PROJ0016-1

BIG DATA PROJECT

Bertrand Cornélusse
Pierre Geurts
Gilles Louppe

Final review

Presentation

Julien Hubar
Pierre Dumoulin
Andreas Duquenne
François Lievens

Master 1 Data science Engineering
Faculty of Applied Sciences

Academic years 2020-2021

Introduction

Today, social networks represent a considerable source of information for assessing, in real time, the impact of various major events and news on the population.

This source of information can be used for many applications:

- In politics, to assess public opinion on different issues
- In the commercial field to assess customer needs or feedback on a newly launched product
- In economics, to predict stock market trends
- And many more...

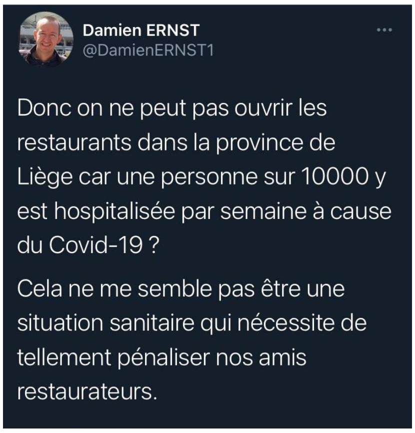
In the context of epidemic management such as for COVID-19, the analysis of data provided by social networks seems to us to be an excellent tool that it can be essential to master. Indeed, this implies the taking of strong sanitary measures, which inexorably lead to an infringement of the individual liberties of the population.

Depending on the form that these measures may take and the way they are presented, it may be difficult to maintain the cooperation of a population that has little knowledge of epidemiology. Social networks can then become the barometer of this fragile cooperation, which it is essential to preserve.

In this project, we will develop a tool to assess the overall impact of the different key announcements concerning the COVID-19 epidemic on the French speaking Belgian population. This project is divided into several phases:

- The creation of a model to solve the subjective problem of classifying the mood of tweets.
- The retrieval and processing of a large amount of data for each day during the epidemic. We chose to scrape tweets quite randomly, without selecting themes or keywords.
- A data analysis phase where we will put together the information we have extracted from the social networks, as well as the scientific data of the covid-19 epidemic, and dates of key announcements concerning the COVID-19 measures in Francophony Belgium

Positivity = 0.23


Damien ERNST
@DamienERNST1

Donc on ne peut pas ouvrir les restaurants dans la province de Liège car une personne sur 10000 y est hospitalisée par semaine à cause du Covid-19 ?

Cela ne me semble pas être une situation sanitaire qui nécessite de tellement pénaliser nos amis restaurateurs.

1 Feelings evaluation

The first part of this project will therefore consist in the implementation of a deep learning model allowing to solve the subjective task of classifying tweets according to their positivity mood. Our work was carried out in two stages which will be detailed in the next parts of this report:

- Identify the most appropriate architecture to solve our classification problem and allow us to obtain the best possible classification performance. Beyond the binary annotation problem, and given the high degree of subjectivity of the task, we plan to use, instead of classes annotation (positive/negative) a "positivity rate", who vary from 0 to 1, and representing the degree of certainty that a tweet is positive.
- Once our architecture is chosen, we need to find a learning set to train it and to evaluate performances of our model.

1.1 Model's architecture: Transformers

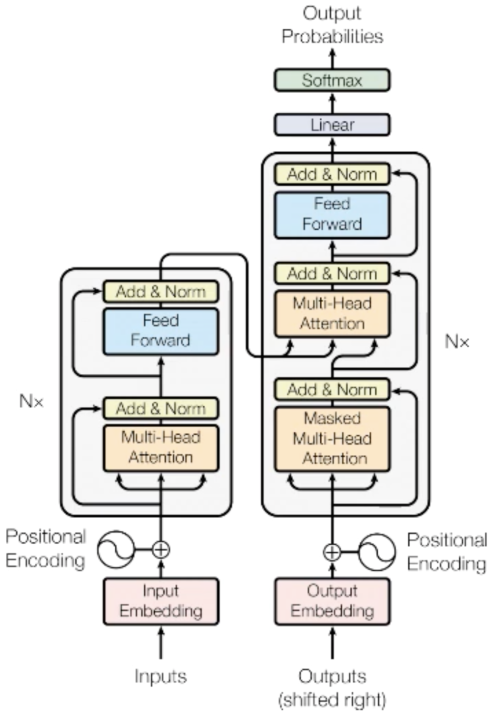


FIGURE 1: General transformer's architecture. [1]

Historically, Natural Language Processing (NLP) task are are performed by recurrent neural networks (RNN). These networks are specialized in sequential data analysis and will read these sequences in the indicated direction. So, in our NLP task, a tweet can be represented as a succession of word's vector representation. Chronologically, the neural network will analyse word's vectors and updating his internal state to, at the end, obtain a final state that we can use for our classification task.

