tweet may lead to the conclusion that the person has the virus. In particular if, the person mentions:

having been tested positive

AND/OR

having at least 2 symptoms of the virus

AND/OR

feeling to be positive to the virus

3.2 CRAFTED TWEETS

Around 150 tweets were labelled as 'Non Relevant' in the three languages. Finding non-relevant tweets was relatively quick and easy. However, we found very few tweets that could be labelled as 'Relevant' in each language. To train a machine learning-based algorithm, a balanced number of 'Relevant' and 'Non Relevant' tweets is usually needed. Therefore, we had to craft positive tweets (e.g. manually write tweets considered as 'Relevant'). Overall, 130 tweets translated in the three languages were crafted. Finally, a balanced number of 150 tweets were fed to the algorithm to train a classifier for each language.

Below is a figure showing the exact quantity of hand-labelled tweets for each language.

Data	FR	DE	IT
Not relevant	152	153	152
Relevant	134	129	135
Total	286	282	287

Figure 8. Number of tweets manually labelled.

3.3 CLASSIFIER ACCURACY: 1ST ITERATION

Using the provided tweets, the algorithm splits the dataset into two parts: 80% of tweets were used for the training process (training set) and 20% of tweets were kept apart for the testing process (test set). After the training, the algorithm had to predict the labels associated to each tweet of the test set. The performance of the algorithm was evaluated using the accuracy metric (ratio of correct number of predictions in %). Below is the classifier's accuracy after both training and testing procedure for each language. Note that the test accuracy is the most important element to consider. However, we must pay attention that this accuracy only reflects performance of the algorithm on the test set and cannot be, in the current context, generalized to all tweets as

it only considers a small subset of tweets; it would require several thousands of manually labelled tweets in order to obtain a representative accuracy.

French classifier Details of the current active model		German classifier Details of the current active model		Italian classifier Details of the current active model	
Type	Result	Type	Result	Type	Result
Model name	model-fr-0001	Model name	model-de-0001	Model name	model-it-0001
Training/test proportion	80.0% / 20.0%	Training/test proportion	80.0% / 20.0%	Training/test proportion	80.0% / 20.0%
Relevant/Not relevant samples	134 / 134	Relevant/Not relevant samples	129 / 129	Relevant/Not relevant samples	135 / 135
Total samples quantity	268	Total samples quantity	258	Total samples quantity	270
Training accuracy	99.5%	Training accuracy	100.0%	Training accuracy	100.0%
Test accuracy	90.7%	Test accuracy	98.1%	Test accuracy	79.60000000000001%
Creation time	31/08/20 15:44:27	Creation time	31/08/20 15:44:27	Creation time	31/08/20 15:44:27

Figure 9. Results of the training of the models.

3.4 CLASSIFIER ACCURACY: 2ND ITERATION FOR THE FRENCH CLASSIFIER

When manually inspecting the tweets classified as 'Relevant', we noticed numerous false positive; therefore, we tried improving/augmenting the training data used by the classification algorithm. Using the labelling tab of the dashboard (see Figure 10), we identified manually the tweets falsely classified as <Relevant> (e.g. tweets classified as 'Relevant' automatically by the algorithm but which actually did not match the criteria that define a tweet as 'Relevant'). By pressing the small <trash> icon on the right of the tweet, the corresponding tweet was automatically added to the manually labelled tweets as 'Non Relevant'.

Figure 10. Tab of dashboard used to detect false positive tweets

The Figure 11 shows the updated number of tweets manually labelled after relabelling around 100 tweets written in French on the dashboard.

Data	FR	DE	IT
Not relevant	255	153	152
Relevant	134	129	135
Total	389	282	287

Figure 11. Updated number of tweets manually labelled.

Then, we retrained the French classifier to generate a new classification model (model-fr-0002). Below are the model's results after the second training process.

