



D 4.1

Methodological Guidelines

A Framework and Methodological Protocol for Work Package 4 – Analysing the Europeanisation and Platformization of Media Representations



Document information

Grant Agreement #:	101004488
Project Title:	EUROPEAN MEDIA PLATFORMS: ASSESSING POSITIVE AND NEGATIVE EXTERNALITIES FOR EUROPEAN CULTURE
Project Acronym:	EUMEPLAT
Project Start Date:	01/03/2021
Related work package:	WP4
Related task(s):	T4.1 – Methodological Guidelines
Lead Organisation:	P5 – FUOC
Author(s):	<p>Jim Ingebretsen Carlson (FUOC)</p> <p>Thomas Niemejer (FUOC)</p> <p>Valentina Latronico (FUOC)</p> <p>Francisco Lupiáñez-Villanueva (FUOC)</p> <p>Andrea Miconi (IULM)</p> <p>Sara Cannizzaro (IULM)</p> <p>Sofie van Bauwel (UGent)</p> <p>Femke de Sutter (UGent)</p>
Status	Final
Submission date:	04/11/2022
Dissemination Level:	Public

Table of Contents

1	<i>Introduction.....</i>	<i>4</i>
2	<i>Theoretical framework of media representations.....</i>	<i>6</i>
3	<i>Data collection</i>	<i>9</i>
3.1	Set-up and programming.....	9
3.2	First data retrieval.....	11
3.3	Filtering of posts and tweets.....	11
3.4	Final data retrieval and sample composition	12
4	<i>Coding.....</i>	<i>14</i>
4.1	Manual coding	14
4.2	Automatic coding	14
4.2.1	Pre-processing of data	14
4.2.2	Predictive modelling	15
5	<i>Analysis of media representation and sentiments of gender and im/migration.....</i>	<i>17</i>
6	<i>Data Management</i>	<i>19</i>
7	<i>Timeline</i>	<i>20</i>
8	<i>References.....</i>	<i>21</i>
9	<i>Appendix 1. Codebook</i>	<i>22</i>
9.1	Introduction.....	22
9.2	Im/migration	22
9.3	Gender.....	29
10	<i>Appendix 2. Guidelines for manual coding.....</i>	<i>35</i>
10.1	Introduction.....	35
10.2	Downloading the excel files	36
10.3	Description of the data collected in the excel files	38
10.4	Manually coding the data	39
10.5	Intercoder reliability check	43
10.6	Overview	43
10.6.1	Training.....	43
10.7	Data for coding for the Intercoder Reliability Check.....	45
10.8	Intercoder reliability calculations.....	45
11	<i>Appendix 3. Ethical approval.....</i>	<i>47</i>

1 Introduction

This document provides methodological guidelines for work package 4, which is concerned with analysing media representations of the two critical issues of im/migration and gender. The guidelines apply to the quantitative analysis embedded in Task 4.2 – Representation of Im/migration in 10 countries and Task 4.3 – Representation of gender in 10 countries. The goal of this work package is to provide an in-depth analysis of media narratives, aiming at detecting to what degree platformization has been changing the representation of gender and im/migration in Europe. The specific focus will be on how platformization affects the process of Europeanisation and how Europe is represented through gender and im/migration. To perform this analysis, we will download social media content in 10 European countries from Facebook and Twitter. This will be done using different Application Programming Interfaces (API) and search queries consisting of a set of keywords related to either gender or im/migration to extract the relevant posts. In addition, one theoretical framework of media representations is developed for each topic. Each theoretical framework comprises several dimensions, or themes, which are commonly encountered in relation to how Europe is represented through the topics in the scientific literature. Comparisons of how frequent the dimensions are in social media posts will constitute the main unit of analysis. The ethical committee at Catalonia Open University (UOC) has approved the proposed research and methods (ethical approval is provided in the Appendix).

The starting point for framing this task is the Europeanisation and Europeanity (E&E) dimension of the Public Sphere. The European public sphere (EPS) approach to E&E focuses on the practices of European citizens, engaging in (allegedly rational) decision-making, providing them with an opportunity to be politically active at a European level. The EPS is also seen as constituted by public discussions on EU (or European) issues in the national media of EU member states (Walter, 2017). Clearly, key European issues are gender equality and im/migration. Through this lens, we aim at answering the following research question: **(RQ1) Are there similar debates about im/migration and gender across Europe - can we find hints of a 'European public sphere' - or is coverage dominated by the national perspective?**

We will analyse social media discussions specifically mentioning some aspects of Europe to answer 0. In this way, we extract European debates for which it is more likely that an EPS exists regarding the topics of im/migration and gender across Europe. However, there may also exist similarities across European countries when studying debates without a European perspective. Moreover, the similarities of debates across Europe may differ depending on whether the perspective is European or non-European. Therefore, we will extend the analysis to include these perspectives by also downloading social media posts not specifically concerned with Europe. Importantly, this provides a baseline comparison in order to assess to which extent there exists an EPS at a European level. By comparing representations of im/migration and gender between posts concerned with Europe and posts not concerned with Europe we can answer the following research question: **(RQ2) Are there similar debates about im/migration and gender across Europe when the perspective is European compared to when it is not?**

To investigate whether platformization changes how Europe is represented through the topics of gender and im/migration, we will compare how representations differ between institutional media and user-generated content namely non-institutional media. We aim at answering the following research question: **(RQ3) How are representations of Europe in relation to gender and im/migration affected by new modes of consumption and production?**

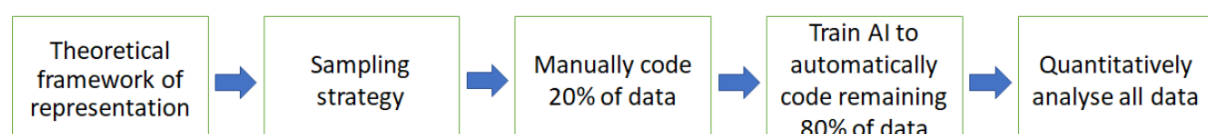
It is common to conduct sentiment analysis when analysing social media conversations (see, e.g., Drus et al. (2019) and Matamoros-Fernández & Farkas (2020)). While representations describe the content of the debates, sentiments provide a sense of the persons' attitudes towards the content. Sentiment analysis can thereby give a deeper understanding of how Europe is represented through the topics of gender and im/migration. Since the topics of im/migration and gender are sensitive topics in general (see, e.g., Malmqvist (2015), Nguyen et al. (2020), Park & Kim (2021), and Öztürk & Ayvaz (2018)), we could expect strong and diverging sentiments regarding the two topics. Consequently, by analysing the sentiments towards the two topics it is possible to assess the level of sensitivity of the topics in Europe and whether sensitivity is similar across European countries. Moreover, it is possible to disaggregate the analysis of sentiments by the dimensions of representations related to Europe to further assess which dimensions generate more sentiments and divergence. We aim to conduct sentiment analysis to answer the following research question: **(RQ4) Are sentiments towards gender and im/migration similar across Europe?**

Importantly, by also analyzing sentiments between debates concerning Europe and debates not concerning Europe, it is possible to assess whether European debates generate more sentiments than other debates. To address this, we will answer the subsequent research question: **(RQ5) Are sentiments different depending on whether debates are European or non-European?**

It is likely that the change in media- production and consumption has been accompanied by a change in the expression of sentiments in debates, specifically when comparing traditional media to user-generated content. Some evidence for this has been provided when investigating other topics (see, e.g., Godbole et al. (2007), Huang et. al (2020), and Kim et. al (2016)). We aim to study this change for the topics of gender and im/migration by answering: **(RQ6) How is the expression of sentiments on the topics of gender and im/migration affected by new modes of production and consumption?**

Figure 1 shows an overview of the methodological framework and the steps that will be performed throughout the process. First, the theoretical framework of representation and gender is developed to provide a basis for analysis and to aid in the design of the sampling strategy. The first step of the sampling strategy is a data collection process that is carried out using search queries consisting of a set of language-specific keywords related to either gender or im/migration that are obtained from the theoretical framework. The data collection will be identical for all 10 countries analysed. Despite using such keywords, many of the posts and tweets may be unrelated to the topics. Therefore, we develop a filtering process to further locate posts that are related to either gender or im/migration, which then will form part of the final sample. Thereafter, each partner country will manually code 20% of the provided data to determine the representations present in the posts and tweets. Using this data as input, an artificial intelligence algorithm will automatically code the remaining 80% of the data. Finally, all data is analysed using quantitative techniques to answer the research questions.

Figure 1. Overview of the methodological framework



Source: Authors' own elaboration

2 Theoretical framework of media representations

We will develop one theoretical framework of media representation for each file on gender and im/migration. The media representations will concern how Europe is represented through these two topics. The dimensions of media representations will be the main unit of analysis. Naturally, neither Facebook nor Twitter provide data on media representations. Therefore, the dimensions of representations will be assessed by manual- and automatic coding.

As previously mentioned, the starting point for this task is the concept of an EPS. Habermas (1974: 49) defines the public sphere, as “A portion of the public sphere comes into being in every conversation in which private individuals assemble to form a public body.” Moreover, the EPS’ materiality is its infrastructure. The EPS consists of interconnected media structures that allow European voices to materially circulate and engage in interactions. In the case of both gender and im/migration, such material infrastructure includes both institutional media and social media.

In this approach, the focus is very much on the degree to which EPS is realised, which is usually established in terms of synchronization of some issues. In this task, we will look precisely at the synchronization of the issues of gender and im/migration. In fact, in regard to im/migration, previous studies already found evidence of a developing European perspective in media coverage of im/migration (Balabanova and Balch 2010: 395). This study compared coverage in the UK and Bulgaria and assumed that the media agenda in each country with their own stakes in the phenomenon would differ, but instead, they found the coverage to be strikingly homogenous, as the Bulgarian media largely mirrored the UK’s media. While the European media content approach to E&E focuses on the material programs that are produced, the representations of Europe approach **focus on whether and how Europe is represented within media content**, which brings in a discursive approach. Together with European media content, this approach forms a (media) bridge between discursive and materialist components, even though this particular approach is tilted towards the discursive side. To circulate and offer themselves for identification, discourses depend on communication platforms but these ones have their specificities that can allow and disallow for discourses to reach particular groups (Carpentier et al., 2021). This approach thus considers how media texts *construct* Europe (and E&E), emphasizing certain features whilst omitting others, and generating contested or partial representations in the process. The construction of E&E through media representations can occur in a wide variety of ways, also relating to, for instance, ethnicity, religion, gender, im/migration, history, eating and drinking, science and technology, arts, music, architecture, and literature. If we take religion as one of the many possible examples, then we find that, for example, Nelsen and Guth (2016) argue that religion plays key role in the production of the idea of Europe.