However, this approach has many problems and limitations:

- The words are read in a specific direction (classically: from left to right). In some cases, words used earlier in a text can have a very different importance depending on the words present later in the sequence.
- The network state is fixed in size and cannot be expanded. Thus, as the length of the sequence increases, this state becomes a bottleneck, and the information read at the beginning is gradually forgotten in favor of more recent entries.
- In terms of computation time, the sequential reading of the inputs also represents a strong limitation because it does not allow a strong operation's parallelization in order to exploit the computation power provided by the GPU's.

Some major improvements have been made by the Long Short-Term Memory cells (LSTM) and bi-directional LSTM in order to obtain better performances on longer sentences and a better sentence understanding by reading sentences in both directions. But despite these improvements, bottleneck problems are only partially solved, and it is still not possible to make strong parallelizations.

Since Parmar et al. [1] in 2017, Transformer's architecture (1) have revolutionized the NLP by far superior performances. This architecture is mainly composed of a succession of attention mechanisms (implemented by simple matrix operations), and simple feed forward layers, allowing the encoder part to read all the given sentence in a parallel way. By this way, the encoder can build a representation of the sentence who extract the main informations about the which extracts the information about the content, allowing to modelize the meaning.

1.2 BERT, RoBERTa, and pre-trained transformers

Building a transformer on NLP's complex tasks requires a long and heavy training on high complexity models. Often, the data needed to train the model on the desired task are in very small quantities, and the understanding of the sentences to be classified requires the model to store a much more global and general representation of the concepts conveyed by the language used.

Thus, to create this necessary representation of the concepts conveyed by a language, as well as the links between them, Google published in 2018 [2] the Bidirectional Encoder Representation from Transformers (BERT) architecture. BERT consists of an heavy Transformer archtitecture trained in a self-supervised way on very large amounts of text. This training consists of simple tasks such as the prediction of hidden parts of texts.

After this heavy pre-training phase, we can then recover the weights from the trained models, removing theses from the decoder part, and adding our own new layers who will be specilized on our classification task.

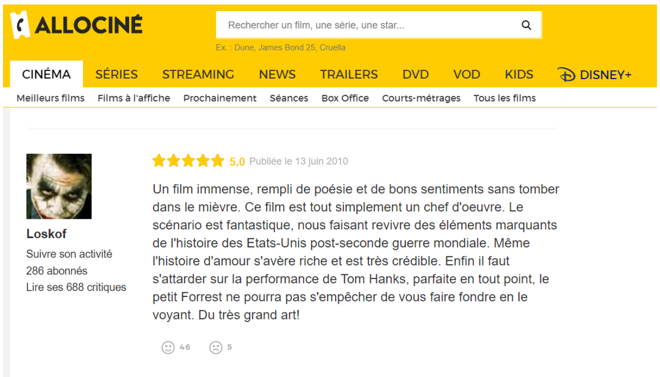
The model we have chosen is based on the RoBERTa architecture. This architecture is an improvement of BERT resulting from Facebook AI research. Since we want to annotate exclusively French tweets, we use CamemBERT [3], a version of RoBERTa pre-trained on French text. This model can be download on the https://camembert-model.fr/ website.

1.3 Training Set

Despite using a pre-trained model, we need a training set to train the layers that we add to the model in order to classify the positivity of our tweets. Unfortunately, there is no available annotated french tweets dataset that we can use for our training. Since we do not have the time to manually annotate a dataset representative of the data we are going to work on, we need to find a public training set with the following characteristics

- To be composed of word sequences of a length comparable to 280 character tweets
- To be in French, consisting of sequences of any language level and including common grammatical and spelling errors.
- Must cover as many topics as possible as we do not select tweets by subject.
- To be balanced between positive and negative moods.

The best solution we found was to use the website AlloCine.fr. This french plateform allow a large number of user to write their own movie reviews and to provide with it a note between zero and five stars. We can so use the text as input in our training set and the note to evaluate if the text is positive or not.



Loskof

Suivre son activité

286 abonnés

Lire ses 688 critiques

★★★★★ 5,0 Publiée le 13 juin 2010

Un film immense, rempli de poésie et de bons sentiments sans tomber dans le mièvre. Ce film est tout simplement un chef d'oeuvre. Le scénario est fantastique, nous faisant revivre des éléments marquants de l'histoire des Etats-Unis post-seconde guerre mondiale. Même l'histoire d'amour s'avère riche et est très crédible. Enfin il faut s'attarder sur la performance de Tom Hanks, parfaite en tout point, le petit Forrest ne pourra pas s'empêcher de vous faire fondre en le voyant. Du très grand art!

FIGURE 2: https://www.allocine.fr/

This dataset therefore allows us to fulfil our objectives concerning the language level and the size of the sequences. However, it has the major drawback that film criticism is still a highly focused topic.