French classifier	
Details of the current active model	
Type	Result
Model name	model-fr-0002
Training/test proportion	80.0% / 20.0%
Relevant/Not relevant samples	134 / 134
Total samples quantity	268
Training accuracy	99.5%
Test accuracy	81.5%
Creation time	22/09/20 13:05:44

Figure 12. Result of the training of the French model with the addition of the False Positive correctly labelled.

We can observe that the Test accuracy decreased by ~10% compared to the previous model. This is a side-effect of the 'Non-Relevant' tweets that were added to the tweets used by the algorithm to learn. The algorithm indeed learns from those tweets and then evaluate its accuracy. As more tweets, that could be considered as ambiguous, were added to the training and testing sets, the algorithm had more chance to make an error. However, when classifying the real collected tweets, the algorithm is probably more accurate in detecting True Positives. Simply speaking, we may say that the tweets used to measure the accuracy were more difficult to classify correctly in the second model hence a decrease in reported accuracy.

Note that this is not easy to explain in a brief paragraph to individuals not familiar with machine learning concepts. What we can say is that we manually observed an improvement in classification with the model-fr-0002 through the results provided by the platform, although there are still too many false positives present with the new model.

4 RELEVANT TWEETS CLASSIFIED BY THE ALGORITHM

4.1 NUMBER OF RELEVANT TWEETS IN SWITZERLAND

For each day, this chart provides the **number** of tweets **localized in Switzerland** and classified as **'Relevant'** by the platform. The machine learning-based algorithm implemented in the platform classified automatically all the tweets collected.

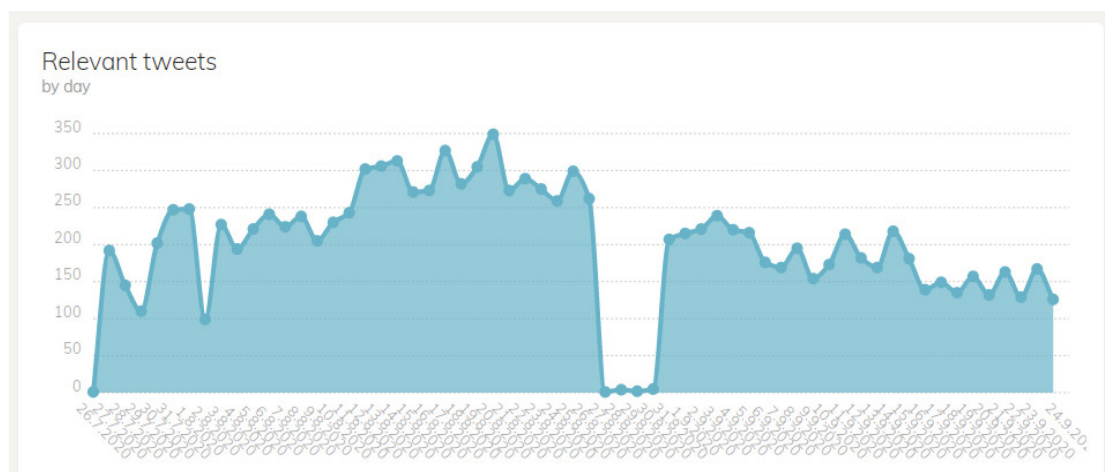


Figure 13. Number of tweets classified as <Relevant> per day.

4.2 MOST USED KEYWORDS IN ALL RELEVANT TWEETS

Below, you will find the **10 most used keywords** in all the collected tweets classified as **'Relevant'** by the platform. It shows the average amount of hits per day and the total number of use for each keyword.

