Protect

The Right to International Protection

Deliverable D7.6 Journal article draft on the content and frame analysis of social media discourses on migration and refugees

Dissemination level:
Public

Lead Beneficiary: Lund University

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 870761.

PROTECT

The Right to International Protection. A Pendulum between Globalization and Nativization?

PROTECT DELIVERABLE No. D7.6
Published by the PROTECT Consortium.

PROTECT CONSORTIUM
The PROTECT Consortium publishes original research on international refugee protection.
The Consortium is composed of:

University of Bergen, University of Catania, Ghent University, Giessen University, Ljubljana University, Lund University, Open University (London), Queen Mary University (London), University of Surrey, University of Stuttgart, Ryerson University (Toronto), University of Witwatersrand (Johannesburg),

To cite this publication:

> Anamaria Dutceac Segesten and Mike Farjam (2022). A content and frame analysis of the social media discourse on refugees and migrants, in the context of the United Nations and the European Union, 2015-2019. PROTECT Deliverable no. D7.6. Bergen: PROTECT Consortium.

A content and frame analysis of the social media discourse on refugees and migrants, in the context of the United Nations and the European Union, 2015-2019

Anamaria Dutceac Segesten and Mike Farjam
Lund University, European Studies, Centre for Languages and Literature

Abstract

In this paper, we perform a cross-platform, cross-language comparison of social media content from Twitter, Reddit and YouTube related to refugees and migrants in the context of two major players in international protection: the United Nations and the European Union. The data covers a period of five years (2015-2019) and is analyzed using a combination of computational methods (word2vec) and computer-assisted content analysis (keyword-in-context) for nine languages. The results show that the most dominant categories of the global cleavage system are the globalists and the nation-statists. The most dominant topics present across the language clusters are *Limiting and anti-migration*, *Global politics, Economy & job-market*, and *Humanitarian and compassionate attitudes* towards migration. While these trends apply to the entire dataset, there are notable differences among language clusters and among the social media platforms, which point out the need to take into account the specific contexts when analyzing the digital discourse on migration.

1. Introduction

The events known as the 'refugee crisis' in the Mediterranean (2015-16), as well as the political agenda of the US president Donald Trump (2016-2020) with its "build a wall" motto, brought to the forefront the issue of the protection of people forced to leave their homes. The United Nations and its agencies are the most important global players in ensuring the protection of these people on the move. In addition, the tiers of refugee and migrant protection extend downwards to the regional (in our case, European) level as well as to the level of the nation state. Beyond the legal provisions in place at these levels, however, the actual implementation of protective measures depends largely on the level of support given by citizens to governmental policies on the assistance of refugees and migrants. In the absence of popular support, the legal texts may retain only symbolic power; their implementation depends on the political will of the decision-makers. In turn, holders of public office and civil servants, as well as civil society actors need the support coming from regular citizens in order to act in accordance with (or even beyond) the letter of the law. This is one of the reasons why charting the public discourse on social media is important: it gives us an idea about the issues that drive the interest of social media users and the lenses through which these topics are seen and interpreted. This paper then asks:

> *Which are the most prevalent frames and topics present in the content of social media posts that brought up the topic of refugees, asylum seekers or migrants over the course of four years (2015-19)?*

To answer this research question, we embark on a large-N comparative study of nine different language clusters, responding to the challenge identified by Lecheler, Matthes and Boomgaarden (2019) and as one of the most "pressing research lacunae for the international scholarship on media and migration": the lack of comparative research. We thereby map the supra-national landscape of discursive pressure that could facilitate or hinder the implementation of the initiatives such as the UN's Global Compact on Refugees and Migration or the EU's Common European Asylum System. Going beyond the language diversity, we introduce also a multiplatform comparison, across three social media with publicly accessible content: Twitter, Reddit and YouTube. Following Bode and Vraga (2018), we believe that cross-platform comparisons capture the broader context of a discourse, and allow for better generalization ability. Moreover, testing theories in a comparative manner increases the scientific value of the findings and allows for possible reconceptualizations.

2. Related literature

2.1. Research on social media and migration

There is, no doubt, a proliferation of research examining discourses on migration in media, of both the traditional and social kind. Part of the reason for the surge in research on this topic lies with the real-world context events such as the refugee crisis in Europe and the increased flux of Latin American migrants towards the United States. In equal measure, the prominence of migration in the literature is also due to its increased politicized nature in the US (Waldinger, 2018) as well as across Europe (Strömbäck et al, 2021; Krzyżanowski, Triandafyllidou, and Wodak, 2018; Van der Brug et al, 2015).

The literature on media and migration can be divided in roughly two strands: on the one hand, a more descriptive one trying to understand the issues, frames, and actors involved in the portrayal of migration as a general phenomenon or of specific crises and, on the other, one focused on media effects, exploring if and how media discourse affects citizens' attitude towards migration or parties' agenda formulation. The present paper situates itself in the first research stream, building on work done on traditional media by Lawlor and Tolley (2017) or Greussing and Boomgaarden (2017) and on social media by Lee and Nerghes (2018) or Heidenreich et al. (2020).

Most related to the research presented in this paper is the strand of literature concerning itself with frame analysis. Frames are understood as lenses through which an activity, event or broader phenomenon are interpreted, by proposing a particular problem definition or causal interpretation (Entman 1993). There appear to be a series of common frames or ways of interpreting migration, and in particular the refugee crisis, across media types. Greussing and Boomgaarden find that in six Austrian mainstream newspapers the most frequent frames in which the refugee crisis was portrayed were the *settlement/ redistribution* of incoming asylum seekers, *criminality* risk posed by them, *economy* (or the economic burden posed by the newly arrived), and *humanitarianism* (desire to assist, especially from the part of civil society). Also present were frames of what the authors called *background/ victimization* (the difficulties encountered by the refugees on their way to Europe), *securitization* (national security and border control), and *labor market integration*.

Going beyond one national case study, Heidenreich et al (2019) gather print and online articles from several news outlets in five European countries (2015-2016) and perform an automated frame analysis. Their findings reveal many similarities with the Austrian case, but also some differences, in particular in the rank ordering by frequency of each frame. Here, the most common lens through which migration was seen across the media outlets was the *economy*, followed by *welfare*, *accommodation* of refugees, and international *humanitarian* aid. *Refugee camps, borders* as well as *national* and *EU politics* were also present.

