EUMEPLAT
European Media Platforms: assessing positive and negative externalities for European culture

Deliverable 4.4

Aggregated Data Analysis Report

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

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 – Exclusion: platformization of Media Representations
Related task(s):	T4.4 – Aggregated Data Analysis
Lead Organisation:	P10 - UNIVE
Author(s):	Debashmita Poddar, UNIVE Fabiana Zollo, UNIVE
Status	Final
Submission date:	10/10/2023
Dissemination Level:	Public

History

Date	Submitted by	Reviewed by	Version
31/08/2023	P10-UNIVE	Steering Committee	V1.0
19/09/2023	P10-UNIVE	Steering Committee	V1.1
07/10/2023	P10-UNIVE	Coordinator	V1.2

Table of Contents

Executive summary

Deliverable 4.4 primarily focuses on conducting a longitudinal quantitative cross-country analysis concerning migration. As outlined in the Declaration of Activities (DoA), our primary objective was to provide an advanced quantitative data analysis to examine the European discourse surrounding the subjects central to WP4, specifically gender and migration. However, due to the complexities of conducting a meaningful gender analysis with the presently available social media data, we have chosen to center our investigation on the topic of migration. Nonetheless, we have enriched this analysis by incorporating an additional dimension, namely, misinformation. In order to facilitate equitable comparisons among different countries, we have relied on third-party independent data sources that employ consistent criteria for assessing the reliability of information across various countries. The results and findings of this analysis are thus complementary to the research covered in the preceding tasks of WP4. They will serve as a foundational basis for discussion during the final workshop dedicated to addressing misinformation, scheduled to take place in Brussels in February 2024. The decision to emphasize migration in our analysis was also influenced by our commitment to major dissemination efforts and policy-oriented activities in this field. Regarding the overarching framework of WP4, we have comprehensively addressed the representation of both gender and migration. This encompassed a simultaneous analysis of the social media discourse across ten countries, spearheaded by the lead team of WP, the Open University of Catalonia (OUC). For our secondary investigation, we have adopted two distinct methodological approaches: 1) Our exploration of the longitudinal thematic aspects of migration is detailed in Deliverable D4.4. The outcomes of this analysis will contribute to the content of a public event centered on the portrayal of migration. This event, co-hosted by UNIMED and IOM, is scheduled to take place in Rome in November 2023. 2) In conjunction with our data-driven analysis, we have undertaken a qualitative examination of the intersectionality between gender and migration discourses in three European languages – namely, Italian, Greek, and Dutch. This is an additional task that was not originally outlined in the DoA. The public conference on gender issues, set to be hosted by ISCTE-IULM in Lisbon on November 20-21, 2023, will provide an opportunity for the initial insights from this study to be presented.

1 Introduction

Misinformation has become a prevalent and concerning phenomenon in today's interconnected world. With the ease of information dissemination through social media and

online platforms, false narratives and misleading content can quickly spread, shaping public debates surrounding migration (Wu et al., 2019). One of the most significant challenges posed by misinformation on migration is its possible impact on public attitudes towards migrants and refugees. False or misleading narratives often depict migrants as criminals, economic burdens, or threats to national security, fostering fear and hostility towards these vulnerable populations (Ruokolainen and Widén, 2020). Such misinformation not only perpetuates stereotypes but also undermines the principles of empathy and compassion that are crucial for building inclusive and cohesive societies. In some cases, misinformation on these issues are intentionally disseminated for political or ideological purposes. Populist movements and some media outlets may exploit fears and concerns surrounding migration to advance their agendas. This deliberate distortion of facts can lead to the polarisation of public opinion and the rise of anti-immigrant sentiments, hindering meaningful and evidence-based policy discussions (Lewandowsky et al., 2020).

In this report, we take into account 112,548 tweets that are related to the topic of migration in selected European countries, specifically France, Germany, Italy, and the UK, over the period of three years (2019, 2020, 2021). We analyse the sources of the tweets and divide them into Trustable and Untrustable categories (see Section 2 for details) for each of the above mentioned countries. First, we investigate the production of content on migration issues, analysing the number of tweets published in each country over the years. Second, we study how the general public engages with such content and what the most discussed topics are in both the Trustworthy and Untrustworthy sources for the four countries. Finally, we will delve into the similarity among retweeters of both Trustworthy and Untrustworthy sources. This analysis will help us gain insights into whether the general public is being exposed to content from both of these sources, thus shedding light on the level of segregation in the public debate around migration within the selected countries.

2 Data Collection

To identify misinformation, we rely on the NewsGuard dataset, which assesses the credibility and transparency of news and information sites based on nine apolitical criteria[1]. NewsGuard categorises sources with a credible score of over 60% as Trustworthy sources, while those with a score below 60% are considered Untrustworthy sources. Untrustworthy sources may propagate misinformation, disinformation, and false news. Starting with the list provided by

[1] https://www.newsguardtech.com/ratings/rating-process-criteria/

NewsGuard, we collected data from the witter timelines of these news source accounts using the Twitter API for academic research[2]. To focus on the migration topic, we used specific keywords in French, German, Italian and English which translates to “migration”, “immigration”, “refugee”, “immigrants”, “migrants” .The complete list of keywords for each language is reported in Table 1. The breakdown of the news sources included in the dataset is reported in Table 2.

Table 1. Keywords used to filter contents on Migration related tweets in selected countries.

