Using stop words in text mining: Immigration and the election campaigns

Francesca Greco[1], Ken Riopelle[2,]Alessandro Polli[3], Julia Gluesing[4]

[1]Sapienza University of Rome, Department of Social Sciences and Economics – francesca.greco@uniroma1.it

[2] Wayne State University, Department of Industrial and Systems Engineering– kenriopelle@me.com

[3]Sapienza University of Rome, Department of Social Sciences and Economics – alessandro.polli@uniroma1.it

[4] Wayne State University, Department of Industrial and Systems Engineering– j.gluesing@wayne.edu

Abstract

In order to understand immigration sentiment and its relationship to other concepts in the Italian general election campaign of 2018 and the European election campaign of 2019, we collected in two corpora all the tweets in the Italian language containing the word "immigration" in the period preceding the vote. Both corpora underwent a sentiment analysis and a stop word analysis using two textual software packages: Linguistic Inquiry and Word Count (LIWC) (Pennebaker et al., 2015) and WORDij. LIWC was originally designed by James Pennebaker to understand how some patients recover from traumatic experiences by writing about those experiences and the emotions associated with it then and afterwards. LIWC consists of a dictionary of words which assesses the percent that they occur in a given text. LIWC analysis provides a measure of positive and negative emotion in the immigration text over time. WORDij is a text analysis program that can compute a Z-Score or the relative proportional test of difference between words and words pairs in two sets of texts. Using an include list of stop words we can determine how these relational words change over time with an emotional valence and Z-score to assess the immigration political debate over time. The paper represents a focus on stop-words, which have been an aspect of textual analysis that is often dismissed yet can be very important to our understanding of relational power.

Keywords: LIWC, WORDij, social media, political debate, immigration.

1. Introduction

Text mining procedures usually are used to analyze textual data, eliminating stop words like conjunctions, articles, and prepositions, and focusing the analysis instead on the "important" words such as nouns, verbs, etc. However, conjunctions and prepositions highlight the relationship among the concepts expressed in the text. This relationship could be of interest in social sciences as it reflects the power distribution, the dependence or the association among the entities evoked in the text.

Gregory Bateson's central message was that relationships are the essence of the living world, and that we need a language of relationships to understand and describe it. One of the best ways to do so, in his view, is by telling stories. "Stories are the royal road to the study of relationships," he would say. What is important in a story, what is true in it is not the plot, the things, or the people in a story, but the relationships between them".[1]

For this paper we are going to examine the story of immigration as told through two corpora of tweets in the Italian language containing the word "immigration" in the period preceding the vote of the Italian political election of 2018 and the European election of 2019. Specifically, we will focus on two aspects of the text: first the sentiment, using the Linguistic Inquiry and Word Count software (LIWC) (Pennebaker et al., 2015), and second, how a small group of "stop-words", or set of relational words, differentiate the Italian 2018 general election from the European election of 2019. In this case we will use Crovitz's 42 relational words he identified as "Take A (one thing) in some relation to B (another thing)." Here we will use "immigration" as one thing and the see how the relational words differentiate the Italian general election of 2018 from the European election of 2019.

Immigration has gained considerable relevance both in the public and political debate nowadays, and, in some cases, it has been associated with a strengthening of nationalist sentiments calling into question European Union membership. Immigration was one of the most relevant topics of the political debate during the electoral campaigns of the last three years in Europe. Part of the political debate during the electoral campaign took place in social media, in which people express their opinions and sentiment on immigration, especially the politically incorrect ones. Therefore, the analysis of social media communication allows for the understanding of people's opinions freely expressed.

In previous studies, the relevance of the immigration discourse in the political debate and its impact on citizen's voting choice was studied through an Emotional Text Mining approach (Greco et al., 2017; Greco, 2019; Greco and Polli, in press). The analysis of the political debate on social media during the French Presidential campaign in 2017 (Greco et al., 2017) highlighted that a negative sentiment prevailed towards immigrants, as only 42% of the messages classified were positive. Among negative sentiments, 35% of messages reflected citizen perception of being invaded or attacked, while only 13% considered immigrants as victims. The *Front National* leader, Marine Le Pen, largely exploited the negative sentiment toward immigrants, representing them as terrorists and invaders (35%) in her campaign. Although the analysis was made only on the messages produced by Twitter and could not be extended to all the voters, it was interesting to note that the *Front National* leader lost the election with 34% of votes on the second round. almost the same amount (35%) of the representation of immigrants as terrorists and invaders.

Immigration was a main topic also in the Italian election campaigns in 2018 and 2019. In the Italian general election of 2018, the political programs were promoted by three sides: the coalition of center-right parties, the coalition of center-left parties, and the *Movimento 5 Stelle* as a third contender. The tone of the election campaign toward immigrants was particularly heated. The center-right coalition and the *Movimento 5 Stelle* claimed an increase of the immigration control closing the border, while the center-left coalition was favorable to reception, protection, and integration policies to manage the immigration flow. On March 4th, 2018, the general election did not have a clear winner resulting in the dissolution of the

[1] Source: http://www.anecologyofmind.com/gregorybateson.html, last accessed on 12/03/2019.

center-right coalition and the set of an unexpected alliance between *Movimento 5 Stelle* and *Lega*, a right-wing party previously belonging to the center-right coalition.

For this reason, the European elections of 2019 in Italy became a crucial test of the general election results of 2018, particularly for the ruling parties: *Lega* and *Movimento 5 Stelle.* Although the campaigns of the Italian political parties also emphasized domestic issues, a main topic of the political debate called into question the immigration issue, criticizing the European Union governance.