To our knowledge, no research has performed a similar type of analysis performed on social media data to ours. Several studies included social media posts in their analyses. However, they all focus either on a specific type of actor within social media or on one specific platform. Among those studies that include a social media component, Ademmer and Stöhr (2019) look at comments left on the Facebook pages of local and regional newspapers in Germany. They identify 100 topics, which they group in three cleavages: GAL/TAN, left-right, and dealignment (cf. Hooghe, Marks and Wilson, 2002). They find that the first cleavage, characterized by an emphasis on culture and identity is clearly dominating the comments studied. The more traditional left-right cleavage between progressive and conservative politics is much less frequent, whereas dealignment, or the lack of a political leaning, is the least prevalent. While providing relevant insights, this study focuses on the microlevel of migration politics and is thus rather limited in scope compared to our undertaking.

In another article related to migration on social media, Heidenreich et al (2020) analyze visibility and sentiment towards migration in the Facebook accounts of political actors across six European countries, between mid-2015 and end of 2017. However, the article does not cover citizen discourse and includes only one social media platform. Conrad (2021) performs a frame analysis on a data set that combines traditional and social media from three countries but focuses only on the Global Compacts and uses a small-N approach. Moreover, he is only interested in the frames employed by populist and right-wing actors.

Thus, our paper is in a position to provide new and important knowledge about the digital public discourse on migration as it 1) includes social media posts across the board (institutional and personal accounts); 2) includes three social media platforms; 3) covers a longer period that allows for the after effects of the refugee crisis to be observed and 4) includes nine different language clusters.

2.2. The global cleavage system

One of the issues with the analyses of media discourses on migration is that they are, for the most part, inductive rather than theory-driven. In our paper, we attempt to complement the content analysis that used inductive frames with a categorization of content in frames derived from cleavage theory. In a general sense, social or cultural cleavages are fault lines dividing societies in distinct categories, which can become politicized (Kriesi et al, 1995). Once political, cleavages are important because they influence party formation and voter preferences. Initially developed by Stein Rokkan (1967, 1970) at the national level, cleavage theory has been adapted to the global level and to the context of refugee protection by Sicakkan (2012, 2016).

In short, Sicakkan contends that there are four different cleavages that separate actors on the international refugee protection scene. Like in the case of national cleavages, the global cleavage system helps identify the various political groups (and their interests) that compete over setting the agenda for international refugee and migrant protection. Each of the groups identified in this manner has its own view over the extent to which these vulnerable groups should enjoy protection, who should be the main provider of the protection and in which way the resulting policies should be implemented or administered. A summary of these positions can be found in the table below.

Table 1: International protection in the global cleavage system

		GROUPS IN THE GLOBAL POLITICAL CLEAVAGE SYSTEM			
		Nativists	**Nation-statists**	**Regionalists**	**Globalists**
NORMS	**What is most worth to protect?**	*Ethnic Belonging* Dispersed nations' rights in terms of ethnic/diasporic identification, language, and territorial belonging	*National Belonging* Citizens' rights and duties in terms of civic culture; states' interests; and the international order	*Regional Belonging* Members' rights and interests in terms of dignity, lives, liberties, and estates in a civil society	*Humanity* Individuals' rights and interests in terms of dignity, lives, liberties, and estates in a civil society
	Is it a duty or charity to protect the refugees?	*No duty to protect others than co-ethnics* Constitutional asylum	*Protection is given as charity, not a duty* Convention, temporary, and constitutional asylum as legal grounds	*Protection is given as an entitlement, not a duty* Convention asylum and subsidiary protection as legal grounds	*Protection is duty and entitlement* Convention asylum as legal grounds
	Minimum Criteria for protection	*Endangering* Endangering by persecution, oppression, assimilation, or non-protection by a state or non-state actors supported by a state	*Persecution* Persecution by a state; or persecution by the majority or non-state actors combined with effective state collaboration	*Persecution* Persecution by a state; or persecution by the majority or non-state actors combined with effective state collaboration	*Non-protection* Non-protection, discrimination, or persecution by a state; persecution by non-state actors combined with states' negligence
	Who is responsible for protection?	*Co-ethnic states* Individual states with historical relations with their diasporas and the states where these diasporic groups reside are responsible.	*Intergovernmental* Individual states primarily, and the international community secondarily have the responsibility to protect.	*Supranational* The regional authorities primarily, and member states, are responsible for protection.	*International* The international community / the international society has the responsibility to protect.
DISCOURSES	**Policies cited in discourses**	*Ethnicization of the refugee problem* Territory and autonomy claims for diasporic groups; population exchanges; unilateral actions such as condemnation and intervention, and bilateral agreements.	*Nationalization of the refugee problem* Focus on root causes; preventive diplomacy, economic relief, forced / voluntary repatriation, military aid, and intervention.	*Regionalization of the refugee problem* Focus on root causes; extensions of sovereignty to stateless communities; regional devolutions; temporary collective protection; creating regional safe zones; externalization; repatriation	*Universalization of the refugee problem* Focus on human rights; individual protection; cooperation across borders; preventive diplomacy; economic aid and relief; voluntary repatriation.
	Where to protect?	In the country of escape, or of asylum	In the country of escape or of resettlement	In or near the country of escape or of origin	In the country of asylum
GOVERNANCE MODES	**How to organize protection?**	*Uni-lateral or bilateral state actions*	*Voluntary unilateral, bilateral or multilateral state cooperation*	*Mandatory state cooperation*	*Global multilateral binding cooperation*
	Governance modes and actors	*State-centric centralist governance* - States - Other states in bi-lateral agreement - Nativist non-state organizations - Ethnic minority organizations in refugee sending countries	*State-centric corporatist governance* - States - Other states in bi- and multi-lateral agreement - National non-state organizations funded by the state - Local authorities	*Region-centric pluralist governance* - Regional organizations (eg. EU) - States - International organizations - Transnational non-state organizations - National non-state organizations - Local authorities	*Global corporate-pluralist governance* - International organizations - Regional organizations - States - Transnational non-state organizations - National non-state organizations - Local authorities

Source: Sicakkan (2021)

The four types outlined above are transformed into discursive frames, and are rejoined also by a fifth category, the market-oriented type. The market-oriented frame sees migrants through an economic frame: either as costs to the welfare state, to the society, or to the state budget, or as benefits to the labor market, to innovation and productivity.

Categorizing the digital public discourse on migration along the five cleavages may reveal the challenges to international protection and may be particularly useful in a comparative setup such as ours. Because this categorization covers national, regional and global levels, it allows us to capture the possible national differences across the nine languages studied.

