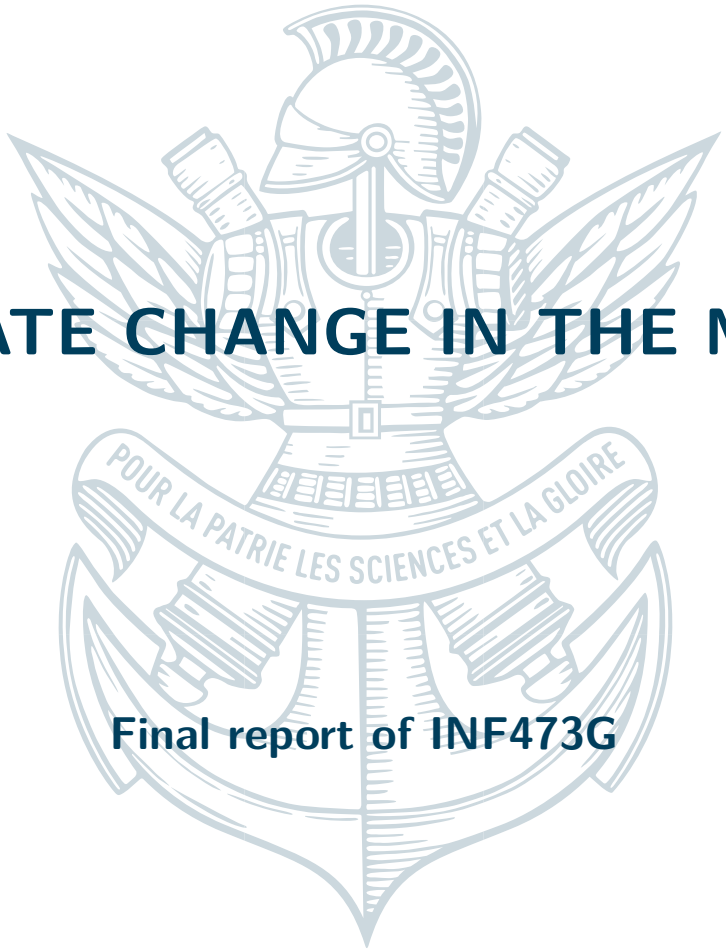


# CLIMATE CHANGE IN THE MEDIA

## Final report of INF473G

18 May 2022

Xiaoyun Ai and Maurício Lima

# TABLE DES MATIÈRES

# 1
# INTRODUCTION

Climate change is a central topic in the society of the 21st century. Indeed, a lot of discussion is being made about the theme. As it's an global issue, it touches a lot of different fronts, such as physics, economics, social and environmental responsibility. Hence, discussions in the media are very dense and involves a lot of data. It seems logical to search for relations to better understand the debates and how climate change is already impacting society.

In the course of this project, two sources of discussions will be analyzed. The IPCC reports and the twitter discussions on **#climatechange**. We have utilized Neural Language Processing (NLP) as a central tool in order to interpret and quantify text data. We also used graph tools to cluster information and to understand relation between these clusters.

# 2
# IPCC REPORTS DATA EXTRACTION AND ANALYSIS

## 2.1 Introduction

We have decided, in a first approach, to tackle the IPCC reports on climate change mitigation. These reports are a reference for governments, institutions and everyone of us when talking about climate change with property. The goal in this part of the project was to construct a knowledge graph to structure information about climate change.

## 2.2 Data Preparation

### 2.2.1 • IPCC reports

This following part can be found in the "Extraction" folder. We took 4 mitigation reports on the IPCC site. The reports are in PDF, so we transformed them into **.txt** with simple online resources.

The next step was to get the most important terms in the reports. For that, we utilized the TF-IDF algorithm in scikit-learn library. For instance, this algorithm allows us to take the most frequent terms in the reports while not considering common words of the English language (the, and, of, ...). We chose the 200 most common words. This code can be found in the **extract-common-words.ipynb** file.

At this point, we realized that common terms such as "climate change" were not being considered because we were searching just for monograms. So we changed a little bit the code in order to search for monograms, bi-grams and finally monograms and bi-grams together. We chose to continue with this last approach for the rest of the project because it seemed more complete. As a cleaning part, we also eliminated words with less than 3 words and numbers. All of these results can be found in the **most-frequent-xxgram-list.json** files, where "xx" can be "mono", "bi" or "monobi" (for the last case).

