UNIVERSITY OF TRENTO

Department of Sociology and Social Research

Master's Degree
in
Data Science

Analyzing Polarization in Climate Change Tweets during COP: A Multi-Layer Networks and Topic Modeling Approach

Alessio Gandelli

Supervisor: Elena Pavan[1]

Co-Supervisor. Matteo Magnani[2]

Academic year 2022/2023

Acknowledgements

And I would like to acknowledge ...

Abstract

This thesis aimed to investigate the polarization of users in the climate change discussion surrounding the Conferences of Parties, with a specific focus on the 26th Conference. Unlike previous studies, this research adopted a topic-by-topic approach using a multi-layer network framework. The objectives were twofold: first, to identify the most polarized topics of COP26 and compare them with a longitudinal study involving COP21; and second, to explore user polarization across different topics and investigate whether users engaging in more topics exhibit higher levels of polarization. The findings revealed that in the most polarized topics, there was an almost equal distribution of users on both sides, indicating a sharp divide in opinions. Notably, this polarization was particularly evident in the Canadian discussion on limiting the use of coal and oil, as well as the air travel debate. Additionally, it was observed that the Canadian discussion on Fossil Fuel was not consistently polarized, as it exhibited complete non-polarization during COP21, confirming the results obtained by other researches. Furthermore, users tended to remain aligned with a specific side of the discussion across multiple topics, although there was no correlation between the number of topics and the polarization score.

Table of contents

Chapter 1

Introduction

1.1 Background

Climate change has been a well-known problem among scientists for a long time. The first paper warning about the effects of the increment of carbon dioxide in the atmosphere is from 1976 [1], while in a study of 1976 [2] it was observed for the first time. In 1988, the World Meteorological Organization Established the Intergovernmental Panel On Climate Change (IPCC) [3], which is the leading organization evaluating climate change up to date. As shown in [4], the researchers' interest in climate change grew significantly after 1990.

"All the News That's Fit to Print"

The New York Times

VOL.CXXXVII . . . No. 47,546 — NEW YORK, FRIDAY, JUNE 24, 1988

Global Warming Has Begun, Expert Tells Senate

Global Warming: Greenhouse Effect?

Average global temperatures through the first five months of 1988. As a baseline, scientists use the global average from 1950 to 1980.

Source: James E. Hansen and Sergej Lebedeff

Sharp Cut in Burning of Fossil Fuels Is Urged to Battle Shift in Climate

By PHILIP SHABECOFF

WASHINGTON, June 23 — The earth has been warmer in the first five months of this year than in any comparable period since measurements began 130 years ago, and the higher temperatures can now be attributed to a long-expected global warming trend linked to pollution, a space agency scientist reported today.

Until now, scientists have been cautious about attributing rising global temperatures of recent years to the predicted global warming caused by pollutants in the atmosphere, known as the "greenhouse effect." But today Dr. James E. Hansen of the National Aeronautics and Space Administration told a Congressional committee that it was 99 percent certain that the warming trend was not a natural variation but was caused by a buildup of carbon dioxide and other artificial gases in the atmosphere.

An Impact Lasting Centuries

Dr. Hansen, a leading expert on climate change, said in an interview that there was no "magic number" that showed when the greenhouse effect was actually starting to cause changes in climate and weather. But he added, "It is time to stop waffling so much and say that the evidence is pretty strong

Drought Raising Food Prices; Inflation Effect Seems Minor

By ROBERT D. HERSHEY Jr.

Cañon Zapata in Tijuana, Mexico, the busiest illegal crossing point.

Fig. 1.1 Front page of NY Times, June 24, 1988

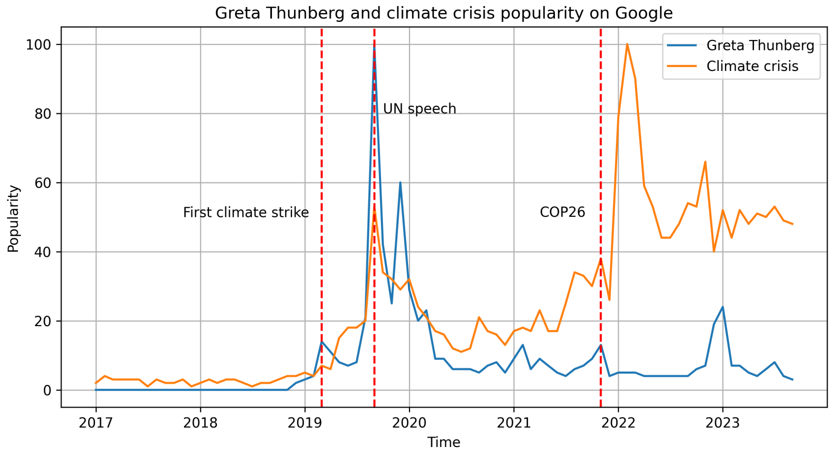

Fig. 1.2 interest on Google over the time of Greta Thunberg and climate crisis

The issue has become mainstream only in the past few years since, partially due to Greta Thunberg [5] and Fridays for Future, people grew more aware of the problem. As we can see in Fig 1.2, only after 2019 people started searching for, and talk about, a 'climate crisis' [1], underlining the urgency with which we should act. It is also true that the strikes caused inconvenience to many ordinary citizens trying to reach their workplaces. As a consequence someone may have developed some hostility toward this activism and, thus, toward climate change. In fact, an increase in polarization has been detected after the strikes [6].

Conferences of Parties (COP) are conferences where the highest political figures of many countries meet and talk about climate emergencies. This serves also to bring the discussion to the masses, increasing awareness on the problem. In particular, Twitter is a place where part of the political debate takes place [7] and political events like this can be studied thanks to the availability of big amount of data.

The increasing polarization poses a significant challenge to the mitigation of the harmful impacts of climate change; for this reason, understanding the root cause has the potential to speed up the reduction of emissions, not having to deal with strong opponents. This can have the effect to save human lives and resources.

Conference of Parties Conferences of Parties are yearly conferences organized by the United Nations where the topic of discussion is climate change; the first was held in 1995 in Berlin, and the ones that ended with a document to ratify are:

[1] we choose this term because it is not descriptive as climate change but focus on the negative version of it

- **COP3**: Kyoto Protocol (1997) It was the first treaty to mandate countries to cut greenhouse gas emissions that was legally binding, but this was true only for developed countries, therefore exluding China and India.
- **COP21**: The Paris Agreement (2015) is a global accord that aims at limiting global warming to well below 2 degrees Celsius (preferably to 1.5 degrees Celsius) compared to pre-industrial levels by requiring all countries to set their own emissions reduction targets. It is considered more effective and inclusive than the Kyoto Protocol because it involves all countries, allows for flexibility in setting emissions targets, and includes a robust system for transparency and accountability.
- **COP26**: Glasgow Climate Pact (2021) is a crucial agreement in the global effort to combat climate change. It is the direct consequence and a refinement of the Paris agreement. It includes significant commitments to address the urgent challenges of climate change, such as phasing down coal usage, increasing climate finance for adaptation, strengthening international cooperation, and supporting countries which are transitioning to low-carbon economies. However, not everyone agrees with the outcomes of the conference: for example, [8] shows how the previous goal has not been achieved and how the roadmap is not clear. While [9] underlines how the politics decisions are not based on the IPCC report, which is the most trusted scientific reference for climate change.

The focus of this work will be on COP26 which is the one that occurred in a context of popular agitation toward the topic. Additionally, a spike in the climate crisis interest has been detected right after it. Then, a comparison between COP21 and COP26 will be done due to its analogies, in fact the Glasgow climate pact exists as a more specific definition of the too general Paris agreement. Both happen in a context of political constrast, during 2017, president Trump of the United States decided to withdraw from the Paris agreement because it would have undermined the US economy. This caused disappoinement to all the others members, especially the developing countries which are the most affected by global warming effects.

This work lays its foundations on the research of Falkenberg et al. [10], who discovered that COP26 was way more polarized than COP21. Using a similar approach, the ideological polarization will be explored topic by topic using cutting edge technologies such as transformers and by exploiting the complexity of multilayer networks.

There is not a universally agreed definition of polarization. Ref [11] help disambiguate the various definitions that fall under the term polarization. It distinguish between sociopolitical, group and individual polarization. Our focus will be on the sociopolitical one which concern the polarization of the influential individuals within the parties, called elite polarization, which is defined as the alignment of the political leader with all the positions of its own party. There is a general agreement between social and political scientist that elites are polarized, while it is unclear if the masses are also polarized in the same way, since the results can depend on the methodology.

In this paper, we will operationalize it using what stated in [12], the same used by Falkenberg, which is: "The most common measure of polarization in the political literature is probably bimodality, which is the idea that the population can be usefully broken down into two subpopulations". In the case at hand, the two sub-populations are pro-climate and climate skeptics.

1.2 Research Questions

Due to the structure of this research, we can now answer a new set of questions related to intra-topic polarization. The first two look at the networks at a macro level focusing on the topology. The most straightforward is the first, RQ1, which aims at inspecting the topics that are driving the polarization of the entire COP26. Secondly, RQ2 wants to identify whether these topics have always been polarized compared to COP21.

Then, we the focus will shift to a micro level with some questions related to the users, similarly to [13] that investigated the users retweet from both sides, but in addition we are interested also in the presence in multiple topics.

In particular, RQ3 looks into whether users are polarized in the same way over the different topics or if there are topics in which they are on the opposite side of the spectrum. RQ4 instead investigates whether the most active users, both in term of number of tweets and the presence on multiple topics, are more polarized than the others.

Summarizing, the questions are:

1. Which are the most polarizing topics discussed on Twitter during Cop 26?
2. How did topics evolve between cop21 and cop26?
3. Is the single-user polarization different across different topics?
4. Does the polarization of users differ depending on whether the users are present in multiple topics rather than just one?

1.3 Structure

This thesis is structured to provide a comprehensive exploration of topic modeling applied unstructured text organized in a multilayer network. It is divided into a series of chapters, each serving a distinct purpose in advancing the understanding of the subject matter.

Chapter 2 touches the state of the art of all the matters used in this work, first exploring the most recent topic modeling techniques with the goal to select the best suited to our scenario; second, a comprehensive review of networks, in particular social networks and multilayer networks, and how polarization is computed.

Chapter 3 presents the dataset along with some statistics about it.

Chapter 4 is the core of this work, the first section covers the evaluation of several topic modeling techniques, using two approaches, unsupervised and supervised, several models have been tested from traditional to the ones based on neural networks. After identifying the best performing model, the network section shows the pipeline that allows to build a multilayer network based on the topics starting from the raw tweets. Finally for each layer it is computed a polarization score using the latent ideology score given to the users.

Chapter 5 show the answer to the research question posed above.

Chapter 6 wrap up everything and show some other research that can be done using the same methodology.

Chapter 2

Related Work

In this Chapter the literature about the main subjects treated in this thesis will be reviewed, starting from the state of the art of topic modeling to the study of networks in particular multilayer networks, and concluding with the mathematical details of the computation of latent ideology score and polarization.

2.1 Topics

Topic modeling is a widely used technique to extract and analyze latent topics from a collection of documents. In this paper dealing with tweets, the focus will be on topic modeling for short text; the challenge rests, therefore, on the limited amount of text available for each document.

In this chapter, we review traditional methods such as LDA and NMF and more advanced techniques like BERTopic, top2Vec, BTM, and GSDMM.

2.1.1 Traditional methods

LDA (Latent Dirichlet Allocation, 2003)[14] is a traditional probabilistic algorithm for topic modeling. It assumes that each document is a mixture of topics, and each topic is a distribution over words. However, LDA struggles with short texts due to the sparsity of the word co-occurrence matrix.

Another classical algorithm based on linear algebra is NMF (Non-negative Matrix Factorization) [15][16]. It assumes that each document is a linear combination of topics, and each topic is a non-negative linear combination of words. It can be used for short texts and as a baseline method.

2.1.2 Advanced methods

Here, the most recent techniques that have been used and tested in different settings will be inspected:

BTM (Biterm Topic modeling, 2013) [17], unlike the traditional methods, does not use the bag of words approach, but it considers the co-occurrence of words to the topic modeling. As a generative probabilistic model, similar to LDA, represents each document as a set of pairs of words (biterms).

GSDMM (Gibbs Sampling algorithm for the Dirichlet Multinomial Mixture model, 2014) [18] is a clustering-based algorithm that uses Gibbs sampling to iteratively assign each document to a topic. It models the distribution of words in each topic using a Dirichlet distribution. GSDMM is designed explicitly for short text from social media, which is typically unstructured and contains slang.

2.1.3 Neural approach

The most advanced techniques (BERTopic [19] and Top2Vec [20]) exploit neural networks, in particular transformers [21], and have a common structure:

1. Vectorization of the document using embeddings
2. Dimensionality reduction
3. Clustering
4. Topic definition

Let us see in detail all these steps:

Embeddings

The first and most crucial step is mapping each document into a vector of numbers easily manipulated by the machine. This can be accomplished in a wide range of ways: from naive one-hot encoding (Recurrent Neural Networks (RNN), Long Short-Term Memory Networks), to more complex transformers. Egger et al. [22] proved that the latter, in particular Bertopic is the best-performing method when using tweets as documents.

A transformer [21] is a deep learning architecture that excels at processing sequential data, such as natural language text. Unlike previous architectures like recurrent neural networks (RNNs) and long short-term memory (LSTM) models, transformers can process all parts of the input data in parallel, which allows for faster training and execution.

The self-attention mechanism works by assigning different weights to different parts of the input. For each token in a sequence, it calculates a score (the attention score) that determines how much to focus on other tokens when encoding that one. The higher the score, the more attention the model places on the token. An interesting feature of the transformer is the ability to map the same word from different contexts into two different points in a multidimensional space. Different words with the same meaning will be mapped in two close points.

Usually, transformers work at a word level. However, in the case at hand, the focus is on a sentence level, so sentence transformers will be used, since they map every document, instead of every word, to a vector.

The goal is to map two documents about the same topic to two multidimensional points that are close according to some notion of distance, usually Euclidean or cosine similarity.

BERTopic traditionally uses a BERT transformer to calculate its embeddings, but the flexibility of the framework allows the use of other embedding methods. The best known are Doc2Vec, Universal Sentence Encoder, and the recent Openai release of their version of embedder [1].