1.4 Training

With a training set of 160.000 movies reviews, we can now launch the training phase of our model. This training consists of fine tuning the new layers after downloading the pre-trained model. The characteristics of our training are as follows:

- Part of the data use like training set: 80%
- Optimizer: Adam
- Loss function: Cross entropy over the two classes classification problem.
- Learning rate: $5 * 10^{-7}$
- Batch size: 15 AlloCine sequences
- Duration: 3 epochs

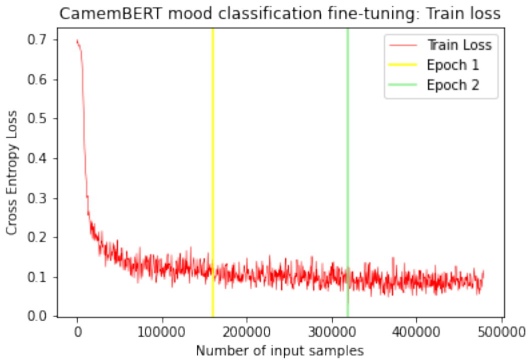


FIGURE 3: CamemBERT fine tuning: Train Loss over 3 epochs

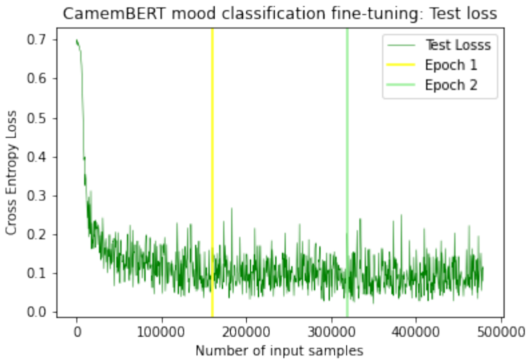


FIGURE 4: CamemBERT fine tuning: Test Loss over 3 epochs

The tracking of our training step are show in the figure 3 for the training loss and in the figire 4 for the testing loss. Each training loss value who is plot represent the average loss over 50 batchs of 15 sequences, and the each testing loss who are show are the average over predictions performed over 10 batchs of 15 sentences from our test set.

During this training process, we had to limit our batches to a size of 15 elements, which was already enough to saturate the 8gb of our GPU's memory. The three epochs were carried out over a period of 8 hours. We have seen that the fine tuning of this transformer is a task that allows high performance to be achieved very quickly, and from little data, as the learning curve shows. However, to get more accurate predictions from our small batches, we should not hesitate to use very low learning rates and let the algorithm run much longer.

2 Collecting data

2.1 Twitter

In order to analyse the sentiment of an important part of the Belgian population we have decided to Work with the social network Twitter. Several choice were open to us.

Firstly, as said, we chose to focus our attention on Belgium, and more specifically on the french community. This choice allows an easier implementation of the predictive model, which would only have to deal with one language. Moreover, the measures taken can differ from the north of the country, and because we are not able to recover as many tweets as we want, reducing the number of people targeted for a given number of tweets will result in a more significant and representative sample.

Secondly, we decided to analyse the general mood evolution of the population. Its means that our implementation aim to recover tweets on all subjects. Not only concerning the covid epidemic. This choice is purely subjective, and we believe that an analyse of the tweets only concerning the epidemic could be very interesting as well. However, because we are limited in the number of tweets we can extract, we were not able to further investigate this way.

2.1.1 Extraction

Fortunately, Twitter comes with an API which is, however, limited in terms of request. Using the REQUEST package from python, as well as a CRON tasks on a Ubuntu server, one can forge http request to recover tweets.

2.1.2 Dataset

At the end of our request budget we have been able to compose a dataset which should be representative of the french Belgian population during the period. Here are some important points of the dataset:

1. Composed of 7 millions tweets

2. Only composed of tweets (no retweets), in french, from Belgium

3. Number of tweets approximately uniformly distributed over the 2020 year (1/1/2020 to 1/1/2021)

3. Can contain special characters, as emojis

2.2 Impacting events

Impacting events related to the COVID-19 pandemic will be used to search after correlation between these events and the general mood of a population(ie. French speakers in Belgium).

We classified impacting events into two categories:

- **Announcement**: Announcements about measure related to COVID-19 that could have had an impact on the feelings of the population.
- **Measure**: Measures related to the COVID-19 that could have had an impact on the feelings of the population.

Note that both announcements and measures can have positive or negative impacts.
Our data have been collected in the following website:

- Wikipedia: Belgium pandemic [4]
- Crisis center in Belgium [5]

2.3 Reproductive rate

The reproductive rate of the COVID-19 virus can be considerate as the average number of infections cases generated by an individual infection. So, a reproductive rate greater than 1 is associated with an expanding epidemic, and an epidemic in recession if it is less than 1.
Our data have been collected in the following website :

- Statista for coronavirus reproductive rate in Belgium [6]

We are able to rely on this source because it obtained its data via "SPF-santé publique/Sciensano" which is the referent site for the COVID-19 in Belgium.