### 2.2.2 • Wikidata relations

This following part can be found in the "Wikidata" folder. We took all of the most common monograms and bi-grams after the cleaning and searched for their possible labeling in the Wikidata data base (Q-codes). This requests can be found in the **monobigrams-labels-dictionary.json** file and the code to do so is in **wikidata-labels.py**.

Now, all of the monobigrams have their wikidata codes. We need, finally, to look for relations between these entities. While looking for 1-layer relations, we didn't find any relation. So we passed to the 2-layer relations. In other words, we searched for relations of second order between all of the monobigram pairs. This request took a really big amount of time. We tried to do so in many ways in order to reduce the time of processing, but the results were always very slow. We also had some issues with exceptions and errors, so we constructed the code in order to ignore it and keep turning independently. This code can be found in the **wikidata-relations.py** file and the result was saved in the **graphe-dictionary.json** file.

## 2.3 Graph construction

The files for the construction of the graph can be found in the "Gephi" folder. The process to construct the graphe was all done with the software "gephi". The file to construct the relations is **ipcc-words-relations.csv**. The file in which we have the Id of each node and the respective modularity class calculated by gephi is **words-modularity-classes.csv**. This part was quite direct, since all of the heavy work has been already done in the last subsection. For instance, the modularity classes calculation of the statistics part of gephi was very useful for finding the clusters in the graph.

## 2.4 Data Analysis

The clusters allows us to categorize the words in themes, which can help us to better understand the information in the reports. It can also help us to understand relations between words, of course. Ultimately, it will help us to make a connection between this set of data and the Twitter one.

The labeling of each cluster is shown in figure 1. It was done by analyzing the words in each cluster defined by gephi for the graph.