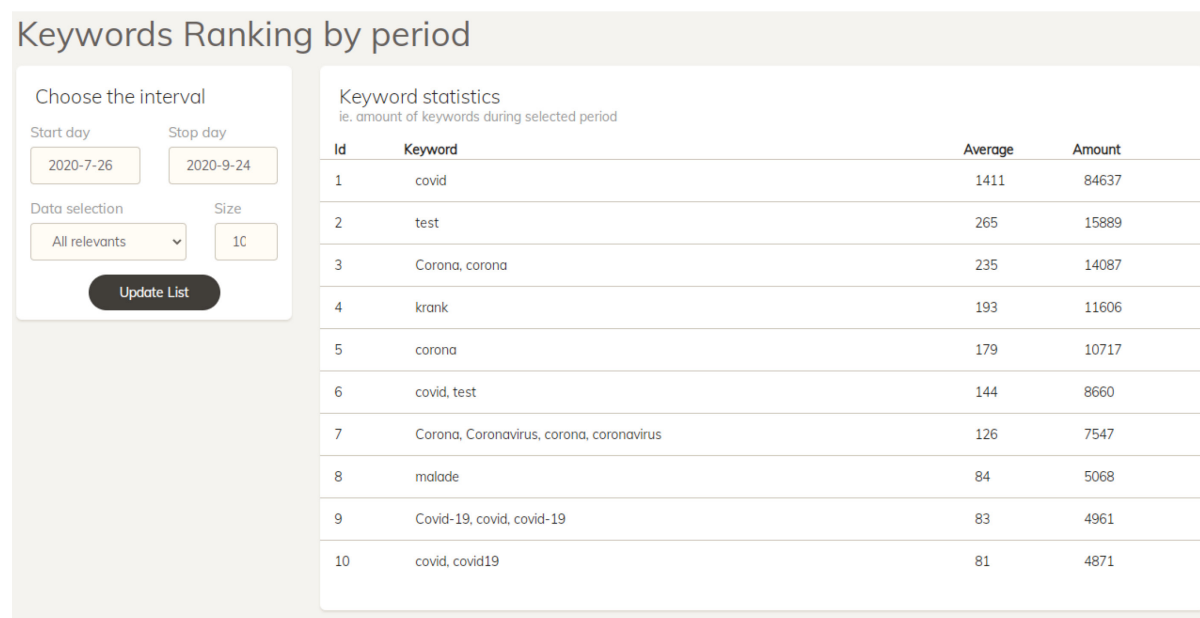


Figure 14. Most used keywords in all tweets classified as 'Relevant'.

4.3 MOST USED KEYWORDS IN SWISS 'RELEVANT' TWEETS

Below, you will find the **10 most used keywords** in all the collected tweets classified as **'Relevant'** and **localized in Switzerland** by the platform. It shows the average amount of hits per day and the total number of use for each keyword.