France	**Germany**	**Italy**	**UK**
Migration	*Migration*	*Migrazione*	*Migration*
Immigration	*Einwanderung*	*Immigrazione*	*Immigration*
Migrantes/Migrants	*Migranten*	*Migranti*	*Migrants*
Immigrants	*Einwanderer*	*Immigrati*	*Immigrants*
Réfugiée/Réfugié	*Flüchtlinge*	*Rifugiati*	*Refugees*

Table 2. Breakdown of the news sources per country and credibility on the topic of migration

Country	**Trustworthy sources**		**Untrustworthy sources**		**Total** *(Covering Migration)*
	General	**Covering Migration**	**General**	**Covering Migration**	
France	187	99	49	18	117
Germany	196	73	25	5	78
Italy	175	113	29	20	133
UK	191	103	22	15	118
Total	749	388	125	58	446

All gathered data is publicly available and data from private accounts is not included in our dataset. The dataset includes all the tweets published by the selected accounts during the period of 01 January 2019 to 11 November 2021. Table 3 shows the breakdown of the data

[2] https://developer.twitter.com/en/products/twitter-api/academic-research

by source category, along with the percentage that represents tweets of each country's posts based on their trustworthiness.

Table 3. Breakdown of the data after filtering for migration issue

Country	Number of tweets	Tweets from Trustworthy sources	Tweets from Untrustworthy sources
France	30,277 (26.86%)	26,248 (26.42%)	3,979 (30.18%)
Germany	7,456 (6.62%)	6,535 (6.58%)	921 (6.98%)
Italy	47,763 (42.44%)	39,581 (39.83%)	8,812 (62.05%)
UK	27,102 (24.08%)	26,999 (27.17%)	103 (0.78%)
Total	112,548	99,363	13,815

3 The public discourse on migration

This section deals with migration content that is being produced on Twitter over the years of 2019 to 2021 in the selected European countries. First, we study content production and engagement for each country. Second, we dive deeper into the most popular topics by year and country, considering both Trustworthy and Untrustworthy sources.

3.1 Content production and engagement

Figure 1 illustrates the level of activity of news sources in each country, measured by their production of tweets over the course of three years. To enable a fair comparison across countries, we use a relative measure by dividing the number of tweets with the number of news sources for the specific country. We notice that, in France, news sources were more active in the final quarter of the year 2019 with a sharp spike during the 2020 start of the pandemic. Germany recorded the highest number of active news sources in the year 2020. Italy's news sources record its highest activity during the last quarter of 2020. The UK has the highest number of active news sources compared to the other three countries. We also notice a sharp increase in content production for both Trustworthy and Untrustworthy news sources from each country during the Covid-19 pandemic era.

Untrustworthy news sources are seen to be relatively way more active than their trustworthy counterparts in every country. The number of tweets produced on the topic of migration decreases gradually at the end of 2020, except for the UK, which maintains its high content production status with a sharp decrease at the end of the first quarter of 2021. France shows

a clear increase in its trustable news content sources starting from December 2020. Germany's news sources are active mostly during the two major waves of COVID-19 in July 2020 and December 2020. Untrustworthy news content dominates Germany's twitter scene. Italy and the UK show a fair balance between the ratio, however, the Untrustworthy score is always higher.