3. Data

We designed a query to download all relevant social media content from Reddit, Twitter, and YouTube between 2015 and 2019 through the services provided by Brandwatch Consumer Research. The query was designed around two central constraints where posts had to include I) at least one of our migration related keywords needed to be in the post (i.e., immigrant, immigration, migrant, migration, refugee, and asylum) and II) references to the EU or UN needed to appear within a 20-word distance to the migration keyword. URLs or mentions of social media accounts associated with the UN or EU or any of their associated organizations (e.g., Frontex or UNHCR) were also considered references to the UN or EU.[1]

Besides English, the query was performed in 12 other major European languages of which nine will be analyzed in this paper because they lead to a number of posts large enough for a valid quantitative analysis. Those languages were English, German, French, Italian, Spanish, Polish, Dutch, Swedish, and Danish. Table 1 provides a per-language overview of the social media posts analyzed by us in the nine languages. We want to emphasize that, because social media conversations are not bound in any way to a national territory, we cannot discuss the findings in terms of countries but only in terms of language clusters. Whereas in traditional media the sources of information are known and are strongly connected to a national system, digital content made and distributed on social media is available for a global audience, with only language being the restriction to access. In our sample, some languages are truly global, used by anyone no matter their mother tongue (English) or very widespread (Spanish, spoken as a vernacular in both Europe and Latin America; French, spoken in Europe and in some parts of Africa). In addition, German is a cross-border language in Europe, as it is the official language in Germany, Austria and Luxembourg. Thus, we want to emphasize that we are capturing in our data a discourse that in some cases is very clearly nation-bound (e.g., Danish, Polish) and in some other cases a global discourse (e.g., English, Spanish).

Because of the relatively large number of posts available in English, French, German, Italian, and Spanish, we chose to analyze posts in these languages separately for, on the one side, Twitter (the most common type of social media in our data) and, on the other side, Reddit and YouTube. This distinction between platforms also makes theoretical sense since Twitter is characterized by relatively short posts, while Reddit and YouTube are platforms in which post length and user interaction is structured relatively similar. Note that all rows in Table 2 include at least 1 million words (after pre-processing) and 10,000 de-duplicated posts. We consider this the absolute minimum data needed for a valid analysis of a separate discourse. The groups/rows in Table 1 will be henceforth referred to as ‘analytical groups’. In total, our analysis includes more than 31,000,000 posts.

[1] The Appendix provides a full list of terms used in the query.

Table 2: Social media posts per analytical group

			Pre-processed ...		
Language	Social medium	N posts	… words per post	… words	Unique users
English	Twitter	21,717,764	17.9	388,747,976	3,001,746
English	Reddit+YouTube	697,490	76.2	53,148,738	235,225
Spanish	Twitter	2,922,920	15.5	45,305,260	685,067
Spanish	Reddit+YouTube	36,947	67.7	2,501,312	13,205
German	Twitter	1,931,743	15.5	29,942,017	163,666
German	Reddit+YouTube	115,245	80.7	9,300,272	30,098
French	Twitter	1,350,283	19.8	26,735,603	205,988
French	Reddit+YouTube	21,302	97.8	2,083,336	8,820
Italian	Twitter	813,261	19.4	15,777,263	116,069
Italian	Reddit+YouTube	10,241	105.2	1,077,353	3,947
Dutch	all	808,555	17.7	14,311,424	69,098
Swedish	all	295,817	18.3	5,413,451	37,853
Danish	all	196,678	16.2	3,186,184	24,160
Polish	all	130,421	20.7	2,699,715	24,818
Total		31,048,667			

4. Methods

In general, our analytical approach is data- and algorithm-driven in order to minimize the amount of human intervention needed for the analysis. This approach was a necessity to analyze the large quantity of data and also in order to deal with the variety of languages analyzed[2]. All posts were pre-processed, which included 1) lower-casing; 2) stemming of words to their simpler form (e.g., simplifying nouns to their nominative singular and verbs with different inflection to a common form); 3) deletion of stopwords (e.g., connectors and pronouns), emojis and syntactic and special symbols; 4) deletion of duplicates by the same user (e.g., from users posting the same comment under different YouTube videos).[3]

As the first step in our analysis, we generated a list of words per analytical group that were ‘descriptors’, that is, words that are representative of the analytical group’s discourse on migration. In order to appear in the list of descriptors, words had to fulfill two criteria. Only if both criteria were fulfilled, words were considered descriptors. First, words needed to be conceptually related to migration. The conceptual relatedness of words was estimated with the help of word2vec models (Goldberg & Levy, 2014) mapping all words used within posts of the analytical group to a common vector space. In such models, the so called ‘cosine’ measure is used as a metric of conceptual distance between word pairs. The first criterion for a descriptor was thus that it needed to be among the 500 closest words in terms of mean cosine distance to the migration related keywords. The second criterion for descriptors was that they needed to appear in proximity to migration-related keywords. To generate a list of words fitting this criterion, we performed a keyword-in-context analysis (Chelvachandran & Jahankhani, 2019) again considering all words in all posts of the analytical group that were within a word-distance of 3 to the migration related keywords.[4] Again, we compiled a top-500 list of words most frequently identified by the keyword-in-context analysis. The final list of descriptors was the intersection between both top-500 lists.

[2] In the Appendix we present a reflection on our failed theory-driven and dictionary-based approach.

[3] The Appendix provides more detail on the pre-processing pipeline and all R-libraries and functions used for each language.

[4] This is the word-distance window *after* stopwords were deleted and this parameter is therefore slightly smaller than the often used 5-word window in such analyses. The Appendix provides additional detail on functions and parameters used in the word2vec and keyword-in-context models.

The resulting lists of descriptors that fulfilled both criteria contained 250-400 words in each of the analytical groups, which was at the upper limit of what could be handled in the following steps of the analysis. Note that our approach to generating descriptors is language agnostic and has no language specific parameters. Only the migration related keywords needed to be translated into the analytical group's language. These words are relatively similar in all languages under investigation and also part of the original query described earlier. We thereby also limited chances of language specific biases in our analysis.

To make descriptors more interpretable in terms of the context in which they were used, we also determined the three most common bigrams (i.e., word pairs) in which a descriptor was used (e.g., for "border" "border control"). We settled with this number of bigrams per descriptor because this in total led to ~1,000 bigrams per analytical group, which was the upper boundary of what could be handled in the following step of the analysis. The list of descriptors and their associated bigrams was then used for a qualitative coding of the discourse within each analytical group. Descriptors were qualitatively coded by (near) native language speakers, familiar with the overall discourse on migration within the specific European countries. Two types of codes were generated: An inductive code type, grouping descriptors into discourse themes (Saldana, 2009) and a deductive code type, matching descriptors to the categories from the global cleavage theory (Sicakkan, 2020). In the following section, we compare distributions of codes between social media types and language groups, identifying particularities and commonalities with respect to actor types engaging in and topics covered within the discourse on migration on social media. As common in qualitative coding (Saldana, 2009), all coding was done by three independent coders and were compared at the end of the coding procedure to ensure inter-coder reliability. In case of conflicting codes, a joint assessment was made in terms of re-coding.

