3rd International Conference on Advanced Research Methods and Analytics (CARMA2020)
Universitat Politècnica de València, València, 2020
DOI: http://dx.doi.org/10.4995/CARMA2020.2020.11655

#immigrants project: the on-line perception of integration

Rosario D'Agata, Simona Gozzo
Departement of Political and Social Sciences, University of Catania, Italy.

Abstract

This paper analyses the content of Twitter's comments during the period covering the last European elections. "#immigrants" is the extraction's keyword in different national languages. With the exception of English and French, whose extraction would be misleading, all of the other languages have been chosen to catch the geographical area of reference. We made sure to extract at least two sentences for each Welfare area. Once the data have been extracted, three different strategies have been used. The first one, dealing with both a qualitative and a quantitative assessment; the second one, analysing automatically the content of the top 10 extracted tweets during the reference period and the third one based on network analysis. Through a deep analysis of the content, three clusters have been identified: the first one dealing with the cultural risks of multiculturalism; the second one (social risks) dealing with the fear of migrants stealing job vacancies and the third one dealing with economic risks. A deep network analysis of Italian and Spanish contexts follows. What emerges is that: communication is extremely heterogeneous; in Italy there unique and duplicated edges prevails; in Spain there are more groups than in Italy, more themes covered and different kind of users and nets.

Keywords: *Big data; immigration; Network analysis; twitter.*

1. Introduction

This work concerns a study carried out by analysing the content of Twitter's comments during the two months close to the last European elections. The extracted tweets contain the headword "#immigrants" in different national languages. The selection of this specific social network platform derives from the decision to identify an instrument oriented to specific forms of communication: that carried by high interest in politics and/or involvement (Laifeld 2018). The detected discussions are conveyed not only to express an opinion, but also to influence public opinion with regard to the issue of integration.

A constraint of the approach implies the exclusion of comments in French and English. This depends on the need to attribute - for the interpretation of the comments - the sentences in a language to a reference area.

The analysis following the extraction of data follows three different comparing strategies allowing to detect complementary information. At the same time, the goal of this study is to assess when the three approaches capture redundant information and when, on the other hand, it is useful to integrate them. In particular, the applied strategies are:

1) A study based on an in-depth qualitative analysis (content) and quantitative (number of underlying links and ego-network structure) assessment. This analysis is centred on the top 10 extracted tweets in the selected language unit and for each week. The number of tweets in Polish and Slovenian languages is very limited. Probably this is due to the preference of other social networks (Surowiec and Štětka 2018).

2) Automatic analysis of the content of top 10 tweets extracted during the reference period

3) A study based on network analysis tools. This choice allows to select – for each week – the whole structure of relational net and the main groups as sub-networks obtained by extracting clusters mutually connected, with higher internal homogeneity and external heterogeneity in terms of links. This stage is referred specifically to Italian and Spanish comments.

2. The on-line comments

The period during which online communication was monitored (weekly) is close to the 2019 European consultations: from April 15 until early June 2019. This extraction is made in order to include mobilization effects on specific elements and / or priorities in communications during the last part of May. The increase in communication in this period could represent the political choice to convey the electoral campaign on issues aimed at either promoting integration or exclusion. This emerges as a strong trait for two Scandinavian areas (Sweden and Denmark), while an increase in communication - although not particularly high

compared to the dynamics that emerged - is recorded in the Mediterranean areas. No effect arises Eastern Europe areas.

What emerges immediately, considering the total number of users for each week, is that the Eastern Europe areas have no bias for communicating via Twitter. Communication in Polish and Slovenian, in fact, involved a few dozen contacts for week.

Table. 1. Number of users within the nets and for the top 10 tweets selected (page-rank) for each week.

Data	DK	FI	DE	IT	NO	NL	PL	PT	SI	ES	SE
15-22 april	528	525	985	9636	1472	1472	15	np	24	17023	2583
23-29 april	492	210	755	11736	1397	1397	26	np	21	18457	2157
30 april - 6 may	512	302	860	9139	1832	1832	42	4176	38	14392	2552
7-13 may	422	279	1129	10348	3093	3093	81	3606	np	17788	2347
14-19 may	371	239	845	11009	2732	2732	61	3189	42	np	2719
20-26 may	739	323	1172	10330	1744	1744	57	4724	13	18189	3108
27 may - 3 june	442	204	879	10792	2105	2105	np	3925	29	17628	3322
Node for top 10 tweets (%)											
15-22 april	33.71	40.38	42.34	34.63	52.51	31.39	73.33	np	58.43	58.43	52.42
23-29 april	36.18	28.10	11.13	41.14	22.96	23.19	65.38	np	30.96	30.96	41.68
30 april - 6 may	29.10	25.83	18.95	30.23	23.28	36.3	52.38	np	47.08	47.08	32.41
7-13 may	19.43	32.97	39.50	33.00	27.76	4.95	2.47	37.02	46.67	46.67	42.05
14-19 may	22.37	28.87	31.12	25.33	22.76	47.25	24.59	21.16	np	np	37073
20-26 may	49.12	45.51	42.41	2.91	20	29.87	29.82	14.39	23.02	23.02	37.19
27 may - 3 june	28.96	40.20	13.31	3.79	19.4	41	np	46.42	37.25	37.25	31.37
Average	31.27	34.55	28.39	24.43	26.95	30.56	41.33	29.75	40.57	40.57	39.26