Dimensionality reduction

Dimensionality reduction is the process of reducing the number of input variables or features while retaining the essential information and preserving the underlying structure of the data. After embedding the documents, we have a set of long vectors; for instance, by using the *all-MiniLM-L6-v2* model, the result is a 384 vector long for each document. Since the goal is to cluster the documents, this algorithm works better with low dimensional data due to the well known curse of dimensionality problem [23]. For this purpose, many algorithms that can reduce the length of the vectors exist. Usually, for clustering, less than ten dimensions are suggested. Some examples are: Principal Component Analysis [24](PCA), which works maximizing the variance of the data along the principal components, t-SNE [25]focuses instead on preserving the local structure of the data in the lower-dimensional space. However, t-SNE is not recommended for clustering or outlier detection as it does not necessarily preserve distances or densities well, UMAP [26], which is particularly known for its scalability, and its ability to preserve global structure, is more computational efficent compared to other dimensionality reduction methods, is suited to reduce the dimensionality of big datasets like the one used in this research.

[1] https://platform.openai.com/docs/guides/embeddings/what-are-embeddings

Clustering

Clustering is an unsupervised machine-learning technique that aims at grouping similar objects together based on their features. The goal of clustering is to partition the data points into subsets, called clusters, such that objects within the same cluster are related to each other in a certain way. There are plenty of clustering algorithms: hierarchical, density-based, distribution-based, and centroid based.

K-means clustering [27] is an iterative algorithm that partitions a dataset into a predefined number of clusters based on the proximity of data points to cluster centroids.

HDBSCAN [28](Hierarchical Density-Based Spatial Clustering of Applications with Noise) is a density-based clustering algorithm that extends the DBSCAN algorithm to handle noise and clusters of different densities. It can automatically determine the optimal number of clusters and allow outliers.

These properties of HDBSCAN make it perfect for clustering tweets, since they are noisy and not all of them should be put in a cluster.

Topic Definition

At this point, it is possible to give a numeric label to each document depending on the cluster; the next step is to give each cluster a set of representative words. Here are the differences between Bertopic and Top2Vec,

Bertopic uses Term Frequency- Inverse Document Frequency (TF-IDF) [29] statistic to find the most relevant word for each topic; more specifically, the terms stand for:

Term Frequency: It's the ratio of the number of times a specific term appears in a document to the total number of terms in that document.

Inverse Document Frequency: The IDF of a specific term can be calculated as the logarithm of the total number of documents in the corpus divided by the number of documents containing the specific term. This gives more importance to rare words in the corpus(the collection of documents)

TF-IDF is the product of these 2.

$$TFIDF(t) = TF(t) \cdot IDF(t) \tag{2.1}$$

Top2Vec, instead, for each cluster, identifies the centroid first, typically calculated as the mean value of every dimension of all the points in the cluster. Then, it gets the n closest words to the centroid.

Topic Representation

After extracting the most meaningful words of each topic, it is useful to give a label, which is a small sentence that can identify in the best way the content of the cluster. This is not an easy task, which is why, traditionally, it has been performed by humans, which is time and energy consuming and not neither does it guarantee getting the best result possible since humans are biased by their opinions [30] [31]. Using Large Language Models, in particular, GPT gives a new opportunity to perform this task quickly, and with good accuracy. The performance in this scenario will be shown in more depth in the methods chapter.

2.1.4 Topic Evaluation

An important step was to evaluate the performance of these models, either by using unsupervised metrics or by setting up a custom experiment using a labeled dataset. The first method uses metrics well known in the field of topic modeling and will be presented here, while the second part can be found in the chapter on methods.

Metrics The coherence and diversity metrics are measures used in topic modeling to evaluate the quality of the learned topics. *Coherence* is a measure of how often words co-occur, 'often' meaning more than would be expected if they were independent. *Diversity* measures the difference between two topics in term of words used.

- **NPMI (Normalized Pointwise Mutual Information)** The NPMI value lies between -1 and 1, where a higher value indicates a higher level of coherence, thus implying better topics. It is calculated as follows: for every pair of unique words, $(\text{word1}, \text{word2})$, in a given topic, the Pointwise Mutual Information (PMI) is calculated as:

  $$\text{PMI}(\text{word1}, \text{word2}) = \log\left(\frac{P(\text{word1}, \text{word2})}{P(\text{word1}) \cdot P(\text{word2})}\right) \tag{2.2}$$

  Where $P(\text{word1}, \text{word2})$ is the probability of both words appearing in a sliding window, and $P(\text{word1})$ and $P(\text{word2})$ are the individual probabilities of each word appearing in a window.

  PMI values are then normalized to ensure that the measure is not overly influenced by the frequency of word pairs:

  $$\text{NPMI}(\text{word1}, \text{word2}) = \frac{\text{PMI}(\text{word1}, \text{word2})}{-\log(P(\text{word1}, \text{word2}))} \tag{2.3}$$

The NPMI coherence of a topic is the average of the NPMI values of all pairs of words in that topic. The NPMI value lies between -1 and +1, where a higher value indicates better topic coherence.

- **Umass coherence**: similar to npmi, for each pair of words, $(\text{word1}, \text{word2})$, the UMass coherence score is calculated as:

$$\text{UMass}(\text{word1}, \text{word2}) = \log\left(\frac{\text{co\_count}(\text{word1}, \text{word2}) + \varepsilon}{\text{count}(\text{word1})}\right) \tag{2.4}$$

  Where $\text{co\_count}(\text{word1}, \text{word2})$ is the number of documents in which the two words co-occur, $\text{count}(\text{word1})$ is the number of documents in which the first word appears, and ε is used to avoid logarithm of zero.

  The UMass coherence of a topic is then the sum of these log values for all pairs of words in the topic. The UMass coherence score can take on any value from negative infinity to zero, with values closer to zero indicating higher coherence.

- **diversity**: In the context of topic modeling, the diversity score is a measure that quantifies how different the topics are from each other. One way to compute this is by looking at the proportion of unique words across all topics.

$$\text{Diversity Score} = \frac{\text{Number of Unique Words}}{\text{Total Number of Words}} \tag{2.5}$$

- **Computation time**: time needed to fit the models.

2.2 Networks

A network, often referred to as a graph, is a data structure that is composed of nodes and ties, which represent the connections between the nodes. This interconnected structure allows for the representation and analysis of relationships and dependencies among various entities. The applications of networks are incredibly diverse and wide-ranging.

One example of a network is a computer network, which consists of interconnected computers that communicate and share resources. Another example is a network of webpages, where each webpage is represented as a node, and the links between webpages form the ties. This type of network is fundamental to the World Wide Web and enables the navigation and discovery of information online. Additionally, networks find applications in logistics and transportation, where they are utilized to optimize routes and streamline the movement of goods and service. These are only few examples in the use of networks, in this thesis the focus will be on social networks, where the nodes represents persons and the tie some sort of interaction between them.

Social Network Analysis is a field that examines the relationships, the interactions, and the structures within a network of individuals or entities. It provides valuable insights into the dynamics, information flow, and influence within social networks. Several studies have applied SNA in different domains [32], such as online social networks, organizational networks, and public health networks. A growing field in this context is the analysis of social media, which completely changed how the research in this area is carried out. Due to the huge amount of data available, it is possible to shift to a data driven methodology, where the data collection is not strictly embedded in the design of the experiment, but it is the starting point on top of which the research is built. Another advantage of this kind of data is that, since it results from a natural observation of actors in a social environment, it is not as biased as self-reported data or as data collected in a laboratory setting. Concerning this, Veltri [33] explains how behavioral data is linked to automatic decisions in the surrounding environments, while the self reported one is more conscious and reflexive and not always the action matches what people say. It must be noted that the data is not completely unbiased, it depends on the goal and the structure that the platform gave to it.

In this scenario, the data is being used to explore the topology of the interaction between users of Twitter; the structure of their interactions is studied to understand if there are some recurring patterns.

2.2.1 Multi-layer Networks

In particular, we will build our network using a Multi-layer networks (MLN) framework[34]. MLN are complex networks that capture multiple types of relationships, or interactions, between nodes. They allow for the representation of different dimensions or contexts in a single framework, providing a more comprehensive understanding of network dynamics. Single-layer networks are, in certain cases, an oversimplification of reality[35] [36].

MLN are often used in biological networks where, due to organism complexity, every biological function is usually influenced by more factors, and modeling with a MLN helps the researcher in studying the interaction between these factors. Another biological use of MLN is epidemiology, where the presence of a certain disease can be strongly influenced by other clinical conditions. For example Kinsley et al. [37] used this framework in veterinary epidemiology to identify the subjects that are more prone to spread a certain disease.

Also in the field of interest for this research, social networks and, in particular, Twitter, several researchers used MLN to structure their study: for instance ref [38] employed it to identify the most central accounts over multiple layers in the discussion of different political candidates, one for each layer. In relation to this, De Domenico [39] mathematically described different ways to compute node centrality on multiple layers, introducing the concept of versatility.

Thanks to the complexity added by the multiple layers, the same users can be seen interacting in the different dimensions, which in our case are the different topics. This implies that for each identified topic In the topic modeling phase, it is possible to observe a network of interactions among users, allowing the study of the users' presence on multiple topics.

2.3 Polarization

As anticipated in chapter 1, it is fundamental for the understanding of this work to comprehend what Falkenberg did in his research [10]. Since the same methods are employed here, we also make the same assumption of the bipolarity of the polarization. In this section, it will be shown how polarization will be computed.

A meaningful metric that gained popularity among social scientists is polarization. The researchers believe that polarization can be harmful for the maintenance of the democratic stability [40]. Thus, understanding the phenomenon is important to develop a solution to it.

Polarization has been used to study the impact of political discussion on social media, especially around US presidential elections [41] [42] . Due to this, attention should be paid when generalizing, especially since US politics are built around two main parties -Democrats and Republicans-, making a bimodal view of polarization the best suited for this case (but not for all). Despite this, an analysis of the polarization over 21 different countries shows that the US is not the only place where it has been detected [43]. Ref [44] show the polarization can have different geometries than the traditional bimodality.

Even though Falkenberg detected an increasing polarization only in the COP26, while in 2015, Williams et al. [45] found the presence of echo chambers around the climate discussion in social media, with a small presence of open forums. In this work, there will be an attempt to connect these two pieces of research, to understand if, by breaking down the discussion into topics, the polarization of specific topics has always been high.

There is not a clear and universally adopted definition of polarization; Bramson et al. [12] tried to summarize it by defining different types of it; the assumption is that there is a measure of 'opinion' for each user:

- **Spread** defines the distance between the two extremes as the breadth of opinion

- **Dispersion** considers the shape of the distribution of the opinions and searches for peaks
- **Coverage** does not look at the shape but at the similarity of opinions within the groups
- **Regionalization** looks at the spectrum not covered between the groups
- **Community fractioning** is the degree to which the population can be broken into subpopulations
- **Distinctness** is defined as "the degree to which the group distribution can be separated"
- **Group divergence** is the opposite of distinctness, how different are the groups
- **Group consensus** shoes how people in the same group have similar opinions
- **Size parity** put relevance on the size of each group

We use the definition of Dispersion polarization that looks into the distribution of beliefs to detect peaks. Falkenberg demonstrated that the assumption of bimodality makes sense when dividing the population into climate supporters and climate skeptics. There is also the assumption that the polarization can be detected using the retweet network.

Latent ideology In order to compute polarization on a retweet network, firstly, a latent ideology for each user is to be estimated, as defined in [46] and adapted for retweets in [42].

Starting from the adjacency matrix of the retweet network, and after some linear algebra, a latent ideology score for each user can be obtained.

The first step is to identify m most retweeted users -from now on *influencers*- , out of the n users, and then build an adjacency matrix $A \in \mathbb{R}^{nxm}$ between users and influencers (where a_{ij} is the number of times user i retweeted influencer j). Now the process will be shown in detail from the matrix to the scores. In this way all the users that did not interact with the top m influencers will be excluded from the analysis.

Firstly, normalize A by the number of retweets:

$$P = \frac{A_{ij}}{\sum_i \sum_j a_{ij}} \tag{2.6}$$

Secondly, get the vector of row and column sums and consider the diagonal matrix:

$$\mathbf{r} \in \mathbb{R}^m, \quad r_i = \sum_i a_{ij} \tag{2.7}$$

$$\mathbf{c} \in \mathbb{R}^n, \quad c_j = \sum_j a_{ij} \tag{2.8}$$

$$R = \begin{bmatrix} \frac{1}{\sqrt{r_1}} & & \\ & \ddots & \\ & & \frac{1}{\sqrt{r_n}} \end{bmatrix} \quad C = \begin{bmatrix} \frac{1}{\sqrt{c_1}} & & \\ & \ddots & \\ & & \frac{1}{\sqrt{c_n}} \end{bmatrix} \tag{2.9}$$

Then, compute the matrix of standardized residuals S:

$$S = R\left(P - (\mathbf{r} \cdot \mathbf{c}^T)\right)C \tag{2.10}$$

Using Singular Value Decomposition (SVD), which is a factorization technique in linear algebra, the standardized matrix can be decomposed into three other matrices. It provides essential geometrical and theoretical insights about linear transformations and it is extensively used in various fields such as data science, engineering, and statistics [47]. Given matrix S, its SVD is written as:

$$S = U\Sigma V^T \tag{2.11}$$

where U is an $m \times m$ matrix whose columns are the orthonormal eigenvectors of AA^T, Σ is an $m \times n$ diagonal matrix whose non-zero elements are the singular values of A, and V^T is the transpose of an $n \times n$ matrix whose columns are the orthonormal eigenvectors of A^TA. The singular values on the diagonal of Σ are typically sorted in descending order. The columns of U and V are called the left-singular vectors and right-singular vectors of A, respectively.

Multiply R and U:

$$X = RU \tag{2.12}$$

Finally, rescale U on $[-1,1]$ and get the user score:

$$score = -1 + 2 \cdot \frac{X - \min(X)}{\max(X) - \min(X)} \tag{2.13}$$

Hartigan's diptest After computing all the users' latent ideology scores, to test the polarization, Hartigan's diptest is used [48].

Hartigan's Dip Test is a statistical test used to determine if a distribution is unimodal. The test works by comparing the empirical distribution function of the data, denoted as $F(x)$, to the unimodal distribution function that minimizes the maximum difference between $F(x)$ and itself, denoted as $G(x)$. The dip statistic D is then defined as:

$$D = \sup_x |F(x) - G(x)| \tag{2.14}$$

Where $\sup_x$ denotes the supremum (least upper bound) overall x, the unimodal distribution function $G(x)$ is chosen so that it minimizes this supremum. In other words, $G(x)$ is the "best" unimodal approximation to the empirical distribution function $F(x)$.