5. Results

We begin reporting the results with the theoretically-driven cleavage coding report. Figures 1 and 2 present the distribution of cleavage categories found in the social media posts from each language cluster. Each bar in the figures represents the percentage of bigrams derived from an analytical group's descriptor list that could be clearly linked with one of the cleavage categories. Examples of such bigrams are "border control" or "human rights" for the category nation-statist and globalist respectively. Figure 1 shows that across the five larger languages within our data and all social media platforms, globalists, nation-statists and regionalists were clearly the dominating categories occurring approximately equally often. Variation between languages exists primarily regarding the regionalists category that was (alongside globalists and nation-statist) most notably relevant within the German discourse. The other two categories (i.e., market-oriented and nativists) seem to only be of marginal relevance within the discourse.

Figure 2 shows the same kind of distribution for the smaller languages. The Polish, and to a lesser extend also the Danish and Dutch, languages clearly deviate from the overall pattern observed. Here, the discourse seems exclusively dominated by the categories of nation-statists and nativists.

A disadvantage of our theory-driven approach to code descriptors under specific cleavage categories was that only 20-25% of descriptors could be clearly associated with a specific cleavage category. This naturally leads to the question of what can be learned from the remaining part of the descriptors. This inductive approach is covered by the qualitative topic analysis, which was based on all descriptors and their associated bigrams.

Fig. 1. Distribution of cleavage categories for the five largest languages. For those languages, Twitter was analyzed separately from the other social media.

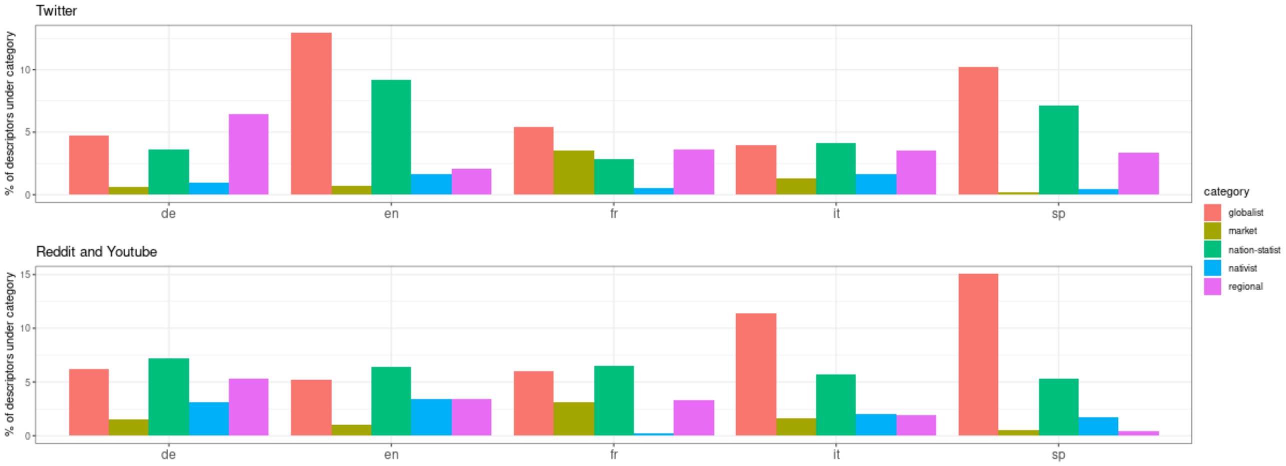
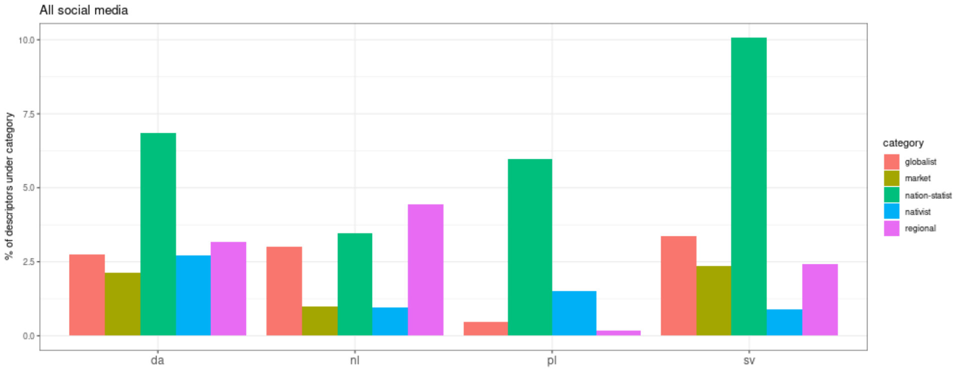


Fig. 2. Distribution of cleavage categories in the four smaller languages in which posts from all social media were analyzed jointly.



Figures 3 and 4 present the distribution of topics identified within each analytical group's discourse. These figures show topics that were common across all language groups. Language specific topics are discussed at the end of this section. We see that a relatively small number of topics are dominating the discourse across the analytical groups. Overall, the most prominent topics were (ordered by their prevalence) 'Limiting and anti-migration', 'Global politics', 'Economy & job-market', 'Humanitarian and compassionate attitudes', and 'EU politics'[5]. Other topics only played a minor role in the discourse.

[5] The Appendix provides an overview of terms associated with each topic.

Fig. 3. Distribution of the nine most common inductively derived themes for the five major languages. For those languages, Twitter was analyzed separately from the other social media.