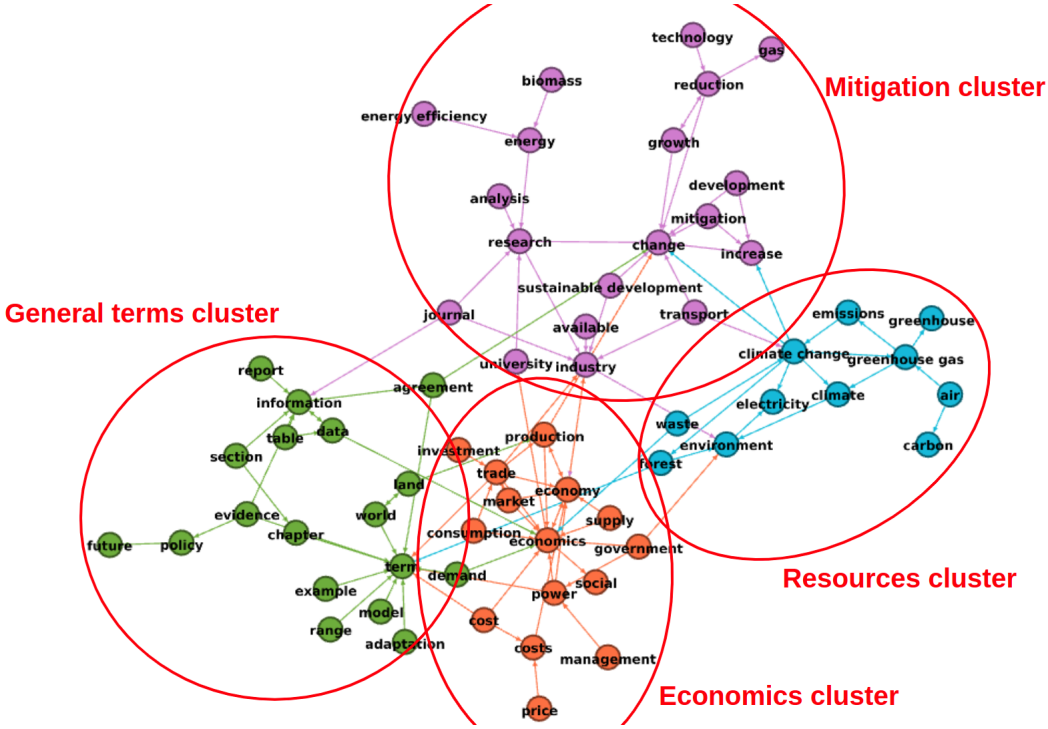


Figure 1 – Classification of each IPCC words cluster

# 3
# TWITTER DATA EXTRACTION AND ANALYSIS

## 3.1 Introduction

The analysis of the IPCC's official reports is clearly insufficient to answer our questions about who is talking about climate change and what is being talked about. First of all this is an official report, drafted by professionals. It does not represent the views that everyone wants to express when talking about climate change at the moment. In reality, people talk about climate change for more diverse reasons, or even just to promote their products. At the same time the official report tries to be objective and neutral, so it does not carry too much emotion. Twitter is an excellent platform for understanding public thoughts on climate change. It has enough active users and anyone can express their thoughts through the platform. Secondly, it has a hashtag feature that allows you to filter information effectively. And there are now very good data mining technical tools for Twitter.

In the next sections, we will present how we did the Twitter data mining and data cleaning processes, and then how we designed and built the graph around our purpose. This is followed by a description of the techniques used to analyze the data collected and an attempt to relate it to the words (topics) frequencies obtained in the previous chapter. We conclude with the difficulties and shortcomings during the process. The code for this parts can be found in the "Twitter" folder of the project.

## 3.2 Data preparation

### 3.2.1 • Data collection

We used a web information scraping technique to collect information. The specific library used is snscrape for python. We crawled all English tweets with **#climatechange** from January 1, 2022 to the day we started collecting data, which was April 23. Because each captured tweet has a different kind, it could be a reply under someone else's tweet, a quote commenting on another tweet, or most often, a user's original tweet. In our project, because the replies to tweets can't be particularly influential or meaningful, we filtered it out when we collected it. Then for each tweet, we record the id and content of the tweet, and if it is a quoted tweet of another tweet, we also record the id of the tweet it references. The number of retweets, quoted, replies and likes of the tweet were also recorded. In addition to that, we have to record the information of the person who sent the tweet. Include his username, user id, user profile description. And if he mentions another user in his tweet, we recorded the relevant user information as well. It is worth noting that the id of each tweet and user as well as username are unique. So this provides natural identification.

The format of the data in the **before_clear_tweets.csv** is as in figure 2 and figure 3.

### 3.2.2 • Data cleaning

After following the setup described in the previous section, a total of 18,095 tweets were collected. For better analysis and visualization, we have to cull the data set. We define the value of a tweet as

| | Unnamed: 0 | user_id | user_name | user_description | tweet_id |
|---|---|---|---|---|---|
| 0 | 0 | 1035554769505378304 | PointerBrampton | Local news, politics and in-depth reporting. T... | 1517654232005943296 |

Figure 2 – The format of tweets data collected I

| content | quotedTweet_id | mentionedUsers | #quote | #retweet | #reply | #like |
|---|---|---|---|---|---|---|
| Doug Ford rushes to start Bradford Bypass ahea... | 0 | NaN | 0 | 1 | 0 | 0 |

Figure 3 – The format of tweets data collected II

the sum of the number of retweets, quotes, replies, and likes of the tweet. Then we did some statistical counting for this "value" of the data set tweets. The results show that there are 50,712 tweets with "value" of 0, 100,002 tweets with "value" between 0 and 10, and 15,573 tweets with "value" between 10 and 20... The number of tweets with a "value" above 200 was only 962. Finally we choose to set the threshold to 200 and delete a tweet if the sum of the number of likes, retweets, quotes and replies is less than 200. This way we ended up with 962 tweets.

In addition to this massive cleaning of the data, there is also a cleaning processing of the collected tweet content later in the NLP processing. This is because the collected tweets are interspersed with unnecessary and even confusing content that can interfere with our deduction. For example, when we need to analyze the information, topics of tweets that being mostly talked, **#climatechange** will be a distraction, because every tweet will carry this hashtag. There are also links, emojis, etc. All are objects that need to be cleaned up.

