EPFL - COM-480

Covid Lexical Field

Figure 1 : main visualization

Sophie du Couédic

Polina Proskura

Introduction

There are a lot of concerning and debating news and discussion about covid. Today, words like "lockdown" or "sanitizer" hit differently than two years ago, when nobody could have an idea of what was waiting for us.

Our work is a visualization of the lexical field related to the term "covid", which originally (we tend to forget that), was no more than a simple disease name. The more global goal is an attempt to grasp the mindset of the population going through such an impactful event. What are their feelings, their main concerns, questions that arise to their mind when they hear the word “covid”?

Data Processing

In the remainder of this report, we will use the words “word” and “topic” in an interchangeable way, as each topic is mapped to a single word in our visualizations.

Our goal visualization is the graph of the topics. For building the graph we need some information about it: the nodes (in our case it is topics), the weight of the nodes (their frequency), what the edge between the nodes means and the weight of that edge. To find all this information we need to process the text data and save the information for the following visualization.

1. Text preprocessing

Since we are interested in topics, we have to find a way to analyze the text.

- To start with, we have to clean the text from the parts we are not interested in. This includes:
    - URL
    - Punctuation
    - Stopwords: words which are used a lot, but do not add anything to the semantics of the sentences and are useless for our case. Like: 'a', 'the', 'ha'.
    - Also, we decided to remove numbers, so we analyze only topics part
- The next part is the lemmatizing of the text. Since we are not interested in sentence structures and only in the topics, we can replace the form of the word with its original form. It has two purposes:
    - Since there might be words in different forms, we will be able to detect them and unite in one topic.
    - We are not interested in grammatically correct sentences (if we are talking about twitter, we may not have grammatically correct sentences), so, we remove the complication as sentence structure and dependence on it. After that we can work with the text as 'bag of words'

For this part of the process we have used the NLTK (Natural Language Toolkit) and RE (Regular expressions) libraries. We applied these techniques to the following datasets.

2. Topics generation

We used three datasets of texts (news and tweets) related to covid :

- Covid fake news Dataset, by Abhishek Koirala
- Covid fake news Dataset, by Sumit Banik *Tweets datasets*
- Covid19 Tweets, by Gabriel Preda

The whole dataset consists of more than 100'000 texts related to covid. We needed such a big dataset to generate a representative set of words related to covid, with accurate frequency of apparition. We initially wanted to eliminate meaningless words such as "small" or "doesnt" but those are really difficult to detect automatically, so we just removed the stopwords using the NLTK toolkit.

3. Word model

We decided to train the word model for our dataset. This idea was following the next purposes:

- We are interested in finding the topics that concern people the most. This means more than just calculation of frequently used words. We are interested in how close the words are to each other and if they belong to the same topic. The embedding of the words into numerical vectors will help to achieve that purpose.
- The numerical vectors are quite easy to visualize, on the plot or on the graph.

As a word model we chose untrained Word2Vec. There were several good reasons for that:

- We have a big dataset. There are a lot of tweets and a lot of news. This amount of the data allows us to achieve quite good results on the training of the model.
- We did not use the pretrained model, because the covid topic is quite a specific theme and there are no good models specializing on it yet. So, we prefer to train our own model.

Also, we decide that the vector length for one word will be 2. It is not the length that can collect a good amount of information for the word, but our final goal is visualization, which will be in 2D. So there were options to calculate the longer vector and apply the dimension reduction techniques after or train the shorter vector. The dimension reduction problem implies some losses of the information, so we choose the second option.

4. Matrix of co-apparition

For the edges of the graph we need to define the link between two words. We decide to do it based cooccurrences.

We also generated a matrix of word co-apparition frequency. A *co-apparition* of two words is defined when the two words appear in the same tweet or news text, but not necessarily next to each other.

After finding the matrix we define the threshold and if the cooccurrences are bigger than that threshold, then there is an edge.

5. Topics classification

We applied our knowledge in machine learning to try to classify the words using the k-means clustering method on the vectors we gained by words' model. The resulting classification is not perfect, but it has some bright groups we can detect. First of all, "covid hygiene" group: "face", "hand", "mask" and "safe" are close to each other. Also, there is a "politics" group with "Trump", "government", "country". Of course, there is a main group with the "covid", 'case" and etc. So, we made the choice to integrate the clustering feature into our main visualization.

6. Tweets news comparison

Also, we were interested in comparison between tweets and news topics. We decided that there are three main categories to compare the texts. They are: regular covid words, like "covid", "corona", "case" and etc. The next category is emotions: "good", "bad", "worst". And the last group is the medical covid terminology group from the words like "pathogen", "influenza" and etc. For the words in each category we calculate their appearances in the tweets and news and normalize it. So, we can compare their usage in tweets and news.