Table 1 shows four dimensions (i.e., Law, People, Culture, and Values) of media representations that are common to the topics of gender and im/migration as well as their corresponding descriptions. Specifically, the definitions of the two topics are the following:

- **Im/migration:** the international movement of people to a destination [country](#) of which they are not natives or where they do not possess [citizenship](#) in order to settle permanently or temporarily.
- **Gender:** the characteristics of femininity and masculinity and the division of humans based on these. This includes, among others, gender identities such as being a man, woman, non-binary, LGBTQ, etc., as well as related discussions on social and cultural roles and behaviours.

Table 1. Dimensions of media representation common to Gender and Im/migration.

	Gender	Im/migration
Law	When the post deals with the legal aspects and rights of gender, and how it describes the specific rights on discrimination based on sexuality, gender, and biological sex.	If the post has to do with the legal aspect of im/migration, and how clearly it describes the specific legal statuses of im/migrants, refugees, and asylum-seekers [as the differences among these statuses are usually not clear at all].
People	When the post is about a person's own experience, or a general experience based on gender (women, men, non-binary and LGBTQIA+ people).	Whether the post is about the im/migrants themselves and their own voice: history, experience, journeys, travel diaries, profession, life conducted both in the country of origin and in Europe.
Culture	<p>Whether the post is about gender in terms of any kind of artistic expression and cultural production; Cultural habits and practices (including daily life); Cultural institutions, including education, the media, science, and the Church; Lifestyle, when related to gender.</p> <p>Posts under this dimension could refer to Artwork/cultural production/media products related to gender issues; Daily life practices and habits connected to gender; Educational practices related to gender issues; Art/cultural centers, educational institutions, scientific institutions, Churches and religious foundations, dealing with gender issues.</p>	<p>Whether the post is about migration in terms of any kind of artistic expression and cultural production; Cultural habits and practices (including daily life); Cultural institutions, including education, the media, science, and the Church; Lifestyle, when related to migration (i.e., multiethnic cities, im/migrants' activities).</p> <p>Posts under this dimension could refer to Artwork/cultural production/media products by/concerning im/migrants; Im/migrants' daily life habits and customs; Educational practices concerning im/migration; Art/cultural centers, educational institutions, scientific institutions, Churches, and religious foundations, dealing with im/migration/im/migrants.</p>
Values	Whether the post is about gender in terms of ideas and beliefs related to gender in/equality, gender im/balance, neutrality/bias, non/discrimination on the basis of gender, in/tolerance, dignity, diversity, freedom (of thought, expression, information, movement, choice), related to gender.	Whether the post is about migration in terms of/Whether the post is about im/migration in terms of Ideas and beliefs related to im/migrant/refugee in/equality, non/discrimination, in/tolerance, dignity, peace, solidarity, diversity, freedom (of thought, expression, information, movement), related to im/migration.

Source: Authors' own elaboration

Some dimensions are believed to be more important for one of the topics only. For example, Identity, New social movements, and Public sphere are expected to be applicable mostly to representations of **gender**:

- **Identity:** Definition for gender, being a man, woman, non-binary, LGBTQ. Is something mentioned and then you can crossbow what is mentioned. This is in terms of gender and sexual identity.
- **New social movements:** Self-organized citizenry including grass-roots social movements and NGOs. Movements that have targeted the structures, cultural practices, and interactional norms that sustain gender inequality. Further, movements that are not oriented specifically around gender issues are also shaped by gender as a central feature of social structure, culture, and everyday life.
- **Public sphere:** When a post is about gender-relevant issues, raised by non-political actors. Particularly, the relationship between citizens and institutions, the involvement in Decision-making, a non-political actor who tries to influence decision-making.

The dimensions of Territory, Institutions, and Interactions & dialogue are believed to be more common for **im/migration**:

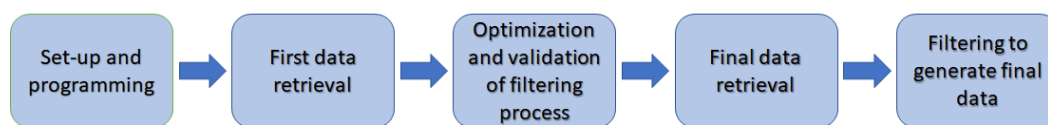
- **Territory:** When the post refers to borders or frontiers being crossed to migrate from one country to another, where at least one of the countries mentioned is a European one. Mentioning of place of departure and place of arrival.
- **Institutions:** When the post is about institutions involved in the field of im/migration regulation, control, governance, and so forth: national institutions, local institutions, European institutions, global institutions, and NGOs.
- **Interactions & Dialogue:** When the post mentions the encounter between im/migrants and natives (hospitality, professional initiatives, hosting, integration, joint activities of any sort).

These dimensions constitute the main unit of analysis and the main part of the manual- and automatic coding.

3 Data collection

The data collection is guided by the theoretical framework defined in the previous section. The steps of the data collection are displayed in **Figure 2**. It consists of the initial set-up and programming, and extraction of the first data using a list of language-specific keywords to perform the search. Using the initial data as input, a filtering process is developed with the aim of extracting the posts and tweets that most likely relate to the topics of gender and im/migration. Once the filtering process is optimized and validated; the second extraction of data will be performed to which the filtering process is applied. Finally, additional fine-tuning of the sample composition will be made to ensure that the analysis can be carried out. The resulting final data will constitute the data for manual- and automatic coding as well as quantitative analysis.

Figure 2. Overview of Data collection



Source: Authors' own elaboration

3.1 Set-up and programming

Initial- and final data will be downloaded in form of posts from Facebook and tweets from Twitter. Posts from Facebook will be downloaded using Crowdtangle¹ and Twitter tweets will be extracted through the Twitter API v2².

APIs are typically provided by social media platforms. For instance, Twitter created and manages its own API, so do Facebook and YouTube. In technical terms, API consists of a component that allows developers to build software for a particular application. APIs facilitate interaction between different software programs and the access to their services. It includes the specification of data structures, protocols, object classes, and runtimes to communicate the consumer with the resources offered by the API. Irs can build new classes or extend existing ones to add new features or functionalities.

Table 2 provides an overview of the APIs that will be used. The Universitat Oberta de Catalunya (UOC) has access to all platforms, notably and academic access to CrowdTangle and Twitter API, which allows for more data extraction and improved analysis of the data.

¹ <https://www.crowdtangle.com/>

² [Twitter API Documentation | Docs | Twitter Developer Platform](#)

Table 2. Overview of APIs

	Facebook	Twitter
Applications	CrowdTangle	Twitter API
Use of keywords	Yes	Yes
Geolocation	Yes, indirect	Yes
Implementation	Manual / <i>RESTful</i> API ³	<i>RESTful</i> API
Results format	.csv / JSON	JSON
Pricing	Free	Free and paid versions

Source: Authors' own elaboration

When considering the extraction of social media data, two important elements stand out regarding the set-up of the search strategy. They relate more specifically to language identification and geo-localization.

- **Data language identification.** Data will be gathered using language-specific keywords on each of the selected countries. In addition, we will limit the extraction to posts in the specific language.
- **Data geolocation.** Limiting the extraction only to language will not be enough since some languages are spoken in several countries. Therefore, it is also important to be able to geolocate the data. APIs can provide geo-located data, although most of the data provided by Twitter is not geolocated, which means that geolocation will exclude a large number of Tweets.

³ Any web service that obeys the REST constraints is informally described as RESTful. Such a web service must provide its Web resources in a textual representation and allow them to be read and modified with a stateless protocol and a predefined set of operations. This approach allows the greatest interoperability between clients and servers in a long-lived Internet-scale environment which crosses organisational (trust) boundaries.

3.2 First data retrieval

Data retrieval will be done by setting up a streaming pipeline for data acquisition through public APIs with the desired set of keywords that will be used to restrict the topics involved in the search. To extract data from each source via the corresponding API, we will be querying the specific keywords of interest (i.e., set of search terms) to download posts and comments related to those keywords. This will be done with the objective of obtaining relevant data for the analysis. To this end, we developed a lexicon with a set of keywords related to the topics of gender and migration. The initial selection of keywords was based on the dimensions arising from the theoretical framework, as well as search queries on Google and Crowdtangle to determine their relevance. Moreover, the keywords related to Europe had previously been established by the EUMEPLAT partners.

When conducting search queries so called Boolean conditions, AND, OR, NOT, are specified. For example, a search query specifying “gender AND feminism” would only return posts and tweets containing at least the two words gender and feminism. For the first retrieval, as much data as possible is downloaded by only applying the OR condition. That is a search query such as “gender OR feminism OR equality OR...”. Consequently, as soon as any one of these keywords is found in a post or a tweet, it is included in the dataset. According to the most recent data (“Digital 2021”, 2021), Facebook has more than 2,7 billion active monthly users worldwide, and YouTube has over 2,3 billion. Twitter has more than 350 million users worldwide. To understand how society expresses itself in social media we will monitor platforms with the aim of identifying keywords about the dimensions being researched and the circulation/distribution of those ones (Cardoso et al., 2021).

The first retrieval of data was done in Spanish and data from Facebook and Twitter were collected from the period between the 1st of March 2022 and the 1st of June 2022 and they included all posts in this time interval.

3.3 Filtering of posts and tweets

Even though we extract data based on keywords, language, and geo-location, the resulting first data comprises a huge amount of data. For example, the extraction of Facebook tweets related to gender contained more than 200 000 posts. It is common that the extraction of social media data contains a lot of posts and tweets which are not related to the specific analysis and, therefore, an additional filtering process is typically applied (see, e.g., FIND REFS). Previous experiences from the EUMEPLAT project as well as an initial review of the data confirms that many posts and tweets are unrelated to the topics of interest. When coding the first data that was retrieved only 25.3% of the top 200 posts supposedly related to gender were actually related to gender. Therefore, we develop a filtering process to obtain as many posts and tweets related to the two topics as possible. This is appropriate since the goal of this exercise is to analyse representations on the topics of gender and migration, and not to provide a general overview of discussions on social media. However, an alternative approach would be to try and filter to obtain as many posts and tweets as possible pertaining to at least one of the dimensions defined in the theoretical framework. However, such a sample would be even less representative of the narratives on migration and gender. Consequently, if we were to focus on such a sample we may run into the issue of analysing narratives of gender and migration which are far from representative of the narratives discussed in Europe.