In terms of technology, the main use is the regular expression operations library in the python standard library. We started by converting all the content to lowercase, then removing hyperlinks in the tweets, removing tagged users from the tweets, removing hashtags related to climate change, removing punctuation in the tweets, and finally converting the emoji in the content to text. In this way we get clean and appropriate content. The figure 4 below shows an example of the content of a tweet before and after it was processed.

```
my_df['content'][57]
```

```
['"When you understand that under capitalism a forest has no value until it\'s cut down, you begin to understand the root of our ecological crises!"🖤\n\n#ClimateCrisis #climatechange https://t.co/zK2b3n0hok',
 'Whales are a key component in ensuring planetary equilibrium.\n\nEach great whale sequesters an estimated 33 tons of CO2 on average, thus playing role against #climatechange \n\nProtecting whales is climate positive \n\n#LoveWhales❤️ #Savewhales🐬 https://t.co/2qk6MOa8yF',
 '"The environment is in us, not outside of us.\nThe trees are our lungs, \nthe rivers our bloodstream, \nand what you do to the environment, \n ultimately you do to yourself"🍃\n#Climatechange #Savenature https://t.co/7GDvv7Knx9']
```

```
my_df['content'][57]
```

```
'when you understand that under capitalism a forest has no value until its cut downyou begin to understand the root of our ecological crises black_heart climatecrisis are a key component in ensuring planetary equilibrium each great whale sequesters an estimated tons of co on averagethus playing role against protecting whales is climate positive lovewhales red_heart savewhales dolphin environment is in usnot outside of us the trees are our lungs the rivers our bloodstream and what you do to the environment ultimately you do to yourself leaf_fluttering_in_wind savenature'
```

Figure 4 – Text before and after cleaning

## 3.3 Graph construction

First we want to use the information already collected to answer the question, who is talking about climate change. This requires creating a graph between users. So we define all the nodes in our graph as users. We are only interested in the fact that there is an implied relationship between the two users. Therefore we create a simpler graph with the relationships for the quotes of one user by another as well as being mentioned by another user.

At the same time, we found that for the tweet of type quotes in our dataset, the original tweet they cited might not have **#climatechange** with them so they were not in the scope of our collection. But we have the original tweet id which is stored as **'quotedTweet_id'**. So we do an exact scrape of the collected **quotedtweet_id** again to get the user information and add it to the list of users that already exist. Finally we store all the user node information we have collected in **user_nodes_new.csv**.

Next, based on the stored information about quoted and mentioned user, we created the edges one by one and stored them in **edge_quoted.csv** and **edge_mentioned.csv**.

## 3.4 Analytical techniques

For the node and edge files we have obtained. We import them into Neo4j and Gephi. Since our aim is to obtain information about the relationships between users, especially in the case of communities or clusters, we mainly use Gephi for the analysis of the data. The clustering is analyzed by the modularity algorithm in the Statistics section. Then we use the filter function and the Layout area to manipulate them and complete the cluster display in the Apppearance area.

## 3.5 Result

### 3.5.1 • User communities

Ignoring some of the small, fine-grained communities, we can see that the users collected can be clearly divided into five categories, as shown in Figure 5.

Among the green ones are the communities represented by Gretha and Aarav, representing the younger generation and the connection between those who produce original content online. In purple are groups of institutions that study climate, mainly the official accounts of the IPCC (Intergovernmental Panel on Climate Change). In light blue are other communities of scientific institutions not specialising in climate change research, and mainly national institutions in the USA. Examples include NASA, NOAA (National Oceanic and Atmospheric Administration). The dark blue and yellow are communities of influential media and individuals and businesses (mainly car companies) in the USA. They are represented by Bill Gates, Elon Musk, CNN, New York Times. In fact, the results are very logical. Each community clearly carries its own label. But at the same time, we can see that these communities are very well connected internally but interact very sparsely with each other. This makes us wary of the existence of "information cocoons". For example, is the younger generation of environmental voices a reliable source of information, and are they providing sufficient, scientific information and analysis, or is it all emotional slogans. And whether businesspeople are speaking out primarily for their own business interests.