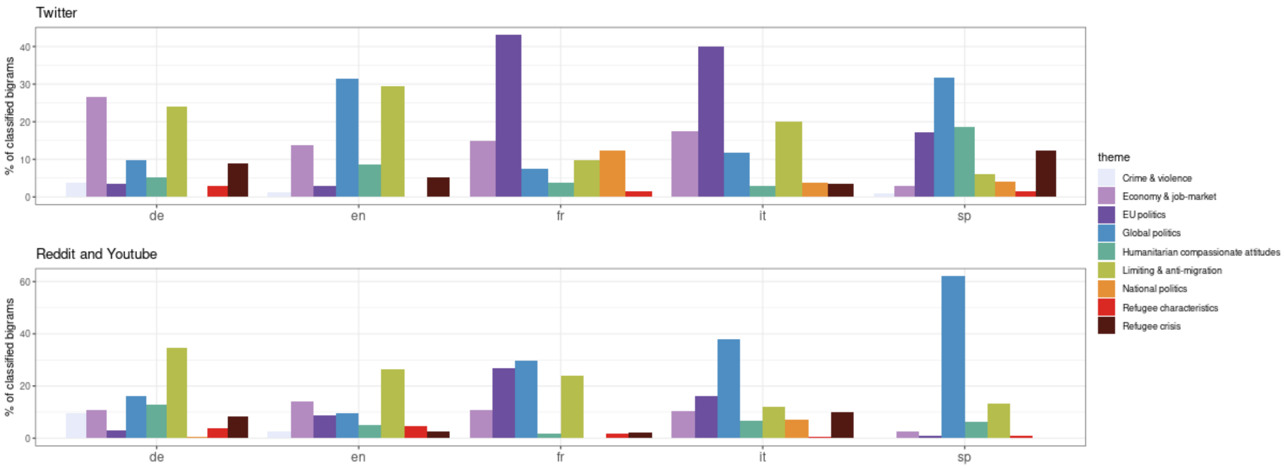
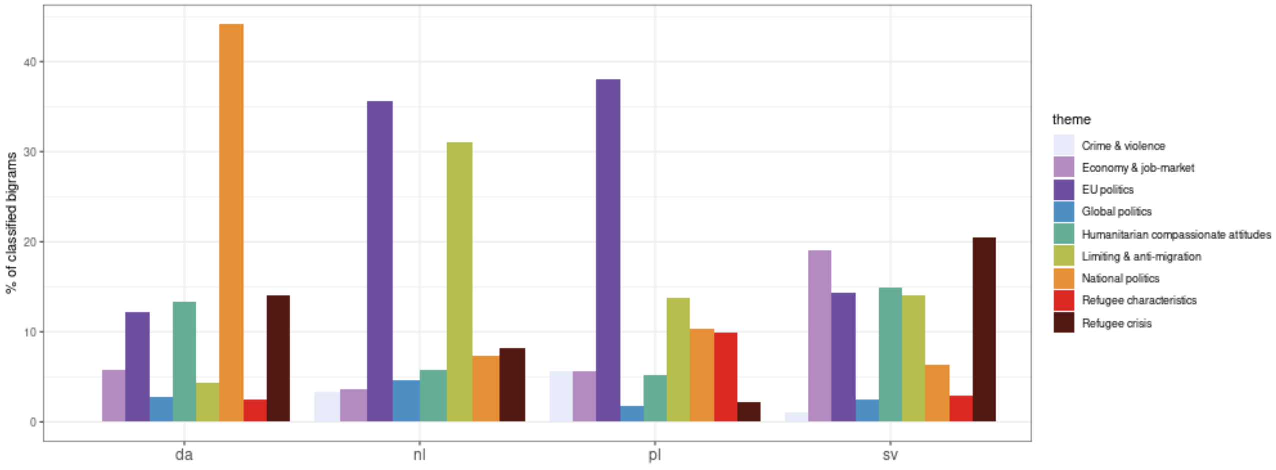


Fig. 4. Distribution of the nine most common inductively derived themes for the four smaller languages. For those languages, all social media were analyzed jointly.



Furthermore, we see that our analysis in terms of cleavage categories and topics also complement each other. The Polish (and to a lesser extent Dutch) discourse, for example, were dominated by nation-statists and, topic-wise, primarily discuss EU-politics, implying that the discourse on EU-politics was primarily anti-EU/EU-critical. Similarly, the Danish discourse which is primarily focused around the topic of 'National politics' is also dominated by the migration critical nation-statist. In other words, we see that the Polish, Dutch and Danish discourses are both critical towards migration, but that for Polish and Dutch the critical view manifested in the context of discussions of EU politics, while in the Danish discourse it resolved around discussions on national politics. In this sense, while the topics are telling us *what* is discussed, the combination between cleavage categories and topics provides insight into *how* and by *whom* a topic is discussed.

To see to what extent our theory-driven cleavage categories and the data-driven topic categories overlap, Table 3 presents a cross-tabulation between the two types of categories.

Each cell shows the number of descriptors found under the combination of categories. We clearly see that each cleavage category is focused around a few unique topics. An exception are the cleavage categories of nativists and nation-statists, that have considerable overlap in terms of topics, with both cleavage categories feeding from topics that often carry an anti-migration sentiment.

Table 3: Cross-tabulation of descriptor categorizations

	Globalist	Marketist	Nation-statist	Nativist	Regionalist
Crime & violence	3	0	25	26	0
Economy & job-market	5	130	4	1	0
EU politics	11	1	30	1	56
Global politics	83	0	5	2	1
Humanitarian compassionate attitudes	129	1	11	3	0
Limiting & anti-migration	8	3	96	68	1
National politics	7	1	36	4	3
Refugee characteristics	2	0	3	43	0
Refugee crisis	11	0	8	0	40

Interestingly, we found that the cleavage *market-oriented* was not common in the English, German, and Italian discourse but that ‘Economy & job-market’ is a prevalent topic. It seems thus that in those analytical groups, economic questions were often mentioned in combination with other questions. Indeed, when looking at posts in those languages with descriptors that fall under the ‘Limiting & anti-migration’ and ‘Humanitarian compassionate attitudes’ topics, we found that argumentation within those topics often included references to the costs of migration or the benefit of migration for the national labor force.

While Figures 3 and 4 primarily show the topics that were common *across* different languages, our coding process also revealed some language specific topics found in only one or few languages. In particular, we found that the topic of within-EU migration and from near other eastern-European countries (e.g., Ukraine and Belarus) was uniquely found in the Polish discourse. Other non-European migration could only be identified as a topic in the English and Spanish data, in particular on migration from Central American countries and Myanmar. This finding is not too surprising since Spanish and English are mainly spoken within non-European countries and Spanish and English posts showed a relatively high number of mentions of UN-related to EU-related keywords[6]. However, this also suggest that (with the exception of Poland) European countries discussed the issue of migration primarily in the context of the 2015 migration crisis and its aftermath. Other remarkable language specific themes were 1) the discussion of legal and technical questions in the context of asylum within the German discourse; and 2) discussions about unaccompanied minors within the Swedish and English discourse.