3 Statistical Aspect

In order to determine whether variations in the average daily positivity curve of the tweets are significant or not, we are using the Student T-test. Indeed, this test allows us to test, according to the desired statistical threshold (we have chosen p=0.05), whether two distributions following a normal distribution differ significantly. We can thus analyse the tweet's positivity distributions of the two days that we want to compared using the following formula:

$$t = \frac{\bar{X}_a - \bar{X}_b}{\sqrt{\frac{S^2}{n_a} + \frac{S^2}{n_b}}}$$

Where S^2 is the covariance of the the two positivity distributions, $\bar{x_a}$ and $\bar{X}_b$ are the mean of the two distributions and n_a and n_b are the number of tweets in theses distributions.

4 Data Analysis

In this section, we will try to link the data from the social networks, the main events and announcements that occurred during the period analysed, and the rate of virus reproduction in the region.

The goal of this part of the project is therefore, on one hand, to evaluate whether the general mood measured by our model on social networks is really influenced by the epidemic and the health measures, and on the other hand, to carry out a critical retrospective study on the measures that were taken and the real effect that they had on the epidemic. This critical step seems to us to be important in order to link the parameters that we used during the first part of the project (effect of health measures on epidemic simulations) with the reality and how to improve her values.

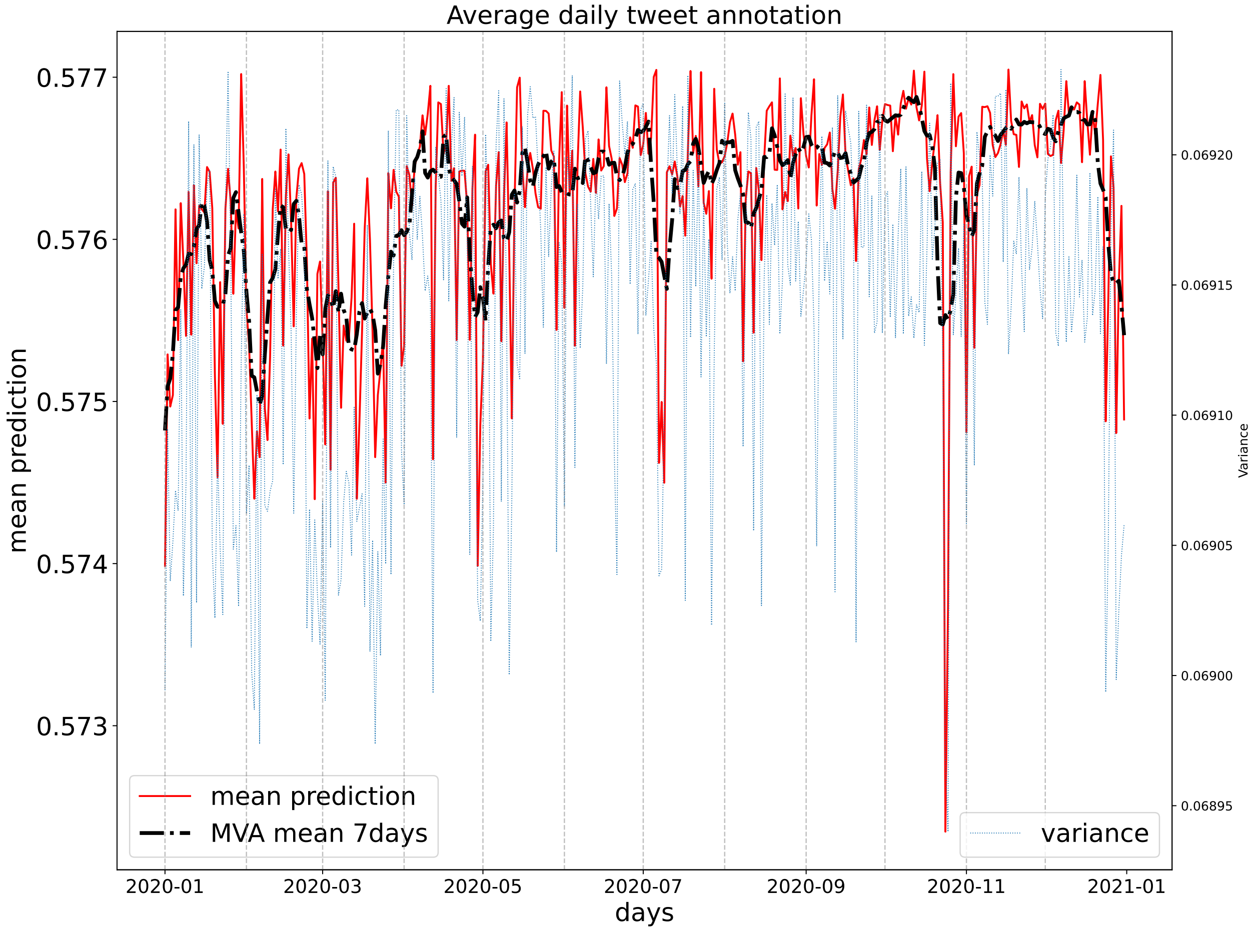


Figure 5: Tweets positivity measurements

First, we can show in the figure 5 the very first result of our tweets annotations are show. This graph present the evolution of the daily "positivity rate" of tweets. Our CamemBERT fine-tuned model output a Softmax activation for our two annotation classes (positive and negative mood). So, we are considering the probability associated to the positive mood class like a "positivity rate", and who is in fact the degree of certainty of the model that the tweet is positive.

