Climate Data Observatory

Monika Rakoczy[1], Robin Lamarche-Perrin[2], Armin Pournaki[3]

[1] Sorbonne Université, CNRS, Laboratoire d'informatique de Paris 6 (UMR 7606), France
[2] CNRS, Institut des Systèmes Complexes de Paris Île-de-France (UPS 3611), France
[3] Max Planck Gesellschaft, Max Planck Institute for Mathematics in the Sciences, Germany

This work is funded in part by the European Commission
H2020 FETPROACT 2016-2017 program under grant 732942 (ODYCCEUS).

Table of contents

Introduction

Climate change is an undeniable truth in the scientific community. However, even though predictions of its impacts have been persisting for years, the debate seems to have gained significant traction in the public sphere only in the last few years. One explanation for this could be the increase of immediate effects of climate change, such as rising temperatures, floods and bushfires. Meanwhile, the increasing interest in the climate change debate could as well have come along with a discursive shift, enhanced by emerging actors on the international scene such as Greta Thunberg. This shift can take place and manifest itself differently in various areas of debate, ranging from political discussions in parliaments to newspaper articles and social media outlets.

Our work contributes to the Climate Change Observatory developed by the Vrije Universiteit Brussel (VUB). The focus of our work is to allow the users, in particular social science researchers and data journalists, to explore data connected to climate change. Our contribution consists of accessible methods and Web services to visualize and analyze the climate change debate on different media spheres.

Data

We consider three large-scale datasets spanning from July 2016 to September 2019:

1. For the perspective of social media, a collection of 80M tweets related to climate change (courtesy of The Digital Methods Initiative, Amsterdam);
2. For the perspective of mass media, 4526 articles published by The Guardian (courtesy of the Vrije Universiteit Brussel);
3. For the political debate, a corpus of 92 287 UK parliamentary speeches (courtesy of the Max Planck Institute for Mathematics in the Sciences, Leipzig).

All documents in these three corpora have been selected by the use of the keyword "climate".

General structure

For each of the dataset, we identify and observe the dynamics of subtopics over time, employing methods ranging from outlier exploration and topic modeling to network representations. This allows us to investigate the possible existence and dynamics of discursive shifts.

In particular, we focus on three aspects of debates:

1. How much is climate change discussed in a given discursive area?
2. What vocabulary and terminology is used to do so?
3. What kind of interaction patterns do we observe in such discussions?

The tools we developed aim to provide analysis of the data in two manners:

- **Topic analysis, or the content (texts) of the discussion:** tweets for Twitter dataset, speeches for Parliamentary dataset, and articles for Guardian dataset. In this part, as a base for our analysis, we use topic modeling methods directly on discussion content to extract main thematics addressed in the documents. This part consists of three functionalities that allow users to analyse discussions and their main topics: topic specificities, topic similarities and topics temporal analysis.

- **Interaction analysis, or the way actors interact with each other:** we provide tools making possible to group the users and visualize their interactions. This part is focused on Twitter data, as it is the only dataset that provides data about information exchange between users. Combining such user interactions with the topics obtained from the content of these exchanges, one can investigate particular topics addressed by groups of users (or communities) and identify important exchange patterns between these groups.