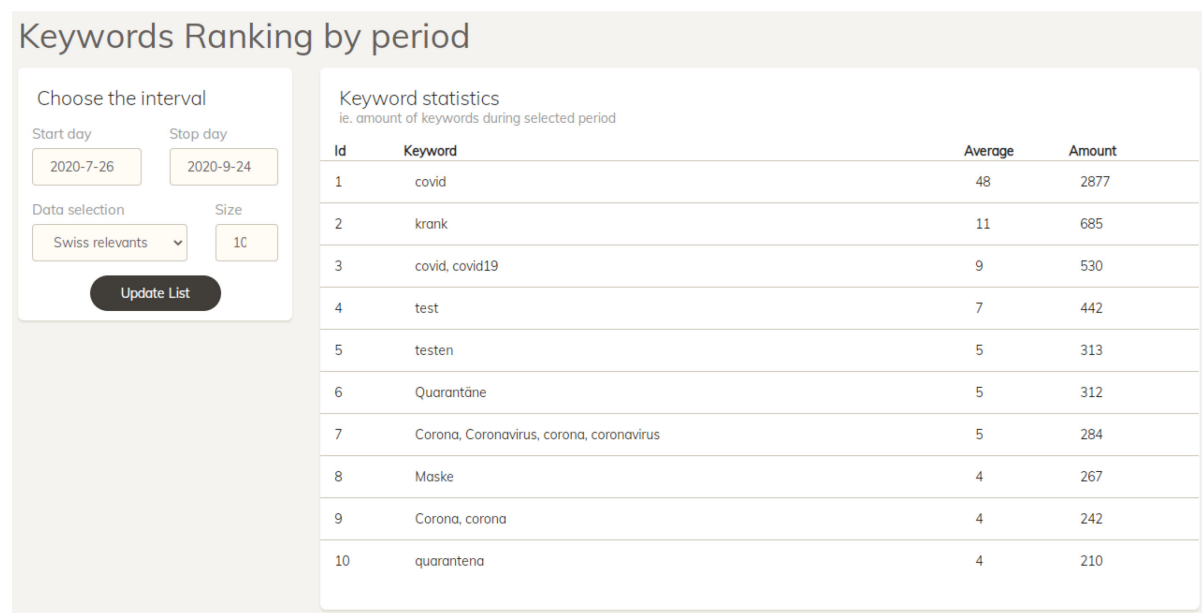


Figure 15. Most used keywords in tweets classified as <Relevant> and localized in Switzerland.

4.4 MAPS

Using the platform, we selected two specific periods (1-5.09 and 19-24.09) and plotted the relevant tweets localized in Switzerland on a Google Map for each period separately. The red markers represent country localized tweets, the yellow markers represent region localized tweets and green markers are city or geolocated localized tweets.

We hoped to be able to distinguish more precisely clusters in specific regions from these maps. Unfortunately, the tweets classified as False positives are probably creating too much noise to clearly highlight and distinguish clusters. Assuming we manage to improve the accuracy of the algorithms and reduce the false positive rate, we could probably identify regional clusters on such maps for different periods.