The main idea of the filtering is to analyse which among the initial set of keywords results in data that are related and unrelated to the topics of gender and migration. From now on we refer to a tweet or post as being “on topic” if it relates to any of the topics of gender or migration and “off topic” if it is

unrelated. We aim to assess the importance of the keywords in generating posts on topic to create a “Relevance score” for the posts. We will validate the Relevance score by manual coding posts to determine whether they are on-topic or off-topic.

The first manual coding of the first data retrieved showed that posts that contained few keywords and keywords that were generic, such as “inclusion”, “identity”, or even “gender”, were often off-topic. However, keywords that clearly referred to the topic such as “gender violence”, “gender equality” or “women’s rights” were always coded as being on-topic. Moreover, the more keywords that a post contained, the more likely it was to be on-topic.

We used these initial insights to create the Relevance score that assigns a score to each post. The idea is that the higher the score a post gets, the more likely it is that it is on-topic and of interest for analysis. For data from Facebook, a post receives a higher score if it:

1. Contains more keywords
2. Contains more non-generic keywords
3. Has a higher sentiment score

The sentiment score measures the degree of positive or negative sentiment in the post. This is done by counting the number of words in the post that are related to positive and negative sentiments respectively. Finally, the sentiment score is greater for a post that has a greater absolute difference in positive and negative words and that includes more positive and negative words. The sentiment score is included in the Relevance score since it is assumed that greater sentiment generates posts that are of more interest to analyse. In addition, part of the analysis will be on the sentiments of the posts. In this way, we increase the likelihood of increasing posts with positive or negative sentiments. Finally, the average Relevance score is typically twice as great for Facebook posts compared to Tweets. To make posts and tweets more comparable, we multiply the Relevance score of Tweets by two.

3.4 Final data retrieval and sample composition

The final data retrieval will be made for the same time interval as WP2. That is, between 1/9 2021 to 30/11 2021. This allows for more comparability between the results generated in the two work packages. The final data retrieval will be done using the same search query as the first data retrieval but translated into all languages relevant languages for the EUMEPLAT project. Data is downloaded from the following countries and languages: Belgium – Dutch, Bulgaria – Bulgarian, Czech Republic – Czech, Germany – German, Greece – Greek, Italy – Italian, Portugal – Portuguese, Spain - Spanish, Sweden – Swedish, Turkey – Turkish. Following the download, the filtering process is applied to the data to generate the datasets which are to be coded manually and analysed.

Table shows the different datasets that will be generated for each country. Two datasets will be extracted for each topic. The datasets differ by whether they concern gender or migration and whether posts contain keywords relating to Europe or not. To decide whether a post or tweet belongs to the “AND Europe” or “NOT Europe” dataset, a set of previously used “Europe keywords” are used. If a post/Tweet contains at least one “Europe keyword”, it will form part of an “AND Europe” dataset and otherwise, it belongs to a “NOT Europe” dataset. Each excel contains both tweets and Facebook posts.

Table 3. Datasets extracted from the final data retrieval and the number of on-topic posts that need to be manually annotated

	Gender	Migration
AND Europe	200	200
NOT Europe	200	200
Total	400	400

Source: Authors' own elaboration

The filtering process is then applied to each dataset to locate posts/tweets that have a higher probability of being on-topic. Among these posts, we keep the 1000 posts with the highest Relevance score and then order them according to the number of interactions. Finally, we will ensure that 50% of the 1000 posts are from institutional media and 50% from user-generated content to be able to make a useful comparison between the two. For the Facebook datasets, it is possible to identify the category of the poster and institutional media posts will come from the categories: MEDIA_NEWS_COMPANY, MEDIA, NEWS_SITE, RADIO_STATION, TV_CHANNEL. For Twitter we will use the list of validated news media provided by EUMEPLAT partners to extract institutional media posts.

4 Coding

Our aim is to analyse how the topics of gender and migration are represented on social media, based on the theoretical framework defined above, as well as the sentiment of the post. In addition, it is important to assess whether the posts and tweets actually discuss the topics of gender and migration. Since Twitter and Facebook do not provide any (or enough) data to assess media representations, sentiments, and the content of the post, manual coding of the data is necessary to be able to carry out the analysis. However, manual coding is time- and resource-demanding, which makes it not feasible to manually code a large number of posts. Therefore, we will manually code a sufficient but smaller sample of posts and tweets and thereafter automatically code additional posts and tweets. In this way, we obtain a large sample of annotated data. An artificial intelligence/machine learning algorithm will be trained and validated on the manually coded data and then the algorithm will automatically annotate the additional posts and tweets.

4.1 Manual coding

The manual coding will be done on individual posts. For each post, coders need to assess:

1. Whether the code is on-topic or off-topic
2. Which classes of representation the post belongs to
3. The sentiment of the post

Naturally, the definition of on-topic differs by gender and im/migration. Moreover, it is not necessary to assess whether a post or a tweet is about Europe. The datasets are distinguished by a set of keywords related to different aspects of Europe. The classes of representations follow the definitions in the theoretical framework. It is possible that a post pertains to several, or none, of the dimensions. The sentiment of the post is assigned as either negative, neutral, or positive. Importantly, it is the sentiment of the post that is to be coded, not the sentiment towards the topic or Europe. A codebook with examples and operational definitions is provided to coders in order to ease coding and ensure alignment in the coding process.

Double coding will be required to further ensure the reliability of the data. The double coding will be required for 20% of the manually coded data. Thereafter, Krippendorff's alpha will be calculated using the double-coded data to assess reliability.

Each partner will receive four datasets and manually code until they find 200 posts that are on-topic in each dataset. Consequently, this will generate 400 posts on-topic for each file on gender and migration and for every country, as shown in [Table .](#)

4.2 Automatic coding

Based on the manual coding of posts on the topics of gender and migration, algorithms will be created for the automatic coding of new posts with Machine Learning techniques.

4.2.1 Pre-processing of data

Data selection: as initial input for the development of the Machine Learning algorithm all manually annotated posts will be used. In the manual coding process, a minimum of 200 on-topic posts will be

required per topic, Europe yes/no and language, as well as a statistically significant number of off-topic posts. The latter will be required as for the prediction of a two-class output (in this case, either on-topic or off-topic), both classes must be present in the input data. Hence, for each set of topic, Europe yes/no, and language, there should be a base input of roughly 300 to 400 annotated posts. However, also unannotated posts come into play. Not only for the mere purpose of annotating them with the algorithm but potentially also to enlarge the data training set as explained in the next paragraph.

Data anonymisation: data will be anonymised as far as required. In accordance with guidelines provided by the UOC ethical committee, we will disassociate any personal information provided in the data collection. Specifically, we will replace the page name (on Facebook) and the Author (on Twitter) with a randomly generated code.

Data cleaning: data will be cleaned as far as necessary. Particularly in the case of posts from both Twitter and Facebook, only those posts that are retrieved by API that have significant keywords in the actual post itself will be used. The APIs may retrieve posts that have the queried keywords in other fields than the actual post itself, for example in linked fields or in the author's description.

Data pre-processing: data will be pre-processed for automatic coding to be possible. This pre-processing is similar to the pre-processing for manual coding. Particularly for the construction of keyword-vectors of those keywords present in the post, the comparison will be done ignoring case and diacritic marks ("é" will be "e", "ä" will be "a", etc.) and lemming and stemming of words will be applied. In the final dataset before modeling, each record will represent a post where the topic, language, and source are identified, as well as having keyword-vectors constructed with regards to both the topic (gender or migration) and Europe, as well as a calculated relevance score, sentiment score and the interaction indicators (likes, shares, etc) that come directly from each platform. Also, data with regard to the author will be available, particularly whether the author represents news media, as well as indicators from the platform, such as the number of followers the author has on the platform.

4.2.2 Predictive modelling

We will develop, train, and validate Machine Learning (ML) algorithms to automatically annotate the data that has not been manually coded. In particular, the algorithms will annotate (label) whether a post is on or off-topic, the dimension of representation it belongs to (if any), and sentiments towards the topic.

Enlarging the training dataset: any algorithm based on a relatively small input set may run the risk of resulting in an algorithm with a relatively low predictive value. As in this case, the initial input set comprises about 400 items, we may run that risk. In order to mitigate this risk, and depending on the actual first results of the algorithm, it should be contemplated to enlarge the input dataset used for training the algorithm with previously unannotated posts. In principle, unannotated posts with a high degree of similarity with an annotated post could be added to the training set. The degree of similarity could be determined by comparing the similarity of keywords and the similarity of the relevance score as mentioned in the retrieval process for the manual coding. Also, Machine Learning techniques could be applied for this purpose, for instance a technique similar to one known as Active Noise Reduction. With this technique, the initial annotated dataset is trained and applied to the unannotated data. Then, those items of the unannotated set with very high (close to 1) or very low probability (close to 0) are added to the initial training set with new labels 1 and 0 respectively, to form a new training set

and to retrain the algorithm with this new set. With this retrained algorithm the remainder of unannotated posts is classified. This process can be done iteratively a number of times until the new training set has a significant number of items, or when easily classifiable items run out. Typically, for this technique Support Vector Machines (SVM) is used as the algorithm.

Modeling: the purpose of the modeling is to predict the target variables as annotated manually. These are the indicators of on-/off-topic, the 6-7 classes of representation, and the sentiment score as perceived by the annotator. The first indicators (on-/offtopic and the classes of representation) are binary 2-class predictors (basically Yes/No), the sentiment score requires a multi-class predictor (having 3 values: Positive, Neutral, and Negative). As such, for each of these predictors, a different model has to be constructed.

The pre-processing of the data already provides some relevant features for the model, such as the keyword-vectors, the relevance score, and the sentiment score. The latter two are constructed based on the input data and as such are different from the similar indicators as annotated manually. In principle, the annotated indicators can be used as features also, whenever that same indicator is not the actual target variable. The set of annotated variables a priori are assumed to be independent of each other.