7. Countries and popularity tracking.

Also, we evaluate the flow on how much every topic appears in the context of the particular country and how popular this topic is (how much people saw is in their twitter). For that we calculate the matrix of cooccurrences (in the context term-country) and its popularity in twitter.

Visualizations

1. Main visualization

The final main visualization displays the 500 most frequently appearing topics related to the word "covid". The frequency of apparition is mapped with the **area** of each node, and the color and width of the edges represent how frequently two words appear in the same tweet to news. The visualization can be seen in figure 1.

Because we had all the necessary data to be able to compare how the topics vary in the news and the tweets, we initially wanted to focus all the visualizations to this aim, as

explained in milestone 2. We also planned to display the individual tweets and news on a circular layout surrounding the topics.

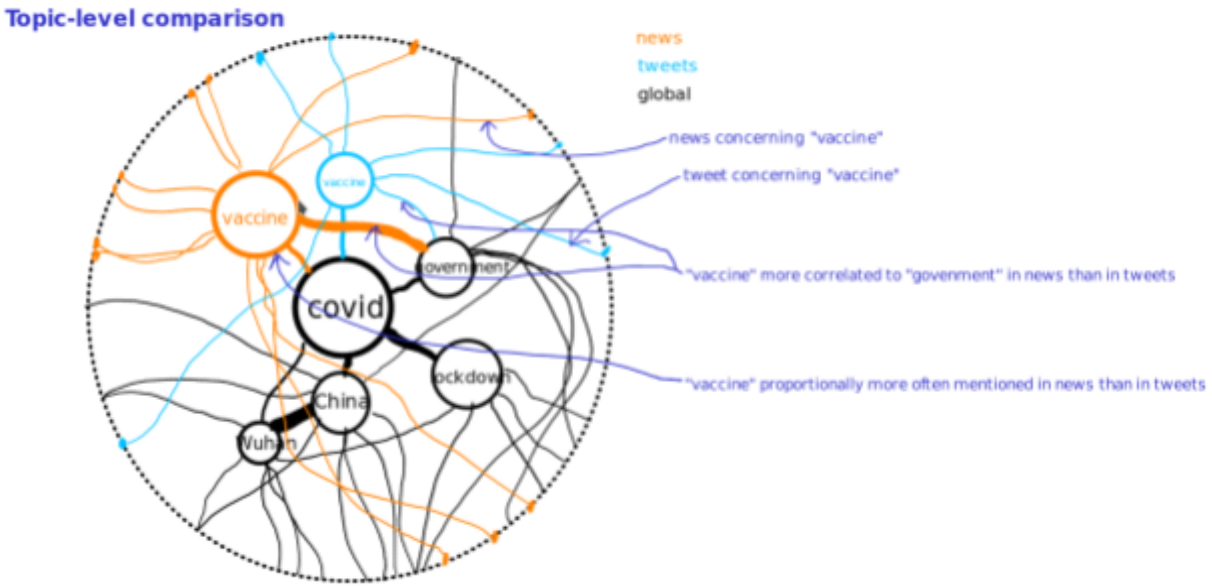

Figure 2 : original sketch main visualization

However, displaying the individual tweets and news would require storing and loading much more data on display, and we then need to trade off with the number of topics, and we decided that it was not central to our visualization.

As for the news and tweets lexical field comparison, the technical implementation of the visualization would also have been much more complicated, involving insertion and deletion of the nodes and positionnement of remaining nodes dynamically, and the library we choose doesn't offer a straightforward way to implement the latter.

We added two additional features in the main visualization, namely :

- A toggleable panel containing some legends and graph manipulations, the manipulations are :
    - Choosing the number of displayed topic
    - Choosing the **display mode**
- The two display modes are 1. Default (all the topics are displayed in the same color) or 2. Classified display, with color assigned for each class produced by our k-mean algorithm.
- Select the nodes (individually or by area) to show the topic.