The null hypothesis of the Dip Test is that the data comes from an unimodal distribution. If the dip statistic D is significantly large, the null hypothesis is rejected and the conclusion is that the data is not unimodal. The p-value of the test is computed by comparing the observed dip statistic to the distribution of the dip statistic under the null hypothesis. This distribution is typically approximated using Monte Carlo simulations.

Chapter 3

Data Description

The content of this chapter will be an overview of the starting data used in this research, as well as some general statistics about it.

3.1 Source of the data

The data considered are tweets collected from the Twitter API containing the hashtags #cop21 and #cop26. For each cop, there are two jsonlines files, one for the tweets and one for the users involved. The fields are the ones stated in the documentation [1]; among the numerous fields, the most relevant for this work are the following:

Field	**Description**
author	The ID of the author
author_name	The username of the author
text	The text of the tweet
date	The creation date of the tweet
lang	The language of the tweet
conversation_id	The ID of the conversation the tweet belongs to
referenced_type	The type of the referenced tweet
referenced_id	The ID of the referenced tweet
mentions_name	The usernames of the mentioned users in the tweet
mentions_id	The IDs of the mentioned users in the tweet

Table 3.1 Description of the fields used of the tweets data

The user's file is used to map the ID of the users to their username, but when this information is not available, the user ID is treated as the username.

3.2 Data Statistics

The data considered originates from 2 COPs: COP21, COP26. All the tweets are in English and without links or image/video content. For this work, an 'original tweet' is a tweet written by a user, meaning that it is not a retweet. Fig 3.1 shows the distribution of the tweets over time for both cops; most of the tweets have been tweeted while the conferences were taking place.

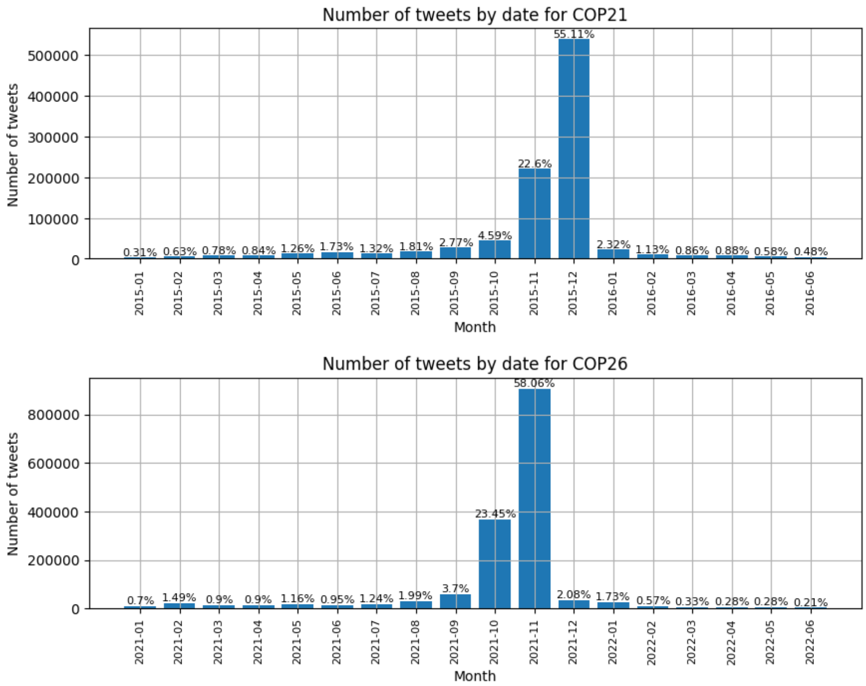

Fig. 3.1 Numeber of tweets by date for cop 21 and cop26

COP21 The tweets span from January 2015 to June 2016, but 77% of the tweets are from November and December 2021; cop26 was held between November 30th and December 12th. In the dataset, 975040 tweets were tweeted by 234389 users, of which only x tweeted an original tweet with at least one retweet; every user tweeted on average 4.16 tweets; the maximum amount of tweets a user tweeted is 9635, while 89% of users tweeted less than five tweets.

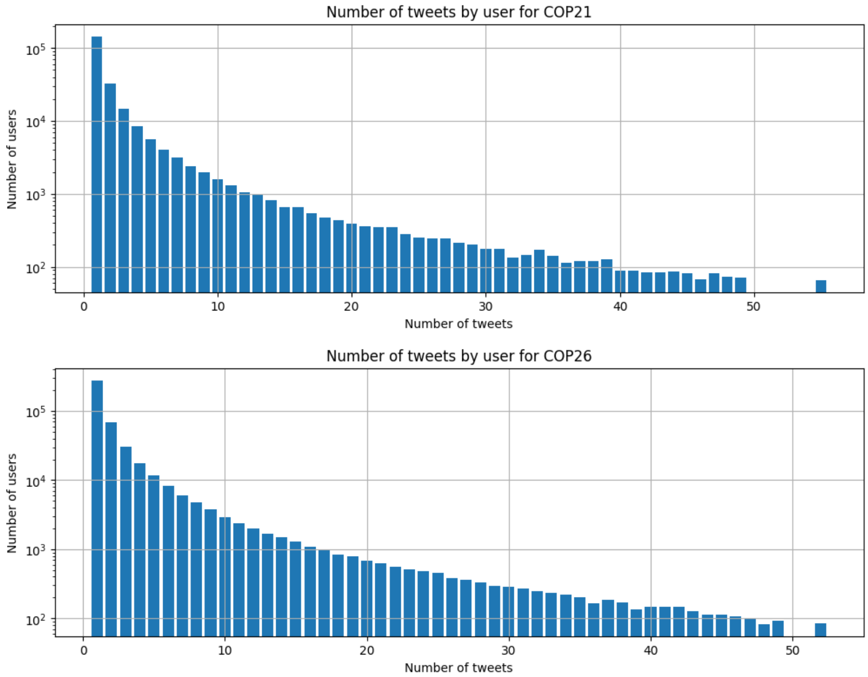

Fig. 3.2 Number of tweets by user for cop 21 and cop26

COP26 The tweets span from January 2021 to July 2022, but 81% of the tweets are from October and November 2021; COP26 was held between October 31st and November 12th. In the dataset, 1558968 tweets were tweeted by 456000 users, of which only 30195 tweeted an original tweet with at least one retweet. Every user tweeted on average 3.42 tweets, the maximum number of tweets a user tweeted is 14267, while 90% of users tweeted less than two tweets.

Fig 3.2 shows how most users tweeted just a few tweets (note that it is logarithmic).

	n_tweets	**n_retweets**	**n_original**	**n_original_with_retweets**
COP21	975 040	562 946	412 094	138 427
COP26	1 558 968	1 191 813	367 155	130 138

Table 3.2 Number of tweets

Fig 3.3 depicts how the 1'558'968 are distributed: 76% of them are retweets generated by only 130k original tweets. It is also worth noting that almost 2/3 of the original tweets have 0 retweets.

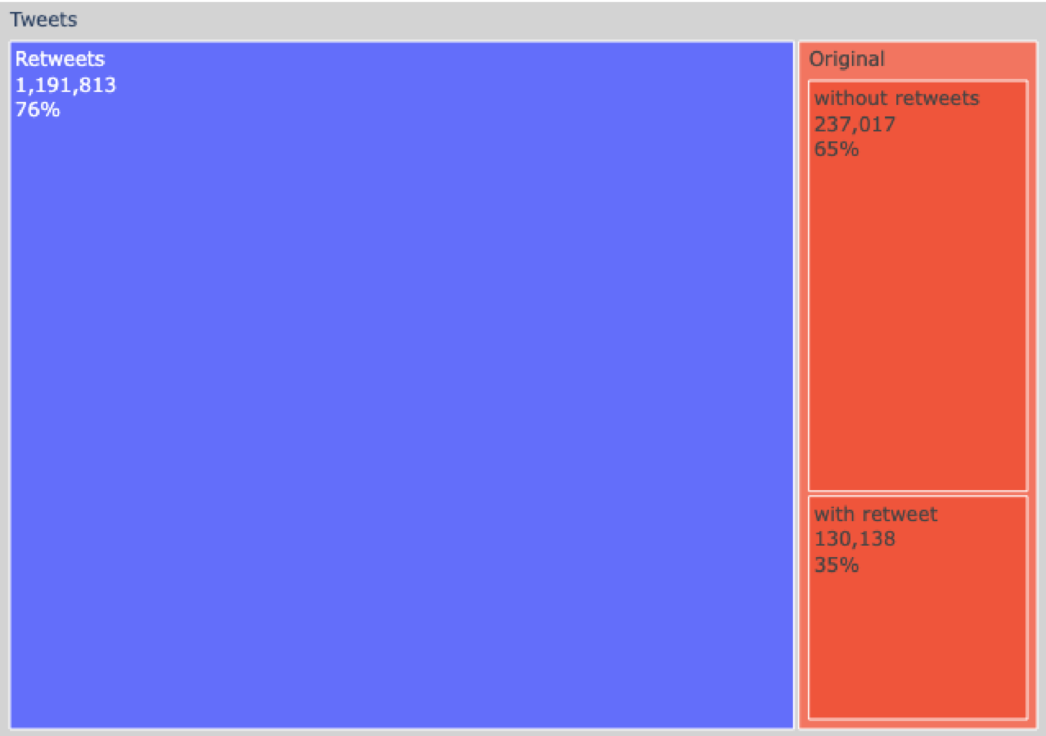

Fig. 3.3 tweets of cop26

Chapter 4

Methodology

This chapter represent the core of the original contribution of this work, starting with the evaluation of the different topic modeling algorithms used in order to choose the one for this research. Then the focus will shift to the creation of the multilayer network and to the polarization calculation.

4.1 Topic Modeling Evaluation

This section presents the evaluation of different models used for tweet labeling. Both unsupervised and supervised approaches were used to evaluate the performance of the models. The evaluation aimed at finding the best-performing model to label tweets accurately. The models included traditional methods (LDA, GSDMM, and NMF) and neural models like BERTopic.

The unsupervised evaluation (4.1.1) assessed the traditional metrics used in this context, such as coherence and diversity scores. The results showed that BERTopic performed better than traditional methods, especially when using all-MiniLM-L6-v2 (BERT) [1] , text-embedding-ada-002 (OpenAI) [2], and tweet_classification [3] embeddings.

The supervised evaluation consisted in building a custom-labeled dataset from scratch, and then in looking at the accuracy of the models. The results showed that BERT and OpenAI were the best-performing models. The section culminates with a summary of the results and a description of the representation used for labeling the tweets.

The models used are both traditional (LDA, GSDMM, NMF), as a reference of the ground truth, and neural, because they seem to be the most accurate; in particular, BERTopic will

[1] huggingface.co/sentence-transformers/all-MiniLM-L6-v2

[2] platform.openai.com/docs/models/embeddings

[3] huggingface.co/louisbetsch/tweetclassification-bf-model

be evaluated with several embedding methods. BERtopic was chosen over Top2Vec both because they are very similar, and because the Python library is more comprehensive and allows for more flexibility.

Evaluating a topic modeling algorithm is challenging due to the lack of objectivity in identifying a topic. In this work, the models were evaluated in two ways: firstly, by employing a widely used unsupervised approach: metrics like coherence and diversity. Then, to validate the results, a supervised evaluation was also carried out by using different datasets built ad hoc for this setting.

4.1.1 Unsupervised

To compare the different models, a library suggested by the creator of BERtopic called OCTIS [49] [50] was used; this allowed the creation of an experiment to measure the different metrics presented in chapter 2: coherence and diversity.

Dataset In this case, the dataset is composed of 1669 preprocessed tweets related to climate change with the hashtag *#cop22*, all in english; the preprocessing phase involved removing retweets, links, punctuation, and the most common hashtags (#cop22, #climatechange #p2). We choose this dataset, which is not the one used in the final analysis because the domain is the same, and the goal analogous, since we have to detect subtopics of a main topic which is climate change. For this reason we can generalize the result to the tweets of the other COPs.

Methods The models used in this evaluation were LDA, NMF, and BERTopic. In the BERtopic case, several embeddings have been tested (all-MiniLM-L6-v2, text-embedding-ada-002, climatebert [51], tweet_classification, USE [52]).

Each model has been fitted several times changing the parameters:

- **number of topics** from 10 to 50 with a step of 5
- **min topic size**: 5 and 15 tweets[4]

Each unique combination of parameters has been fit three different times; then, we took the mean value of the three computations.

Results The results show that BERtopic performs way better in these tests than the traditional methods. In comparison, the best Bertopic embeddings are mini, OpenAi, and tweet

[4]only for bertopic

classification. The experiment demonstrates how the *min_topic_size* value of 5 is too small, thus the results will be gathered considering a value of 15.