Other features will be constructed for the modeling. In particular, NLP techniques will be applied to deconstruct the core text of the post, through what is known as Word Embeddings. Word Embedding are numerical representations of a text, which can be used more optimally by the Machine Learning algorithm. Word Embeddings can be broadly classified into two categories:

- 1) Frequency-based Embedding
- 2) Prediction-based Embedding

Under the first category, three types of vectors can be constructed: Count Vector, TF-IDF Vector (Term Frequency, Inverse Document Frequency), or Co-Occurrence Matrix. Under the category of prediction-based embedding, we find Continuous Bag-of-Words (CBOW) and Skip-Grams.

Any of these techniques may be used for the purpose of implementing Word Embeddings. It should be noted that when constructing these vectors, potentially all words of the post may be used, not just the set of pre-defined keywords themselves. Once the feature space has been constructed, the actual training of the algorithm will be done. For a two-class or multi-class prediction, a number of different algorithms is available, among others: Random Forest, Logistic Regression, Gradient Tree Boosting or XGBoost. Each has their set of parameters and hyperparameters that will be finetuned for optimal performance.

The performance of the algorithm in the first instance will be measured by optimizing for AUC (Area Under the Curve) of the ROC-curve (which stands for Receiver Operating Characteristic). This measure is the most commonly used way to evaluate the performance of ML algorithms. Also, the so-called F1 score, which is the harmonic mean of precision and recall, will be used to optimize for the threshold between output classes. The training of the algorithms by its nature is an iterative process in order to get the best performance.

The aforementioned ways of measuring performance, as well as the mentioned types of algorithms may be subject to change according to need and if required. Validation of the algorithm will be done by feeding the algorithm previously unseen data and checking the output manually for selected languages (at least Spanish). This should serve to test whether the algorithm holds more generally and whether the training set may have had imbalanced data.

5 Analysis of media representation and sentiments of gender and im/migration

Quantitative techniques will be applied to answer the research questions and analyse the coded data. In particular, regression analysis and statistical tests will be applied to assess whether statistically significant differences exist in the variables of interest. In addition, the descriptive analysis will be provided to broaden the analysis. Both the manually and automatically annotated data will be used in the analysis. The same analysis will be applied to both gender and im/migration. The outcome variables used for the analysis will mainly be the dimensions of representations and sentiments.

The first research question (**RQ1**): Are there similar debates about migration and gender across Europe - can we find hints for a 'European public sphere' - or is coverage dominated by the national perspective?, will be answered by comparing the dimensions of representation individually, but also grouped, across the European countries. The grouping will be made by creating a binary variable that takes the variable 1 if the content of a post/tweet is coded as being represented by at least one of the dimensions from the theoretical framework, and 0 otherwise.

The second research question (**RQ2**): Are there similar debates about migration and gender across Europe when the perspective is European compared to when it is not?, will also be answered by comparing the dimensions of representation individually and grouped, but at a first step within each country separately. Thereafter, the within-country-difference is compared between countries.

Representations will be compared between institutional media and user-generated content to answer the broad third research question (**RQ3**): How is Europe in relation to gender and migration represented and how are the representations affected by new modes of consumption and production?

We follow a similar structure when answering RQ4 – RQ6. Sentiments will be compared across European countries to answer the fourth research question (**RQ4**), and the fifth research question (**RQ5**) will be answered by comparing sentiments of European to non-European debates. Sentiments will be compared between institutional media and user-generated contents to answer the sixth research question (**RQ6**).

The following question will be addressed when writing the reports:

1. Representations of the topic
 - a. What dimensions of representation of the topic are most frequently talked about?
 - b. Is there a difference in representation between institutional media and user-generated posts?
 - c. To what extent do the posts relate to the European dimensions of representation?
 - d. Is there a difference in how the topic is represented when the perspective is European compared to non-European?
 - e. Do the representations change over time?
2. Sentiments
 - f. What are the predominant sentiments toward the topic in your country?
 - g. Is there a difference between institutional media and user-generated posts?
 - h. Which dimensions of representation generate the most negative and positive sentiments?
 - i. Do the sentiments differ if posts concern Europe compared to when they do not concern Europe?
 - j. Do the sentiments change over time?

3. Comparison of Gender and migration

- k. Are there differences between the topics of gender and migration when the posts concern Europe compared to when they do not concern Europe, for:
 - (a) The classes of representation
 - (b) The sentiments of the posts
 - (c) How representation and sentiments differ by type of poster and time

Each partner is responsible for providing a country report of 5 – 10 pages, for each topic, answering the questions above by the **28th of February 2023**. FUOC (Spain) will provide each partner with the necessary tables, graphs, and statistical tests/analyses to be able to do this. This ensures a harmonized analysis and saves the overall workload. Furthermore, partners are responsible for illustrative cases. Specifically, illustrative case studies are well-known for their descriptive nature. The aim is to explain the details relating to a particular subject matter reporting examples to let the reader understand the situation that is being described. This type of research methodology relies on concrete results and also applies quantitative data to gain a deeper insight into the topic. In addition, FUOC will write a between-country report of 15 – 20 pages that analyses the differences between the 10 countries.

6 Data Management

The data will be collected from public Facebook pages and Twitter accounts using the APIs Crowdtangle and Twitter API v2. No personal data from the users will be downloaded other than that which is publicly available through the APIs. We abide by the terms, conditions, and privacy policies of Twitter and Facebook. Therefore, the collected data will be managed in agreement with GDPR EU 679/2016. In particular.

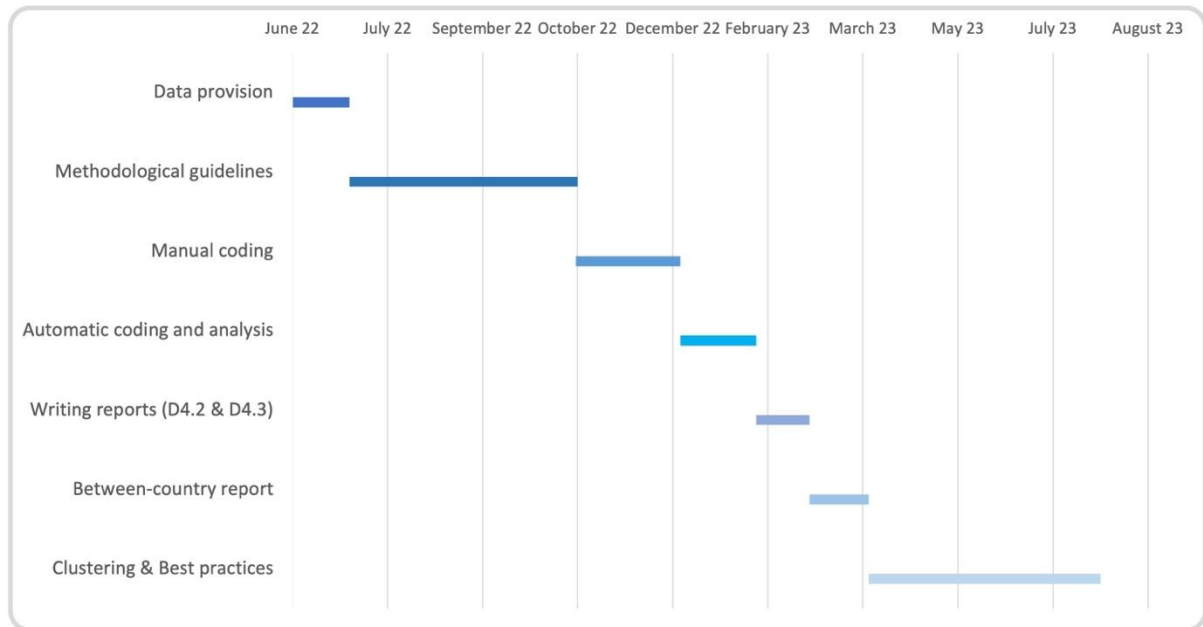
- No genetic data are or will be collected for the research.
- No biometric data are or will be collected for the research.
- No data concerning health are or will be collected for the research.
- No children are or will be involved in the research.

Data and results will be stored in the restricted area of their project website and available solely to authorized partner researchers. Public display of parts or all the public data will be subject to rules regarding the dissemination of knowledge contained in the EUMEPLAT project. Each member of the EUMEPLAT project has accepted the Ethical Guidelines which are the foundation of all project works. Data destination is EU27 plus Turkey. Data will be stored in a shared cloud drive available to the partners and created specifically for this purpose. Access to data on the storage drive is subject to authentication. In compliance with the GDPR EU 679/2016 law, data will be shared only among the participants to the project; they are therefore closed access.

7 Timeline

Figure 3 shows the timeline of WP4.

Figure 3. Timeline of WP4



Source: Authors' own elaboration

8 References

- DRUS, Z., AND H. KHALID. (2019): "Sentiment Analysis in Social Media and Its Application: Systematic Literature Review," *Procedia Computer Science*, 161, 707–14.
- G. CARDOSO, C. ÁLVARES, J. MORENO, R. SEPÚLVEDA, M. CRESPO, & C. FOÀ, D2.1-A *Framework and Methodological Protocol for analyzing the platformization of news*.
- Godbole, N., Srinivasaiah, M., & Skiena, S. (2007). "Large-Scale Sentiment Analysis for News and Blogs". *Icwsn*, 7(21), 219-222.
- HUANG, Y., J.-C. THILL, H. ZHANG, X. YU, C. ZHONG, D. LI, AND W. XU. (2020): "Sentiment analysis for news and social media in COVID-19," *Proceedings of the 6th ACM SIGSPATIAL International Workshop on Emergency Management using GIS*, , 1–4.
- KIM, E. H.-J., Y. K. JEONG, Y. KIM, K. Y. KANG, AND M. SONG. (2016): "Topic-based content and sentiment analysis of Ebola virus on Twitter and in the news," *Journal of Information Science*, 42, 763–81
- MALMQVIST, K. (2015): "Satire, racist humour and the power of (un)laughter: On the restrained nature of Swedish online racist discourse targeting EU-migrants begging for money," *Discourse & Society*, 26, 733–53.
- MATAMOROS-FERNÁNDEZ, A., AND J. FARKAS. (2021): "Racism, Hate Speech, and Social Media: A Systematic Review and Critique," *Television & New Media*, 22, 205–24.
- N. CARPENTIER, M. HROCH, S. CANNIZZARO, A. MICONI, & V. DOUDAKI, *Towards an operational definition of Europeanity and Europeanisation*, in *D1.6-Europeanisation: operational definition*.
- NGUYEN, T. T., S. CRISS, E. K. MICHAELS, R. I. CROSS, J. S. MICHAELS, P. DWIVEDI, D. HUANG, ET AL. (2021): "Progress and push-back: How the killings of Ahmaud Arbery, Breonna Taylor, and George Floyd impacted public discourse on race and racism on Twitter," *SSM - Population Health*, 15, 100922.
- PARK, S., AND J. KIM. (2021): "Tweeting about abusive comments and misogyny in South Korea following the suicide of Sulli, a female K-pop star: Social and semantic network analyses," *El Profesional de la información*,
- ÖZTÜRK, N., AND S. AYVAZ. (2018): "Sentiment analysis on Twitter: A text mining approach to the Syrian refugee crisis," *Telematics and Informatics*, 35, 136–47.
- WALTER, S. (2017) *EU Citizens in the European Public Sphere: An Analysis of EU News in 27 EU Member States*. Wiesbaden: Springer.