Component functionalities

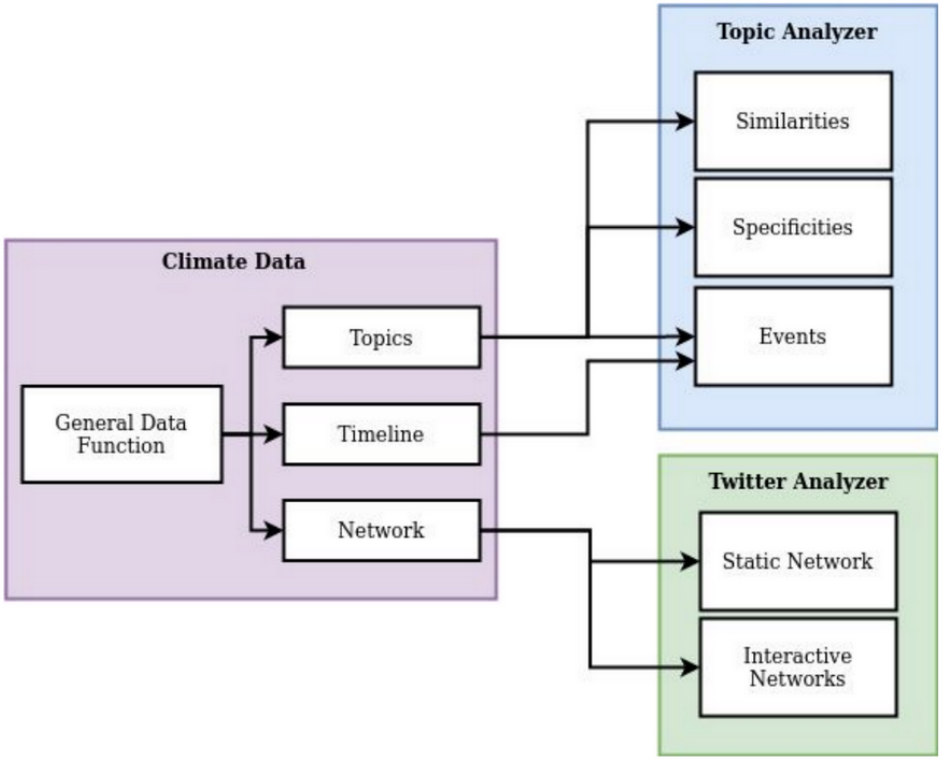

Fig. 1 Overview of component functionalities

Our component consists of the following functionalities:

- **Climate Data**
    - The **General Data Function** can globally provide metadata about stored documents such as author, date of publication, possible interactors (when a document constitutes an interaction with someone else, such as a retweet), number of words or characters, depending on the input parameters. It can also provide detected topics for each document[1]. This general function is hence a tool for designing specific requests depending on the type of information one wants to study.

    - **Topics** returns the lexical distributions of pre-learned topics (probability that a given word belongs to a given topic), as well as following metadata: corpus name, topic name, number of documents belonging to that topic, number of words in these documents, and their ids.

    - **Timeline** returns the evolution through time of the number of documents or number of words, optionally dividing them by topics. It hence includes the following metadata: corpus name, date, topic name, number of documents, number of words, and optionally document ids.

    - **Network** returns an interaction graph where nodes are authors and links are documents corresponding to information exchanges between these authors. Currently, this function is only available for the Twitter dataset where interactions correspond to retweets: the author posts a retweet of a tweet previously published by the interactor. These documents can also be divided by topics.

- **Topic Analyser**
    - **Similarities** - In order to study similarities between topics, a measure of relative entropy (Kullback-Leibler divergence) is used. For each pair of topics, this measures quantifies information discrepancy when assuming these topics are similar. The result is then used to gather topics in similarity groups, ending up with a topic network where the thickness of links between two given topics indicate their similarity (the thicker, the more similar).

[1] Due to high computation time, topics were acquired beforehand and are used as pre-trained information. We applied Latent dirichlet allocation (LDA) to each corpus with 1-gram words, $\alpha = 0.1$, and $\beta = 0.05$. For both Guardian articles and Parliamentary speeches, we retrieved $k = 10$ topics, while for Twitter we retrieved $k = 9$ topics.