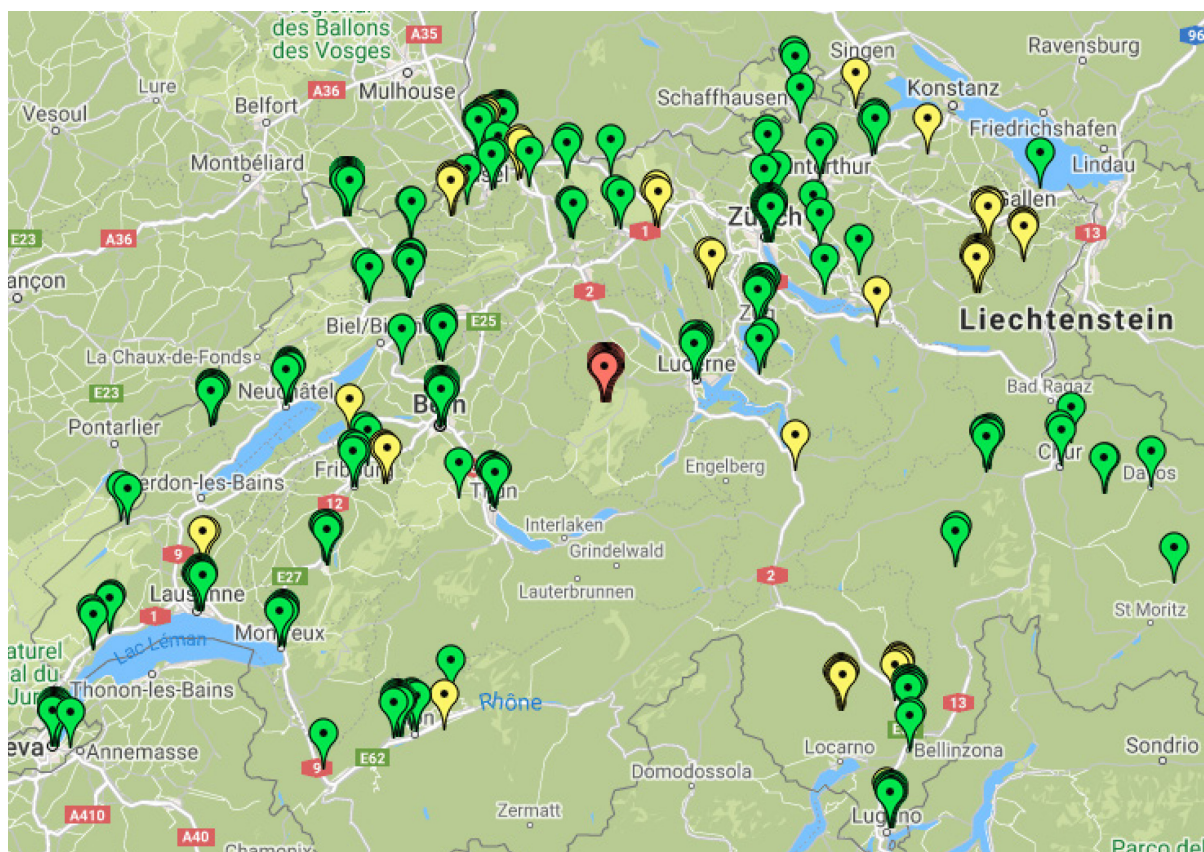


Figure 16. Map showing relevant tweets in Switzerland between September 1st and September 5th.

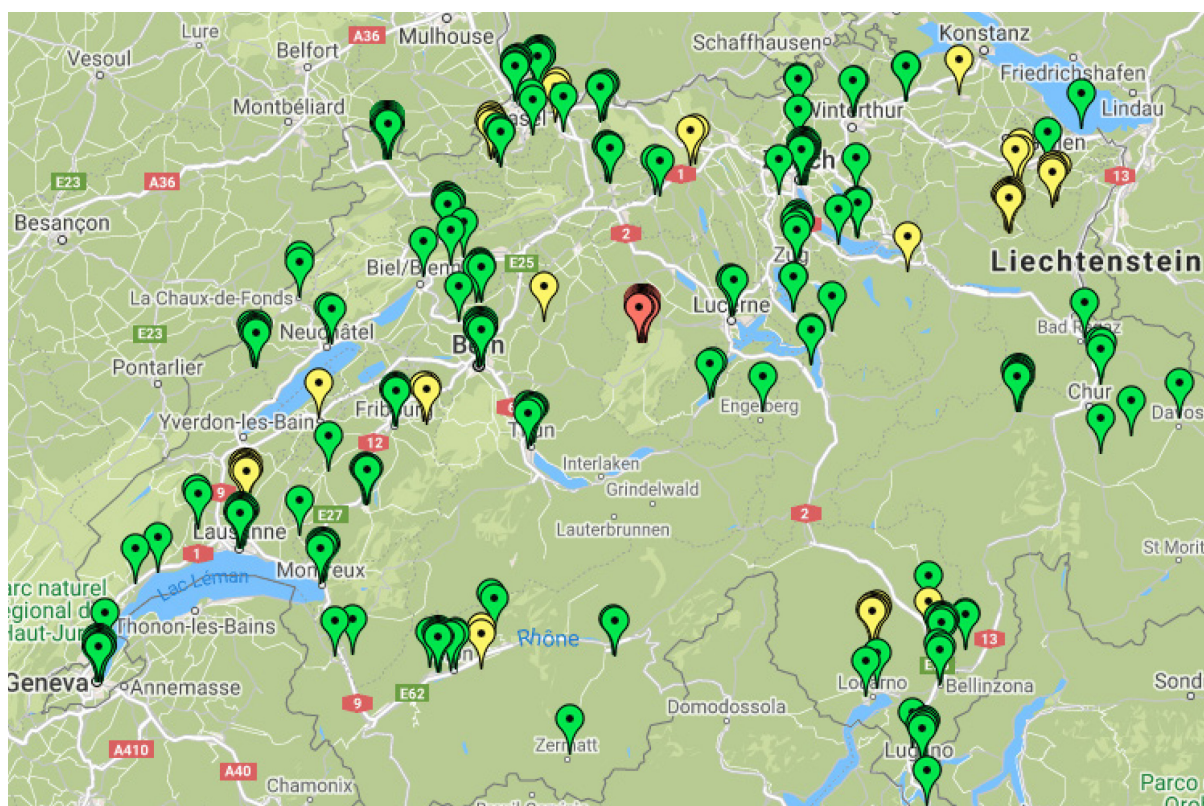


Figure 17. Map showing relevant tweets in Switzerland between September 19th and September 24th.