Fig 4.1 shows the value of all the metrics with a different number of topics for the traditional methods and the best-performing neural one (OpenAI)

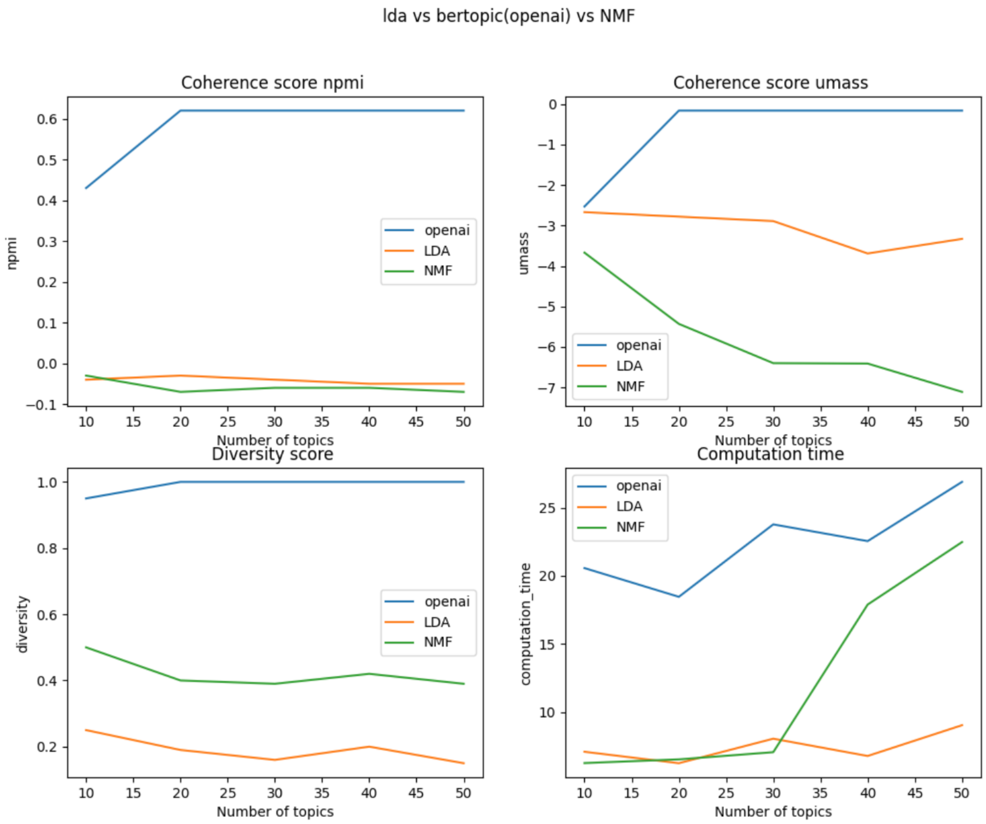

Fig. 4.1 coherence and diversity for LDA, BERTopic, and NMF

However, Hoyle et al. [53] explained how these metrics are not very meaningful for the evaluation of these models due to the lack of validation for neural models, therefore these results should be taken with a grain of salt.

From this evaluation, it can be concluded that BERTopic's topic size is better bigger than smaller, especially if many documents are available. Topics of Bertopic are way more diverse than the ones of LDA and NMF and, within the topics, the most relevant words are more semantically related.

model	npmi	umass	diversity
tweet_classification	0.62	-0.16	1
openai	0.58	-0.63	0.99
climatebert	0.20	-2.28	0.83
U.S.E	0.20	-4.35	0.89
BERT	0.20	-5.46	0.97
LDA	-0.04	-3.07	0.19
NMF	-0.06	-5.80	0.42

Table 4.1 all the models tested in the unsupervised evaluation

4.1.2 Supervised

Considering the result of the unsupervised evaluation, another method should be used to validate what was found. In this case, two ad-hoc datasets were created to see how the models perform in a real-case scenario.

Dataset The first step for the supervised phase was the data collection. In this case, specific datasets were packed to test the chosen models. The datasets were selected based on the trending topic on Twitter at the time (March 2023). The first dataset is simpler and contains very different topics, it follows that it should be effortless enough to cluster the documents. In contrast, the second dataset is trickier because it includes only politics-related tweets, including some which overlap with more hashtags related to US politics.

- **simple**: 1093 labeled tweets of 5 different topics identified by a hashtag [5]
- **politics**: 1492 labeled tweets of 7 politics-related hashtags [6]

For both datasets, we used two different versions: with and without hashtags. The reason beheind this was to avoid the risk of the model to cluster based on the hashtags.

The tweets have been extracted using twarc2, getting only English tweets and excluding retweets. Only the trending topic at that time (March 2023) was used because Twitter API did not allow access to data before 48h.

Metrics In order to evaluate the topics, it was necessary to define some accuracy metrics, which is not a straightforward task because, by using BERTopic, there is no set number of topics apriori, but it has to figure it out by itself. After running the model, both the known

[5] #Bitcoin, #stormydaniels, #UkraineRussianWar, #SaudiArabianGP, #climatechange

[6] #IndictArrestAndConvictTrump, #kabul, #BidenHarris2024, #KamalaHarris, #taiwan, #belarus, # stormydaniels

topic (the hashtag) and the inferred one (a number) are established, making it possible to create a confusion matrix between the two sets. To map the inferred topic to the known one, the inferred topic with the highest value was considered. In the case at hand, the error can be detected by combining this value with another metric: min_topic_size.

The metrics defined are the following:

- **Accuracy**: for each known topic, take the biggest inferred topic of the same row of the confusion matrix and divide by the number of tweets in that topic, fig 4.4 can be taken as reference.
- **Accuracy no outliers**: in the Bertopic case, the label -1 refers to outliers. Therefore, to achieve it, compute the same as for accuracy but without counting the outliers.
- **Min_topic_share**: same as for accuracy but in the opposite direction, after having computed it for all of inferred topics, the minimum is considered. This is helpful to detect when the accuracy is considering the wrong topic, which could happen when the inferred topic size is less than the actual topic one (so one inferred topic contains tweets from multiple topics, and then this number is low)

Parameters max_df is used to remove the terms that appear too frequently; a value of 0.95 means remove the terms that appeared in more than 95% of documents, while min_df is the opposite; in this case, with the value being an integer, the parameter refers to the minimum number of documents a term should be in to be considered.

$Alpha$ is a parameter that influences the number of clusters that will be created; low alpha results in many clusters with single words, while high alphas results in fewer clusters with more words. For these 3 the default settings were used.

ngram_range defines the number of consecutive words to be considered: for example, a value of (1,2) tells the method to consider single words and bigrams (two consecutive words), but including threegrams was too computationally expensive.

A min_topic_size of 50 was made because 200 tweet for each hashtag were inspected and, considering that some of them will be classified as outliers, a size of 50 was a safe trade-off in order not to risk that a topic nor the size were too small since, from unsupervised evaluations, it was concluded that the bigger the size, the better.

```
BERTopic: (nr_topics = 'auto', min_topic_size = 50)

NMF: (max_df = 0.95, min_df = 3, ngram_range = (1,2))

GSDMM: (alpha = 0.1, min_df = 0.1, n_iters = 30)
```

Simple Dataset Results At first, the evaluation was on the simple dataset with hashtags. As Fig 4.2 shows, base (all-MiniLM-L6-v2) and OpenAi obtained almost a perfect score for each topic. At the same time, climatebert seems to have a great accuracy but a low mean topic share, this is a signal that something is wrong and the heatmap should be inspected.

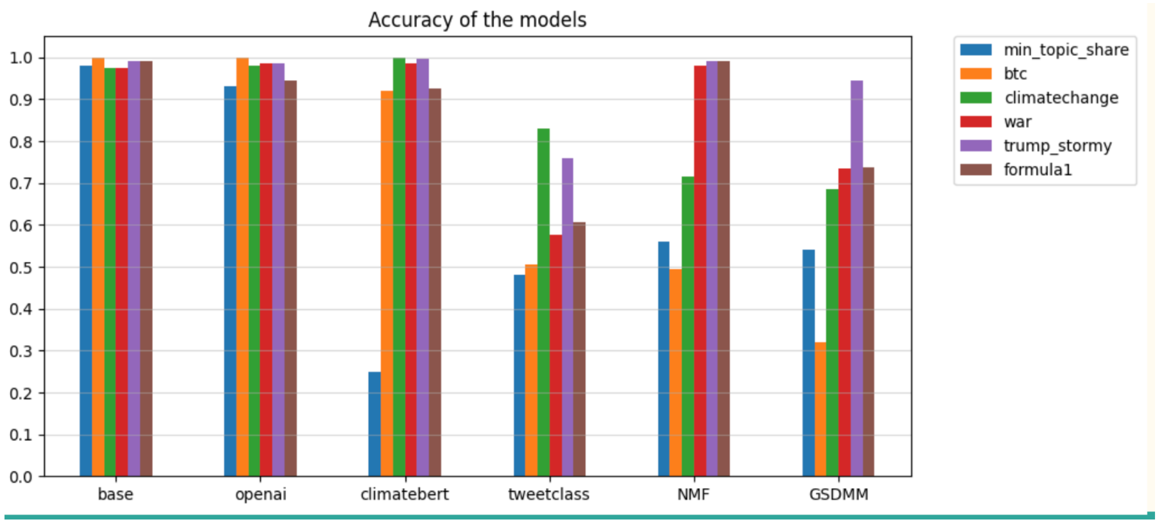

Fig. 4.2 All models accuracy simple with hashtags

In fact, it is clear from 4.3 that even though the accuracy is very good, climatebert has some difficulties in dividing the topics, putting almost all the tweets in the same inferred topic. While the first two are performing very well as expected, the same is not true for the others. Climatebert puts almost all the tweets in topic 0, thus being able only to find the formula1 tweets and not the climatechange one, as it is designed to do. That's the reason why the decision was to remove the models that were not performing well in the simplest case, with the exception of NMF to use as ground truth.

Fig 4.4 shows how BERT and OpenAI performed in the simple dataset but without the hashtags and, in particular, how BERT tends to find more outliers than OpenAI. Overall, both get a good performance.

An interesting feature of Bertopic is the ability to visualize the different topics in 2-dimensional space; Fig 4.5 shows the document distribution of OpenAi after reducing the dimensionality of the embeddings .

Politics dataset results The politics dataset is clearly more difficult to evaluate, but with the hashtags, it is still doing a good job. Fig 4.6 4.7 shows heatmap and topic distribution for the politics dataset with hashtags.

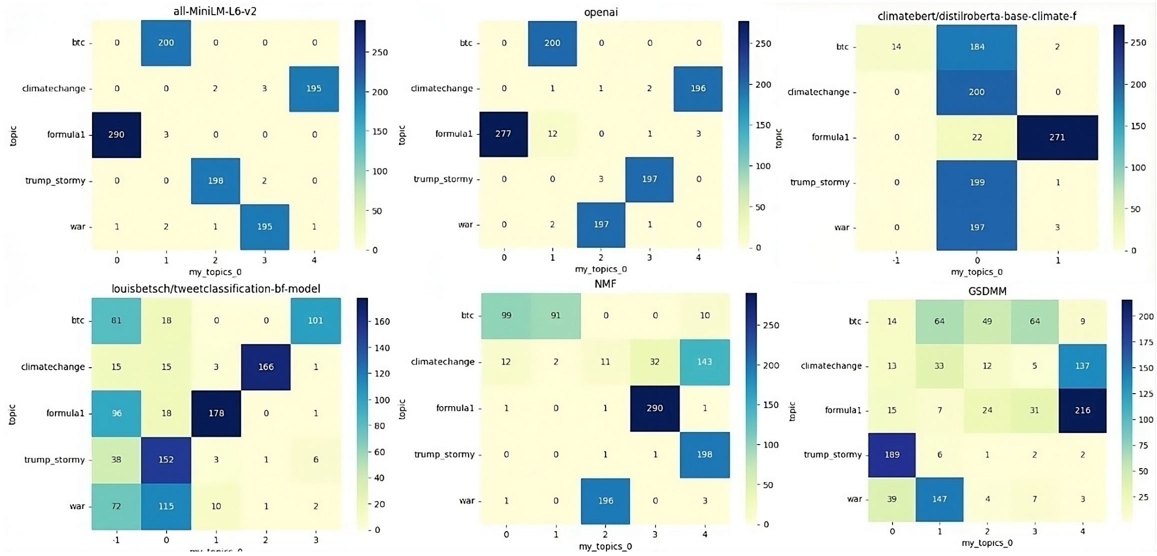

Fig. 4.3 Heatmap comparison of the different model with the simple dataset with hashtags

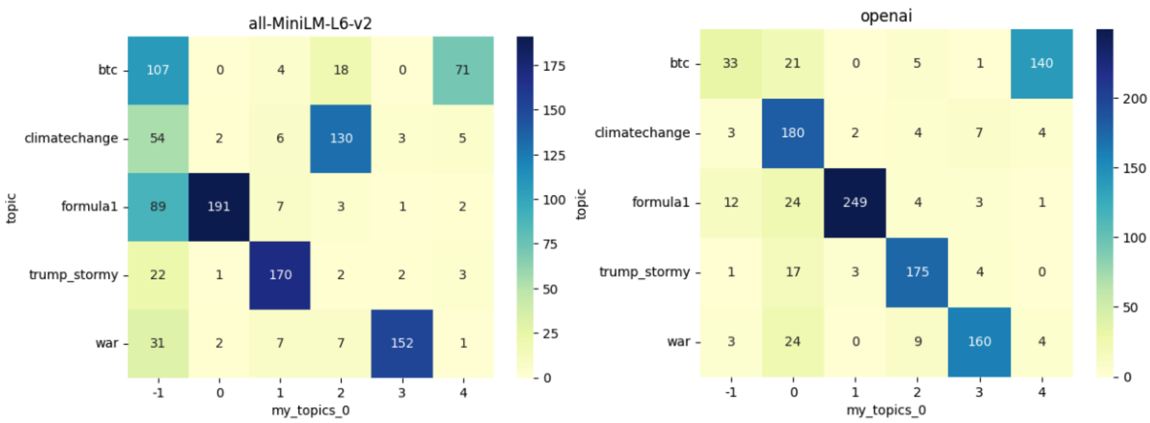

Fig. 4.4 Heatmap comparison of mini and OpenAi of the simple dataset without hashtags

Both BERT and OpenAi are creating two topics from the Taiwan case. OpenAi merges two topics, which is coherent with the fact that the two hashtags related to Trump are also related to the same event(#IndictArrestAndConvictTrump and #stormydaniels).

To validate the results, the algorithm was run 100 times and, most of the time, for BERT the min topic share was 0.9, which means it got the correct number of topics and classified them in a good way.

In the case without hashtags, OpenAi and BERT put in a single cluster all the tweets related to American politics, both also understanding that Kamala's tweets were about something else.