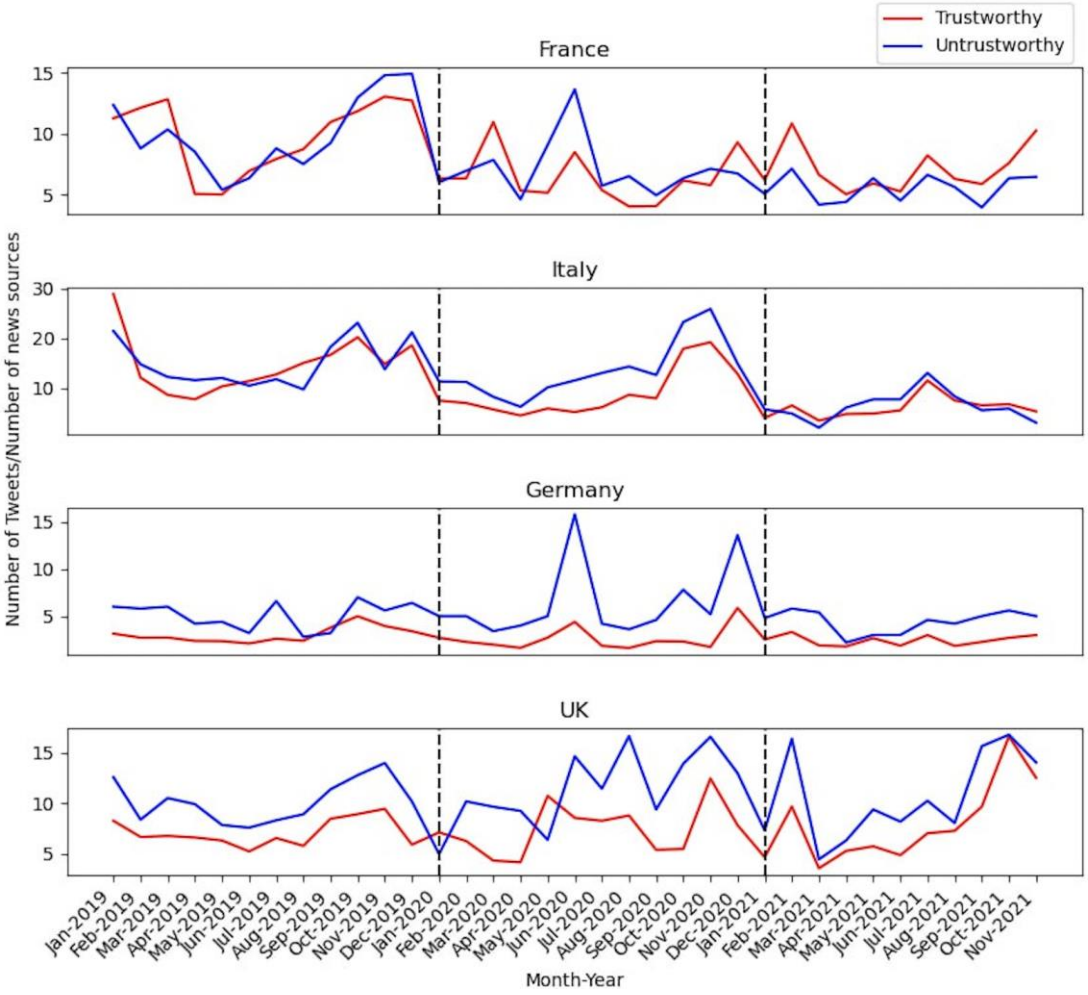

Figure 1. News production over time. For each month, we show the total number of tweets produced in each country divided by the total number of news sources in that country grouped by the credibility of the news sources. Dashed lines indicate transition to the next year.

Figures 2 and 3 show the Complementary Cumulative Distribution Functions (CCDFs) of the number of Likes, Quotes, Replies, and Retweets received by the tweets of Trustworthy and Untrustworthy sources. In this way, we can analyse how users interact with content produced by each country's news sources, both Trustworthy and Untrustworthy. The top two engagement factors are attributed to Likes and Retweets followed by Replies and Quotes.

We notice that in the UK Trustworthy sources receive the highest amount of overall engagement in all the four categories. This can be attributed to the amount of content the UK generally produced that can be seen in Figure 1. France is the next country with the second highest interactions under the Trustworthy category, followed by Italy and Germany. Germany seems to lag behind the Quotes category, and their highest form of interaction is via Likes.

More variations can be observed in Figure 3, which shows the engagement factors under the Untrustworthy news sources. Italy has the highest number of interactions overall in all the four categories, followed by Germany, which shows nearly identical structure with its Trustworthy version, Likes being its primary form of interaction. We notice that in 2019, France was engaging a lot with retweets, however the interactions were reduced when compared to its Trustworthy counterparts. The UK shows a huge reduction in the engagement factor, especially in 2019, considering that Brexit was the most debated topic for them during that year. The engagement increased in 2020 due to the pandemic and seemingly decreased with the advent of the year 2021.

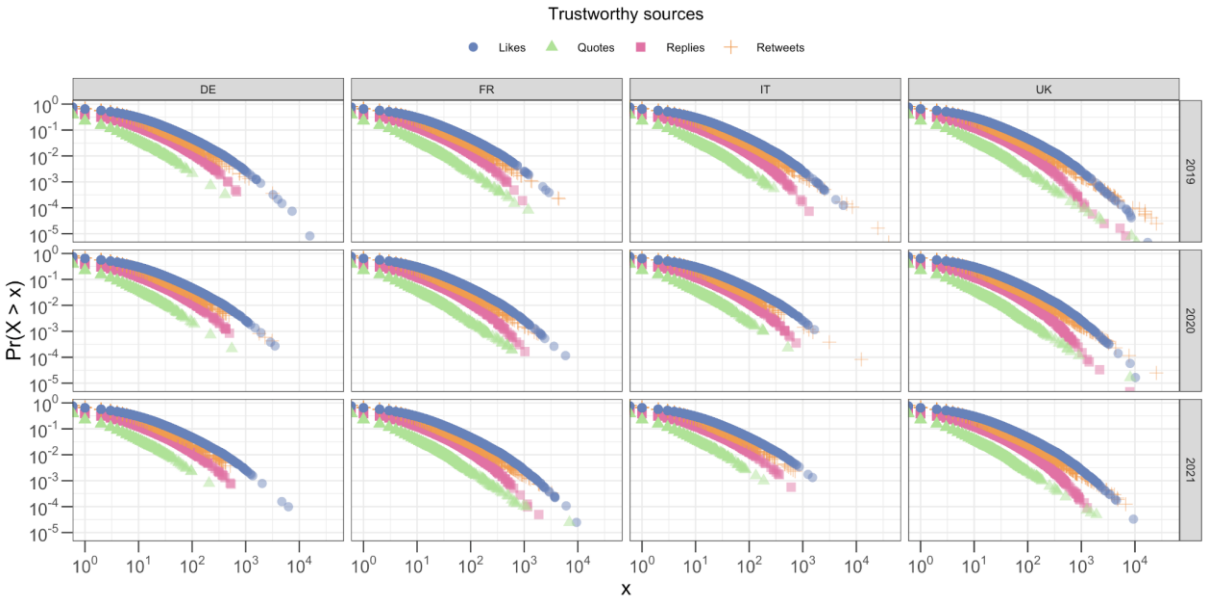

Figure 2. Complementary Cumulative Distribution Functions (CCDFs) of the number of Likes, Quotes, Replies, and Retweets received by tweets of Trustworthy sources for each country over the three years.