The framework which was used for the main visualization is *Cytoscape.js*. This is a special framework, which allows us to work with the graphs and its different features and properties. We also used

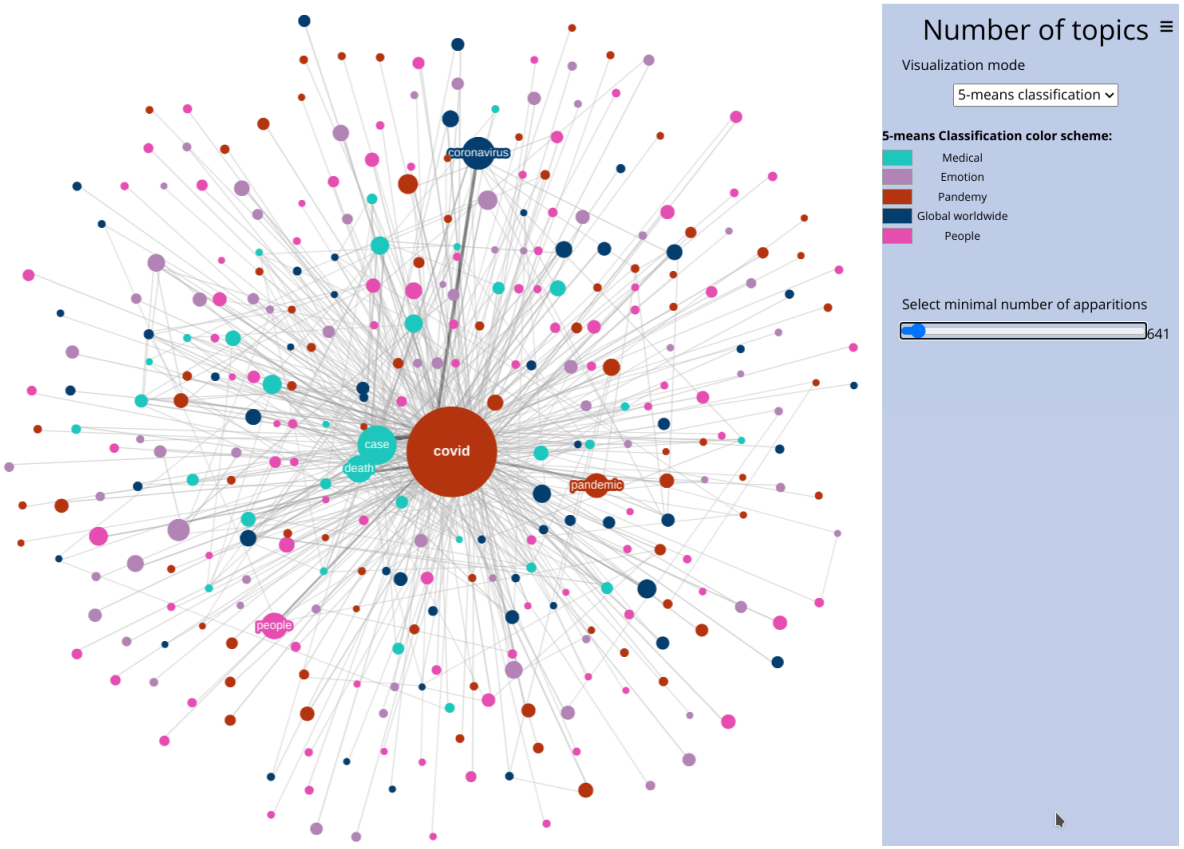

Figure 3 : main visu, additional features

2. Tweets-news comparison

For each category (emotions, regular words and terminology) we chose 10 most popular words. For each of them we calculate frequencies in the tweets and in the news and normalize them. The color of the plot represents to which category the data belongs (blue for the tweets and orange for the news). The bigger the bar on the plot, the more frequently people mention the topic.

The total visualization of the double-sided **barchart** build using *d3 framework*. As an extent, we implemented the drop-down menu updates, so the users are comfortable with choosing the category they are interested in.