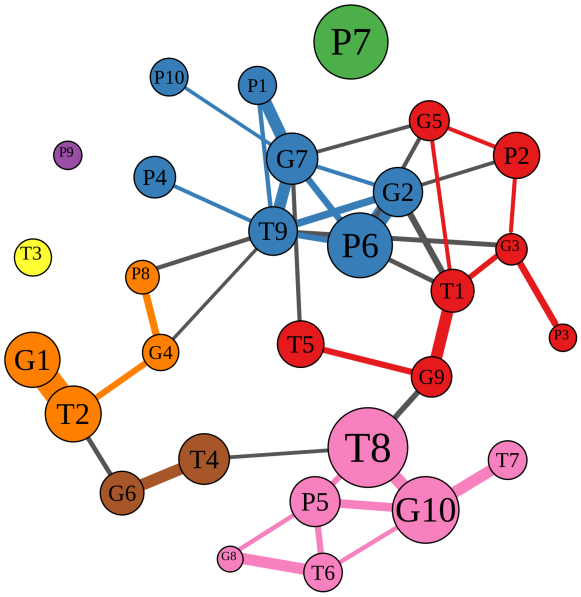

Fig. 2 Exemplary result of the similarity function

- **Specificities** - This function allows identifying and measuring lexical specificities within one particular group of similar topics.Applying Principal Component Analysis (PCA) on lexical distributions, using words as variables and topics as observations, the resulting two first components highlight main specificities of selected topics.

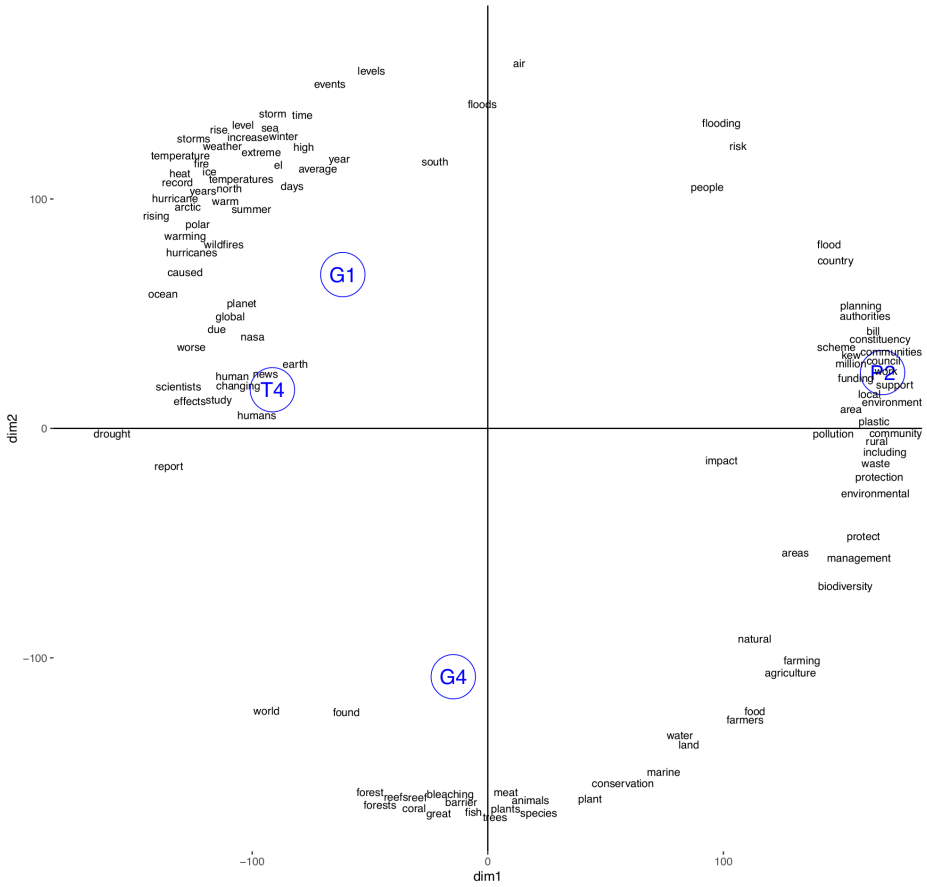

Fig. 3 Exemplary result of the specificity function

- **Events** - Topics popularity changes, as well as identification of real life events, is obtained by using distribution of topics probability among documents through time (aggregated by days, weeks, months or years). This method allows the user to define parameters of an event: its length and the minimum probability threshold used to detect it. As a consequence, users have the possibility of observing, distinguishing and, focusing on particular peaks, thus identifying real life events that are addressed by a given sphere of debate. Result consists in the distribution of topics probability through time with highlighted events (as defined by the user).