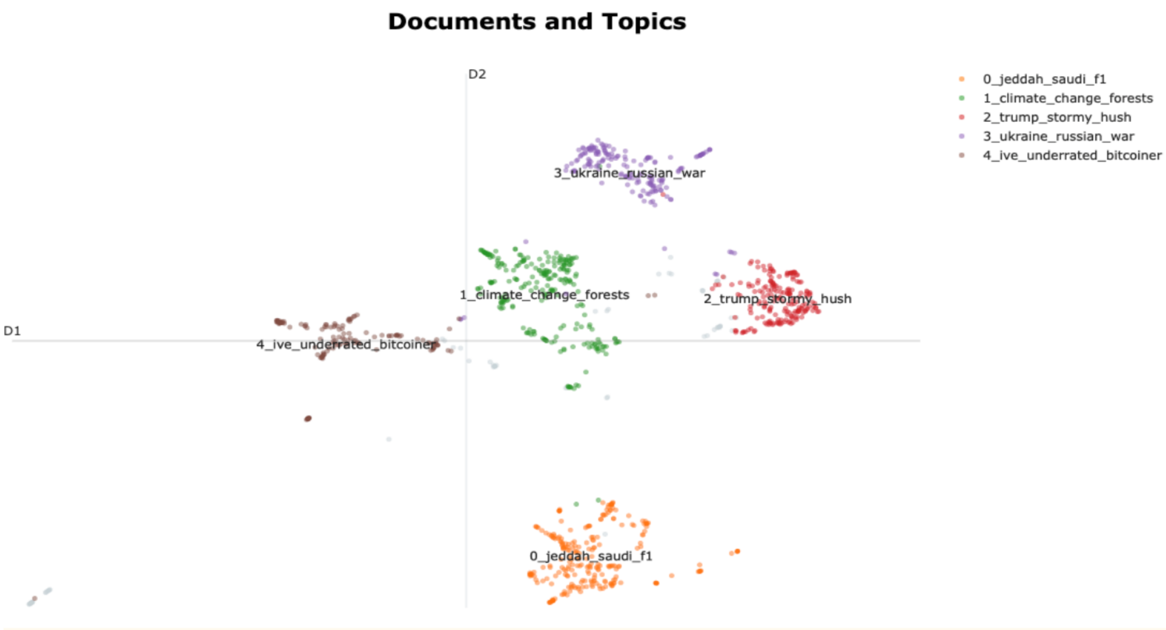

Fig. 4.5 docs representation of simple dataset for openai

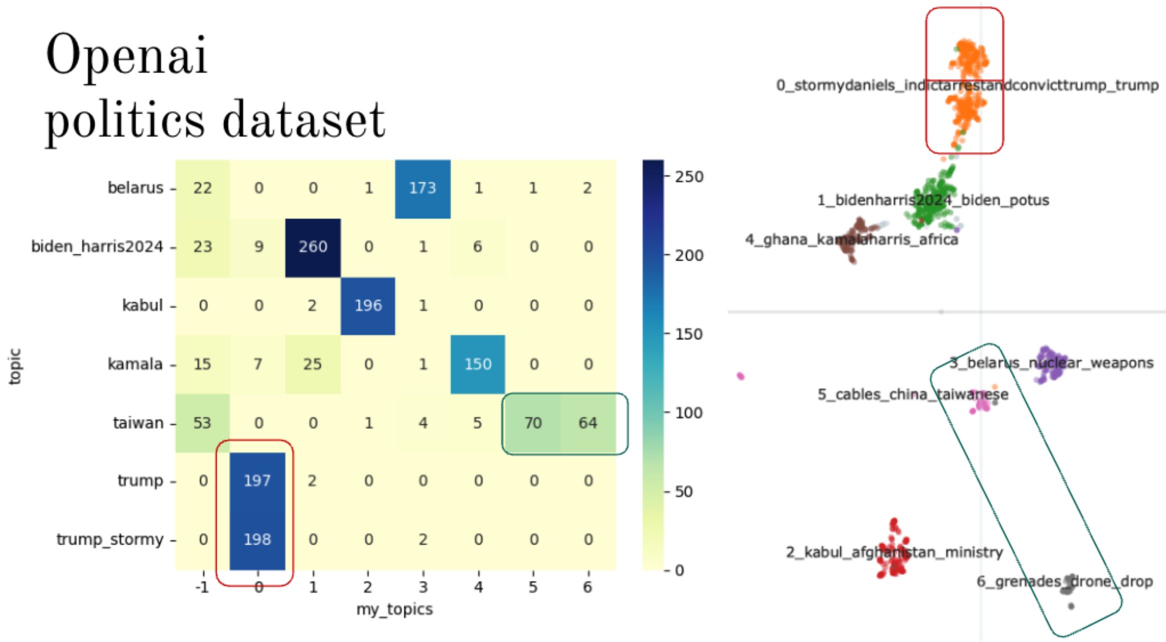

Fig. 4.6 Heatmap and documents representation of the politics dataset with hashtags evaluated with openai

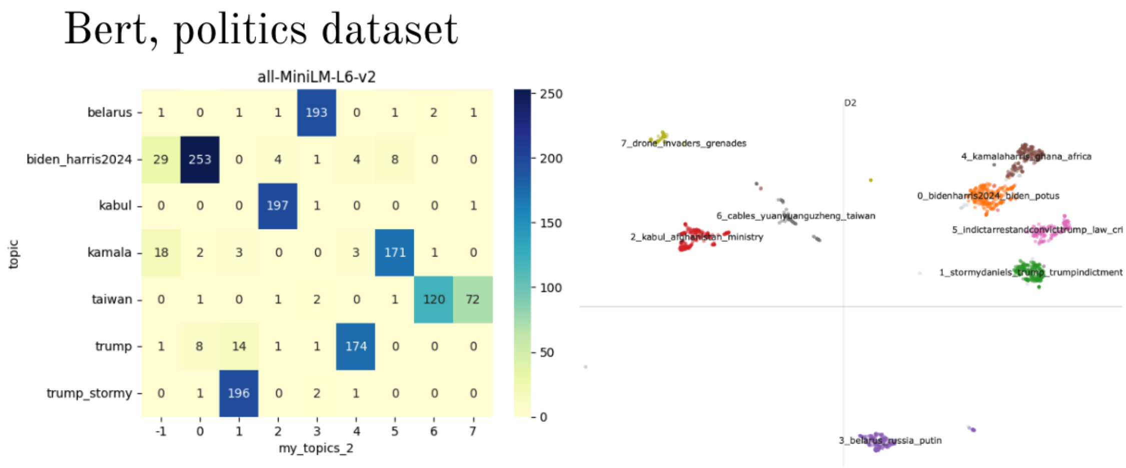

Fig. 4.7 bert heatmap and document viz politics with hashtags

Topic representation The last step is giving a meaningful label to the clusters created, so that it can be seen if the openai API works well; in particular, the model named *gpt-3.5-turbo* was used, which worked surprisingly well; table 4.2 contains the label generated for both simple and political datasets.

Fig 4.6 as reference. Note that the discussion under the #taiwan hashtag is divided into two different topics, as depicted in the document representation.

The prompt used is: *you are a tweet labeler, you are given representative words from a topic and three representative tweets, give more weight to the words, given all this information give a short label for the topic (max 10 words), starts all with topic:*

Conclusion Tab 4.1 shows the result of unsupervised evaluation, while Tab 4.3 shows the supervised. Through the supervised evaluation, it is demonstrated how the results of the unsupervised one were not completely true; this helped to discard some models and confirm the hypothesis that the neural model with Bert Embedder and Openai was the best performing model. There is still a significant difference between the two: Bert is open source and can be run locally, while Openai is not free and can only be used through API.

4.2 Multilayer Network

In this section, the analysis will dive into all the unexplored paths, starting from Falkenberg's work. In particular, the same polarization analysis will be conducted on a topic level: instead of computing it at a full network level, a retweet network was created for each topic, so the

simple dataset	**label**
#Bitcoin	Cryptocurrencies
#stormydaniels	Trump’s hush money payment to Stormy Daniels.
#UkraineRussianWar	Ukraine-Russia conflict
#SaudiArabianGP	F1 Saudi Arabian Grand Prix 2023
#climatechange	Forests and Climate Change
politics dataset (openai)	
#IndictArrestAndConvictTrump	Stormy Daniels controversy
#stormydaniels	Stormy Daniels controversy
#kabul	Suicide bombing near foreign ministry in Kabul
#BidenHarris2024	Politics and Leaders
#KamalaHarris	Kamala Harris official visit to Ghana and Africa
#taiwan	Tensions between China and Taiwan over undersea cables cut
#taiwan	use of small drones for warfare’

Table 4.2 labels generated using GPT API both for simple and political dataset

model	accuracy	topic share
BERT	0.84	0.86
OpenAI	0.83	0.85
NMF	0.78	0.69

Table 4.3 recap of supervised evaluation

topics that were driving the polarization of cop emerged. Furthermore, the aim was also to explore how the polarization of topics evolved over time.

By reading the previous section, the concept of topic modeling is known, and so is the way the main models perform. The goal is to create a multilayer network where every layer represents a topic.

In order to do so, a Python library was developed, which can be used as a toolbox, starting from the tweets fresh out from the official API of Twitter. The design is modular and can achieve different goals. In fact, even if the interest is only in the retweet network of the users (nodes are users, ties are retweets), this framework can be used to perform different tasks, but to the purpose of this study the tasks highlighted are:

- Clustering tweets according to their topic
- Giving a meaningful label to the clusters
- Creating the retweet network (global and multilayer version)

The steps are independent, so, for example, it is possible to create the network without the need to run the topic modeling part.

Steps Even though some steps are avoidable, the natural and minimal pipeline follows these steps:

1. from JSON to a tabular format
2. label each tweet with a topic
3. create multilayer retweet network

4.2.1 Process input

The first step consists of the transformation of the JSON objects into tabular data to optimize the space and handle the data in an easier way with pandas. This is also helpful to save space; in the case of COP26, a shift is made from a 14 GB JSON to a less than 2 GB CSV, since most of the fields are not relevant to this study.

In this process, all the tweets with attachments and not in English are discarded. The tweets are divided into multiple dataframes: one for original tweets - i.e., the ones that the author actively writes-, and one for the retweets.

At the end of this stage, a CSV and pkl file are saved in case somebody needs the tweets in tabular data. If the script is re-run and these files exist, they will be loaded instead of running the process again.

4.2.2 Topic modeling

As extensively discussed in chapter 2, in this segment of the pipeline the tweets can be labeled using Bertopic, with the possibility to choose the embedder; the one used in this research is *all-MiniLM-L6-v2*, the one evaluated in the previous section.

This step is the most computationally expensive; for this reason, to avoid redundancy, the topic modeling has been run only using original tweets.

After this step, all the original tweets are labeled with a numeric topic, and then the label has been propagated to all the retweets so that the entire dataset is now labeled with a topic.

At this point, the OpenAI API can be used to give a meaningful label to the topics, while beforehand, the labels were just made of the most relatively frequent words of the topic. By using the langchain library, it is possible to structure an employable prompt. The model was assigned the words identified with TF-IDF and 3 representative tweets sampling a subset of the documents in each topic, and calculated on the cosine similarity between TF-IDF representations.

This is the one used:

I want you to act as a tweet labeler, you are given representative words from a topic and three representative tweets, give more attention to the words, all the tweets are related to climate change and COP, there is no need to mention them, detect subtopics. start with "label:" and avoid hashtags, which is a good short label for the topic containing the words [words]?, here are 3 tweets to help you: first = "tweet1", second = "tweet2", third = "tweet3

Similarly to the previous stage, the labeled dataset is saved in the cache folder both in CSV and pkl. The model and the labels are saved, too.

4.2.3 Network

In this phase, after labeling the tweets, a retweet network will be created for each topic, gathering it all together in a multilayer network.

In the process of the creation of the network, there are retweeted tweets that do not have the original one, so they are discarded.

The last step is the creation of a multilayer network using the multinet library developed by Uppsala University; Every layer is a retweet network of a specific topic. It was built starting from the subset of tweets of a specific topic, the unique users are the nodes and, if user A retweets user B there will be a tie $A \rightarrow B$. The network is directed and weighted with the number of retweets. Fig 4.8 is a simple example of a network with 2 layers.

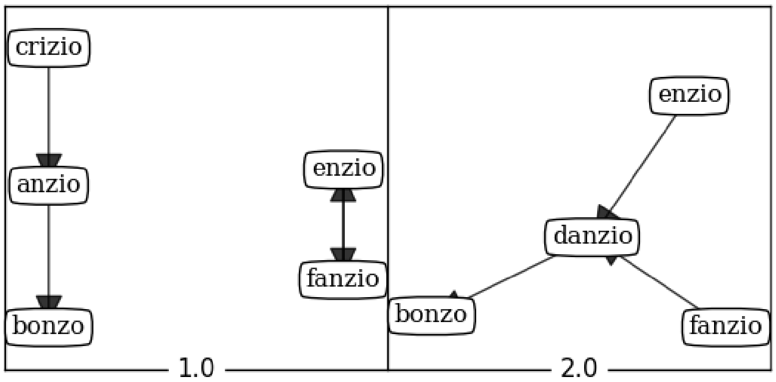

Fig. 4.8 Example of multilayer retweet network

The network then has been filtered by removing the outlier layer, labeled with -1.

COP26 network For COP26, 70 layers were detected, the average number of users per layer is 8127, min is 15, max is 108829,

COP21 Network For COP26, 36 layers were detected, the average number of users per layer is 10123, min is 696, max is 142062,

4.3 Polarization

At this point, for each layer, it is possible to compute for each user a latent ideology score, and then, using Hartigan's diptest, a polarization value can be assigned to each topic. More details on how this is computed are dealt with in the related works. In this process, there are some parameters that can be adjusted: the number of influencers - defined as the most retweeted users-, and n - representing the minimum number of retweets a user should have made to an influencer to be considered-.

The ideology score is not computed on all the users, but when selecting the influencers, the users are limited to the ones that retweeted at least n times those influencers.

In order to have enough data to be analyzed, the values set were $n\_influencers = 100$ and $n = 2$. The hartigan diptest requires a minimum of nodes to be statistically significant and, since at this point the layer are filtered, all the layers with a p-value of the diptest higher than 0.05 have been discarded.

COP26 The number of networks goes from 70 to 30 . The total influencers considered in the analysis of COP26 is 1698. 22302 users have an assigned score for cop26, with an average of 1311 actors per topic, with a min of 151 and a max of 7764.

COP21 the number of networks went from 36 to 8. The total number of influencers considered in the analysis of COP21 is 392. 8020 users have an assigned score, with an average of 2058 actors per topic, with a min of 35 and a max of 7524.

Tab 4.4 presents a summary of the starting networks

Description	**COP21**	**COP26**	**COP2x**
Initial topics	36	70	54
Final topics	8	26	29
Influencers scored	392	1557	1559
Users scored	8020	22161	33312
Mean users/topic	2058	1311	1766
Min users/topic	35	151	123
Max users/topic	7524	7764	14504
diptest	1	0.07	0.05

Table 4.4 Summary of Latent ideology

4.4 Logitudinal analysis

In order to see topic polarization over time, we need to run the topic modeling with all the tweets. Still, there are too many, so instead of taking the original tweets of COP21 and COP26, we only take the original with at least one retweet which are around 1/3 of the total but are the one needed to create the rest of the network.