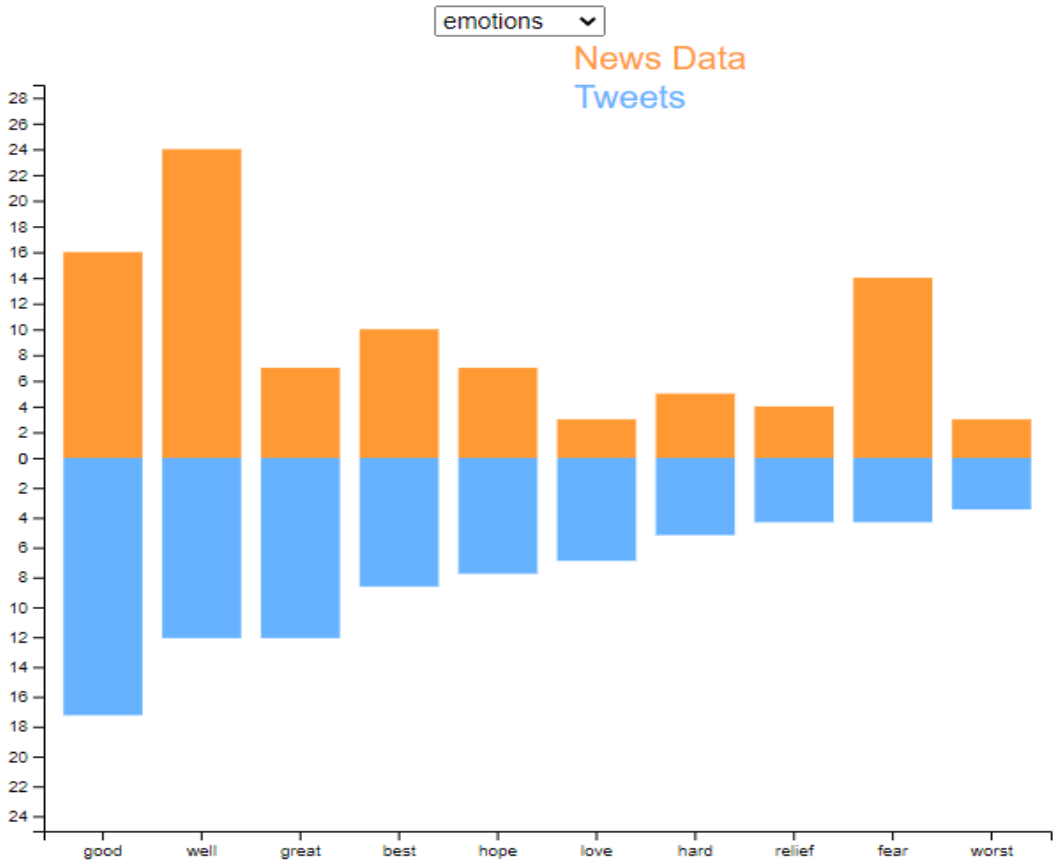

Figure 4 : original plans bi-chart

3. Tracking the country flow

For this part of visualization we chose the **Sankey diagram** from the *d3 framework*. The left part nodes are the main covid topics, the middle ones are for the most popular countries in tweets and the right one if for popularity of the tweets.

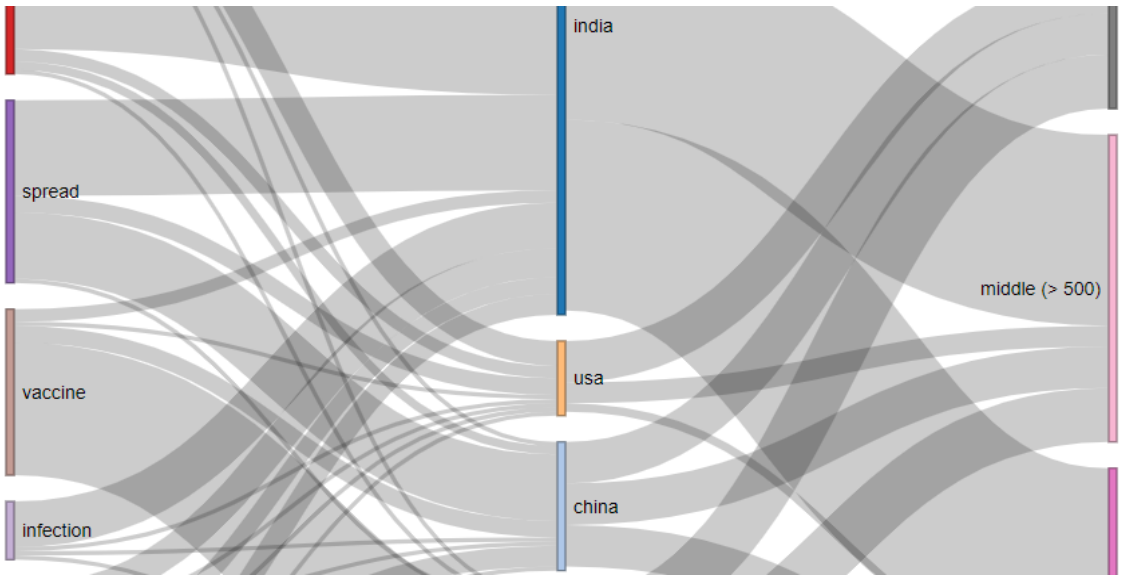

Figure 5 : original plans end-to-end graph

We realize the planned part for that visualization, but we have not had the time for the bigger extent. The possible extension would be animation and the ability to choose the countries you are interested in the most.

Peer assessment

Sophie did the main visualization and overall look of the website.

Polina did data processing, barplot part and Sankey diagram.

Conclusion

The work on the website turned out really interesting. The resulting website looks comfortable and informative.

Possible future improvements are:

- More clear website navigation (between its part and categories)
- Expansion of the features and topics: maybe the section with articles about covid and its analysis.

Thank You for Your attention!