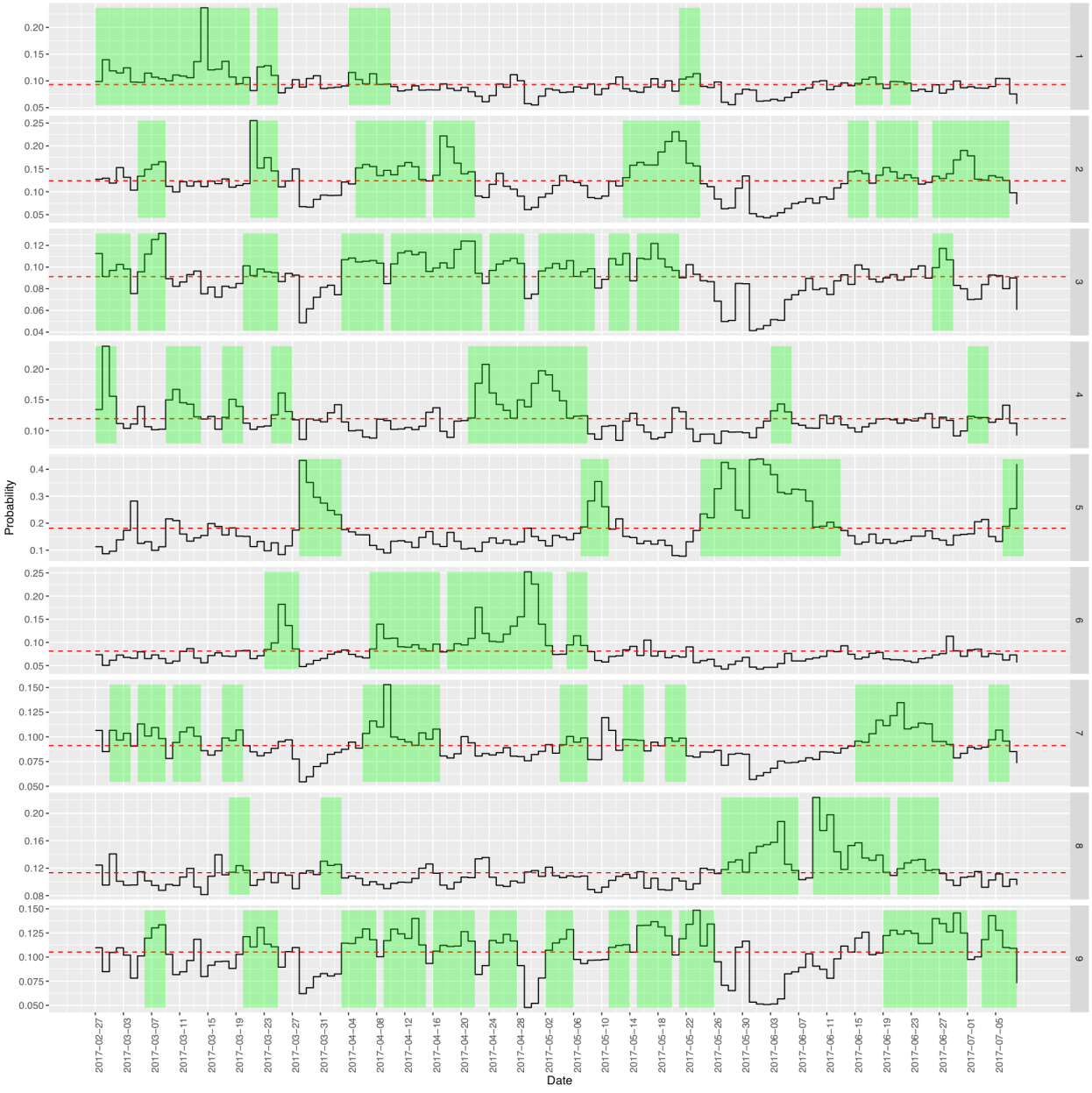

Fig. 4 Exemplary result of events function

- **Twitter Analyzer**
  - **Static network** - this part allows users to get a structural overview of the debate by showing the Twitter communities based on mutual retweets and their interconnection. It gives insight into the content of the inner-community debate by showing the inner-community retweets' topic distribution as pie charts. We consider this a distant-reading approach that combines the analysis of the overall interaction network with the overall textual content. As a result, static network visualization is obtained, where nodes are communities and directed links are drawn from community i to j if a Twitter user in i retweets a user in j. The width of arrows is proportional to the number of retweets between two communities. The node size is proportional to the number of retweets within one community.