9 Appendix 1. Codebook

9.1 Introduction

This document provides a detailed description of how to manually code the columns of WP4. Since the columns to be coded differ by the topics of gender and im/migration, the codebook will treat each topic separately.

From now on, we refer to a Facebook post or Twitter Tweet as a post.

Unit to be coded:

You should only assess the context of the **TEXT** of the post.

Examples below include pictures, and link texts, that are vital to understanding the dimensions.



Columns:



Each column can be given any of the allowed values regardless of the answers in the other columns.



IMPORTANT: No cells can be left blank. All the cells of the codebook must be filled with an allowed value.



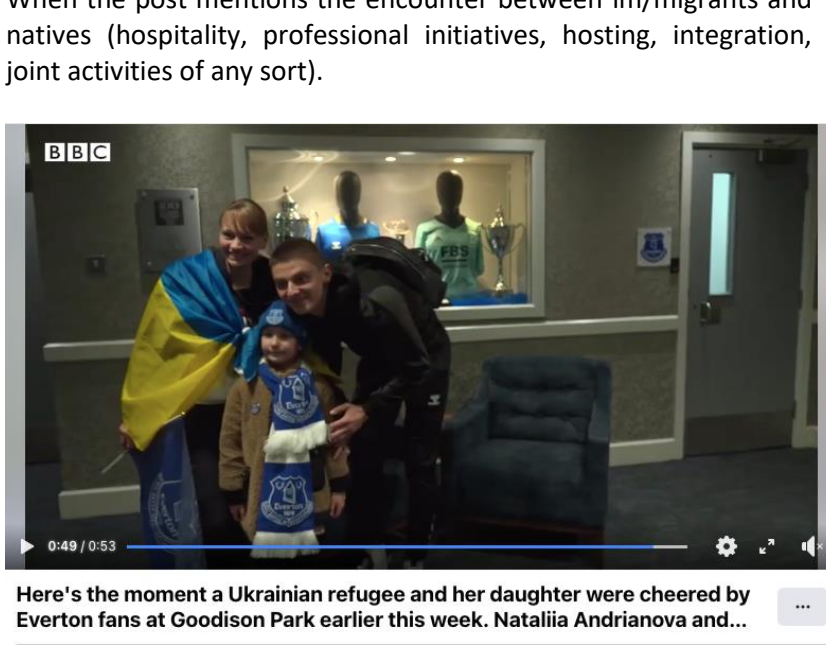
9.2 Im/migration




Column	Values	Description
On topic	YES: 1 NO: 0	A post is on topic if the context of the text is directly or indirectly related to human im/migration. I.e., the international movement of people to a destination country of which they are not natives or where they do not possess citizenship in order to settle permanently or temporarily. (See below for examples of posts on topic) If the context of the text is not directly or indirectly related to human im/migration, then the post is off topic . Examples of off topic posts could be posts talking about migration of animals, migration related to IT (data migration, software migration etc.), etc.
Law	YES: 1 NO: 0	When the post has to do with the legal aspect of im/migration, and how clearly it describes the specific legal statuses of im/migrants, refugees, and asylum-seekers [as the differences among these statuses are usually not clear at all].

		<p>🇬🇧🇷🇺🇰🇪🇺</p> <p>The deal is signed. If you come to the UK illegally you will be relocated to Rwanda.</p> 
People	<p>YES: 1</p> <p>NO: 0</p>	<p>When the post is about the im/migrants themselves and their own voice: history, experience, journeys, travel diaries, profession, life conducted both in the country of origin and in Europe.</p> <p>A heartwarming story</p>  <p>INDY100.COM ✓</p> <p>Refugee toddler excited to receive same doll she left in Ukraine from charity</p>

Culture	<p>YES: 1</p> <p>NO: 0</p>	<p>Whether the post is about migration in terms of Artistic expression and cultural production (of any kind); Cultural habits and practices (including daily life); Cultural institutions, including education, the media, science, and the Church; Lifestyle, when related to migration (i.e., multiethnic cities, im/migrants' activities);</p> <p>Posts under this dimension could refer to Artwork/cultural production/media products by/concerning im/migrants; Im/migrants' daily life habits and customs; Educational practices concerning im/migration; Art/cultural centers, educational institutions, scientific institutions, Churches and religious foundations, dealing with im/migration/im/migrants.</p> <p> London Migration Film Festival @LONMigFilmFest</p> <p>We're delighted to reveal the poster for this year's #LondonMigrationFilmFestival! Thanks to R&F Mo for the inspiring design.</p> <p>Save the dates & keep 🗓️ on our socials... Full programme coming soon!</p> <p>#migration #immigration #thisislondon #noborders #nowalls #migrationart #cinema</p>  <p>Genesis Cinema and 9 others</p> <p>1:01 pm · 17 Oct 2022 from London, England · Twitter for Android</p>
Values	<p>YES: 1</p> <p>NO: 0</p>	<p>Whether the post is about migration in terms of/ whether the post is about im/migration in terms of: Ideas and beliefs related to immigrant/refugee in/equality, non/discrimination, in/tolerance, dignity, peace, solidarity, diversity, freedom (of thought, expression, information, movement), related to im/migration.</p>



		<p>On April 13, 1985, while the Swedish Neo-Nazi The Nordic Realm Party was having demonstration in Växjö, Sweden.</p> <p>The woman who hit skinhead was Danuta Danielsson, a Polish immigrant whose mother had survived Majdanek concentration camp.</p> <p>#ww2history #ww2veteran #ww2pictures #ww2 #veteranowned #veterans4vetera #veterans #veteran #veteransupport #community #heritage #archaeologynews #archaeologist #archaeology #archaeological #archaeologylife</p> <p>I ❤️ 🇵🇱</p>  <p>fb@Historia, o której się nie mówi</p>
		<p>Yuliia Korzun was welcomed by her British sponsor after travelling 2,000 miles from her home in southern Ukraine</p>  <p>NEWS.SKY.COM ✓</p> <p>Ukraine war: 'Life was worth living that day' - emotional scenes as refugee arrives at Birmingham Airport</p>
Territory	<p>YES: 1</p> <p>NO: 0</p>	<p>When the post refers to borders or frontiers being crossed in order to migrate from one country to another, where at least one of the countries mentioned is a European one. Mentioning of place of departure and/or place of arrival.</p>





		
Institutions	<p>YES: 1</p> <p>NO: 0</p>	<p>When the post is about institutions involved in the field of im/migration regulation, control, governance, and so forth: national institutions, local institutions, European institutions, global institutions, and NGOs.</p> 
Interactions & Dialogue	<p>YES: 1</p> <p>NO: 0</p>	<p>When the post mentions the encounter between im/migrants and natives (hospitality, professional initiatives, hosting, integration, joint activities of any sort).</p> 
Sentiment	<p>POSITIVE: 2</p>	<p>When the sentiment of the post is predominantly positive. Words such as glad, happy, good, better, etc. appear in the post. If the sentiment is more positive than negative, it should be coded as positive.</p>



		<p>"I am so glad that through this play, I can finally tell our story to the Hungarian public."</p> <p>On #WorldTheatreDay, check out the story of Abouzar and his son through a theatre production about refugee experiences.</p>  <p>UNHCR.ORG </p> <p>Iranian father and son re-enact their story for Hungarian theatre goers Iranian refugee Abouzar and his son, Armin, chose to stay in Hungary despite bei...</p>
	<p>NEUTRAL: 1</p>	<p>When the sentiment of the post is predominantly neutral. Typically, the post conveys facts or describes a story without any positive or negative sentiments.</p> <p>On April 13, 1985, while the Swedish Neo-Nazi The Nordic Realm Party was having a demonstration in Växjö, Sweden.</p> <p>The woman who hit skinhead was Danuta Danielsson, a Polish immigrant whose mother had survived Majdanek concentration camp.</p> <p>#ww2history #ww2veteran #ww2pictures #ww2 #veteranowned #veterans4veterans #veterans #veteran #veteransupport #community #heritage #archaeologynews #archaeologist #archaeology #archaeological #archaeologylife</p> <p>I ❤️ 🇵🇱</p>  <p>fb@Historia, o której się nie mówi</p>



	<p>NEGATIVE: 0</p> <p>When the sentiment of the post is predominantly negative. Words such as sad, bad, worse, disappointed, miserable, etc appear in the post.</p> <p>This week - the 'battle bus', harassing migrants and the police harassing them.....of this is what they spend their entire week doing they must have very sad lives. They probably did more than this too. I may spend some time on here, but I have a life outside of it. Imagine being so full of hate you spend every waking minute focusing on the things that you hate. Sad.</p> <p>- Khaleesi.</p>
--	---