Then the two dataset have been merged, we will refer to this dataset $COP2x$ and the same process described below has been done.

Chapter 5

Results

Let us start drawing a big picture of the different topics that emerged after the latent ideology analysis.

Fig 5.1 summarizes the size of the networks and the diptest results -the polarization- for each topic. The first thing to note is the fact that the highest polarization corresponds to the smaller topics, with the exception of topic 14; For the first two topics the polarization is significantly higher than the others. The average diptest score for COP26 is 0.069.

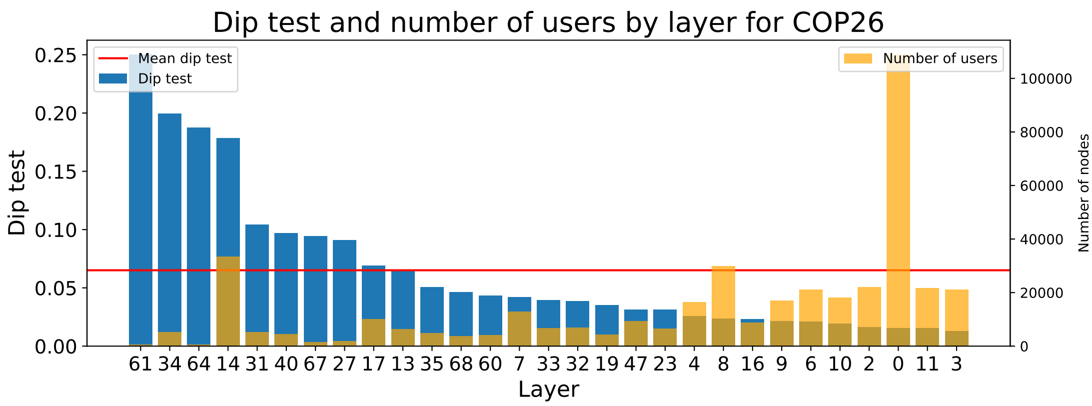

Fig. 5.1 Dip test and number of users for cop26 topic by topic

5.1 RQ1 Most Polarized Topics

Fig 5.2 demonstrates the most and the least polarized topics. Every network represents the retweet network of the 100 biggest influencers; in the leftmost plots there is the full network, while in the rightmost only the influencers are present. In the most polarized topics it can be clearly seen how the influencers are almost equally split between the two poles. Falkenberg,

user	**tweet**
User1	This meme could just as easily apply to Canada. Trudeau's willingness to destroy our economy to the benefit of others is akin to cutting off our noses to spite our faces!%.
User2	#COP26 Maybe some people are still fooled by Justin Trudeau and his dishonest climate change stories, but there are plenty of us here in Canada who are not. Look into the truth about the Lytton fire. It won't come out of Justin Trudeau's mouth
User3	Capping emissions in the country while exporting oil, gas and coal out of the country. Hypocrisy.

Table 5.1

in his work, identified a majority of pro climate users and a minority of climate skeptics but, looking at a topic level, the two groups are equally split and the networks that present the majority-minority dichotomy are the least polarized. A notable example is topic 34: "Canada Climate Change goals", which refers to the decision of Canada's prime minister to cap gas and oil emissions. This announcement caused much disagreement, displayed in Tab 5.1 with some random tweets against the decision, User1's argument is that this will destroy the economy; while User2 is instead generally against all the decision of Trudeau, as stated in her biography : "Lover of gardening, antiques and anyone who wants to see the end of the Trudeau government." This follows the typical elite polarization pattern, where political exponents strictly adhere to their party policies. In this case, it is not a political party but a politically-aligned individual, which is, in some way, forced to follow her self-imposed guidelines in her biography to avoid cognitive dissonance [54]. Looking at the right side of 5.2 we can see how the influencers, whom we can use as a proxy to the elite, tend to be more polarized than the 'mass'. This is especially evident in topic 14, related to air travel, where the influencers are lined up, while the normal users are more distributed in the spectrum.

It is interesting to note in Fig 5.3 the distribution of the tweets of each topic over time during the COP, with the dotted line marking the start and end date of COP26. The most polarized had interest only for a few days, quickly losing interest. The opposite happens in the least polarized topics where the discussion is distributed over a longer timespan.

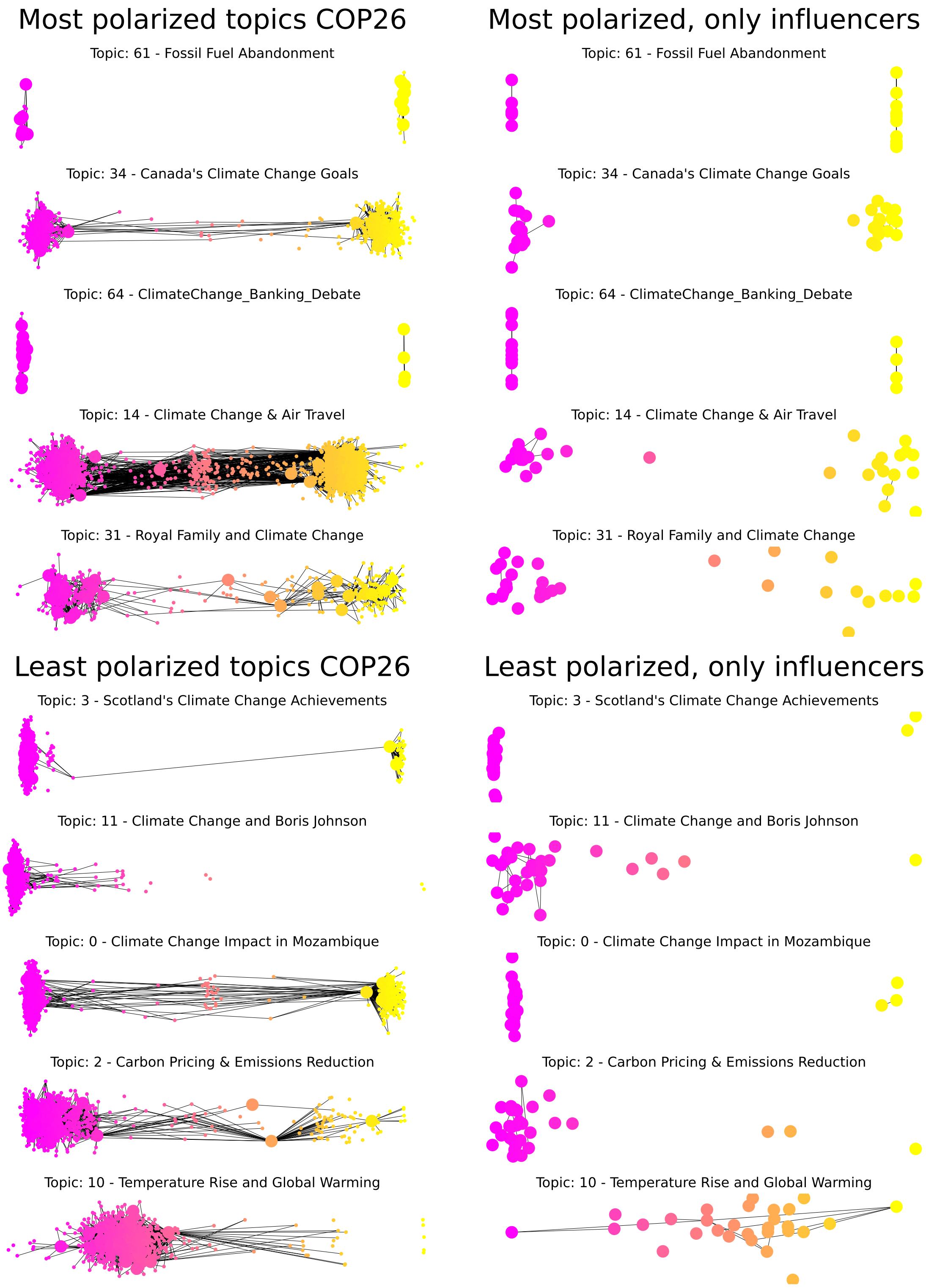

Fig. 5.2 Most and Least polarized topics in cop 26; on the right side the full network, on the left side the influencers network.

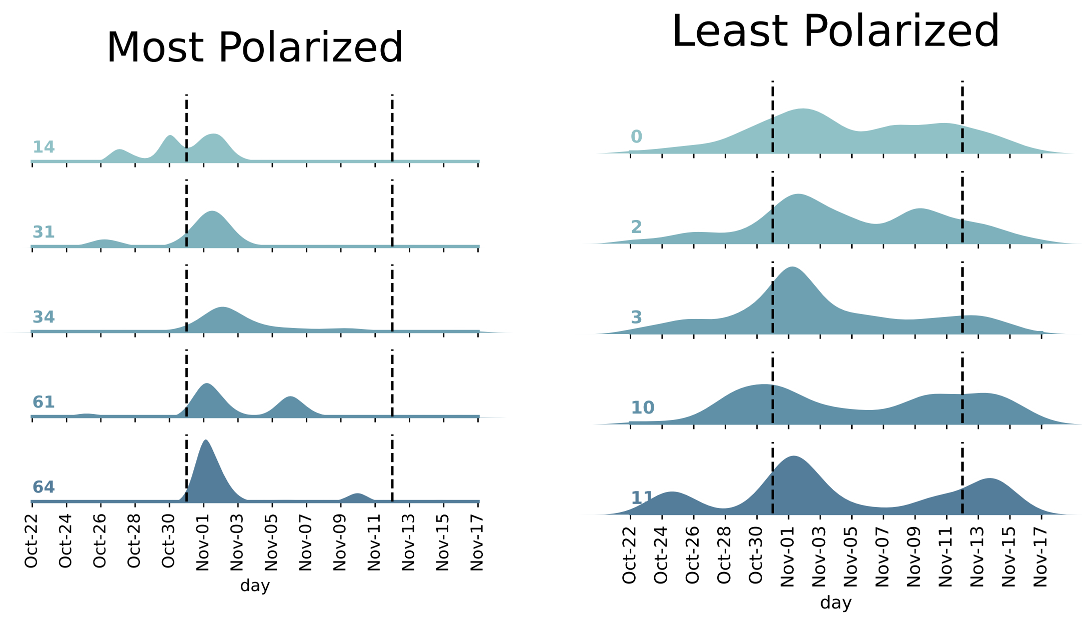

Fig. 5.3 Distribution of tweets belonging to the most and least polarized topics during the days around COP26

5.2 RQ2 Longitudinal analysis

In this section a comparison is made about the polarization between COP21 and COP26 and, to do so, it was necessary to run the topic modeling together. In Fig 5.4 the diptest results are shown, and a similar trend to that of the cop26 is evident: the biggest topics are less polarized, with the exception of topic 9. Fig 5.5 helps understand the share of the tweets between cop21 and COP26. It is also interesting to see that topic 9, which is the most polarized, is composed mostly by tweets from COP26, being therefore aligned with the literature. Now the focus of this work will shift to some topics, creating the network of retweets for both COPs, which will then be compared. The computation of this analysis is possible only for topic 1,3 and 12, which are the ones that are polarized and have enough tweets in both cops to run latent ideology.

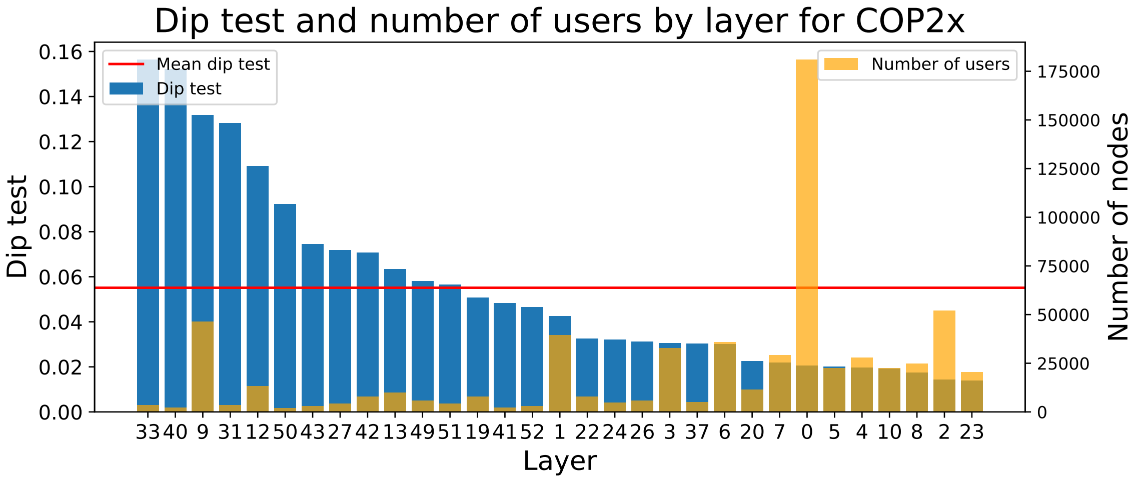

Fig. 5.4 Dip test result for COP2x, containing both tweets of COP21 and COP26

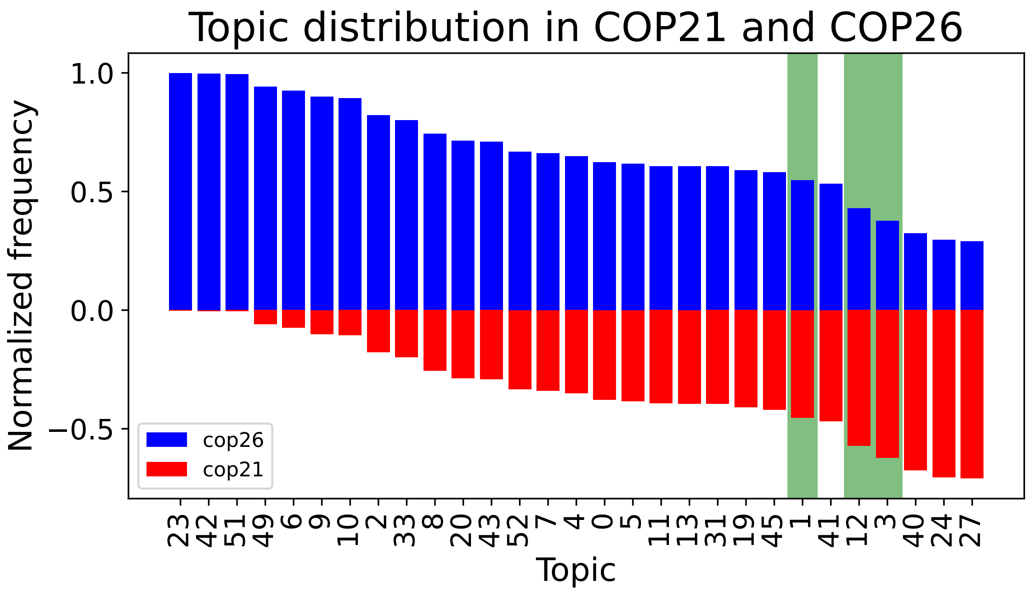

Fig. 5.5 The share of the tweets, topic by topic, between COP21 and COP26

Fig 5.6 shows the results of this analysis. It is worth noting how topic 12 is the topic dealing with Canadian fossil fuels, discussion that was present in both COPs, but with a very different level of polarization. Overall, the results confirm the hypothesis that COP26 is more polarized than COP21, but these are just three topics, thus the same analysis with more data should be run to confirm it.