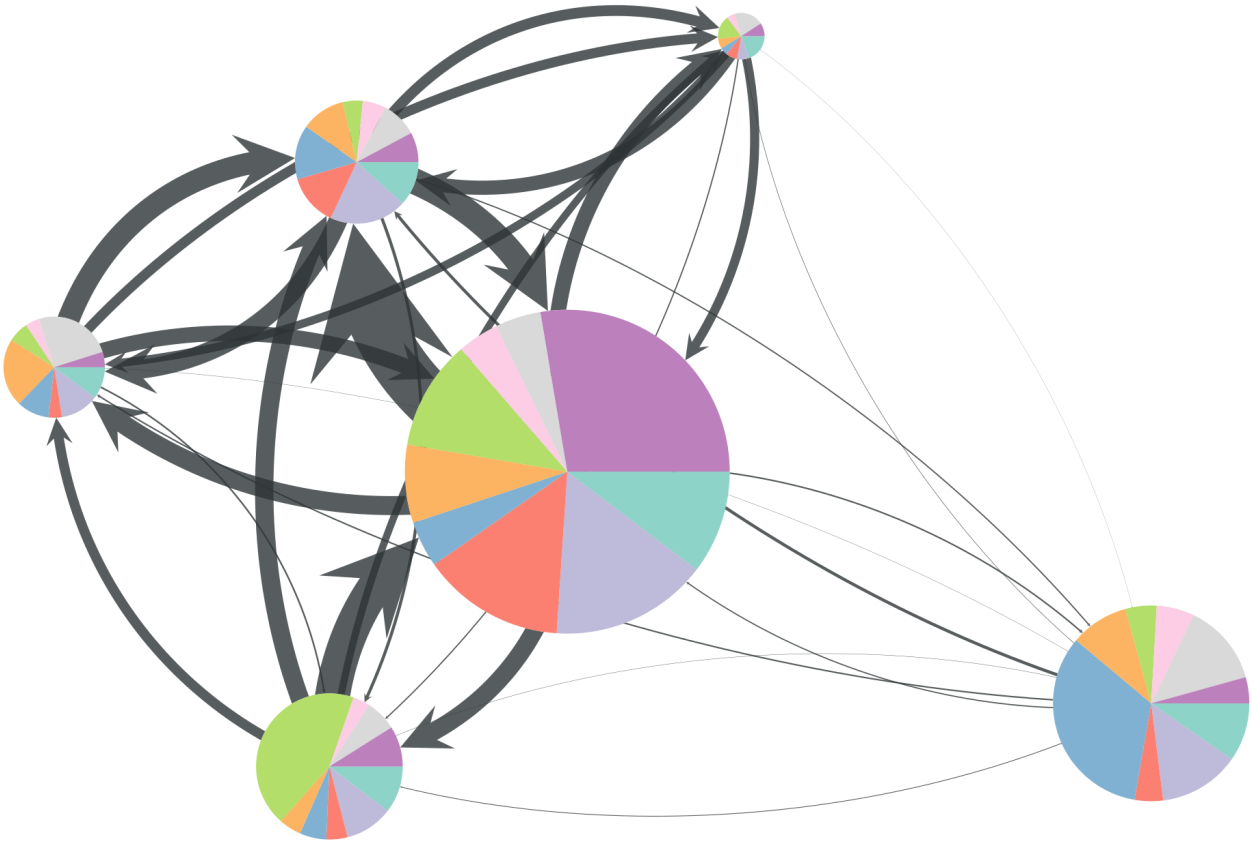

Fig. 5 Exemplary result of static network function

  -

- **Interactive Network** - allows users to get a closer look at a specific event in the debate by presenting an interactive retweet network, where they can explore the content that was shared during this event. It gives insight into the main actors during that event, their relations and the content they promoted. We consider this a close-reading approach that combines the analysis of a specific part of the interaction network with the textual content from this event. As output, user obtains an interactive network visualization of the Twitter data from the selected date range using the twitter explorer framework, where nodes are users and directed links are drawn from user i to j if i retweets j. A click on a node displays their content in the dataset (if the tweets are still publicly available, therefore complying with Twitter's visual guidelines).