6. Discussion and conclusion

We use three axes in formulating the discussion of our findings, which can be summarized as follows. First, looking at the results concerning the *global cleavage system*, we see that, for the large languages, there are two main cleavages represented: the globalist and the nation-statist. Content-wise, the positions of these two agendas are not entirely aligned in terms of refugee and migrant protection, but they are not necessarily at odds with one another either. The absence of a prevalent nativist political group may signal that the migration issue is not pushing

[6] The Appendix Figure A1 presents the distribution of EU- and UN-related keywords across the 9 languages.

parties to adopt more radical positions. Those who want to uphold and implement the Geneva Convention and the Global Compacts for Refugees and Migration appear to be quite vocal on social media, whereas those emphasizing national/ethnic identities are less present. However, the picture is rather different for the smaller languages. Here the nation-statists are by far the largest group, followed by regionalists and globalists. This could be explained by the demographics of social media users in these countries, which may overrepresent those with a clear national focus in their attitudes toward migrant and refugee protection. It could also be the case that institutions, MEPs, and NGOs accounts tend to express themselves more frequently in languages with larger audiences. Even in the case of smaller languages, though, nativists are a minority. The extent to which this opinion distribution is a case of social media population bias or more representative for the speakers of these languages can only be explored in a comparison with survey results.

Second, the results about the main topics present in the social media content revealed that the most discussed subjects across all languages were *Limiting and anti-migration*, *Global politics*, *Economy & job-market*, *Humanitarian and compassionate attitudes*, and *EU politics*. This distribution of topics fits rather well together with the distribution of cleavage categories identified in Figures 1 and 2. As shown in Table 3, the topics Humanitarianism and Global politics have considerable conceptual overlap with the 'globalist' cleavage category. The topic *Economy & job-market* overlaps with the market-oriented cleavage category, and the topic *EU politics* is strongly related with the regionalist cleavage. However, some other subjects picked up by our inductive analysis span across cleavages, such as *Limiting and anti-migration*; in this case, nativists, nation-statists agendas could include positions in favor of limiting refugee and migrant protection, with the difference being more as to which authority is invoked (the state, the market, the national/ethic group). The *Refugee Crisis,* and *National politics* are subjects not entirely related to the cleavages, but that give us an additional insight into the nature of the online discussion surrounding the issue of migration.

Comparing our findings with those of previous research on the topic of migrant discourse in the media, we see largely a continuity, with a refinement perhaps, between the earlier studies and our own. Greussing and Boomgaarden (2017), Heidenreich et al (2019) and Ademmer and Stöhr (2019) all identify categories that are similar to our *Economy, Refugee Crisis, Humanitarian and compassionate attitudes* and *Criminality*. Probably the most significant refinement is evident in our results' sensitivity to national contexts.

While the general trends mentioned above are valid, there are some significant national divergences. English and Spanish, the most globally used languages in our set, are also clusters of high focus on global politics. Italian, French and Polish discourses include many references to EU politics, whereas Danish is overwhelmingly displaying references to National politics. Most of the language clusters have the topics of *Limits to migration* or clearly *anti-migration* attitudes as the second most frequent topic, but even here there are exceptions: for Spanish, French and Danish this category ranks third or lower. To explain these differences, further investigations in each national context need to be performed. However, we can already draw the conclusion that national discourses vary substantively and that generalizations should be made with caution outsides the most dominant themes.

Third and final, the *comparison across social media platforms* for the five largest languages in the dataset reveals some uneven trends. For all of the studied clusters, we record differences between Twitter on the one hand and Reddit and YouTube on the other. In Germany, for example, the *Economy* topic is much more popular on Twitter compared to the other two platforms, whereas the levels of the anti-migration discourse stay similar. In English, Global politics is the most popular topic on Twitter but that topic is hardly discussed on Reddit and YouTube. EU politics is the most popular issue for French and Italian and quite important also for Spanish – on Twitter. On Reddit and YouTube, the EU practically disappears, being replaced as the most dominant topic by Global politics.

To explain these differences, one can look at the platform structures and/or at their different demographics. Twitter is an elite medium, where many institutions and organizations have accounts. One explanation for the predominance of, for example EU politics over other topics is that its presence may be driven by the content generated by EU institutions' public communication. Reddit and YouTube, in contrast, are not typical places for institutional representation but more for regular individuals engaging in (usually, anonymous) commentary.

On a similar line, the socio-demographic composition of the platforms' user base may also explain some of the observed differences. Because it is largely non-anonymous, we know more about Twitter's demographics: Twitter is the digital home of political and media elites; its individual users tend to belong to an educated and more affluent segment of the population. Reddit and YouTube tend to be populated by users keen on anonymity, so accurate information about who uses the platform needs to be obtained via separate surveys, leading to less available data. More information about the distinct demographics of each of the platforms in each of the language contexts would help to explain the variations in discourse topics.

Overall, our study has provided descriptive answers to the question of the content and frames of social media posts on refugees and migrants. It has done so in a cross-language and cross-platform analysis that spans a period of five years, including the aftermath of the refugee crisis in Europe. The insights revealed here need to be taken further in future studies that will test alternative hypotheses to explain our language and platform differences as well as the rank ordering of cleavage frequencies.

List of references

Ademmer, Esther, and Tobias Stöhr. *The making of a new cleavage? Evidence from social media debates about migration*. No. 2140. Kiel Working Paper, 2019.

Bode, Leticia, and Emily K. Vraga. "Studying politics across media." *Political Communication* 35, no. 1 (2018): 1-7.

Chelvachandran, N., & Jahankhani, H. (2019). A Study on Keyword Analytics as a Precursor to Machine Learning to Evaluate Radicalisation on Social Media. In *2019 IEEE 12th International Conference on Global Security, Safety and Sustainability (ICGS3)* (pp. 1-7). IEEE.

Conrad, Maximilian. "Post-truth politics, digital media, and the politicization of the Global Compact for migration." *Politics and Governance* 9, no. 3 (2021): 301-311.

Entman, Robert M. "Framing: Towards clarification of a fractured paradigm." *McQuail's reader in mass communication theory* (1993): 390-397.

Goldberg, Y., & Levy, O. (2014). word2vec Explained: deriving Mikolov et al.'s negative-sampling word-embedding method. *arXiv preprint arXiv:1402.3722.*

Greussing, Esther, and Hajo G. Boomgaarden. "Shifting the refugee narrative? An automated frame analysis of Europe's 2015 refugee crisis." *Journal of Ethnic and Migration Studies* 43, no. 11 (2017): 1749-1774.

Heidenreich, Tobias, Jakob-Moritz Eberl, Fabienne Lind, and Hajo Boomgaarden. "Political migration discourses on social media: a comparative perspective on visibility and sentiment across political Facebook accounts in Europe." *Journal of Ethnic and Migration Studies* 46, no. 7 (2020): 1261-1280.

Hooghe, Liesbet, Gary Marks, and Carole J. Wilson. "Does left/right structure party positions on European integration?." *Comparative Political Studies* 35, no. 8 (2002): 965-989.

Kriesi, Hanspeter, Ruud Koopmans, Jan Willem Duyvendak, and Marco G. Giugni. *New social movements in Western Europe: A comparative analysis*. Routledge, 2015.