The opposite trend, namely that of a particularly widespread and pervasive communication, is mainly registered in the Mediterranean Europe (Portugal, Italy and, above all, Spain) and in Sweden and Netherlands. The most relevant network comments cover only part of the entire communication flow. However, this selection permits an in-depth qualitative analysis on every single twit for each language, i.e. more specific information about the sources of information – not easy to catch (Tab. 2), the meaning of the tweet and its purpose. What has

been detected, when possible, is that news published either in online newspapers or private comments prevail, especially in Sweden, Spain and Finland, while Italy is characterized by an unusually high proportion of institutional comments by both parties and politicians (the social and web campaign conducted by the main governing parties is widespread).

Further information refers to the content of the tweets (Tab 3). All comments and posts refer to prejudices of some kind against immigrants and yet some specificities emerge quite clearly. Scandinavian areas, for example, are particularly sensitive to the issue of subsidies to immigrants, which are specifically considered a major problem in Denmark.

Table 2 - Source of information on twitter.

Country	Parties or politics	Trade Unions	Newspapers	Private users	Total
Denmark	7	0	10	5	22
Finland	3	5	7	12	27
Germany	5	2	10	2	19
Italy	21	2	1	11	35
Norway	3	0	8	3	14
Netherlands	3	0	8	11	22
Poland	0	0	1	1	2
Portugal	3	0	5	8	16
slovenia	0	0	4	2	6
Spain	6	0	2	13	21
Sweden	*3*	0	11	46	60
Tot	***54***	***9***	***67***	***114***	***244***

Tab. 3. Themes in tweets (%).

Country	Against prejjudices	Against immigrant subsidies	Prejudices	Analysis of prejudices and immigration argument	Against parallel comminities that do not communicate	Against institutions or politicians	Against journalist	Integration and work
DK	11,9	25,4	31,3	23,9	3,0	1,5	3,0	0,0
FI	11,6	2,9	24,6	47,8	2,9	5,8	0,0	4,3
DE	6,0	9,0	26,9	44,8	0,0	4,5	0,0	9,0
IT	11,8	1,5	33,8	25,0	1,5	26,5	0,0	0,0
NO	17,9	6,0	17,9	38,8	0,0	16,4	0,0	3,0
NL	1,6	8,1	32,3	32,3	0,0	24,2	0,0	1,6
PL	10,0	0,0	20,0	62,5	0,0	5,0	0,0	2,5
PT	16,0	2,0	8,0	46,0	0,0	12,0	0,0	16,0
SI	11,4	2,3	25,0	31,8	0,0	11,4	0,0	18,2
ES	13,6	4,5	31,8	18,2	2,3	25,0	4,5	0,0
SE	8,3	8,3	36,7	8,3	0,0	30,0	5,0	3,3

Moreover, unlike what is often believed, Scandinavian areas are among the main ones to express prejudices and fears, despite many comments aimed at "understanding" and solving problems related to the migration phenomenon. The most tolerant of the Scandinavian areas seems, in this sense, to be Finland, while the one less inclined to welcome and understand is Denmark. On the other hand, the Mediterranean areas are mainly characterized by institutional criticism (political, social, cultural). Concerning the subjects of the tweets (Tab. 4) we notice that *Political institutions* are present in 45.5% of Dutch tweets and *Parties* in 31.8% of Danish ones. In Spain, most of the tweets refer to *National Community.* In Germany and Portugal, almost one out of two tweets, concerns *immigrants* in general and in Sweden 1 out 5 *urban violence* related with immigrants.

At a later stage, a qualitative analysis has been carries out, analysing the 'sound' of the tweets (Tab. 5). The most positive evaluations seem to be twitted in Finland (14.5%) and Portugal (14.0%). On the contrary, the most negative ones appear in Netherlands (58.1%) and in Spain (57.8%).