4.5 EXAMPLES OF TWEET CLASSIFIED AS 'RELEVANT' BY THE PLATFORM

In this section, we show sample tweets that were classified as 'Relevant' at different dates and that match the eligibility criteria defined at the beginning of the project. Figure 18 shows that different people reported being positive to the COVID-19 after a test or reported having some of the symptoms of the list defined by the FOPH. These tweets were posted at different dates in September.

61	24/9/2020	@Carlito__34	Le frère de mon collègue il a dit à son patron qu'il faisait le test covid pour pas allé travailler, il va faire le test : POSITIF. Méchant retour de flamme 😞	👁	✓	🗑
30	24/9/2020	@sibyllejouteau	Fin du TDF : 20 septembre Début de RG : 27 septembre Entre les deux : rien, absolument rien Moi 🍷 attraper le covid cette semaine là 😞	👁	✓	🗑
108	23/9/2020	@cjemelin	@gaelhurlimann Et la meilleure : résultat du test reçu ce jour : ni positif, ni négatif, il faut refaire le test 😞 //answer to : J'ai été testé positif au Covid-19. Je devais recevoir des codes à entrer dans l'app SwissCovid dans les deux heure... https://t.co/HngBovogKT	👁	✓	🗑
112	23/9/2020	@gaelhurlimann	J'ai été testé positif au Covid-19. Je devais recevoir des codes à entrer dans l'app SwissCovid dans les deux heure... https://t.co/HngBovogKT	👁	✓	🗑
131	23/9/2020	@rousselbenoit7	Positif au Covid putain la flemme	👁	✓	🗑
137	23/9/2020	@ALittlePoppy	Svp le médecin Twitter !! Ça fait plusieurs jours qjebjai des symptômes covid, pas encore de résultats au test mais bon.ajd la fièvre a baissé mais je suis essoufflée sa maman pour rien, dès le réveil, j'ai l'impression d'avoir couru. Ça veut dire c'est bientôt fini ?	👁	✓	🗑
28	22/9/2020	@saintechnik	J'ai encore un chouilla de fièvre mais quasi plus rien, je refuse de reprendre du Doliprane je verrai plus tard mais c vrrmt infâme ça me fait me sentir trop mal d'avaler ce truc là	👁	✓	🗑
76	21/9/2020	@laeticia_plt	Je dois refaire le test du COVID 🚫	👁	✓	🗑
85	21/9/2020	@2Fast4_U_	Bon, actuellement en suspicion Covid (fièvre de mort, toux, gêne respiratoire, migraine, courbatures etc et) Le seul truc que le compagnon de ma mère trouve à dire quand je me lève : "Ah bah t'es pas allé en cours toi !"	👁	✓	🗑
13	3/9/2020	@Lilibitina	@BentElGharb Moi la respiration ça va. Mais les autres symptômes je les aient et je dors comme quand je suis malade olalala Je suis sortie qu'une fois en 1 mois pour te dire 😞 //answer to : @Lilibitina On est ensemble mdr, des courbatures partout + du mal à respirer + fatiguer pour un rien, je commence à me poser des questions //answer to : Dites vous que ça fait un mois que je suis clouée au lit dû à un "rhume" mais je me demande si c'est vraiment un rhume que j'ai 😞	👁	✓	🗑
14	3/9/2020	@BentElGharb	@Lilibitina On est ensemble mdr, des courbatures partout + du mal à respirer + fatiguer pour un rien, je commence à me poser des questions //answer to : Dites vous que ça fait un mois que je suis clouée au lit dû à un "rhume" mais je me demande si c'est vraiment un rhume que j'ai 😞	👁	✓	🗑
28	3/9/2020	@NassimAkaNassim	Y'en a deux elles ont le Covid dans ma classe Insh'Allah ils nous mettent en quarantaine	👁	✓	🗑
5	16/9/2020	@vileans	devinez qui a été placé en quarantaine parce qu'elle va se faire tester au covid :)))	👁	✓	🗑
6	16/9/2020	@leiedio	@SegoleneGrini Oui ! J'avais peur d'être positive parce que je suis tombée très malade en l'espace de 24h, allez viens on trinque tchin ! 🍷😞	👁	✓	🗑
16	16/9/2020	@TinaT14497787	J'ai dû sur ordonnance faire cette semaine un test. Il s'est déroulé en extérieur avec un horaire de passage (RDV depuis une plate forme). Souci aujourd'hui on a des tests mais pénurie de révélateur ! Quelle organisation, toujours à la rue. #Punchline	👁	✓	🗑