Krzyżanowski, Michał, Anna Triandafyllidou, and Ruth Wodak. "The mediatization and the politicization of the "refugee crisis" in Europe." *Journal of Immigrant & Refugee Studies* 16, no. 1-2 (2018): 1-14.

Lawlor, Andrea, and Erin Tolley. "Deciding who's legitimate: News media framing of immigrants and refugees.*" International Journal of Communication* 11 (2017): 25.

Lecheler, Sophie, Jörg Matthes, and Hajo Boomgaarden. "Setting the agenda for research on media and migration: State-of-the-art and directions for future research." *Mass Communication and Society* 22, no. 6 (2019): 691-707.

Lee, Ju-Sung, and Adina Nerghes. "Refugee or migrant crisis? Labels, perceived agency, and sentiment polarity in online discussions." *Social Media+ Society* 4, no. 3 (2018): 2056305118785638.

Lind, F., Eberl, J. M., Heidenreich, T., & Boomgaarden, H. G. "When the Journey Is as Important as the Goal: A Roadmap to Multilingual Dictionary Construction." *International Journal of Communication* 13 (2019): 21.

Lipset, Seymour Martin, and Stein Rokkan. *Cleavage structures, party systems, and voter alignments: an introduction*. Free Press, 1967.

Rokkan, Stein. "Nation Building, Cleavage Formation and the Structuring of Mass Politics", *In Citizens, Elections, Parties*, edited by S. Rokkan, Oslo: Universitetsforlaget, 1970.

Saldana, Johnny. *The coding manual for qualitative researchers.* London: SAGE Publications (2009).

Sicakkan, Hakan G. "Trans-Europeanising public spaces in Europe." *Javnost -The Public* 19, no. 1 (2012): 5-26.

Sicakkan, Hakan G. "European state building, top-down elite alliances and the national media." In *Integration, Diversity and the Making of a European Public Sphere*. Edward Elgar Publishing, 2016.

Sicakkan, Hakan G. Conceptualizing the Right to International Protection: A Cleavage Theory Approach. PROTECT Deliverable no. D1.1. Bergen: PROTECT Consortium, 2021.

Strömbäck, Jesper, Christine E. Meltzer, Jakob-Moritz Eberl, Christian Schemer, and Hajo G. Boomgaarden, eds. *Media and Public Attitudes Toward Migration in Europe: A Comparative Approach*. Routledge, 2021.

Van der Brug, Wouter, Gianni D'Amato, Joost Berkhout, and Didier Ruedin. *A framework for studying the politicisation of immigration*. London: Routledge, 2015.

Waldinger, Roger. "Immigration and the election of Donald Trump: Why the sociology of migration left us unprepared… and why we should not have been surprised." *Ethnic and Racial Studies* 41, no. 8 (2018): 1411-1426.

Appendix

Query

These were the keywords used for the English query. Two constraints needed to be met in order for a post to be in our data: 1) The post had to include at least one of the migration keywords; 2) it had to include at least one of the EU or UN keywords within a 20-word distance from the migration keyword, or include EU or UN hashtags/urls. In the list below, keywords are written as regular expressions and only with lowercase letters to account for variations in terms of spelling, grammar, etc.

- Migration related keywords[7]: refugee*, migra*, immigra*, asylum*
- EU keywords: eu\'?s?\'?, eu(ropean)? ?union\'?s?\'?, eu(ropean)? ?comm?iss?ion\'?s?\'?, eu(ropean)? ?parli?ament\'?s?\'?, eu(ropean)? ?cou?ncil\'?s?\'?, council ?of ?the ?eu(ropean)? ?(union)?\'?s?\'?, council ?of ?ministers?\'?, eu(ropean)? ?court ?(of justice)?(ofjustice)?\'?s?, ecj\'?s?,frontex\'?s?, easo\'?s?, europol\'?s?, eu-lisa\'?s?, echo, eu_echo, eurightsAgency, eulisa_agency, eu_commission, europarl_en, eucourtpress, eurmigrforum, emnmigration
- EU urls: europa.eu, ec.europa.eu, frontex.europa.eu, easo.europa.eu, fra.europa.eu/en, europol.europa.eu, eulisa.europa.eu, euagencies.eu, consilium.europa.eu/en/european-council/
- EU hashtags: #TheEU*, #EU, #EuropeanUnion*, #TheEuropeanUnion*, #EuropeanCommission*, #TheEuropeanCommission*, #EuropeanComission*, #TheEuropeanComission*, #EuropeanCommision*, #TheEuropeanCommision*, #EUCommission*, #FortressEurope*, #Frontex*, #EASO*, #EURightsAgency*, #Europol*, #EULISA_agency*, #Europarl*, #EuropeanParliament*, #TheEuropeanParliament*, #EUParliament*, #EuropeanCourt*, #EuropeanCourtofJustice*, #EUCourtofJustice*, #EUCouncil*, #EuropeanCouncil*, #TheEuropeanCouncil*, #EurMigrForum*, #EMNMigration*
- UN keywords: (the)?united ?nations\'?, (the)?u\.?n\.\'?s?, (the)?global ?compact\'?s?, u ?n ?h ?c ?r\'?s, un(itednations)foundation\'?s, journal_un_onu, un_photo, unfpa, unpol, unitednationsjo, unnniversity, undp, un_cted, unvolunteers, unescap, unwebtv, unandageing, unindia, unoosa, usambun, un_bih, germanyun, unpeacekeeping, unfccc, unstats, usun, uncdf, un_hrc, ungei, unesco_mepp, unglobalpulse, unhabitat, unops, unctad, iom_usa, undpkaz, unphilippines, un_ukraine, wfp_mena, unenvoysyria, ochaafg, uniccanberra, undpthailand, wfp_europe, unicmanila, un_ovra, unhcrcanberra
- UN urls: unfoundation.org, un.org, unric.org

Packages, libraries, and parameters used

All analysis in this paper was performed with R 4.1.2. The packages *word2vec* and *quanteda* were used for the word2vec and key-word-in-context (kwic) analysis. For the word2vec models, 25-dimensional word vectors were used and a minimum word count of 5 was required for words to be considered in the model. As a robustness check, we ran separate analyses with +/- 50% of the word2vec parameter values and found that results were very similar with different parametrization (i.e., cosine distances of word pairs to the migration related keywords were correlating with at least r = 0.9). For the kwic analysis, a word window of 3 around the

[7] The * symbol is a placeholder for any continuation of a word.

migration related keywords was used (after deletion of stopwords) and to identify the most common bigrams around the migration related keywords that word-window was increased by 1.