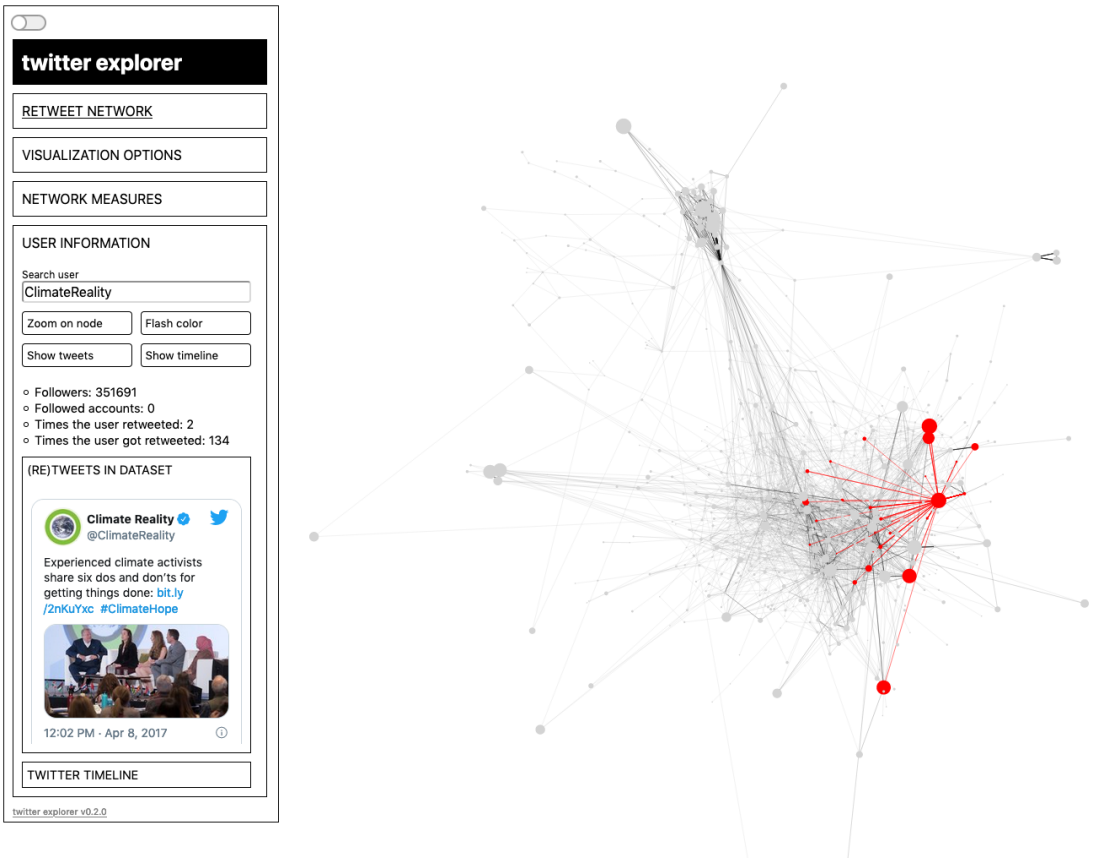

Fig. 6 Exemplary result of interactive network function

Interface

Interface is coherent with VUB Climate Observatory design. Pipeline includes:

- **Step 1** - Selecting dataset (Twitter, Guardian, UK Parliamentary) including metadata and pre-trained topics for each dataset (Twitter - 9 topics, Guardian - 10 topics, UK Parliamentary - 10 topics)
- **Step 2** - Selecting particular Topic Analysis method:
    - Similarity (see Similarities in Component functionalities)
    - Specificity (see Specificities in Component functionalities)
    - Events (see Events analysis in Component functionalities)
- (from: Events) **Step 3** - in the step 2 user is able to select particular time period from the data that is of interest (i.e. by targeting particular event); therefore in Step 3 the user is able to export the data in .csv including the texts (tweets/articles/speeches) that can be than used as an input to UVB Climate Observatory tab; moreover, a possible extension involves directly connecting the UVB Climate Observatory to Step 3, in which case Step 3 would consist of text analysis method selection from UVB Climate Observatory

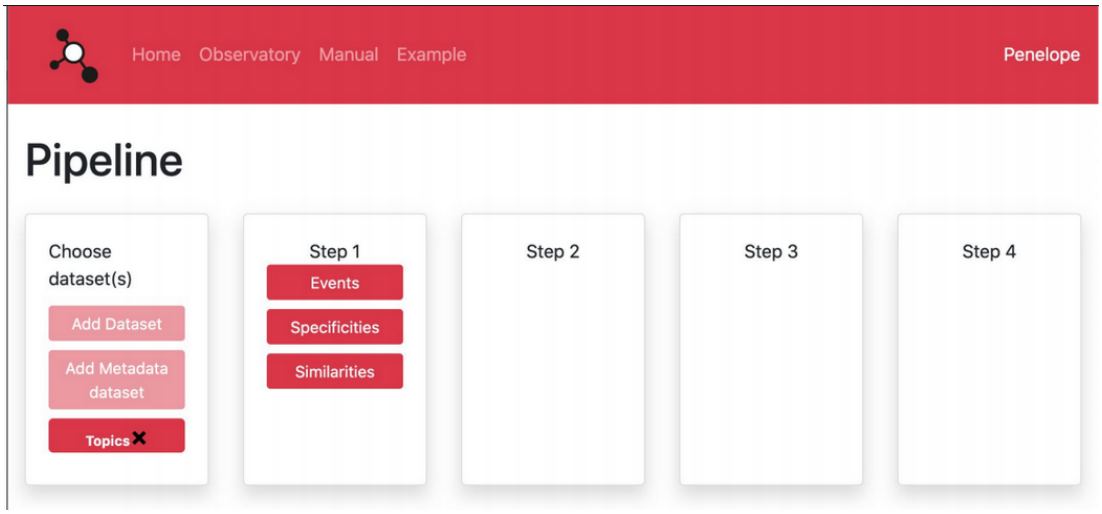

Fig. 7 Example of Climate Data tab available in Penelope interface

Relation to conflict and opinion dynamics

Our work directly relates to opinion dynamics, as it involves tools aimed at observation and analysis of debates. Tools provided in our work allow users to track changes taking place in a global discussion about climate change. The tool and its methods is designed for data exploration, such as trend observation, investigation of actors involved in information exchange, general topics of interest of climate change discussion, etc. Consequently, it can be useful for navigating more precise research hypotheses in social, political, or communication sciences, that can then be tested with close reading of selected documents, such astext tools provided by VUB. Furthermore, the tool can be useful for helping in opinion formation and evolution, as they enable users to compare different debate sources in terms of particular topics.

In particular, users are able to grasp main ideas in climate change debate in each of researched media spheres: political, classical news media, and social media. Moreover, by utilizing tools for observing similarity and specificity of lexical differences between each of the spheres, users can deepen their familiarity in the subject, as well as their recognition of nuances between each spheres. Additionally, observation of popularity in time of each topic, and ability of event recognition provided in our work enables users to navigate, observe and recognize different conflicts amongst important actors involved in the climate change debate. Importantly, our work also consists of visualization tools which provide users with an easy way of navigating and understanding interactions between climate change debate participants. Visualization tool is also able to incorporate topic analysis results, which leads to a broader overview of the climate discussion presented to end users.

All of the described tools facilitate deeper understanding of the subject by particular users, which directly influences the basis on which they form their opinions, and, as a consequence, allows for easier and more careful opinion formation and evolution.