9.3 Gender




Column	Values	Description
On topic	YES: 1 NO: 0	<p>A post is on topic if the context of the text is directly or indirectly related to human gender.</p> <p>I.e., when a post is about the characteristics of femininity and masculinity and the division of humans based on these. This includes, among others, gender identities such as being a man, woman, non-binary, LGBTQ, etc., as well as related discussions on social and cultural roles and behaviours.</p> <p>If the context of the text is not directly or indirectly related to human gender, then the post is off topic.</p> <p>For example, if a post is about grammatical, or animal gender.</p>
Law	YES: 1 NO: 0	<p>When the post deals with the legal aspects and rights of gender, and how it describes the specific rights on discrimination based on sexuality, gender, and biological sex.</p> <p>And contraception or gay marriage could be next.</p>  <p>FT.COM </p> <p>FT View: Reversing Roe vs Wade would shatter women's rights</p> <p>There is no easy way to undo a ruling that would hobble access to abortion</p>
People	YES: 1 NO: 0	<p>When the post is about a person's own experience, or a general experience based on gender (women, men, non-binary and LGBTQIA+ people).</p>

		 <p>I'm Non-Binary And My Wife Is 3ft 2in</p>
Culture	<p>YES: 1</p> <p>NO: 0</p>	<p>Whether the post is about gender in terms of artistic expression and cultural production (of any kind); cultural habits and practices (including daily life); cultural institutions, including education, the media, science, and the Church; Lifestyle, when related to gender.</p> <p>Posts under this dimension could refer e.g. to Artwork/cultural production/media products related to gender issues; Daily life practices and habits connected to gender; Educational practices related to gender issues; Art/cultural centers, educational institutions, scientific institutions, Churches and religious foundations, dealing with gender issues.</p> <p> Cuppy 15 april · 🌐</p> <p>Quality Education and Gender Equality has ALWAYS been an integral part of my brand so naturally, the Cuppy Foundation focuses on these SDG Goals 🌟</p> <p>I am super proud and honored to partner with Lagos State Government Office of Sustainability Development Goals as a Champion for the Youth Alliance to further amplify the 4th and 5th goal and inspire the younger generation. 🌍📚📱 #CuppyCares #SDGS2030 #SDG4 #SDG5</p> <p>Vertaling weergeven</p> <div>   </div>

<p>Values</p>	<p>YES: 1</p> <p>NO: 0</p>	<p>Whether the post is about gender in terms of Ideas and beliefs related to gender in/equality, gender im/balance, neutrality/bias, non/discrimination on the basis of gender, in/tolerance, dignity, diversity, freedom (of thought, expression, information, movement, choice), related to gender.</p> <div data-bbox="539 409 1362 846">  </div> <p>Gender-based Rights and Responsibilities in Quran Watch it on YouTube: https://youtu.be/tiLLzHUL8WY</p> <p>Despite the ban, British Cycling says they're committed to inclusivity.</p> <div data-bbox="539 1019 1337 1809">  </div>
<p>Identity</p>	<p>If YES: 1</p> <p>If NO: 0</p>	<p>Definition for gender, being a man, woman, non-binary, LGBTQ. Is something mentioned and then you can crossbow what is mentioned. This is in terms of gender and sexual identity.</p>

		<p>The singer came out as non-binary last year ❤️</p>  <p>VT.CO</p> <p>Demi Lovato Quietly Updates Their Pronouns On Instagram</p> <p>Demi Lovato has quietly updated the pronouns they wish to be addressed by. T</p>
New social movements	<p>If YES: 1</p> <p>If NO: 0</p>	<p>Self-organized citizenry including grass-roots social movements and NGOs. Movements that have targeted the structures, cultural practices, and interactional norms that sustain gender inequality. Further, movements that are not oriented specifically around gender issues are also shaped by gender as a central feature of social structure, culture, and everyday life.</p> <p>"It was a historic night for women in film. Jane Campion, the revered New Zealand-born Australian director, became the third female film-maker ever and the second in a row to win best director in the 94-year history of the Academy Awards.</p> <p>Director Sian Heder won the night's top gong, taking home best picture and best adapted screenplay, for Coda – just her second feature-length film and the first win for Apple. And earlier in the night, Ariana DeBose became the first queer woman of colour to win best supporting actress. At just 20, Billie Eilish won for best original song, and costume designer Jenny Beavan picked up her third Oscar for her work on Cruella.</p> <p>But in the end a night of momentous achievements was overshadowed by an act of violence between two men"</p> 
Public sphere	<p>If YES: 1</p> <p>If NO: 0</p>	<p>When a post is about gender relevant issues, raised by non-political actors. Particularly, the relationship between citizens and institutions,</p>

		<p>the involvement in Decision-making, a non-political actor who tries to influence decision-making.</p> <hr/> <p>A woman ran onto the red carpet at the Cannes Film Festival, to protest against sexual violence in Ukraine.</p> <p>https://bbc.in/3LtXkFL</p> <hr/>
Sentiment	<p>POSITIVE: 2</p>	<p>When the sentiment of the post is predominantly positive. Words such as glad, happy, good, better, etc. appear in the post. If the sentiment is more positive than negative, it should be coded as positive.</p> <p>Happy #NonBinaryDay! Today and everyday we celebrate the wide range of people worldwide who identify as non-binary. You are valid, and beautiful, and we love you! 💜💜💜💜</p> 
	<p>NEUTRAL: 1</p>	<p>When the sentiment of the post is predominantly neutral. Typically, the post conveys facts or describes a story without any positive or negative sentiments.</p>

		<p>The legislation would ban discussing sexual orientation and gender identity in primary school classrooms.</p>  <p>BBC.COM </p> <p>White House slams new Florida 'Don't Say Gay' law</p>
	<p>NEGATIVE : 0</p>	<p>When the sentiment of the post is predominantly negative. Words such as sad, bad, worse, disappointed, miserable, etc appear in the post. When the sentiment is more negative than positive, it should be coded as negative.</p> <p>Mae Martin described the BBC story as "bad journalism" which is "contributing to a culture of hysteria that makes life scarier for trans/ non-binary people in this country".</p>  <p>PINKNEWS.CO.UK </p> <p>Mae Martin joins protest at BBC headquarters against infamous anti-trans article</p>

10 Appendix 2. Guidelines for manual coding

10.1 Introduction

For WP4 we need you to manually code Facebook posts and Twitter tweets about the topics of gender and im/migration.

For each country, we have prepared 4 excel files on which you will do the manual coding. The four excel files differ by the topic, either gender or im/migration, and the perspective, which is either European or non-European. The names of the excel files follow the following structure: COUNTRYCODE_Topic_Perspective.

Table displays the four excel files that are provided to each country and a short general description.

Table 4. Datasets to be manually coded for each topic

Name	Description
COUNTRYCODE_Gen_EUR	Posts and tweets about gender with a European perspective
COUNTRYCODE_Gen_noEUR	Posts and tweets about gender with a non-European perspective
COUNTRYCODE_Mig_EUR	Posts and tweets about im/migration with a European perspective
COUNTRYCODE_Mig_noEUR	Posts and tweets about im/migration with a non-European perspective

Source: Author's own elaboration

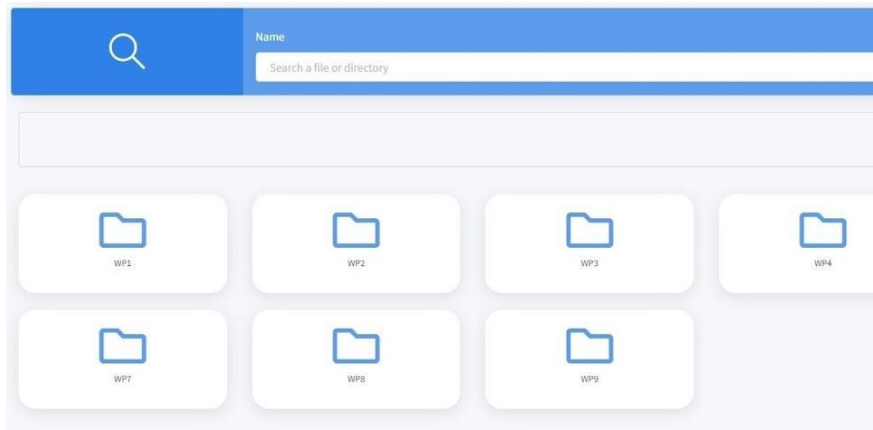
10.2 Downloading the excel files

To download the excel files, you need to log into the EUMEPLAT restricted area:

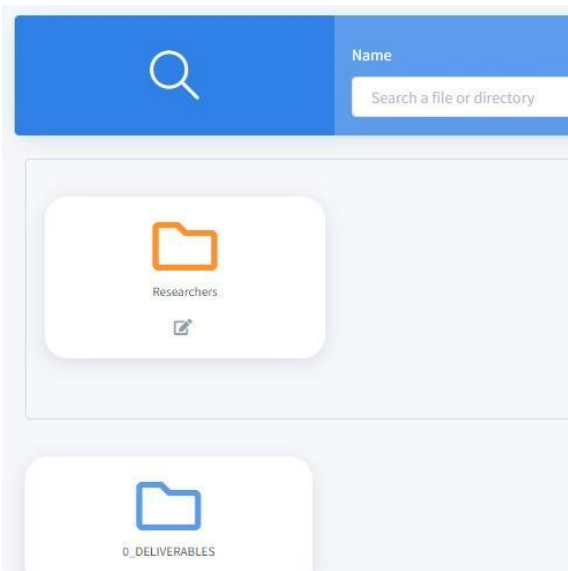
<https://restrictedarea.eumeplat.eu/>

Thereafter, you need to carry out the following steps:

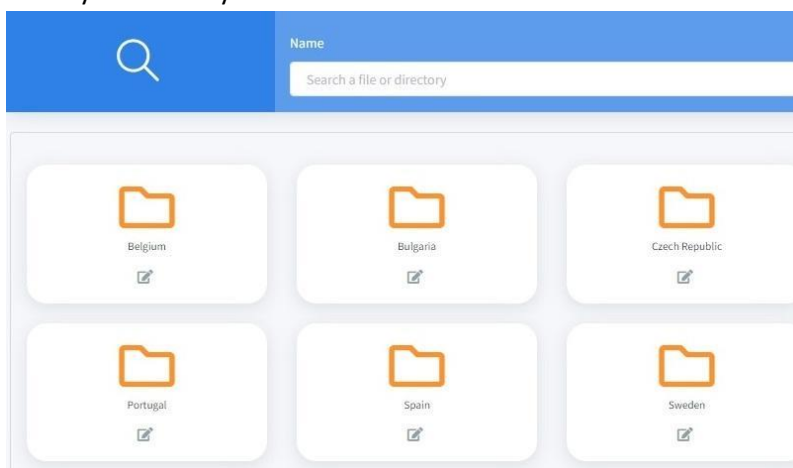
- Go to the folder named WP4.