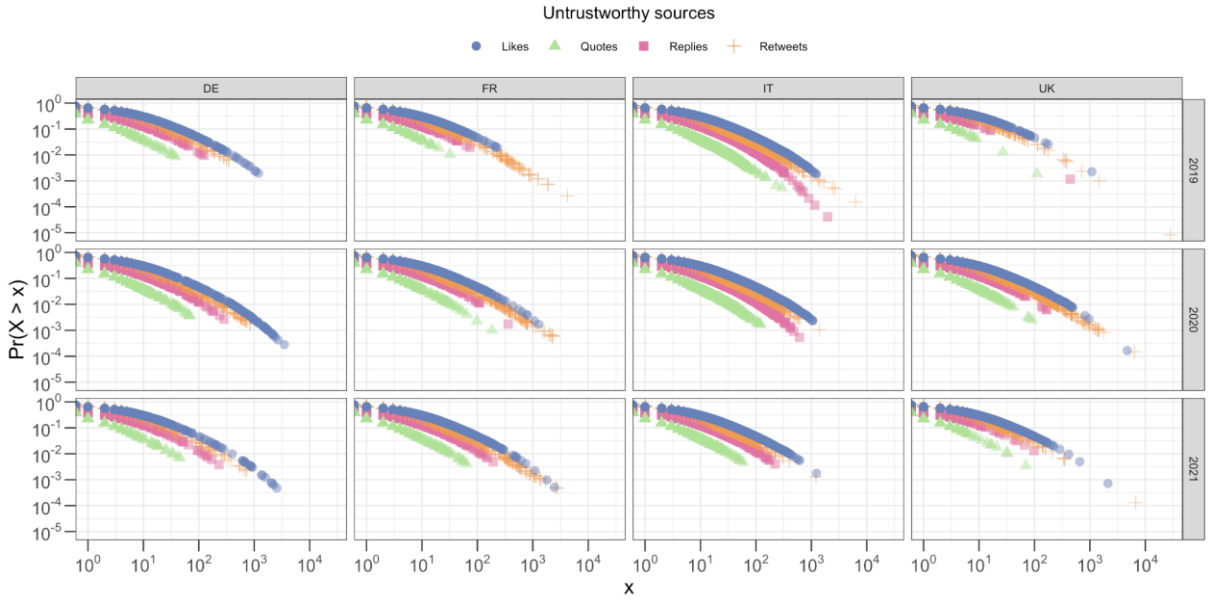

Figure 3. Complementary Cumulative Distribution Functions (CCDFs) of the number of Likes, Quotes, Replies, and Retweets received by tweets of Untrustworthy sources for each country over the three years.

To compare Trustworthy and Untrustworthy engagement, we performed the Kolmogorov-Smirnov test on the CCDFs for each type of reaction (Likes, Quotes, Replies, and Retweets) and for each country. Figure 4 illustrates the p-values associated with these tests. All distributions are statistically distinct, except for "Quotes" in 2019 and "Replies" in 2020, which exhibit a higher degree of similarity.

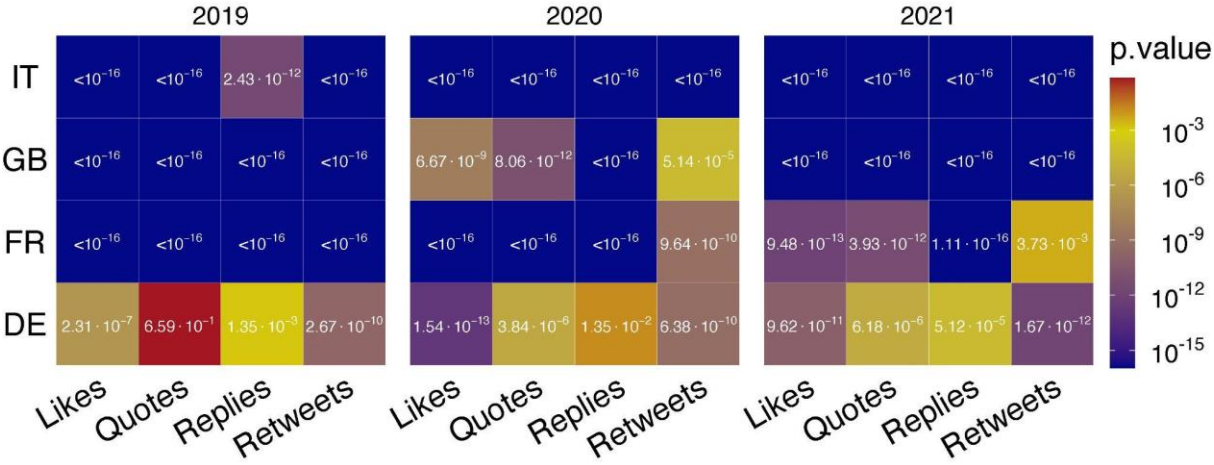

Figure 4. P-values for the Kolmogorov-smirnov test on CCDF engagement distributions between Trustworthy and Untrustworthy sources.

3.2 Migration topics from 2019 to 2021

In this section, we aim to explore the most debated topics within each country and assess the corresponding level of engagement, determined by the sum of reactions divided by the number of tweets associated with each topic. To achieve this, we employ BERTopic (Grootendorst, 2022), a natural language processing (NLP) technique that leverages the BERT (Bidirectional Encoder Representations from Transformers) language model (Devlin et al., 2020) for topic modelling within a collection of text documents. BERTopic combines the capabilities of BERT's contextual word embeddings with topic modelling algorithms to extract coherent and meaningful topics from textual data.

Traditional topic modelling methods, such as Latent Dirichlet Allocation (LDA) (Blei et al. 2003) and Non-Negative Matrix Factorization (NMF) (Lee and Seung, 1999), typically rely on statistical patterns in word co-occurrences to identify topics within a document collection. Conversely, BERT is a pre-trained transformer-based model that learns contextual embeddings for words within a sentence, enabling it to capture more nuanced semantic relationships. The advantages of employing BERTopic include its capacity to capture subtler semantic associations between words, resulting in more precise and coherent topics. Furthermore, BERTopic eliminates the need for explicitly specifying the number of topics, as required in some traditional methods.