Tab. 4 – Subjects of tweets (%)

	Political institutions	***Parties***	***Associations***	***National community***	***Immigrants***	***Extremists (violence)***	***Poverty -ghetto***	***Immigrants (urban violence)***
DK	13,64	31,82	0,00	9,09	28,79	6,06	3,03	7,58
FI	26,09	4,35	5,80	15,94	34,78	0,00	1,45	11,59
DE	14,93	8,96	7,46	8,96	49,25	0,00	0,00	10,45
IT	27,94	22,06	10,29	17,65	13,24	0,00	0,00	8,82
NO	11,94	22,39	5,97	22,39	20,90	1,49	1,49	13,43
NL	43,55	17,74	1,61	9,68	20,97	1,61	0,00	4,84
PL	5,00	20,00	5,00	17,50	35,00	0,00	0,00	17,50
PT	17,65	7,84	5,88	13,73	50,98	3,92	0,00	0,00
SI	28,89	11,11	0,00	22,22	33,33	0,00	0,00	4,44
ES	26,67	15,56	0,00	37,78	4,44	4,44	0,00	11,11
SE	28,33	18,33	5,00	16,67	10,00	1,67	0,00	20,00

Finally, we focused on the content of tweets related with internal of external policy. It is interesting to notice that while in Scandinavian area the content focuses on *Internal Policy* (94.5% average percentage), in Portugal 2 out 3 of tweets speak about foreign policy as well as 3 out 10 tweets relived in Germany.

Tab. 5 – Evaluations of immigrants and Content of tweets.

Country	Positive	Negative	Neutral	Internal policy	foreign policy
Denmark	1,49	41,79	56,72	96,97	3,03
Finland	14,49	24,64	60,87	98,55	1,45
Germany	4,41	23,53	72,06	70,15	29,85
Italy	7,35	41,18	51,47	85,29	14,71
Norway	5,97	29,85	64,18	97,01	2,99
Netherlands	0,00	58,06	41,94	80,65	19,35
Poland	10,26	28,21	61,54	89,74	10,26
Portugal	14,00	16,00	70,00	33,33	66,67
slovenia	0,00	37,78	62,22	82,22	17,78
Spain	11,11	57,78	31,11	86,67	13,33
Sweden	3,33	53,33	43,33	93,33	6,67

Main tweets' comments

The Clustering procedure carried out on the content of top-10 tweets selected for each week allowed identifying the main thematic cores. From that, three different clusters emerged. The results are analysed looking at the chi-square value and the frequency of each lemma within each cluster and within the tweet (defined as "elementary context"). These analyses allow to name and define the different identified clusters.

The first cluster (*cultural risks*) includes tweets aimed at questioning the multi-cultural approach. Comments from Finland and Sweden are mainly associated to this thematic group. One of the main issues is the of inability to integrate, in particular with regard to the differences in language, habits and culture. However, anti-racist comments in German and Portuguese are also related to this cluster.

The second cluster (*social risks*) is also the one in which most part of comments converges and includes many phrases, especially in Danish (but also in Polish), that use the terms as "work" and/or "people".

The third cluster (*economic risks*) is the one that is numerically less consistent and it refers to those comments that include terms such as Euro, Europe, immigrants, illegal immigrants, foreigners. This cluster is the least specific because it is associated with comments from all countries (although the association is stronger with Eastern Europe).

Tab. 7. Weight of clusters identified by elementary contexts analysis.

CLUSTER 1	341	31.31%
CLUSTER 2	455	41.78%
CLUSTER 3	293	26.91%

The clustering effects

In this step relational structures are analysed through network analysis tools (Borgatti *et al.*, 2013), reconstructing the graphs for each week, then extracting those groups with greater internal homogeneity and external heterogeneity (looking at both unique and reciprocal links). On the other hand, the frequent display of same tweets, domains and information probably implies a conformity of perspective among users. We can verify this hypothesis looking at the main content of the tweets for each sub-net. The analysis, carried out for each week, aims to assess precisely the presence of *communities* (although weak and iridescent) that share perspectives and materials. Furthermore, the analysis shows a very high heterogeneity which is reflected in the detection of a very high number of groups per week (extracted by applying the algorithm Clausette-Newman-Moore). As a discriminating criterion we agreed to take the presence of at least 1000 nodes in the network. In such a way, the reference groups get considerably reduced (Tab. 8). This choice does not produce a lack in information but avoids redundancies and "noise" due to self-referential or isolated nets. Besides, some differences in either area, contents and weeks emerge.