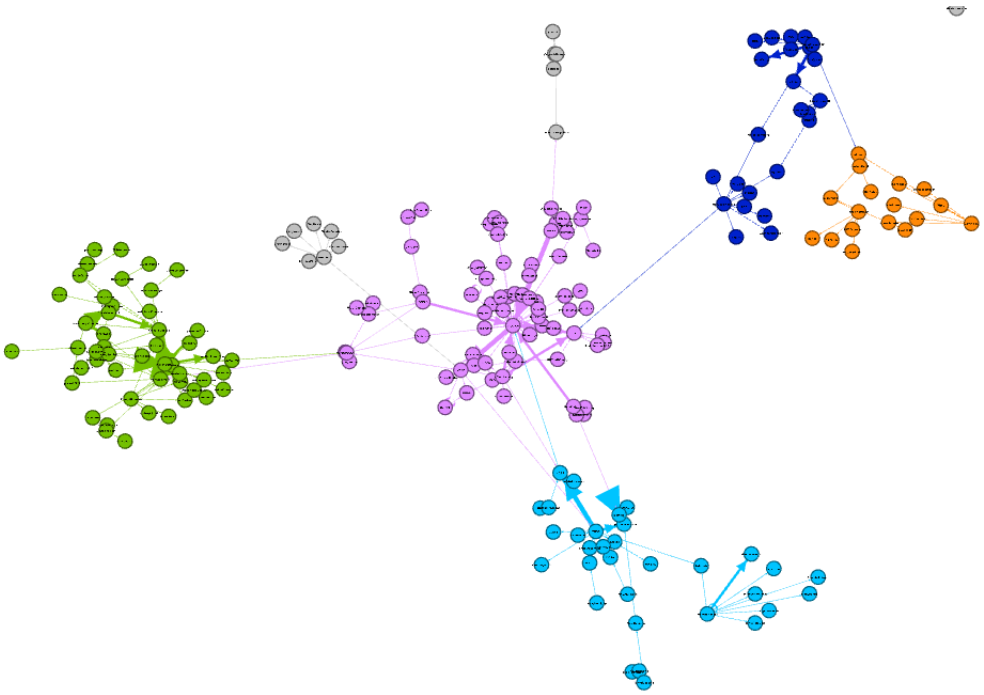

Figure 5 – The user communities

### 3.5.2 • difficulties and defects

The first problem encountered had to do with large-scale data processing. I initially wanted to try to keep as much data as possible to analyze the user community. So only those data where the sum of the count of those indicators was 0 were removed. However, I did not expect that this would result in a very poor visualization due to too many data, so I started to reduce the data set heavily. Other difficulty encountered is still mainly in NLP processing. At first time, my idea was to first analyze the user's self-description by using NLP and assign labels to the user such as organization, politics, business, etc. However this technology was too complicated to implement.

As for the defects, first of all, for data collection, several more tags with similar meanings can be collected to increase the scope, such as **#climatecrisis** and **#globalwarming**. For data cleaning criteria, it is rough to use the sum of likes, retweets quotes and replies directly, a linear combination of these indicators would be better. For user content classification, the processing techniques of nlp can be refined even further.

# 4
# FINALS RESULTS

With the information collected and analyzed from the two parts, we were able to connect the two structures two make more deeper and detailed analysis. In a first approach, we used the modularity labeling of the words clusters in the IPCC report to analyse the content of the discussions in Twitter for each cluster. We fixed the structure as in Twitter graphe with the labeling of IPCC.

## 4.1 NLP process

Here we tried to answer, what was said, the question. After we get the results of the text analysis of the official IPCC report, we want to classify the users by their post content topics. So for each user's content we compared it with four topic categories got from IPCC analysis using spacy's semantic similarity analysis. The most similar category was selected as the topic category for this user. Notice that the fourth cluster of topics does not have a special meaning, but just some very broad words of exposition, so it leads to the vast majority of users falling into this category. So we also did a similarity matching classification after removing the fourth category of topic clustering. In addition to this a sentiment analysis of each user's tweet content was also done.

## 4.2 Interpretation

As we see, there is a lot of null data, because our algorithm labels the user that made the Twitter, and not the one that is tagged. We thought that this is a more rational approach. For instance, we cannot label a person just for being cited in other tweet. So we decided to eliminate the null data.

We realize that the cluster "General Words" takes most of the space. This is reasonable, as everybody uses general words for everything. But as this cluster doesn't really represents the discussion, we eliminated it in order to obtain more constructive results. The final result can be see in figure 6.

This final result shows us that discussions are mostly focused in mitigation. The "Gretha and Aarav cluster" focus more in the economics part. This is very interesting because this cluster is representing the young voice in the debate. We see that they re already engaged in the economics impacts on climate change, which shows how constructive this theme is for the society ! An other cluster that focus in economics is the "American vehicles and influential people cluster", but this is more comprehensive.