Within this context, topic modelling can offer valuable insights into the nature of news discussed in tweets posted by both Trustworthy and Untrustworthy news sources. Utilising BERTopic, we generate a set of keywords for each topic, which are subsequently used to label each topic with informative titles.

Starting with the year 2019, we observe that the most tweeted topic in France under Trustworthy news sources is the increase of immigrants, whereas under the Untrustworthy news sources, it’s rescuing immigrants followed up by the invasive nature of migration. For Germany, Trustworthy news sources dominate the tweets mostly talking about Islamic refugees. For Untrustworthy news sources, the most discussed topic was the criticism over the video leaked of Merkel allowing 1 million refugees in Germany back in 2015. Italy’s most discussed topic on trustworthy sources include immigrants and ships, which is the most common way of entering Europe with Italy serving as the base. Untrustworthy sources dominate their tweet pool with topics related to the negative talks by Salvini on migration and the unending political justifications. Lastly, the most talked about topics under both the news sources for the UK is about the refugee crisis that took place after the declaration of Brexit.

Post Brexit analysis concluded that the single strongest issue prompting the UK to leave the EU were Immigration policies.

Unsurprisingly, the year 2020 dominates with the topic of COVID-19 in every country under all news sources. According to the EU commission migration statistics[3], Italy observed an 154% increase in border crossings mainly in Lampedusa when compared to the 2019 numbers around the same period. Most of the topics discussed are generally about the illegal refugees from Afghanistan, Turkey and North Africa from Untrustworthy news outlets.

The year 2021 focused more on the migration crisis caused due to the impending COVID-19 measures and border controls. One central topic for all four countries includes the situation in Afghanistan where the Taliban gained power in August 2021, causing a huge displacement inside the country. Women rights were revoked causing the Afghan women to take refuge in the EU.

Figure 5 also reveals that the similarity in topics discussed among the four countries each year is remarkably consistent. In 2019, the predominant topic was Brexit, while 2020 was dominated by the coronavirus pandemic. In 2021, the most significant news centred around the Taliban takeover of Afghanistan.

[3] https://ec.europa.eu/commission/presscorner/detail/en/ip_21_232

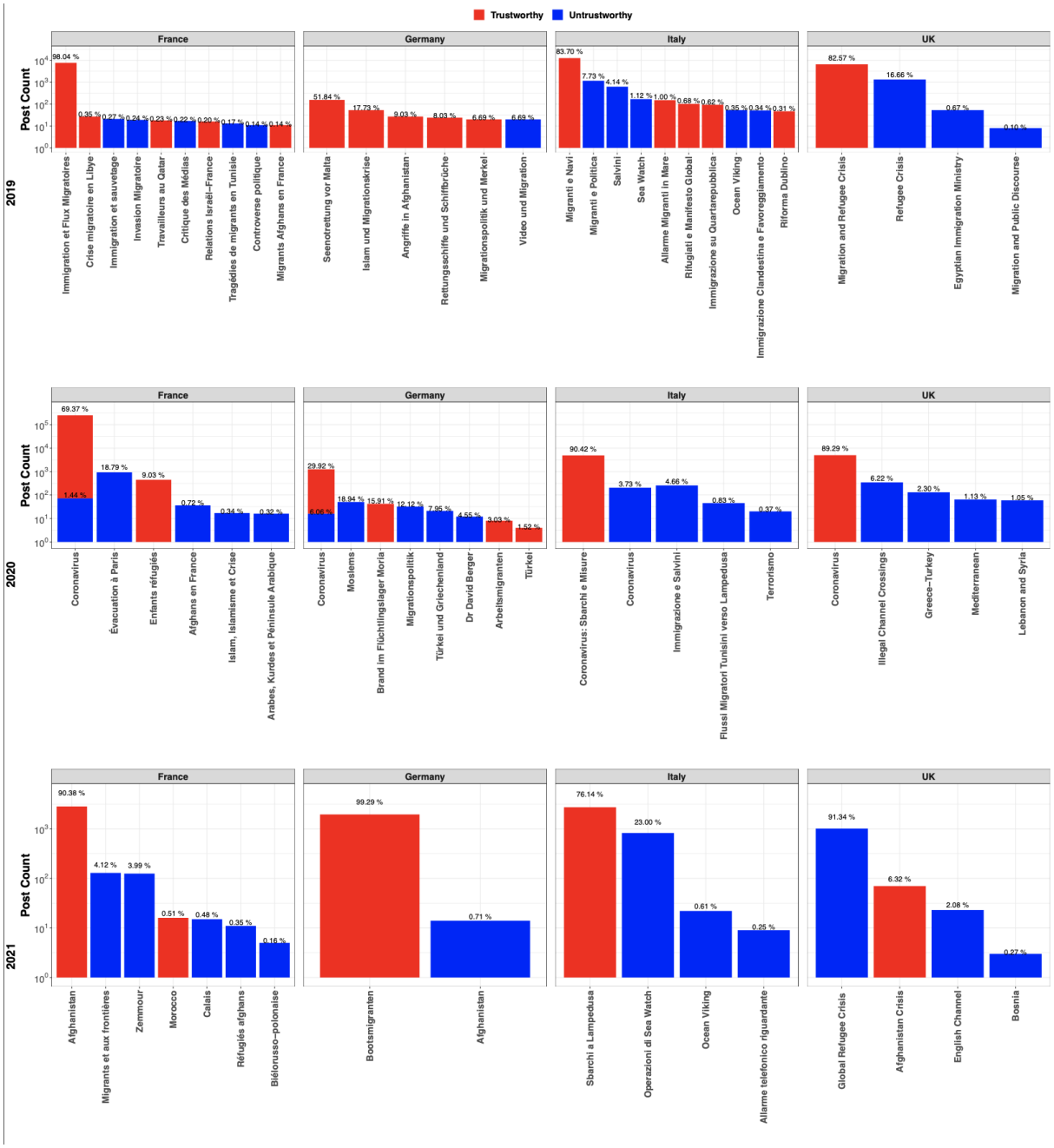

Figure 5. Most debated topics and their percentage of discussion for each country in 2019, 2020 and 2021 divided by the credibility of news sources.