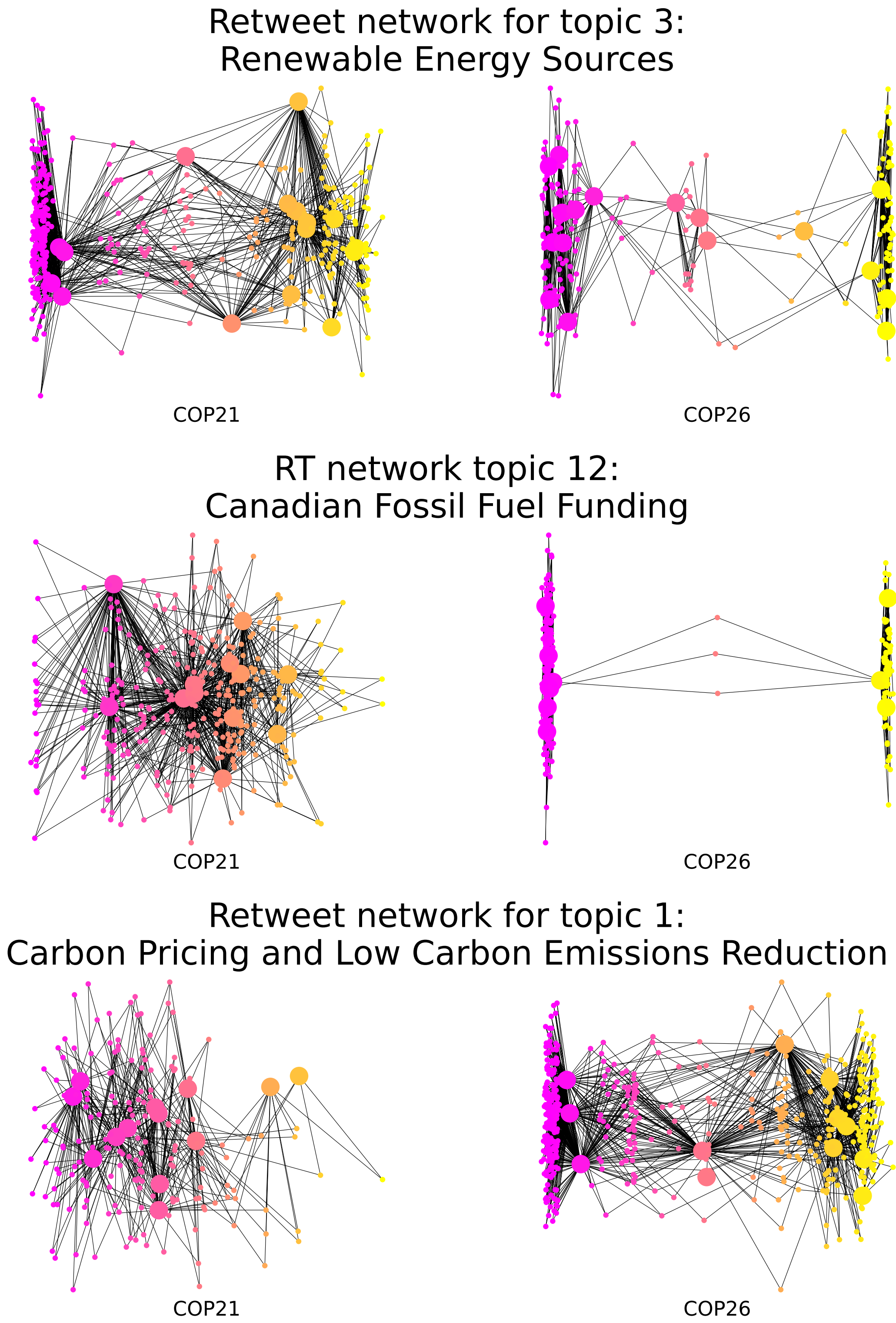

Fig. 5.6 A comparison of the retweet network about the same topic for COP21 and COP26

5.3 RQ3 User polarization among different topics

After computing the polarization score for all users, it can now be analyzed whether the users are polarized in the same way among all the topics they were active in.

The number of users involved in this analysis is 22161, active in 26 topics. Most of them (16141) were only active in one topic, while the maximum is 23, and the average is 1.53 topics per user.

Then, the average and the standard deviation of the score was computed for each user present in more than 1 topic. This value is higher for the users that are present in both sides of the spectrum, allowing for the identification of the degree to which users tend to be monopolar.

Fig 5.7 shows how the distribution of the average score for every topic aggregated together. This matches the global results of Falkenberg, where a majority is present on the 1 side versus a minority in the 1 side.

In the distribution of the std(standard deviation) we can see how there is a strong tendency to stay in the same side of the spectrum.

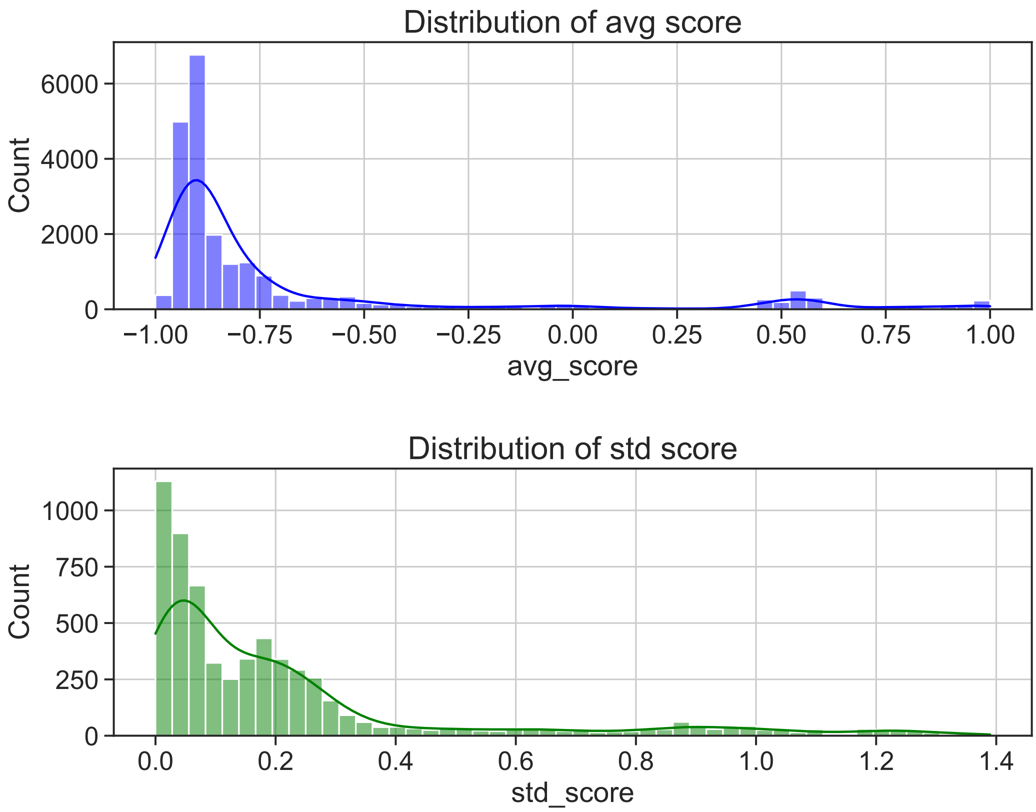

Fig. 5.7 distribution of the mean and the standard deviation of users' score

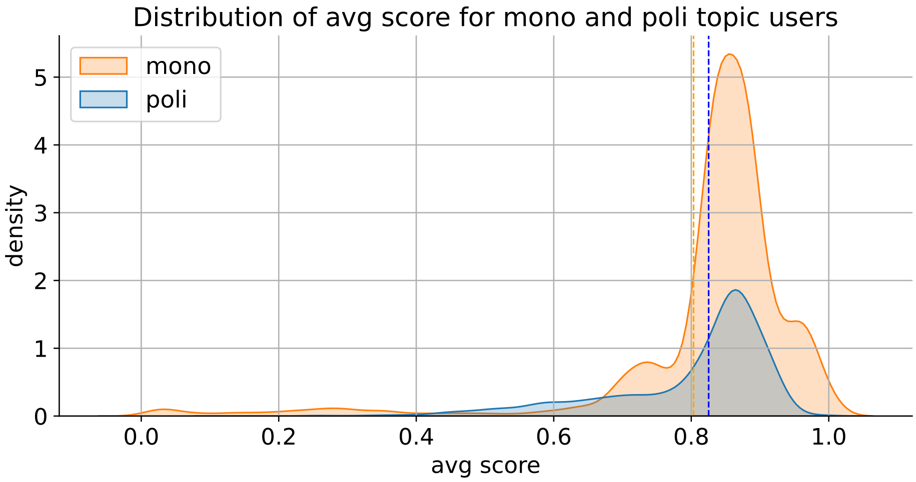

Fig. 5.8 A comparison of the absolute value of the average score both for users that are present in one topic(mono) and on multiple (poli)

5.4 RQ4 Polarization of mono vs poli topic users

Out of the 22 thousand users, most of them (16 thousand) are present only in one topic; out of the 6000 present in more than one topic, half of them are present in only 2. Fig 5.8 does not show a big difference in the ideology score between mono and poli users (0.8 for monotopic users, 0.82 for politopic). The computation of this score considered only the absolute value.

Chapter 6

Conclusion

The goal of this research was to investigate the polarization of users in the climate change discussion around the Conferences of Parties, specifically focusing on the 26th Conference. Unlike similar studies, this analysis was conducted topic by topic using a multi-layer network framework. The research is aimed at answering several questions: first, identifying the most polarized topics of COP26 and comparing them with a longitudinal study involving COP21; second, examining user polarization across different topics and exploring whether users who engage in more topics are more polarized.

The first part of the analysis is concerned with the evaluation of the topic modeling techniques, using both traditional metrics such as coherence and diversity both an ad-hoc experiment. The results showed that the models exploiting neural networks architectures, in particular Bertopic, performed better than the traditional ones.

The second part was the actual analysis of the multilayer networks. The results revealed that in the most polarized topics, both sides had nearly an equal number of users, indicating a sharp divide in opinions. This was particularly evident in the Canadian discussion on limiting the use of coal and oil, as well as in the air travel debate. Additionally, it was observed that the Canadian discussion on Fossil Fuel was not always polarized: in COP21, for instance, it was completely non-polarized, which aligns with Falkenberg's findings.

The users tended to stay on the same side of the discussion across multiple topics, but there was no correlation between the number of topics and polarization score. It is important to note that this thesis only addressed a limited set of questions using the presented methodology, focusing solely on the retweet network. However, the reply and quote networks have different properties that can be further explored, especially the reply network, which allows for the study of direct discussions between users.

The use of transformers in the topic modeling step enabled highly accurate results, which previously would have required extensive manual work (annotating tweets). The use of a multilayer network framework prevents information loss by consolidating all data into a single network. Moreover, this framework has the potential to extend beyond the study of climate change and Twitter, and can be applied to any scenario where actors engage in discussions on multiple topics.

A limitation of this work is that only English tweets have been considered and while they constitute a significant portion of the total, conducting the same analysis with other languages may yield different results. This is particularly relevant for regional languages such as Italian. Additionally, adjusting the methodology by incorporating domain knowledge about polarization geometry[44] could further enhance the findings.

Further research should investigate the growth of polarization resulting from new environmental activism that inherently exhibits polarization in its methodology, such as blocking streets or defiling monuments. These acts undoubtedly generate conversation, but it remains to be seen whether these discussions occur within echo chambers or involve individuals from opposite ends of the spectrum.

References

[1] Syukuro Manabe and Richard T. Wetherald. Thermal equilibrium of the atmosphere with a given distribution of relative humidity. 24(3):241–259. ISSN 0022-4928, 1520-0469. doi: 10.1175/1520-0469(1967)024<0241:TEOTAW>2.0.CO;2. URL https://journals.ametsoc.org/view/journals/atsc/24/3/1520-0469_1967_024_0241_teotaw_2_0_co_2.xml.

[2] Charles D. Keeling, Robert B. Bacastow, Arnold E. Bainbridge, Carl A. Ekdahl, Peter R. Guenther, Lee S. Waterman, and John F. S. Chin. Atmospheric carbon dioxide variations at mauna loa observatory, hawaii. 28(6):538–551. ISSN 00402826, 21533490. doi: 10.1111/j.2153-3490.1976.tb00701.x. URL http://tellusa.net/index.php/tellusa/article/view/11322.

[3] F.W.G. Baker. First session of wmo–unep intergovernmental panel on climate change (ipcc), held in geneva, switzerland, during 9–11 november 1988. *Environmental Conservation*, 16(2):185–186, 1989. doi: 10.1017/S0376892900009127.

[4] Rafael M. Santos and Reza Bakhshoodeh. Climate change/global warming/climate emergency versus general climate research: comparative bibliometric trends of publications. 7(11):e08219. ISSN 2405-8440. doi: 10.1016/j.heliyon.2021.e08219. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8571708/.