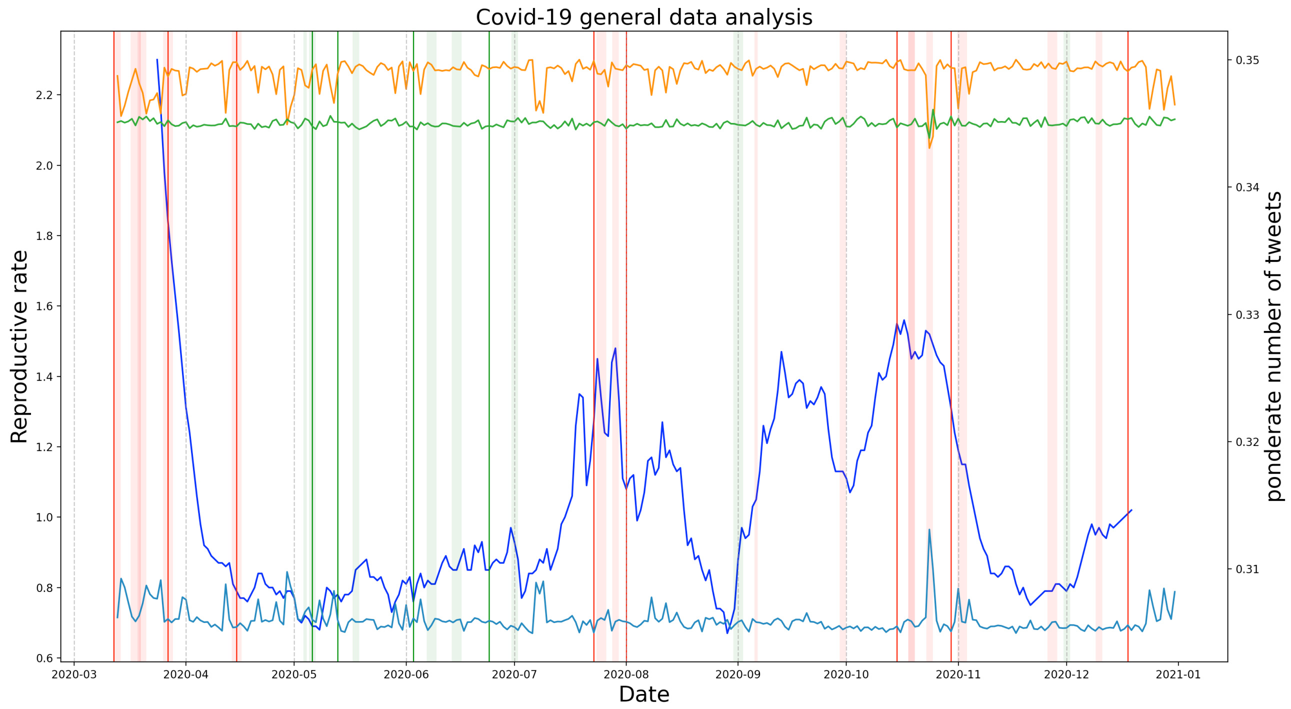


Figure 6: Legend:
●: viral reproductive rate
●: relief measure / announcement
●: restrictive measure / announcement
●: Proportion of positive tweets
●: Proportion of negative tweets
●: Proportion of neutral tweets

In the figure 6, we are presenting the following information's:

- The reproductive rate of the COVID-10 virus in Belgium are from SPF-santé publique/Sciensano.
- For the curves who are showing data from our mood analysis model, we have decide to spare it in three curves: The "Positive", "Neurtral" and "negative" curve. Each represents the proportion of tweets, whose positivity rates are respectively upper than 0.75 for positive, lower than 0.25 for negative and between are neutrals.
- On this graph, we have drawing vertical lines who are following this rules:
  - The thin bars are associated with the announcement of health measures during national safety councils.
  - The wide bars with lower opacity correspond to the date at which the measures are applied.
  - The green color is associated with relaxation of measures, while the red colour is associated with restrictive measures.
  - The width of the lines is an assessment of the extent of the measures taken.