4 Network Analysis

A practical method for exploring and visualising whether users/accounts share a common narrative is to examine the similarity in the news sources' retweeters sets. Such an analysis can be performed by analysing the clustering of nodes representing the news sources and the

edges representing retweeters' similarity. There are two kinds of clustering, content-based, where the semantics of the data is taken into account or structure-based, where the structural information of the data is used to form the clusters. In this report, we use structure-based clustering using the network architecture, and to compute the similarity measure we use cosine similarity.

Cosine similarity takes into account the angles between the list of retweeters for each news source. Each of these lists of retweeters is a vector pointing in a certain direction. If the vectors are pointing in the same direction (meaning they are similar), the cosine similarity will be closer to 1. If they are perpendicular to each other (indicating dissimilarity), the cosine similarity will be closer to 0. If they are somewhat related but not exactly the same, the cosine similarity will be somewhere in between. Cosine similarity is calculated using the dot product of the two vectors (the sum of the products of their corresponding scores) divided by the product of their magnitudes (lengths).

The dataset includes a list of news sources, their Tweet IDs, and the retweeter IDs of those tweets. Using this data, we create a matrix where columns represent the news sources (both Trustworthy and Untrustworthy) and rows represent the retweeters. Each cell of this matrix is filled with a value denoting the number of times a retweeter retweeted content posted by a news source. For example, if a retweeter A retweeted 3 tweets of the news source B, the cell AB in the matrix will contain the value 3. We use this matrix to build a network where nodes are news sources that are connected though edges based on their retweeters' cosine similarity (i.e. the edge weight). Then, we use the median edge weight of the network as a threshold to keep the strongest connections. In other words, if the weight of an edge is less than the median, that edge is discarded from the network.

Figure 6 shows the graphs for each year in the four countries. Trustworthy news sources are marked as a red triangle node, while the blue circular nodes depict the Untrustworthy sources. For visualisation purposes, the thickness of the edges connecting the nodes is proportional to the level of similarity of the retweeters between the news sources. Thus, when two nodes share similar retweeters, they are connected by an edge. The higher the similarity among their retweeters, the thicker the edge connecting them becomes.

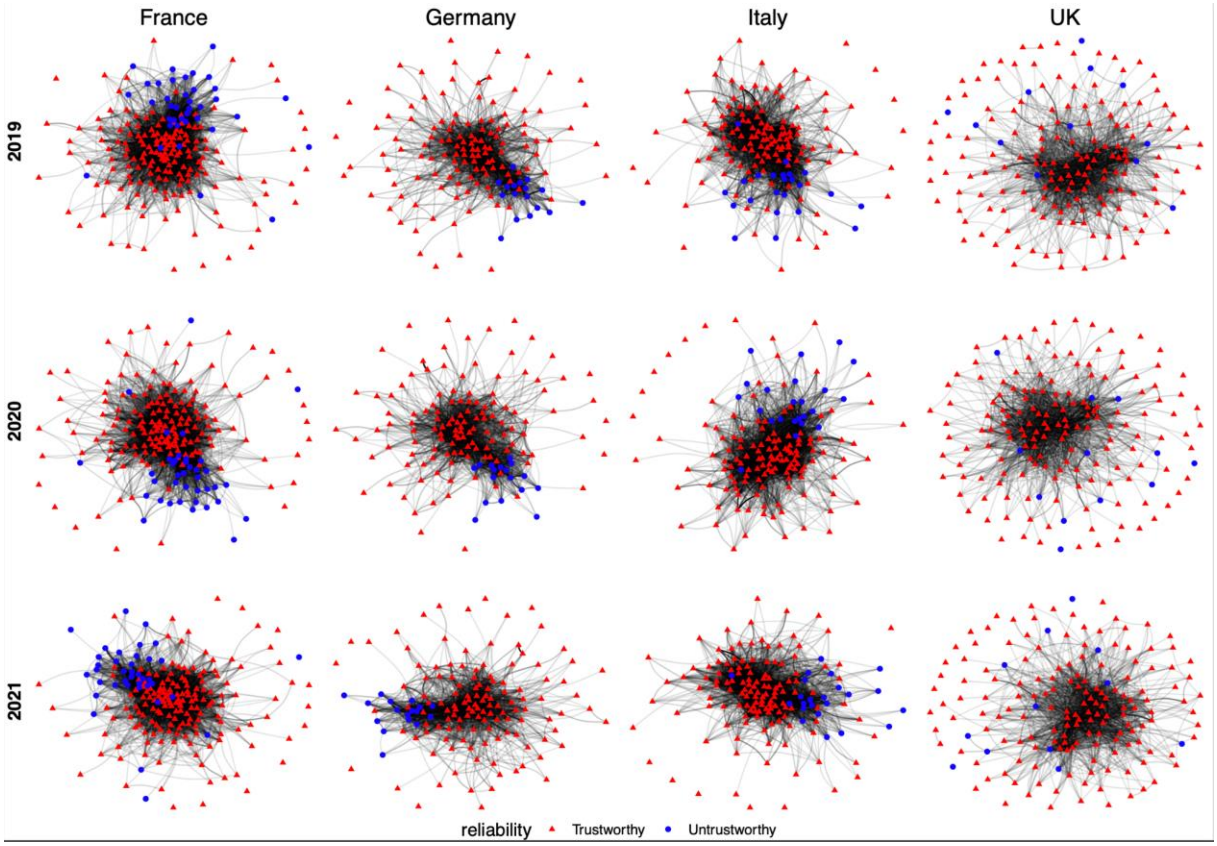

Figure 6. Network Analysis using cosine similarity for each country over the period of three years.