The above descripted procedure was mainly applied to the Italian and Spanish contexts, which demonstrated to be similar under many points of view: belonging to the Mediterranean context, lack of integration, transitional measures mainly oriented towards temporary reception of immigrants, with actions often defined as "emergency measures", mirroring in both areas a high level of prejudice towards immigrants.

Tab. 8 - Vertices for weeks in Italy and Spain (averages).

	Group	Vertices	Unique Edges	Edges With Duplicates	Total Edges	Self-loops
weeks			**SPAIN**			
First	4	1870	2178	136	2314	30
Second	3	1665	1792	56	1848	54
Third	4	1656	2336	191	2528	65
Forth	2	3346	4070	81	4151	28
Fifth	3	1443	2700	503	3203	126
Sixh	3	2096	3169	456	3625	106
Seventh	5	1587	2229	279	2508	36
Weeks			**ITALY**			
First	2	2064	3654	907	4561	295
Second	3	2298	3836	778	4614	235
Third	4	1203	1834	480	2315	143
Forth	3	1728	2883	1046	3929	257
Fifth	1	1736	3251	974	4225	679
Sixth	2	2236	4261	1162	5423	410
Seventh	3	1532	2403	845	3248	224

This analysis shows a greater network presence of Spanish users, although communication in Italian is pervasive. In addition, the peak of connected subjects who interact in Spanish is recorded during the fourth week while Italian users prevail in the second and sixth weeks. The number of subjects in the networks is higher in Spain while the number of links (especially in the duplicate proportion) is higher in Italian, as well as self-loops. The Italian nets appears, therefore, more redundant than the Spanish ones, which are more heterogeneous (more homogeneous groups of larger entities). Further information can be referred to the content of tweets. Indeed, these structural differences can probably be better understood by looking at the type of communication underlying them and, in particular, by analysing the sources (institutional, private, academic, political, etc.) and the users' goal (Yang et alii 2018). A similar operation was carried out through the qualitative analysis of the top 10

tweets. In this case, however, the selection becomes fully automated on the basis of Urls, Hashtag, words, etc. in each identified group.

Unlike the previous one, which selected the top 10 tweets of the week, the analysis proposed in this section refers to groups that emerged from the entire extracted network. The selection is made subsequently by importance of the single tweets with respect to the number of times it is selected or displayed and analysing the entity of the overall relational flows. The network structure identifies homogeneous sub-networks with respect to dynamics that are reconstructed, specifically, taking into account the structure of edges (Pfeffer, 2018). This procedure allows to analyse a very high number of nodes and links by identifying homogeneous networks.

Moreover, each group is homogeneous with respect to the content of the information conveyed within it because the links are given by the same information flows. Therefore, if a sub-network is more homogeneous, this depends on the fact that within it the nodes contact and send each other the same contents (or comments on them). In this way, it is therefore possible to establish a relationship between links, network structure, communication content and heterogeneity. What emerges considering the structures of networks is that in Italy and Spain:

- communication is extremely heterogeneous (this leads to a loss of information that is as relevant as you choose to select only the main networks, eliminating the noise from micro-communications or self-loops);
- in Italy there are more main unique edges and edges with duplicates (so there is more propensity to communication but even more redundancy in communication)
- in Spain there are more groups than in Italy, that is there are more themes, probably different kind of users and nets.

3. Conclusions

The information on the structure of networks has a relative value if it is not related to the content of the main tweets. The presence and the combination of different methods, with both analysis qualitative and quantitative, is the main novelty of the work. The qualitative step refers to messages that have a higher page-rank index in the network. The analysis of the content of the single tweets, furthermore, allowed to identify the type of senders. Finally, the use of the NA permitted to identify the structure of the links that, together with the content of the messages, allows further considerations in a more accurate comparative perspective.

References

Borgatti S., Martin E. &Johnson J. (2003), *Analysing Social Networks*, Los Angeles, Thousand Oaks, CA, London, Sage Publications.

Laifeld P. (2018), *Discourse Network Analysis. Political Debates as Dynamic Networks*, Victor J.N. & Montgomery A.H. (Eds) the Oxford Handbook of Political Networks, Oxford Un Press, 301-325.

Pfeffer J. (2018), *Visualization of Political Networks*, Victor J.N. & Montgomery A.H. (Eds) the Oxford Handbook of Political Networks, Oxford Un Press, 277-299.

Surowiec P. & Štětka V. (2018), Social Media and Politics in Central and Eastern Europe, Routledge

Yang S. & Gonzàlez-Bailòn S. (2018), *Semantic Networks and Applications in Public Opinion Research*, Victor J.N. & Montgomery A.H. (Eds) the Oxford Handbook of Political Networks, Oxford Un Press, 326-353.