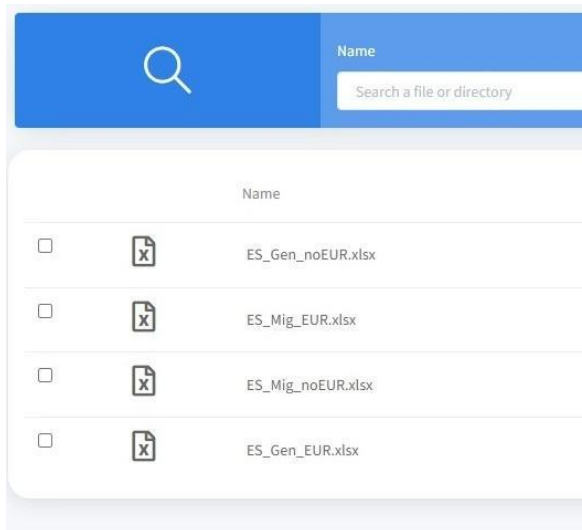
- Go to the folder named Researcher.



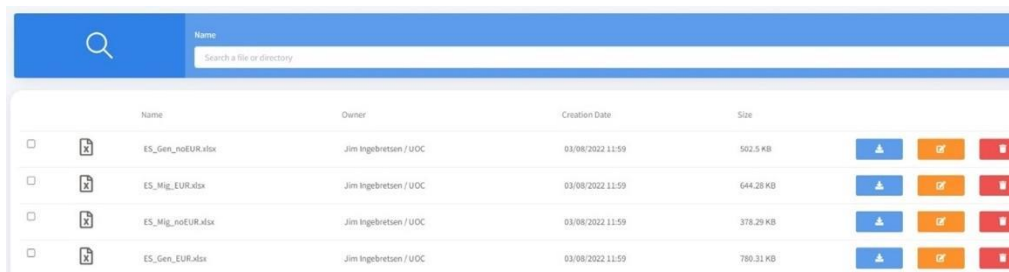
- Go to your country's folder.



- Once inside your country's folder, you will see the four excel files.



- Download each excel file by pressing the blue button on the right.



10.3 Description of the data collected in the excel files

Each excel contains both posts from Facebook and tweets from Twitter as well as posts and tweets from both institutional media and non-institutional media (user-generated). Each excel contains a maximum of 1000 posts and tweets.

The excel files contain columns with collected, and sometimes calculated, data from Facebook and Twitter as well as columns with no data. The columns with no data are the ones that need to be manually coded.

Table shows the columns with collected, and sometimes calculated, data from Facebook and Twitter. These variables should **NOT** be modified during the manual coding.

Table 5. Columns with data collected, and sometimes manipulated, from Facebook and Twitter

Name	Description
lang	Language of the posts and tweets
topic	Topic of the posts and tweets. Either im/migration or gender
source	Source of the posts or tweets. Either Facebook (FB) or Twitter (TW)
location	Country where the post or tweet was posted
created_at	Data and time the post/tweet was created
author_id	A hashed identifier
followers	The number of followers at the time the post/tweet was created
category	The category of the poster. Exists a large number of possible categories for Facebook, but only MEDIA or NO-MEDIA for Twitter
is_newsmedia	Indicator variable taking the value 1 if the post/tweet was created by an institutional media and 0 otherwise
likes	Number of likes for the post
replies	Number of replies for the post on Twitter and Number of comments on Facebook
retweets	Number of retweets for the post on Twitter and number of shares on Facebook
sentiment_score	A score assessing the sentiment. Calculated by FUOC.
interaction_abs	Total number of interactions. The sum of all possible interactions provided by either Facebook or Twitter.

ontopic_score	A score assessing the degree of on topic. Calculated by FUOC
numElements_europe	Number of keywords relating to Europe. Calculated by FUOC
numElements_theme	Number of keywords relating to the topic. Calculated by FUOC
text	The text of the post/tweet
URL	The URL to the post/tweet
Sep	A separator to introduce a blank space

Source: Author's own elaboration

10.4 Manually coding the data

Each excel contains a number of columns that need to be manually coded.

Table displays the columns that need to be manually coded for each excel file.

Note that some columns are the same for both topics, and some are topic specific. How to manually code these columns is explained in detail in the codebook, which can be found in the EUMEPLAT restricted area: <https://restrictedarea.eumeplat.eu/>

Table 6. Columns to be manually coded by gender and im/migration

	Gender	Im/migration
On topic	X	X
Law	X	X
People	X	X
Culture	X	X
Values	X	X
Identity	X	
New social movements	X	
Public sphere	X	
Territory		X
Institutions		X
Interactions & dialogue		X
Sentiment	X	X

From now on, we refer to posts from Facebook and tweets from Twitter as **social media posts**.

1. When coding you should **ONLY** assess the context of the **text** of the social media post.

We will carry out **intercoder reliability checks** (*See next section for details of the reliability check*).

You are **done coding** when you have:

- a. Passed an **intercoder reliability check** calculating Krippendorff's alpha on the 20% of the total items.
 - b. Manually coded a total of **200 social media posts** that are **on topic** in **each of the 4 excel files**.
2. In total, you need to find 800 posts on-topic.
 3. A social media post is **on-topic** if it relates to the topic in the name of the excel file (either im/migration or gender). Otherwise, the social media post is off-topic. Details are provided in the codebook.
 4. Importantly, you **DO NOT** need to assess whether the social media post concerns Europe or not.
 5. For each post that is **on topic**, you need to manually code all columns that require manually coding (the columns with no data).
 6. For each post that is off-topic, you only need to code the column On topic.
 7. The coders should code the social media posts in order of appearance in the excel file.
 8. For the intercoder reliability check (*see next section for details*), 20% of the data should be manually coded by two coders in each excel file. Therefore, the coding proceeds in a number of stages:
 9. **Stage 1:**
 - a. Both coders are given training data to practice and ensure alignment in coding (*see next section*).
 - b. When the training has generated a satisfactory level of alignment, the first coder manually codes until he/she has coded 40 posts on-topic in each excel file. 160 posts in total.
 - c. Once the first coder has coded 40 posts on topic in an excel file, the second coder is provided these 40 posts, not saying whether they are on or off topic. The second coder manually **codes all columns that require coding** (the columns with no data), regardless of whether the second coder finds them to be on or off topic. This

is repeated for all excel files until the second coder has coded all 160 posts coded by the first coder.

- d. Thereafter, an intercoder reliability check is carried out (see next section).
- e. If the intercoder reliability check results in a Krippendorff's Alpha greater than equal to 0.67 for all variables/columns, the team has passed the intercoder reliability check and can proceed to Stage 2. Otherwise, they proceed to "Stage 1 repeated".

10. Stage 1 repeated:

- a. Both coders are given additional training data to practice and ensure alignment in coding (see next section).

Stage 1 repeated follows the same steps as Stage 1 **with one difference:**

- b. When the additional training has generated a satisfactory level of alignment, the first coder manually codes a **NEW set of posts** until he/she has coded 40 posts on topic in each excel file. In other words, the first coder does **NOT code any posts that were coded in any previous stage**. In total 160 posts.
- c. Once the first coder has coded 40 posts on topic in an excel file, the second coder is provided these 40 posts, not saying whether they are on or off topic. The second coder manually **codes all columns that require coding** (the columns with no data), regardless of whether the second coder finds them to be on or off topic. This is repeated for all excel files until the second coder has coded all 160 posts coded by the first coder.
- d. Thereafter, an intercoder reliability check is carried out.
- e. If the intercoder reliability check results in a Krippendorff's Alpha greater than equal to 0.67 for all variables/columns, the team has passed the intercoder reliability check and can proceed to Stage 2. Otherwise, they proceed to "Stage 1 repeated" again.

- 11. Stage 1 repeated** is repeated until the team passes the intercoder reliability check. Naturally, if too much discrepancy is found in some column/variable across all teams, revision of that column/variable may be necessary.

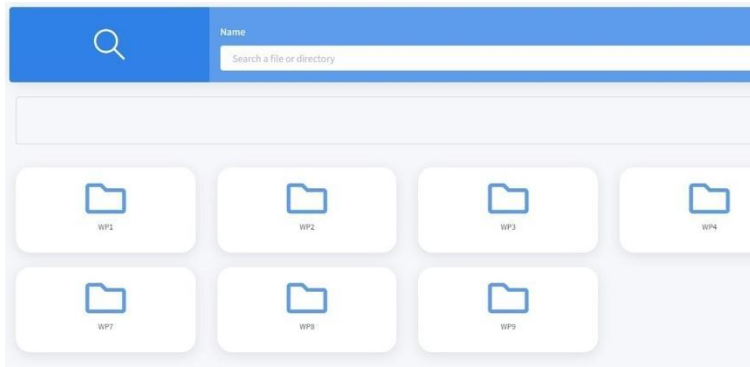
Note that for the column sentiment we can use a lexicon approach if it does not pass the check, but this is not applicable to the other columns.

12. Stage 2:

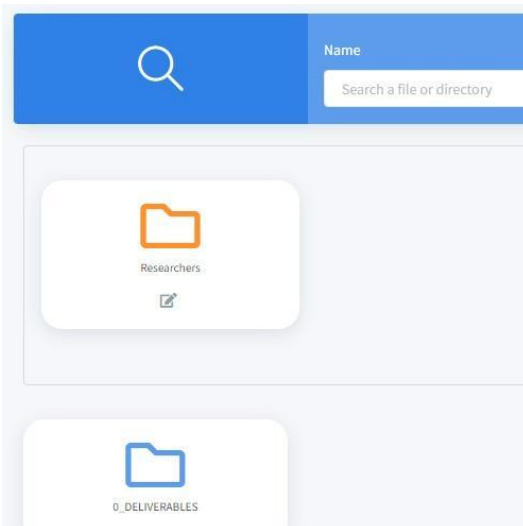
- f. One (or both) coders complete the coding until they have found a total of **200 posts on-topic in each of the four excel**.
- g. If necessary, the coders manually **re-code** the social media posts used in Stage 1 and Stage 1 repeated.
- h. If you do not find 200 social media posts on topic in an excel file, you code all social media posts in that excel file **AND** code additional posts in the other excel files until you have coded 800 posts that are on topic in total.
- i. If 800 social media posts on-topic cannot be found across all four excel files, you code all posts in all excel files.