Figure 18. Various tweets in French classified as Relevant by the platform

Figure 19 shows that we can target the relevant tweets that match the eligibility criteria among other tweets that can be considered as 'False Positive' (classified as 'Relevant' by the algorithm but not matching the criteria to be considered as 'Relevant'.

The last relevant tweets
ie. COVID positive cases

Id	Date	Author	Text	Link	CH	False?
1	16/9/2020	@mkaff_	Rien que la samedi je maquette trois feux dans le 16e pr un mariage ça va me prendre trop de temps à cause du Covid j'ai grv grv la flemme			
2	16/9/2020	@Sir_Senku	@alishirogaya En vrai on sait pas, c'est peut être le Vandenreich qui a créé le corona 🤔🤔			
3	16/9/2020	@ReDeZoK3	@sanscomics @MacS1ms @PlayStationFR Rien avoir avec le Covid-19 il son toujours sortir en avance ailleurs que en Europe regarde la PS4 //answer to : @MacS1ms @PlayStationFR T'as entendu parler du COVID ?			
4	16/9/2020	@phg_17	Covid, blessures, suspensions, tout est contre nous en ce moment... Mais on lâche rien, jusqu'au bout !!! #PSGFCM #AllezParis			
5	16/9/2020	@vileans	devinez qui a été placé en quarantaine parce qu'elle va se faire tester au covid :)))			
6	16/9/2020	@leiedio	@SegoleneGrini Oui ! J'avais peur d'être positive parce que je suis tombée très malade en l'espace de 24h, allez viens on trinque tchin ! 🍷🤔			
7	16/9/2020	@oximozyt	Aujourd'hui -Le dentiste m'a endormi la moitié du visage tellement son truc était puissant, je suis allé faire des courses, fait 1 live, 3 vidéos le tout en étant malade (et je dois rec encore 1 vidéo et tout monter) je suis KO			
8	16/9/2020	@selim_sivad	Pour moi un rhume ou le corona c'est la même ! Quand je suis malade, je suis bon à rien, anéanti et proche du décès ! Pensée à ceux qui combattent vraiment la maladie... avec un courage que je n'aurai jamais.			
9	16/9/2020	@bsmssdt	@yxns91 W'ALLAH TES UN MALADE JAI RIEN DIS SALE FC 1er degrés allez ntm			
10	16/9/2020	@Cellolmonde	@TheWamuu @RamzaBehoulve @BAG_OFSP_UFSP Mais cet hiver on risque d'avoir une multitude d'autres urgences qui vont se mêler au covid, et devoir peut-être faire des choix difficiles.			
11	16/9/2020	@lmgvcy	je suis malade mais si je sèche demain ils vont croire que c psq j'ai la flemme de venir			
12	16/9/2020	@loic_charrier	@BeninouOM A un quiproquo, à toujours une solution , dis que ton collègue est rendu du au covid, ses enfants allé pas bien , à toujours une excuse			
13	16/9/2020	@ActuSuisse	Fauchon ferme 2 de ses 3 magasins à Paris, plombé par le Covid et les gilets jaunes https://t.co/6azaMaSeCf			
14	16/9/2020	@chapi_chapot	Pas d'autre personnalité politique qui doit se faire tester pour suspicion de Covid ? Ben, c'est quoi cette #moitié de #manipulation pour faire croire à une recrudescence du Covid ? Même dans la manipulation, ils sont mauvais... 🤔🤔🤔			
15	16/9/2020	@GelTank	Tek tank gel est un distributeur de gel hydro-alcoolique. Il a été créé par de jeune entrepreneur dans le cadre du covid-19 https://t.co/Bvg5o1qptx			