Starting with France for all the three years, we observe that the Largest Connected Component (LCC) –i.e., the largest set of nodes connected to one another– is composed of both Trustworthy (red nodes) and Untrustworthy (blue nodes) sources. This suggests that the retweeters in France have a mixed content consumption, with users engaging in contents posted by both Trustworthy and Untrustworthy sources. Germany also has a similar structure with France, however, we observe that the network is not as dense as that of France (edge density: 0.20 vs 0.25 in all three years), due to a much lower number of sources as compared to France or Italy (Table 2 for reference). Italy has a more densely connected network (edge density 0.27), as the previous two countries and exhibits a similar structure of retweeters being exposed to contents from the two kinds of sources.

The networks of the UK for 2019, 2020 and 2021 have the most different structure as compared to France, Germany and Italy. We observe an increase of isolated nodes and nodes with thicker edges, thus showing little to no similarity between the retweeters of those sources, if any. We also notice that the most concentrated part of the network consists mainly of red nodes, thus depicting that retweeters are not engaging the same way for Trustworthy and Untrustworthy news sources. We may notice that the network is not as dense as for the other

countries (0.13). This may suggest that users generally consume content from a narrow set of popular Trustworthy and Untrustworthy sources, tending to ignore most other Trustworthy news sources.

France, Germany and Italy exhibit similar structure in the graph amongst themselves over the three consecutive years. We can notice that the LCC of all the networks in Figure 6 is composed mainly of red nodes (81% for France, 88% for Germany, 85% for Italy, 93% for the UK), thus concluding that the information ecosystem is dominated by Trustworthy sources. Moreover, clustering analysis (Figure 7) revealed the presence of highly connected groups of Trustworthy nodes suggests that most of the audience tend to consume mainly content from these news outlets. However, the fact that clusters of Untrustworthy nodes are found in the largest connected component of the networks implies the presence of some users with a mixed diet.

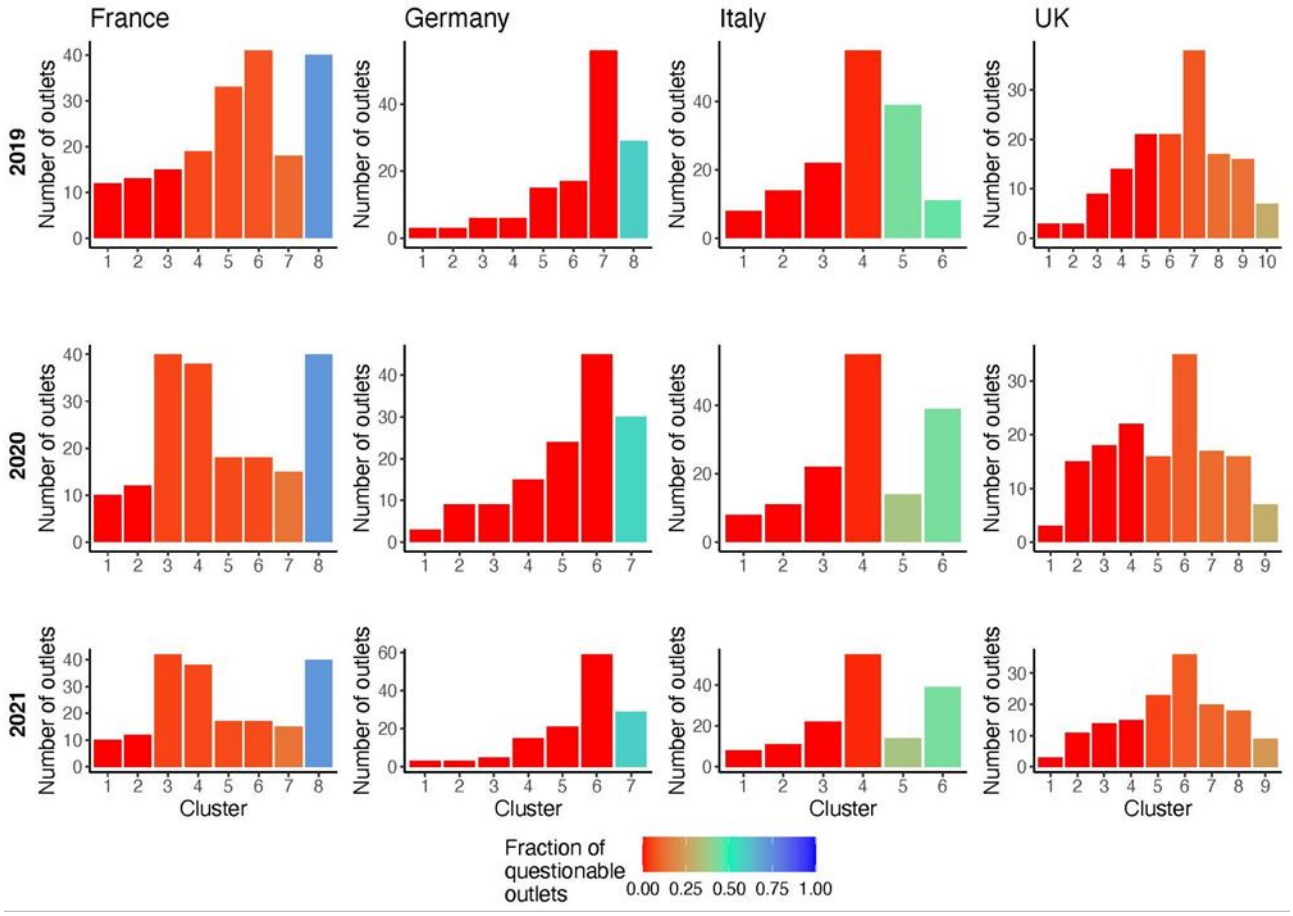

Figure 7: Cluster analysis with Louvain algorithm of cosine similarity networks. Colours represent the fraction of questionable nodes in each cluster.