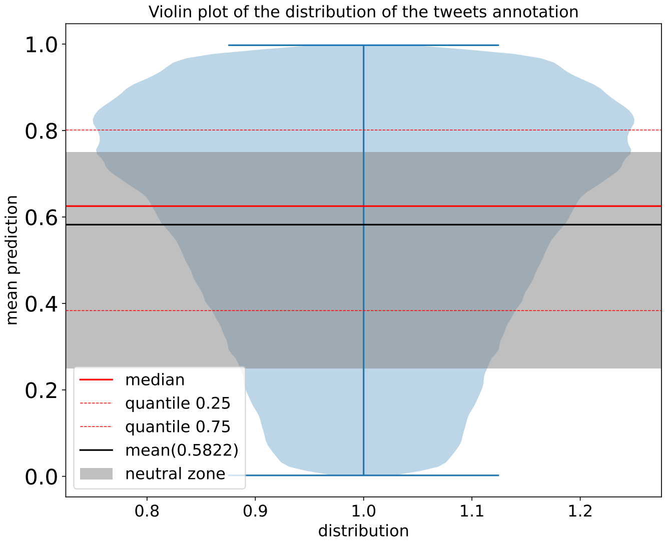


FIGURE 7: Tweets positivity distribution

If we look at the distribution of annotations of our tweets, we can see that the distribution of the positivity of the tweets is globally unbalanced to the profile of the negative tweets, with an average positivity around 0.6 for the period studied.

In the next part of this report, we will analyze more precisely data by smaller time steps.

4.1 Data from 12/03 to 01/05

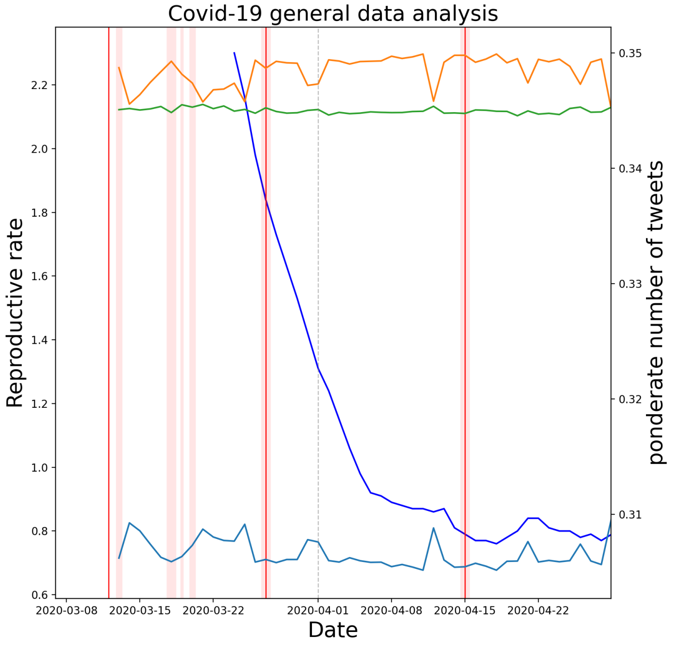


FIGURE 8: Data from 12/03 to 01/05

Epidemic's event during this time step (fig. 8 are the following:

- 12/03/20: [ANNOUNCEMENT] Closure of schools, discos, cafes and restaurants and the cancellation of all public gatherings for sports, cultural or festive purposes
- 13/03/20: [MEASURE] Putting into practice of previous measures
- 18/03/20: [ANNOUNCEMENT] Lock-Down
- 19/03/20: [MEASURE] Limitation to the Belgian coast
- 20/03/20: [MEASURE] Closure of the borders except for freight and the return from abroad
- 27/03/20: [ANNOUNCEMENT] Lockdown extension
- 27/03/20: [MEASURE] Lockdown extension until April 19
- 15/04/20: [ANNOUNCEMENT] Measure extension(including lockdown)
- 15/04/20: [MEASURE] Extension of the measures until May 3

We can see in the figure a drastic decrease in the viral reproduction rate. This decrease can most probably be explained by the strong measures taken at the beginning of this period.

In this early period of the epidemic, we can see a fairly high variance in the daily average positivity of tweets. These daily variations can be considered statistically significant, according to Student's T-test. However, we could not link them to specific health announcements in Belgium. However, this period has been busy with various international news items concerning COVID-19.

4.2 From 01/05 to 21/07/20

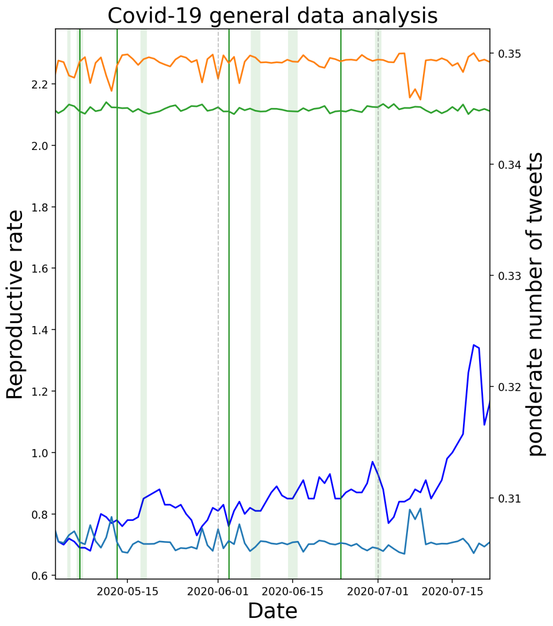


FIGURE 9: Data from 01/05 to 21/07

Epidemic's event during this time step (fig. 8 are the following:

- 04/05/20: [MEASURE] Progressive lockdown lifting (Phase 1A)

- 06/05/20: [ANNOUNCEMENT] Progressive lockdown lifting (Phase 1B)
- 06/05/20: [MEASURE] Progressive lockdown lifting (Phase 1B)
- 13/05/20: [ANNOUNCEMENT] Progressive lockdown lifting (Phase 2)
- 18/05/20: [MEASURE] Progressive lockdown lifting (Phase 2)
- 03/06/20: [ANNOUNCEMENT] Progressive lockdown lifting (Phase 3)
- 08/06/20: [MEASURE] Progressive lockdown lifting (Phase 3)
- 15/06/20: [MEASURE] Reopening of the borders to EU
- 24/06/20: [ANNOUNCEMENT] Progressive lockdown lifting (Phase 4) + second wave virus risk announcement
- 01/07/20: [MEASURE] Progressive lockdown lifting (Phase 4)

During this period, it was mainly release measures that were taken. Unsurprisingly, after the simulations we carried out during the first quarter, these measures were accompanied by a re-increase in the reproduction rate of the virus.

There seems to be no relation between the peak of negativity around 7 July and the covid measurements. Even if this is not trivial to infer the consequences of this peak, we assumed that the COVID-19 is not responsible of this mood variation. Regarding to the tweet trends it seems that it should be related to an intense football match between Juventus and AC Milan.
(Or perhaps our algorithm trained on film reviews was affected by the death of Ennio Morricone...)

4.3 From 17/07 to 01/10/20

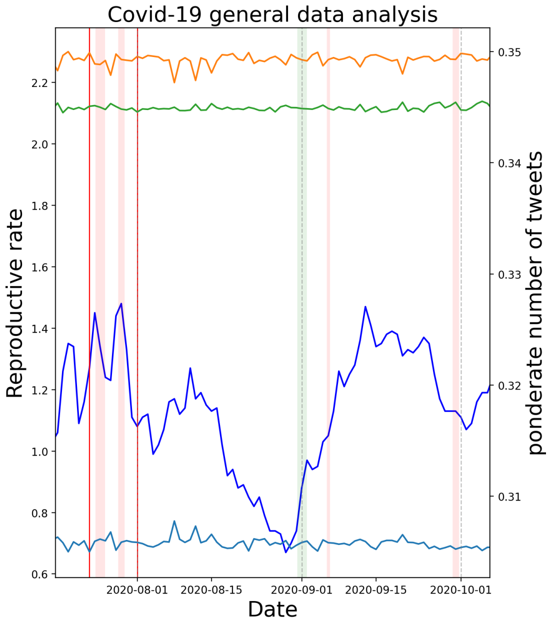


FIGURE 10: Data from 17/07 to 10/01

Epidemic's event during this time step (fig. 10 are the following:

- 23/07/20: [ANNOUNCEMENT] Progressive lockdown lifting cancelled (Phase 5)
- 25/07/20: [MEASURE] Mask wearing becomes mandatory on shoping street, market and flea market
- 29/07/20: [MEASURE] Reduction of the social bubble size
- 01/09/20: [MEASURE] Code yellow and resumption of lessons for all students regardless of their ages + relaxation of certain measures
- 06/09/20: [MEASURE] Restriction of the social contact and closure of the bars at 11PM
- 30/09/20: [MEASURE] Numerical tracing of the exposed people

A very interesting aspect of this period is the introduction of masks wearing and their positive effect on the viral reproduction rate, which we had already observed during our simulations in the first phase of the project.

The second big change during this period is the start of the new school year. This is accompanied by a sharp increase in the viral reproduction rate, before it stabilises. This may mean that the application of sanitary measures inside schools required a period of adaptation in order to be fully effective.