Two remarks are important here. The first one is that it is not very surprising that mitigation is the center of discussions, since we utilized the IPCC reports on mitigation. Maybe using more IPCC reports can be a good way to generalize his result and eliminate the bias. Nevertheless, it shows that our results are coherent even after merging 3 different data bases ! The second remark is that we don't actually have that much of users in the graph to make generalizations. The "American vehicles and influential people cluster" have just 7 nodes after the cleaning process for example.

In the next graph, we utilized the sentiment analysis for the twitter discussions to quantify the sentiment of each class of words.

The result is shown in figure 7.

In a analysis, we can infer some sentimental points of each cluster in the graph. The mitigation cluster is the most positive one, which make sense. Indeed, talking about mitigating problems is

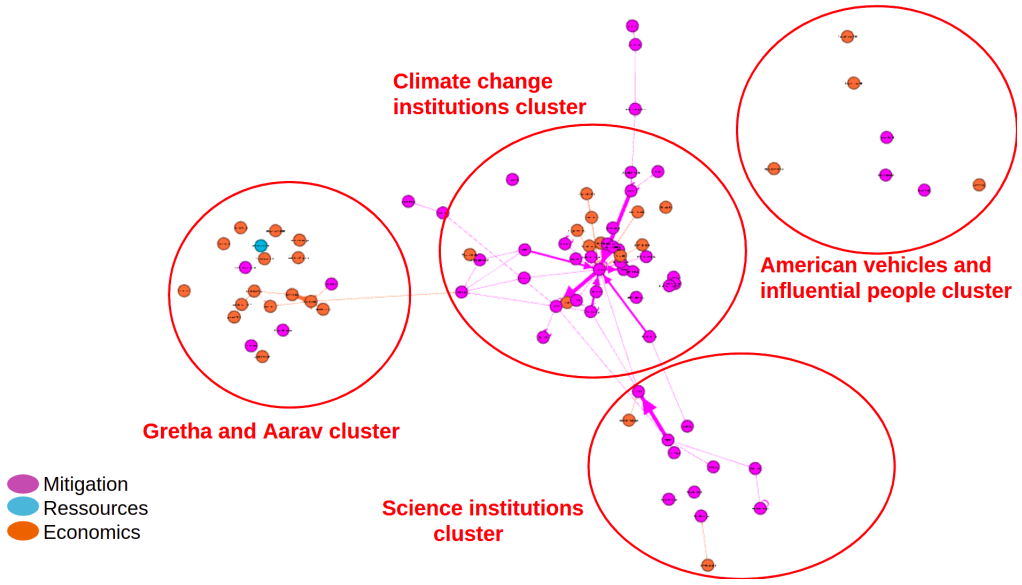


Figure 6 – The twitter interactions with principal IPCC clusters labeling and not-null filters

usually a good thing, nothing very new here. The resources cluster is still positive, but a bit less. In fact, talking about the resources on earth should be less pleasant, since it's a critical situation. The economics cluster is the less positive in debates (but still greater than 0). This can be explained by the fact that economy is maybe the most affected sector by the sustainable development. All of these conclusions have a logic explanation, but still, they are not so obvious to infer without this analysis, which corroborates to the debate. The very last cluster is in general, positive, but that's because, as we already said, the debate is positive in general, as we have general words, they will follow the mean of the data set.

Similarly, we can perform sentiment analysis on user communities based on their content of tweets. The results are shown in Figure 8. We can see that overall, there is a generally positive sentiment. This is in line with our previous analysis, as most of the tweets are related to mitigation, which is relatively one of the topics with the most positive sentiment. But at the same time, each user community is slightly different. With the exception of the professional climate change institution community, which appears to have a lot of positive sentiment because of its emphasis on mitigation, the most positive community is in the influence and business communities. This can be explained by the fact that people in this community want to make their products or themselves look better by talking about the environment with a positive tone. At the same time, the ""Gretha and Aarav" community has the largest share of negative sentiment in the group. This is also consistent with their choice to spread inflammatory videos and messages on the Internet, as well as their choice to strike in protest of actions such as the climate crisis.

In conclusion, these results shows us from a new angles how the debate about climate change is structured in the media. By aggregating the information in graphs, we can form clusters, which helps us to classify the information. By analyzing each of the many graphs plotted in this report, we understand better the relations between this clusters and how information is seen, utilized and discussed in the media.