Because we had to analyze different languages, a separate pre-processing pipelines was needed for each language. Stopword lists for English, German, Spanish, French Italian, Swedish, and Dutch were taken from the *snowball* package. Since that package did not provide reliable stopword lists for Danish and Polish, stopwords for those languages were taken from the *stopwords-iso* list in R. For all languages the text_tokens function from the *corpus* package were used for stemming of words, except for Polish were such a stemmer was not available and words were thus processed in their raw form.

Failed dictionary approach

Initially, we planned to analyze all posts using a dictionary method. The intent was to have a theory-derived list of words describing each of the 5 cleavage categories and then get a probability distribution per post corresponding to the percentage of words in the post matching the categories in the dictionary.

For this reason, and based on the Global Cleavage System (Sicakkan 2012, 2016), a dictionary was designed that contained per cleavage category the same number of words that best described the actors, content, and topics that we expected to be central in social media posts associated with the respective cleavage categories. Summarized, the theory-driven dictionary-based approach led to a number of problems that we could not solve and, hence, we went for the data-driven identification of descriptors as outlined in the main text. Hoping that others can learn from the problems we faced, we describe some of them here:

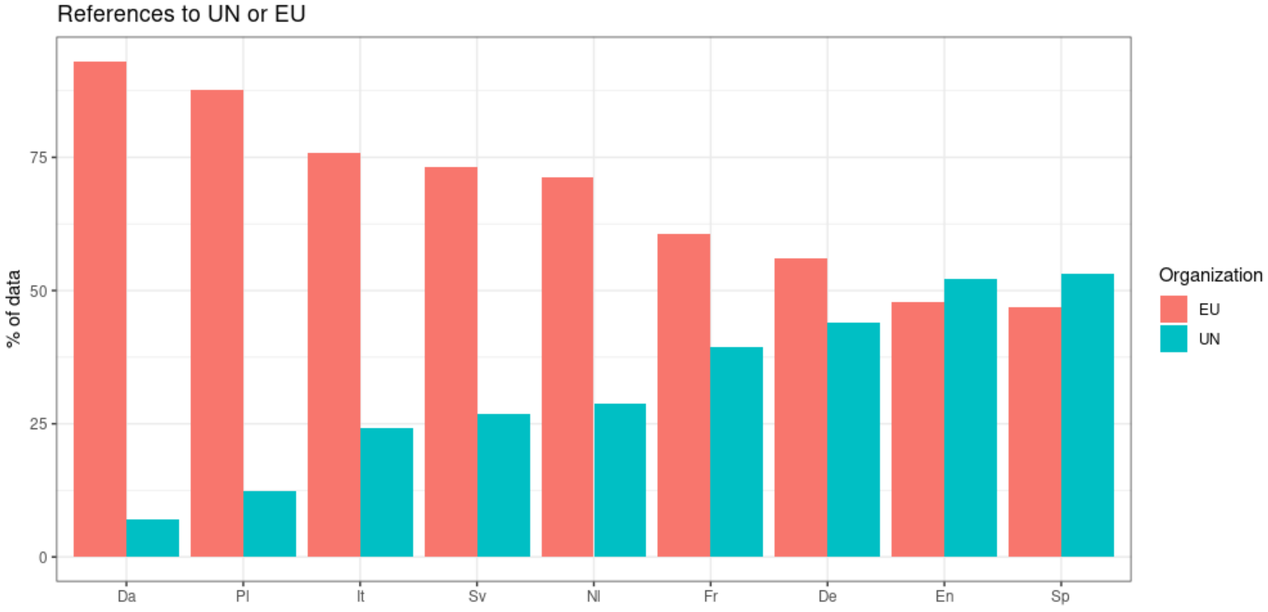
1. A fundamental problem was that language on social media is far less structured and formal than that found in traditional media. We also found that many social media posts heavily rely on quotes from other posts/media, making up a large part of the overall post. Often, these are then ridiculed and discussed with a sarcastic/ironic tone that simple dictionaries of the kind we intended to use cannot detect.
2. Since language was less formal on social media, theoretically meaningful distinctions between terms like “asylum seeker”, “refugee”, and “migrants” played hardly any role in the real discourse on social media. We, therefore, decided to instead use a bottom-up approach, starting from the language that is really used on social media and use our theoretical framework instead as a conceptual lens through which the used language is interpreted.
3. Social media posts are relatively short units of texts and as such dictionaries, even when containing hundreds of words, produce relatively few hits per posts. Any post classification based on a dictionary method is therefore extremely sensitive to the in- or exclusion of individual words in the dictionary.
4. Individual words were hardly able to capture the more nuanced distinctions made in our theoretical framework. When trying to increase the specificity of terms by adding constraints (e.g., by specifying the context in which they needed to appear) this further reduced the number of matches between expressions in the dictionary and the posts (as discussed in point 3). Increasing the sophistication of our dictionary therefore made the classifications even more volatile.
5. Previous research has found that dictionary-based methods on multilingual data work best when the data is translated first and the dictionary in one language then is applied to all translated data (see Lind et al. (2019) for a similar dictionary-based approach applied to text from legacy media on the topic of migration). Due to the large amount of data, we were dealing with, posts would have to be machine translated. First, this is,

as of 2021, still a very costly process (in our case more than $10,000). Second, because we are dealing with social media posts where informal language is common and because we included small languages for which algorithms are often not optimized, we were worried about differences of algorithm performance across languages, potentially leading to a bias in our analysis.

Actor distribution

The query on which our data is based specified that besides words related to migration, a post needed to mention either a reference to an UN- or EU-related organization. The figure below provides an overview of how common mentions of these two types of organization where within the different languages analyzed in the main text. As discussed in the main text, mentions of UN-organizations were most common within Spanish and English.

Fig. A1. Distribution of EU- and UN-related keywords in posts across the analyzed languages.



Topics

Below we provide an overview of the nine most commonly identified topics across languages. Per topic, we show a (not exhaustive) list of descriptors/bigrams commonly associated with each topic. Note, although the list presented here includes only English terms, in our analysis, topics were identified separately per language and by (near) native speakers of each language familiar with the specific discourse on migration within a language.

Crime & violence: plunder, gang, violence, rape, crime
Economy & job-market: wage, skill, labor market, tax, billions
EU politics: EU organizations, Schengen, Dublin agreement, European countries
Global politics: refugee/migration accord/convention, UN organizations, resettlement, international
Humanitarian & compassionate: escape, persecution, protection, vulnerable

Limiting & anti-migration: deport, illegal, quota, invader, migrant flood
National Politics: national party names and politicians
Refugee characteristics: Moslem, Muslim, different, Islam, culture
Refugee crisis: Mediterranean Sea, refugee boat, camps and places (e.g., Lesbos and Moria)