4.4 From 10/10 to 31/12/20

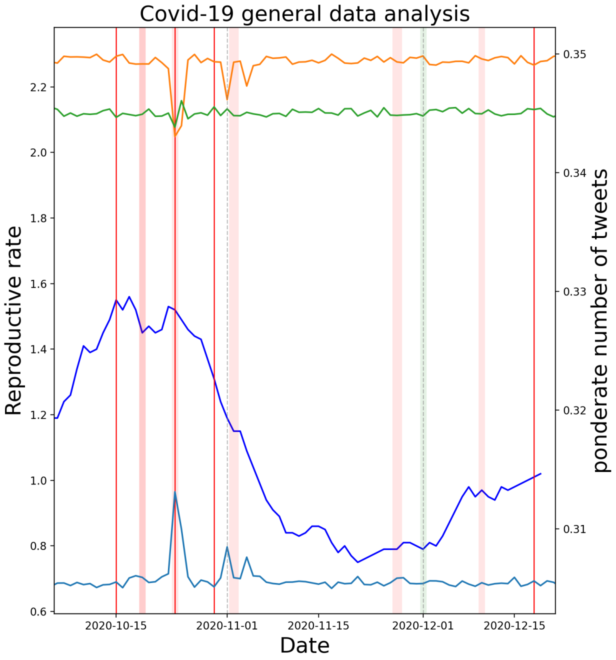


Figure 11: Data from 10/10 to 31/12

Epidemic's event during this time step (fig. 10 are the following:

- 15/10/20: [ANNOUNCEMENT] Orange code for the schools
- 19/10/20: [MEASURE] Orange code for the schools

- 19/10/20: [ANNOUNCEMENT] Curfew (00:00 - 5:00) + closure of the restaurants and bars + 1 person close contact
- 19/10/20: [MEASURE] Curfew (00:00 - 5:00) + closure of the restaurants and bars + 1 person close contact
- 24/10/20: [ANNOUNCEMENT] Extension of the curfew (22:00 - 6:00)
- 24/10/20: [MEASURE] Extension of the curfew (22:00 - 6:00)
- 30/10/20: [ANNOUNCEMENT] Lockdown from Nov 02 until Dec 13
- 02/11/20: [MEASURE] Lockdown until Dec 13
- 27/11/20: [ANNOUNCEMENT] Extension of the lockdown until January 31
- 27/11/20: [MEASURE] Extension of the lockdown until January 31
- 01/12/20: [MEASURE] Reopening of the non-essential business
- 10/12/20: [ANNOUNCEMENT] Extension of the curfew until January 15
- 10/12/20: [MEASURE] Extension of the curfew until January 15
- 18/12/20: [ANNOUNCEMENT] No relaxation for the Christmas holidays + threat of fine and control during the celebrations.

During this period we can see several peaks of negativity in the tweet trends. The first big spike is clearly a consequence of the curfew, containment and social bubble reduction measures. The second one is consecutive to announcements of reinforcement of these measures. This allows us to highlight the certain impact of liberticidal measures on the general mood of the population. However, despite this poor public reception, the measures seem to have had a strong impact on the viral reproduction rate immediately after their implementation.

From a statistical point of view, we can notice that these movements in the curve of positivity of the tweets, despite their relatively small magnitude, are significant ($p=0.05$), according to the T-test who return a value of 2.18 between the date of 24 October and the previous day, and a value of -1.51 for the 1st November.

Since health measures don't change a lot with the arrival of the end-of-year period, we can also deduce that the cooperation of the population seems to be deteriorated during festivities. Indeed, the viral reproduction rate increases monotonously from the 1st of December to the end of the period.

5 Conclusion

It was very interesting to make the link between the predictive approach to an epidemic that we did during the first semester, and the retrospective analysis of real data. We realised that quantifying the effect of various health measures against the spread of a virus is difficult, but above all that this effect can vary with time and depends strongly on the cooperation provided by the population. Thus, when deciding on these measures, it is important to measure this cooperation in order to predict the real effect of these measures. An example of this is the growth of the epidemic curve during the end-of-year holidays.

During this work, we have chosen to analyse the general feeling of the Belgians. We therefore selected tweets in a completely random way in order to preserve the general distribution of discussed topics on the platform. In this way, we were able to detect certain disturbances that were clearly associated with collective frustration about the epidemic's measures. However, it may be interesting for anyone who want to go deeper into this subject to look at a more selective analysis of the tweets. Indeed, some of the disturbances captured during the beginning of the epidemic are more difficult to relate to COVID-19 because of the presence of interference from major sporting, cultural and political events.

Bibliography

[1] Niki Parmar Jakob Uszkoreit Llion Jones Aidan N. Gomez Lukasz Kaiser Illia Polosukhin Ashish Vaswani, Noam Shazeer. Attention is all you need. 30:205–210.

[2] Kenton Lee Kristina Toutanova Jacob Devlin, Ming-Wei Chang. Bert: Pre-training of deep bidirectional transformers for language understanding.

[3] Pedro Javier Ortiz Suárez Yoann Dupont Laurent Romary Éric Villemonte de la Clergerie Djamé Seddah Benoît Sagot Louis Martin, Benjamin Muller. Camembert: a tasty french language model.

[4] https://fr.wikipedia.org/wiki/pandémie_de_covid-19_en_belgique.

[5] https://centredecrise.be/fr/news?page=2.

[6] https://www.statista.com/statistics/1109371/coronavirus-reproduction-rate-in-belgium/.