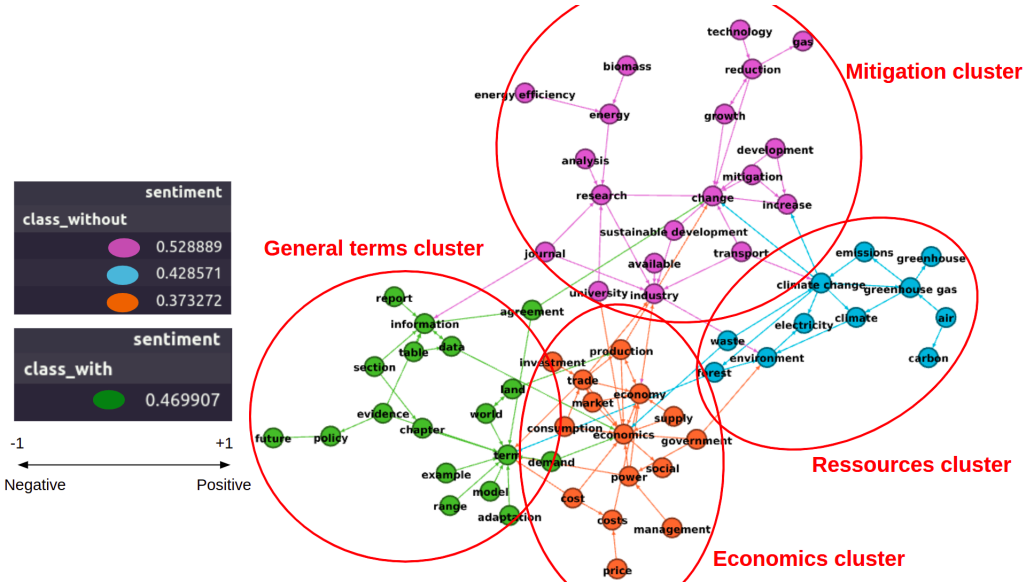


Figure 7 – The IPCC most frequent words graph sentiment analysis

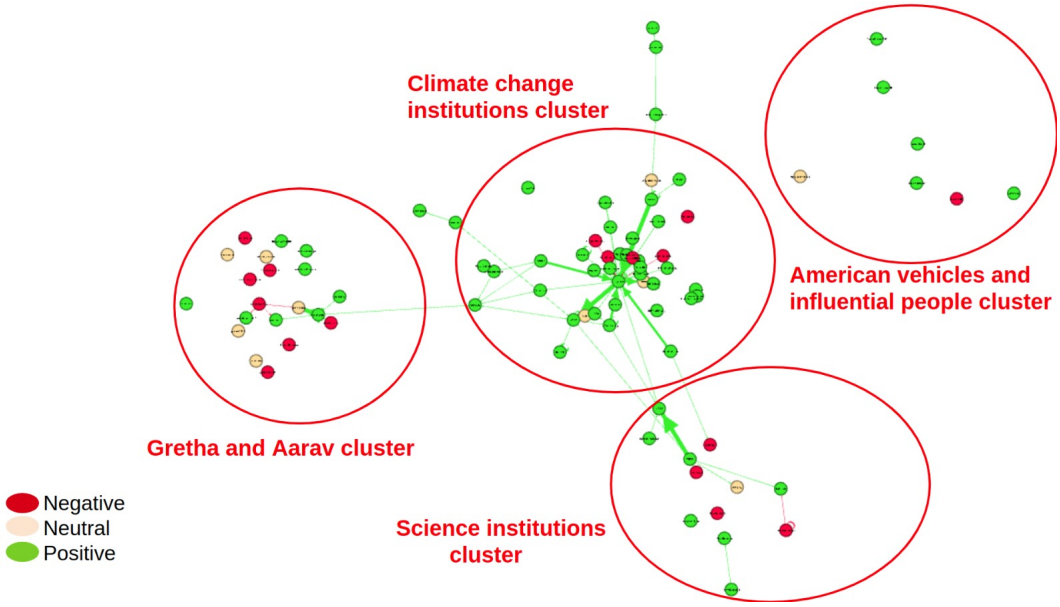


Figure 8 – The user communities sentiment analysis