5 Conclusions

In this report, we have conducted a cross-country analysis to gain a deeper understanding of how the topic of migration is discussed by both Trustworthy and Untrustworthy news sources, as well as how it is consumed by users, in selected European countries. Our analysis encompassed the examination of news source activity related to this topic over a three-year period and delved into the ways in which users engage with the content from these sources. The analysis provides valuable insights into how specific information is consumed and disseminated within each respective country. Furthermore, we studied the specific topics that were being discussed and identified key issues of each year and the concerns related to migration within these countries. Lastly, we employed network analysis to assess whether the public is being exposed to the content from both Trustworthy and Untrustworthy news sources. Indeed, the lack of exposure to diverse sources can contribute to an increase in polarisation, potentially leading to the formation of echo chambers. This, in turn, can facilitate the rise of radicalised or misinformed opinions.

We would like to acknowledge certain limitations in this analysis:

- *Misinformation Classification*: Due to the vast volume of data generated on Twitter and the quantitative approach utilized in this analysis, conducting a manual inspection of the content is not feasible. Consequently, we are unable to evaluate each individual tweet produced by a specific source to determine its trustworthiness. To address this challenge, we rely on NewsGuard, an independent third-party organization, to assess whether content is considered trustworthy or not. NewsGuard's classification is at the source level, meaning that the trustworthiness of a tweet is determined by the credibility of the source that produced it. Therefore, content from a trustworthy news source will be labeled as such, and vice versa.

- *Country Selection*: This analysis focuses on European countries where NewsGuard's data is available. This approach enables us to make meaningful comparisons across countries using consistent definitions and classifications for misinformation. However, it is important to note that this choice limits our analysis to public debates surrounding migration within these specific countries.

- *Social Media Platform Selection*: Twitter, as one of the major social media platforms, plays a significant role in hosting socio-political debates. Furthermore, at the time of data collection, Twitter provided access to its historical data through an official API, making this analysis feasible. Nevertheless, we acknowledge that Twitter may not be

entirely representative of the entire population, and populations can differ across social media platforms and countries.

- *Methods*: To analyze the topics being discussed, we employ BERTopic. However, it's worth noting that BERTopic is limited to handling textual content only, lacking the capability to process multimodal inputs that include images and videos alongside text. Analyzing content containing multimedia elements would require additional models in conjunction with BERTopic, which were beyond the scope of this analysis.

Despite these acknowledged limitations, we believe that this analysis offers a valuable overview of the migration debate over recent years in the selected European countries. It also includes an examination of how misinformation has covered this issue. Such an overview is crucial for developing solutions aimed at mitigating the impact of polarisation and the proliferation of radicalised viewpoints within the general public, and at increasing exposure to high-quality news content.

To effectively combat misinformation, a comprehensive approach is imperative. This involves promoting media literacy and critical thinking education, which empowers individuals with the skills to differentiate between reliable information and misleading content. Equipped with this knowledge, individuals can form more informed opinions on specific issues.

To make progress in this direction, collaborative initiatives that involve academic scholars, governments, civil society organizations, and tech companies can play a pivotal role in creating a safer online environment and disseminating accurate information across various topics. Furthermore, policymakers can leverage evidence-based research and data to base their decisions and develop tailored communication strategies and solutions.

Misinformation related to migration poses significant challenges to public attitudes, policy discussions, and social cohesion. Addressing this issue requires a collective effort to promote media literacy, support independent journalism, and establish transparent and evidence-based policy-making processes. By countering misinformation, societies can foster greater understanding and empathy towards migrants and refugees, and work towards inclusive and humane migration policies.

6 References

Blei, D., Ng, A., Jordan, M.I., 2003. Latent Dirichlet Allocation. Journal of Machine Learning Research 2003, pp. 993–1022.

Devlin, J., Chang, M., Lee, K., & Toutanova, K., 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. ArXiv, abs/1810.04805.

Grootendorst, M., 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. https://doi.org/10.48550/arXiv.2203.0579

Lee, D.D., Seung, H.S., 1999. Learning the Parts of Objects by Non-negative Matrix Factorization. Nature 401, pp. 788-791

Lewandowsky, S., Smillie, L., Garcia, D., Hertwig, R., Weatherall, J., Egidy, S., Robertson, R.E., O'connor, C., Kozyreva, A., Lorenz-Spreen, P., Blaschke, Y. and Leiser, M., Technology and Democracy: Understanding the influence of online technologies on political behaviour and decision-making, EUR 30422 EN, Publications Office of the European Union, Luxembourg, 2020, ISBN 978-92-76-24088-4, doi:10.2760/709177, JRC122023.

Ruokolainen, H., Widén, G., 2020. Conceptualising misinformation in the context of asylum seekers, Information Processing & Management, 2020.

Wu, L., Morstatter, F., Carley, K.M., Liu, H., 2019. Misinformation in social media: Definition, manipulation, and detection. ACM SIGKDD Explor. Newsl., 21,pp. 80–90

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