[5] Anandita Sabherwal, Matthew T. Ballew, Sander Van Der Linden, Abel Gustafson, Matthew H. Goldberg, Edward W. Maibach, John E. Kotcher, Janet K. Swim, Seth A. Rosenthal, and Anthony Leiserowitz. The greta thunberg effect: Familiarity with greta thunberg predicts intentions to engage in climate activism in the united states. 51 (4):321–333. ISSN 0021-9029, 1559-1816. doi: 10.1111/jasp.12737. URL https://onlinelibrary.wiley.com/doi/10.1111/jasp.12737.

[6] Falkenberg. Climate polarization is increasing on twitter. 12(12):1094–1095. ISSN 1758-6798. doi: 10.1038/s41558-022-01533-z. URL https://www.nature.com/articles/s41558-022-01533-z.

[7] Pew Research Center. Politics on twitter: One-third of tweets from u.s. adults are political.

[8] Naveen Kumar Arora and Isha Mishra. COP26: more challenges than achievements. 4(4):585–588. ISSN 2523-8922. doi: 10.1007/s42398-021-00212-7. URL https://doi.org/10.1007/s42398-021-00212-7.

[9] Juan Layna and Leandro Altamirano. Promises that don't work? COP26 and the problems of climate change. 8(3):118–129. ISSN 2413-8053. doi: 10.17351/ests2022.1377. URL https://estsjournal.org/index.php/ests/article/view/1377.

[10] Max Falkenberg, Alessandro Galeazzi, Maddalena Torricelli, Niccolò Di Marco, Francesca Larosa, Madalina Sas, Amin Mekacher, Warren Pearce, Fabiana Zollo, Walter Quattrociocchi, and Andrea Baronchelli. Growing polarization around climate change on social media. *Nature Climate Change*, 12(12):1114–1121, December 2022. ISSN 1758-6798. doi: 10.1038/s41558-022-01527-x. URL https://www.nature.com/articles/s41558-022-01527-x.

[11] Magdalena Wojcieszak. *Polarization, Political*, pages 1–7. John Wiley Sons, Ltd, 2016. ISBN 9781118541555. doi: https://doi.org/10.1002/9781118541555.wbiepc168. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118541555.wbiepc168.

[12] Aaron Bramson, Patrick Grim, Daniel J. Singer, William J. Berger, Graham Sack, Steven Fisher, Carissa Flocken, and Bennett Holman. Understanding polarization: Meanings, measures, and model evaluation. 84 (1):115–159. ISSN 0031-8248, 1539-767X. doi: 10.1086/688938. URL https://www.cambridge.org/core/journals/philosophy-of-science/article/abs/understanding-polarization-meanings-measures-and-model-evaluation/FFACA58EFC4A6E421107020290BF592C.

[13] Kiran Garimella and Ingmar Weber. A long-term analysis of polarization on twitter, 2017.

[14] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3(null):993–1022, March 2003. ISSN 1532-4435.

[15] Daniel Lee and H. Sebastian Seung. Algorithms for Non-negative Matrix Factorization. In *Advances in Neural Information Processing Systems*, volume 13. MIT Press, 2000. URL https://papers.nips.cc/paper/2000/hash/f9d1152547c0bde01830b7e8bd60024c-Abstract.html.

[16] Da Kuang, Jaegul Choo, and Haesun Park. Nonnegative Matrix Factorization for Interactive Topic Modeling and Document Clustering. In M. Emre Celebi, editor, *Partitional Clustering Algorithms*, pages 215–243. Springer International Publishing, Cham, 2015. ISBN 9783319092591. doi: 10.1007/978-3-319-09259-1_7. URL https://doi.org/10.1007/978-3-319-09259-1_7.

[17] Xiaohui Yan, Jiafeng Guo, Yanyan Lan, and Xueqi Cheng. A biterm topic model for short texts. In *Proceedings of the 22nd international conference on World Wide Web*, WWW '13, pages 1445–1456, New York, NY, USA, May 2013. Association for Computing Machinery. ISBN 978-1-4503-2035-1. doi: 10.1145/2488388.2488514. URL https://doi.org/10.1145/2488388.2488514.

[18] Jianhua Yin and Jianyong Wang. A dirichlet multinomial mixture model-based approach for short text clustering. *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 233–242, August 2014. doi: 10.1145/2623330.2623715. URL https://dl.acm.org/doi/10.1145/2623330.2623715.

[19] Maarten Grootendorst. BERTopic: Neural topic modeling with a class-based TF-IDF procedure, March 2022. URL http://arxiv.org/abs/2203.05794. arXiv:2203.05794 [cs].

[20] Dimo Angelov. Top2Vec: Distributed Representations of Topics, August 2020. URL http://arxiv.org/abs/2008.09470. arXiv:2008.09470 [cs, stat].

[21] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. URL http://arxiv.org/abs/1706.03762.

[22] Roman Egger and Joanne Yu. A Topic Modeling Comparison Between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts. *Frontiers in Sociology*, 7:886498, May 2022. ISSN 2297-7775. doi: 10.3389/fsoc.2022.886498. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9120935/.

[23] Michael Steinbach, Levent Ertöz, and Vipin Kumar. The challenges of clustering high dimensional data. In Luc T. Wille, editor, *New Directions in Statistical Physics: Econophysics, Bioinformatics, and Pattern Recognition*, pages 273–309. Springer. ISBN 9783662089682. doi: 10.1007/978-3-662-08968-2_16. URL https://doi.org/10.1007/978-3-662-08968-2_16.

[24] Andrzej Maćkiewicz and Waldemar Ratajczak. Principal components analysis (PCA). 19(3):303–342. ISSN 0098-3004. doi: 10.1016/0098-3004(93)90090-R. URL https://www.sciencedirect.com/science/article/pii/009830049390090R.

[25] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008. URL http://jmlr.org/papers/v9/vandermaaten08a.html.

[26] Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. Umap: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3(29):861, 2018. doi: 10.21105/joss.00861. URL https://doi.org/10.21105/joss.00861.

[27] Xin Jin and Jiawei Han. K-means clustering. In Claude Sammut and Geoffrey I. Webb, editors, *Encyclopedia of Machine Learning*, pages 563–564. Springer US. ISBN 9780387301648. doi: 10.1007/978-0-387-30164-8_425. URL https://doi.org/10.1007/978-0-387-30164-8_425.

[28] Leland McInnes, John Healy, and Steve Astels. hdbscan: Hierarchical density based clustering. 2(11):205. ISSN 2475-9066. doi: 10.21105/joss.00205. URL https://joss.theoj.org/papers/10.21105/joss.00205.

[29] Data mining. In Anand Rajaraman and Jeffrey David Ullman, editors, *Mining of Massive Datasets*, pages 1–17. Cambridge University Press. ISBN 9781107737419. doi: 10.1017/CBO9781139058452.002. URL https://www.cambridge.org/core/books/mining-of-massive-datasets/data-mining/E5BFF4C1DD5A1FB946D616D619B373C2.

[30] Ishan Misra, C. Lawrence Zitnick, Margaret Mitchell, and Ross Girshick. Seeing through the human reporting bias: Visual classifiers from noisy human-centric labels. pages 2930–2939. URL https://www.cv-foundation.org/openaccess/content_cvpr_2016/html/Misra_Seeing_Through_the_CVPR_2016_paper.html.

[31] Luke Haliburton, Sinksar Ghebremedhin, Robin Welsch, Albrecht Schmidt, and Sven Mayer. Investigating labeler bias in face annotation for machine learning. URL http://arxiv.org/abs/2301.09902.

[32] Stephen P. Borgatti, Ajay Mehra, Daniel J. Brass, and Giuseppe Labianca. Network analysis in the social sciences. 323(5916):892–895. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.1165821. URL https://www.science.org/doi/10.1126/science.1165821.

[33] Giuseppe Veltri. *Digital Social Research*. 10 2019. ISBN 9781509529315.

[34] M. Kivela, A. Arenas, M. Barthelemy, J. P. Gleeson, Y. Moreno, and M. A. Porter. Multilayer networks. 2(3):203–271. ISSN 2051-1310, 2051-1329. doi: 10.1093/comnet/cnu016. URL https://academic.oup.com/comnet/article-lookup/doi/10.1093/comnet/cnu016.

[35] Dominik Traxl, Niklas Boers, and Jürgen Kurths. Deep graphs—A general framework to represent and analyze heterogeneous complex systems across scales. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 26(6):065303, 06 2016. ISSN 1054-1500. doi: 10.1063/1.4952963. URL https://doi.org/10.1063/1.4952963.

[36] Zaynab Hammoud and Frank Kramer. Multilayer networks: aspects, implementations, and application in biomedicine. 5(1):2. ISSN 2058-6345. doi: 10.1186/s41044-020-00046-0. URL https://bdataanalytics.biomedcentral.com/articles/10.1186/s41044-020-00046-0.

[37] Amy C. Kinsley, Gianluigi Rossi, Matthew J. Silk, and Kimberly VanderWaal. Multilayer and multiplex networks: An introduction to their use in veterinary epidemiology. 7. ISSN 2297-1769. URL https://www.frontiersin.org/articles/10.3389/fvets.2020.00596.

[38] Nhut-Lam Nguyen, Ming-Hung Wang, Yu-Chen Dai, and Chyi-Ren Dow. Understanding malicious accounts in online political discussions: A multilayer network approach. *Sensors*, 21(6), 2021. ISSN 1424-8220. doi: 10.3390/s21062183. URL https://www.mdpi.com/1424-8220/21/6/2183.

[39] Manlio De Domenico, Albert Solé-Ribalta, Elisa Omodei, Sergio Gómez, and Alex Arenas. Ranking in interconnected multilayer networks reveals versatile nodes. 6(1):6868. ISSN 2041-1723. doi: 10.1038/ncomms7868. URL https://www.nature.com/articles/ncomms7868.

[40] Jennifer McCoy, Tahmina Rahman, and Murat Somer. Polarization and the global crisis of democracy: Common patterns, dynamics, and pernicious consequences for democratic polities. *American Behavioral Scientist*, 62(1):16–42, 2018. doi: 10.1177/0002764218759576. URL https://doi.org/10.1177/0002764218759576.

[41] Michael Conover, Jacob Ratkiewicz, Matthew Francisco, Bruno Goncalves, Filippo Menczer, and Alessandro Flammini. Political polarization on twitter. 5(1):89–96. ISSN 2334-0770. doi: 10.1609/icwsm.v5i1.14126. URL https://ojs.aaai.org/index.php/ICWSM/article/view/14126.

[42] James Flamino, Alessandro Galezzi, Stuart Feldman, Michael W. Macy, Brendan Cross, Zhenkun Zhou, Matteo Serafino, Alexandre Bovet, Hernan A. Makse, and Boleslaw K. Szymanski. Shifting Polarization and Twitter News Influencers between two U.S. Presidential Elections. 2021. doi: 10.48550/ARXIV.2111.02505. URL https://arxiv.org/abs/2111.02505.

[43] Noam Gidron, James Adams, and Will Horne. Toward a comparative research agenda on affective polarization in mass publics. URL https://papers.ssrn.com/abstract=3391062.

[44] Tommaso Radicioni, Tiziano Squartini, Elena Pavan, and Fabio Saracco. Networked partisanship and framing: A socio-semantic network analysis of the italian debate on migration. 16(8):e0256705. ISSN 1932-6203. doi: 10.1371/journal.pone.0256705. URL https://dx.plos.org/10.1371/journal.pone.0256705.

[45] Hywel T. P. Williams, James R. McMurray, Tim Kurz, and F. Hugo Lambert. Network analysis reveals open forums and echo chambers in social media discussions of climate change. 32:126–138. ISSN 0959-3780. doi: 10.1016/j.gloenvcha.2015.03.006. URL https://www.sciencedirect.com/science/article/pii/S0959378015000369.

[46] Pablo Barberá. Birds of the Same Feather Tweet Together: Bayesian Ideal Point Estimation Using Twitter Data. *Political Analysis*, 23(1):76–91, January 2015. ISSN 1047-1987, 1476-4989. doi: 10.1093/pan/mpu011. URL https://www.cambridge.org/core/journals/political-analysis/article/birds-of-the-same-feather-tweet-together-bayesian-ideal-point-estimation-using-twitter-data/91E37205F69AEA32EF27F12563DC2A0A.

[47] Gene H Golub and Christian Reinsch. Singular value decomposition and least squares solutions. *Numerische Mathematik*, 14(5):403–420, 1970. URL https://ieeexplore.ieee.org/document/1102314.

[48] J. A. Hartigan and P. M. Hartigan. The Dip Test of Unimodality. *The Annals of Statistics*, 13(1):70–84, March 1985. ISSN 0090-5364, 2168-8966. doi: 10.1214/aos/1176346577. URL https://projecteuclid.org/journals/annals-of-statistics/volume-13/issue-1/The-Dip-Test-of-Unimodality/10.1214/aos/1176346577.full.

[49] Silvia Terragni and Elisabetta Fersini. OCTIS 2.0: Optimizing and Comparing Topic Models in Italian Is Even Simpler! In Elisabetta Fersini, Marco Passarotti, and Viviana Patti, editors, *Proceedings of the Eighth Italian Conference on Computational Linguistics, CLiC-it 2021, Milan, Italy, January 26-28, 2022*, volume 3033 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2021. URL http://ceur-ws.org/Vol-3033/paper55.pdf.

[50] Silvia Terragni, Elisabetta Fersini, Bruno Giovanni Galuzzi, Pietro Tropeano, and Antonio Candelieri. OCTIS: Comparing and optimizing topic models is simple! In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 263–270. Association for Computational Linguistics, April 2021. URL https://www.aclweb.org/anthology/2021.eacl-demos.31.

[51] Nicolas Webersinke, Mathias Kraus, Julia Anna Bingler, and Markus Leippold. ClimateBert: A Pretrained Language Model for Climate-Related Text, December 2022. URL http://arxiv.org/abs/2110.12010. arXiv:2110.12010 [cs].

[52] Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. Universal sentence encoder. URL http://arxiv.org/abs/1803.11175.

[53] Alexander Hoyle, Pranav Goel, Denis Peskov, Andrew Hian-Cheong, Jordan Boyd-Graber, and Philip Resnik. Is Automated Topic Model Evaluation Broken?: The Incoherence of Coherence, October 2021. URL http://arxiv.org/abs/2107.02173. arXiv:2107.02173 [cs].

[54] Leon Festinger. *A Theory of Cognitive Dissonance*. Stanford University Press, Redwood City, 1957. ISBN 9781503620766. doi: doi:10.1515/9781503620766. URL https://doi.org/10.1515/9781503620766.