Figure 19. A pair of True Positive <Relevant> tweets amongst <False Positive> tweets.

4.6 ESTIMATION OF CLASSIFICATION

The table below provides an estimation of the tweets collected over the period of 60 days considered in this document. The table provides a summary of the total number of tweets collected by the platform, the number of tweets localized in Switzerland, the total number of swiss tweets classified as 'Relevant' and finally our rough estimation of the total number of tweets that should be considered as 'Relevant'. It shows that less than 0.2% percent of the swiss tweets are relevant in reality. The algorithm successfully filters about 85% of the 'Non Relevant' swiss tweets but still misses ~14% of 'Non Relevant' tweets

Table 1. Summary of tweets collected and processed by the platform.

Total tweets collected	2'800'000
Tweets localized in Switzerland	73'734
Swiss tweets classified as Relevant	12'355
Estimation of Swiss tweets really <Relevant>	~1000

5 CONCLUSION

The current platform provides interesting insights on the mentions of the COVID'19 on Twitter social network and many opportunities to analyse the data. Trends and tendency can be observed over different periods of time, most referenced keywords can be identified and quantified, the localisation of tweets can be displayed on a map, etc.

The current algorithm used to classify 'Relevant' and 'Non Relevant' samples is probably not powerful enough to identify correctly only the True Positives (= true 'Relevant') due to the noise, heterogeneity of the tweets as well as the quantity of data it has to process. It is however important to note that the current models already successfully filter many of the collected tweets; it successfully rejects about 85% of the 'Non Relevant' tweets; although probably ~14% are still not correctly rejected. In order to better understand those accuracies, we performed a qualitative analysis by reading all tweets classified as 'Relevant' for a few selected days and manually identified the ones that matched our initial criteria. Out of the 200 swiss tweets daily classified as 'Relevant', about 5 to 15 of tweets could really be considered as 'Relevant'. This indicates that we are probably looking for ~5-20 'Relevant' tweets out the 1250 collected daily from swiss individuals.

In order to improve the accuracy in identifying on the True Positives Relevant' tweets, we plan, in the next steps, to investigate more advanced algorithms for natural language processing (NLP). In order to integrate them easily in the current platform, we need to update the platform and notably enable the possibility to choose from different algorithms.

Finally, the collected data (Swiss tweets mentioning COVID-related keywords), which are being stored on our servers, may be used in many different ways in future projects to perform a posteriori analysis. We notably plan to investigate the timely evolution of the sentiments of the population on the COVID by analysing the valence of the collected tweets. We plan to perform such investigations through student projects in the future

6 CONTACT

If you have any question or remark regarding the data presented in this report, please feel free to contact one member of the development team using one of the addresses below:

- Simon Ruffieux, project coordinator: simon.ruffieux@hefr.ch
- Quentin Meteier, collaborator: quentin.meteier@hefr.ch
- Omar Abou Khaled, professor: omar.aboukhaled@hefr.ch