13. When you are done coding an excel file you upload it to the EUMEPLAT restricted area by taking the following steps:

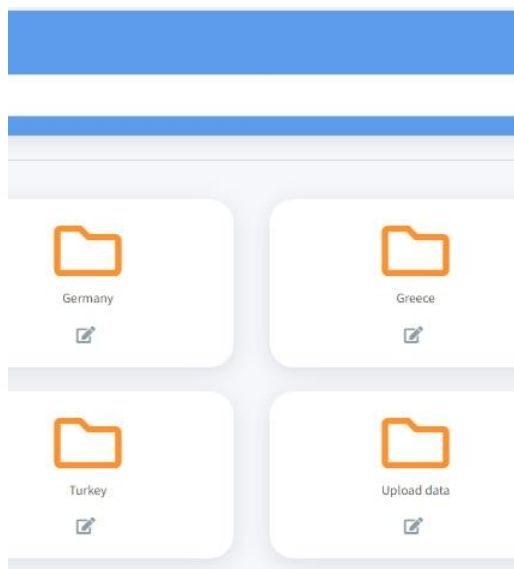
- Go to the folder named WP4.



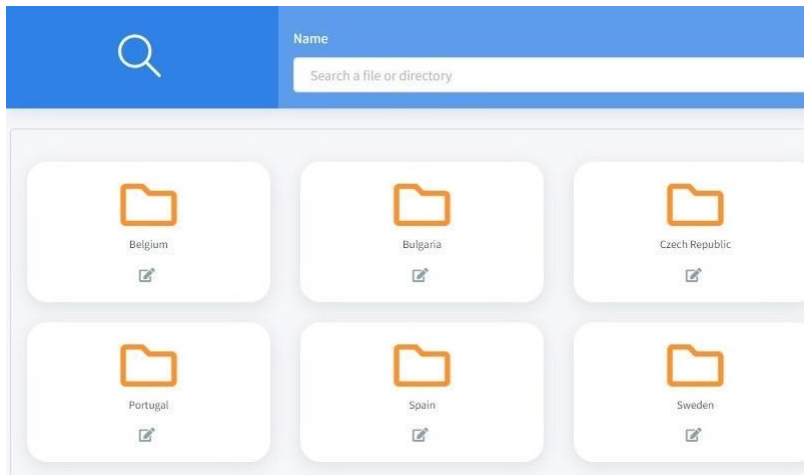
- Go to the folder named Researchers.



- Go to the folder "Upload data".



- Go to your country's folder.



- Once inside your country's folder, you upload the excel file(s).

14. The deadline for providing Stage 1 intercoder reliability check is set up if option A will be the selected one.

15. The deadline for providing all 800 manually coded posts is: **22/12/2022**

10.5 Intercoder reliability check

To be done with the coding, an intercoder reliability check needs to be passed by the team. While the general structure of the intercoder reliability check was outlined in the previous section, details are provided in this section. The structure follows very closely that of WP2 and we are consequently extremely thankful for all the work by the teams from ISCTE (Portugal), CU (Czech Republic), and others on which most of this is based.

10.6 Overview

Each partner/team needs to have minimally 2 coders.

One team member will act as trainer/coordinator. This person can be the main coder or the second coder or simply maintain the role of the trainer/coordinator [depending on time availability and resources]. The trainer/coordinator introduces the coding project, the codebook, and the datasets to the other coder/s.

10.6.1 Training

1. All involved coders code independently a few posts in selected datasets across platforms and dimensions, e.g.:
 - 3 posts in GEN_EUR
 - 3 posts in GEN_noEUR
 - 3 posts in MIG_EUR
 - 3 posts in MIG_noEUR

The trainer/coordinator will distribute the posts to the coder/s of the team. These posts can be selected from the last posts in each excel file that have not previously been used in any training.

2. a. The trainer/coordinator compares the trial coding and identifies differences in coding, in each question.
b. In a consultation meeting, led by the trainer/coordinator, the differences in coding are discussed among the coders. Reasons of coding differently and solutions on how to deal with the discrepancies are discussed and decided upon.

General issues in coding [irrespective of national particularities] are addressed with the FUOC (Spain) team. National particularities, instead, are discussed and dealt with internally by each team. A collaborative document where clarifications in coding are clustered, compiled by the trainer/coordinator, and regularly updated, will facilitate the process.

3. After the trial coding and consultation [stages 2-3] one, two, or more rounds of trial coding are repeated, depending on the needs, following the same procedure [code independently a small number of posts across datasets, compare coding, discuss and consult, update clarifications]. It is important, even though the coders discuss and coordinate, that they code independently.

10.7 Data for coding for the Intercooder Reliability Check

The data for the intercoder reliability check differs depending on which stage of the coding procedure the team is in, as explained above. Specifically:

Stage 1:

1. When the training has generated a satisfactory level of alignment, the first coder manually codes until he/she has coded 40 posts on-topic in each excel file, starting from above. 160 posts in total.
2. Once the first coder has coded 40 posts on-topic in an excel file, the second coder is provided these 40 posts, not saying whether they are on or off-topic. The second coder manually **codes all columns that require coding** (the columns with no data), regardless of whether the second coder finds them to be on or off-topic. This is repeated for all excels until the second coder has coded all 160 posts coded by the first coder.

Stage 1 repeated:

3. When the additional training has generated a satisfactory level of alignment, the first coder manually codes a **NEW set of posts** until he/she has coded 40 posts on-topic in each excel file. In other words, the first coder does **NOT code any posts that were coded in any previous stage**. In total 160 posts.
4. Once the first coder has coded 40 posts on-topic in an excel file, the second coder is provided these 40 posts, not saying whether they are on or off-topic. The second coder manually **codes all columns that require coding** (the columns with no data), regardless of whether the second coder finds them to be on or off-topic. This is repeated for all excels until the second coder has coded all 160 posts coded by the first coder.

Providing the second coder with 40 posts once the first coder has finished coding these is done to speed up the process.

10.8 Intercooder reliability calculations

After the first coder and the second coder code independently the 160 posts, the trainer/coordinator calculates the intercoder reliability levels, **for each question separately**.

The coefficient that we will be using is Krippendorff's alpha (applicable to any number of coders and adjusts itself to small sample sizes). The admissible level of the agreement will be 0.67, for each question.

We will not use percentage agreement as it tends to overestimate reliability and may be misleading. Specifically, you should:

Copy and paste into an excel sheet the coding results of the two coders for each question side by side, one next to the other. For example, copy and paste the coding results of coder 1 for question 1 in column A, then the coding results of coder 2 for question 1 in column B. Then, the coding results of coder 1 for question 2 in column C and the coding results of coder 2 for question 2 in column D, etc. You continue until you paste the coding results of the two coders in the same excel sheet, side by side, for the 12 dimensions in total. In the end, the excel sheet will have 24 filled columns and 160 rows.

It is important that this excel file does not contain any other information [e.g., column titles, the numbering of columns or rows, etc.].

After you have pasted all the coding results of the two coders, save the excel file you have created, also as a **CSV file** [you can find this option if you go to your excel file and select 'save as']. Remember to keep an **xls** version of the file.

To calculate the Krippendorff's alpha, we use: <http://dfreelon.org/utis/recalfront/recal2/>

Deen Freelon, Ph.D.

Associate professor, Hussman School of Journalism and Media, UNC-Chapel Hill

ReCal2: Reliability for 2 Coders

ReCal2 ("Reliability Calculator for 2 coders") is an online utility that computes intercoder/interrater reliability coefficients for **nominal** data **coded by two coders**. (Versions for [3 or more coders working on nominal data](#) and for [any number of coders working on ordinal, interval, and ratio data](#) are also available.) Here is a brief feature list:

- Calculates four of the most popular reliability coefficients for nominal data: percent agreement, Scott's Pi, Cohen's Kappa, and Krippendorff's Alpha.
- Can calculate reliability for multiple variables at a time
- Accepts any range of possible variable values
- Results should be valid for **nominal data coded by two coders** (other uses are not endorsed, and accurate results are not guaranteed in any case — trust but verify!)

If you have used ReCal2 before, you may submit your data file for calculation via the form below. If you are a first-time user, please read [the documentation](#) first. (Note: failure to format data files properly may produce incorrect results!) You should also read ReCal's [very short license agreement](#) before use.

<input type="button" value="Choose file"/>	No file chosen	<input type="button" value="Calculate Reliability"/>
--	----------------	--

Click on 'Choose file'. Select the csv file of the two coders' coding results. Then click 'Calculate Reliability'.

Download the results and send them to flupianez@uoc.edu

11 Appendix 3. Ethical approval



Evaluation by the Ethics Committee of the UOC

Exp.: CE22- PR23

Dr. Marta Aymerich, president of the Ethics Committee of the Universitat Oberta de Catalunya

CERTIFIES

That the Committee has evaluated the addenda submitted by Francisco Lupiáñez Villanueva for the project already approved that is entitled “EUMEPLAT: European Media Platforms: assessing positive and negative externalities for European Culture”, and considers that

- The ability of the researchers and their collaborators, and the facilities and resources available are adequate to carry out the study.
- The established experimental protocol ensures the integrity and dignity of the participants.
- The protocol is adequate to the objectives of the study and the possible risks and discomfort for participants are adequate given the expected benefits.
- The procedure for obtaining informed consent of participants, including the information sheet, and the procedure for the recruitment of subjects are adequate.
- The researchers of the project will ever respect the obligations derived from the Organic Law 3/2018 on Personal Data Protection and Digital Rights, General Regulation on Data Protection (UE) 2016/679 and the current complementary legislation.

Having met on July 26, 2022, and having considered the ethical implications concerning human experimentation and the processing of personal data, this committee APPROVES the execution of the aforementioned project.

For the record, I sign this document in Barcelona, July 26, 2022.

Signed:

VIDsigner code: F4A1E215E49040B885...

A green rectangular box containing a signature. Below the box, the name 'Marta Aymerich Martínez' is printed in a small, black, sans-serif font.

Dr. Marta Aymerich,
Av. Tibidabo, 39-43
08035 Barcelona – Spain
Tel. +34 93 253 23 00
Fax +34 93 417 64 95

Get in touch

 info@eumeplat.eu

 www.eumeplat.eu



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101004488