The general dissatisfaction of Eurozone rules and the need for greater solidarity divided the political parties into two camps: those in favour of a revision of the Dublin regulation for a common immigration policy based on solidarity with fair redistribution and division of responsibilities among the European Union countries, and those emphasizing the need to hinder immigration by protecting external borders, increasing effective repatriations and opposing the redistribution of immigrants. The two ruling parties, *Movimento 5 Stelle* and *Lega*, were not aligned on the immigration policies. The *Movimento 5 Stelle* was favorable to the revision of the Dublin regulation and the redistribution of immigrants in line with the *Partito Democratico*, the center-left coalition's main party of the general election 2018, while the *Lega* focused its propaganda on the need to close the border, reinforce control and improve repatriations.

The election of the 73 Italian members of the European Parliament was held on May 26th, 2019, in Italy. The electoral round saw a strong affirmation of the *Lega* and a sharp fall of the *Partito Democratico* and the *Movimento 5 Stelle*. The overall winner of the elections was the *Lega*, the right-wing ruling party, who secured only 6% of votes in the 2014 European elections, tripling its consensus in the 2018 general election (17%) and doubling it (34%) the following year in the 2019 European election.

2. Literature review

The literature review chronicles Crovitz's 42 relational words over a span of 45 years from 1934 to 1979 and the four influential authors: C.K. Ogden, Karl Drucker, H.F. Crovitz and Karl Weick. In 1934 C.K. Ogden published his book, *The System of Basic English*, which was designed to help others learn English with just 850 words that even a child of six would know. In the Preface Ogden states: "The purpose of this volume is to present in a connected, and as far as possible, a complete form the System of Basic English for English-speaking readers" (Ogden; 1934, p. V).

Basic English emphasized visual learning in four areas: first, with 200 of the 600 Things classified as "Picturable Things";[2] second, 21 of 100 Operatives were Directives that could be visualized; third, all 850 words could be viewed at-a-glance on a single sheet of paper; and fourth using a word wheel or "Panopticon" enables the user to see all the words together (*ibidem*, p. 305). This is the original source from which Crovitz compiled his list of 42 relational words and cited in his published work of 1967, *The Form of Logical Solution*, and

[2] Ogden Basic English reference links: 850 words (http://ogden.basic-english.org); I. Introductory (Questions & Answers) (http://ogden.basic-english.org/sbe110.html#1); Basic English: A General Introduction with Rules and Grammar (http://ogden.basic-english.org/booksum1.html); The ABC of Basic English (Table of Contents) (http://ogden.basic-english.org/islabc.html); Word Pictures of 21 Directives (http://ogden.basic-english.org/wordpic2.html).

later in his 1970 book: *Galton's Walk*. Crovitz discovered there was a very simple heuristic to problem solving that Karl Drucker had missed because of Drucker's fixation on "obstacles to problem solving that caused him to miss those paths which do succeed (*ibidem*, p.95). Crovitz in his book: *Galton's Walk* Chapter 8 "The Relational Algorithm" goes through twelve of Drucker's example problems using the 42 relational words as a demonstration of the power of this heuristic[3]. Crovitz states:

> Action solves problems - it is the immediate cause of the arrival at the desired situation. Thought or chance may be a remote cause. The theoretical point is this: the basic feature of all possible solutions to a problem is that set of all possible actions that might mediate between the given and the desired solution (*ibidem*, p.96).

Crovitz explains:

> A heuristic question is whether there might be a set of words that are particularly worthy of thought, when the goal is to discover or invent. A well-known formal characteristic of many creative problems is that they often consist of taking two things in a new relation to each other. How many relations exist in which things may be taken? A potentially useful list of such relations follows from a few successive simplifying assumptions. (*ibidem*, p. 461).

Table 1 is Crovitz's 42 relation-words derived from Ogden's Basic English (*ibidem*, p.462).

Table 1 - Crovitz's 42 Relation Words

about	at	for	of	round	to
across	because	from	off	still	under
after	before	if	on	so	up
against	between	in	opposite	then	when
among	but	near	or	though	where
and	by	out	our	through	while
as	down	no	over	till	with

Weick in his 1979 book, *The Social Psychology of Organizing*, continues Ogden's visual theme in his section titled: "Visualize Organizations as Evolutionary Systems" as a heuristic for problem solving and creativity (Weick, 1979).

Weick states:

> An especially useful device to aid in solving problems is the relational algorithm (Crovitz 1967, 1970). This algorithm is a miniature evolutionary system and it is this resemblance that makes it a suitable exhibit of how evolution operates inside an organization (*ibidem*, p. 252-253).

Evolutionary systems are creative systems, and creativity usually means putting old things into new combinations and new things into old combinations.

[3] The twelve examples are: 1. X-Ray Problem, the Clock Problem, 2. the River Problem, 3. the Hanging Rope Problem, 4. the Four Triangles Problem, 5. the Door Problem, 6. the Mountain-Climbers Problem, 7. the Problem of the Circle in the Square, 8. the Gimlet Problem, 9. the Box Problem, 10. the Pliers Problem, 11. the Weight Problem, and 12 the Paperclip Problem.

In either case, novel relations between pairs of things are the essence of creativity. It was Crovitz's genius to see that there were exactly 42 relational words in Basic English (a word-list of 850 words) that could be used to relate two items.

The prototype sentence for depicting an idea is "Take one thing in relation to another thing". Or, more sparsely, "Take one thing […] another thing." One at a time, the 42 relational words (for example, *about, across, after, where, while, with*) are inserted in the brackets to see if they solve the problem (*ibidem*, p. 252-253).

Weick suggests users construct a word-wheel and put two problem concepts on discs with the 42 relational words between them and spin it, to put "the components of your problem into new relationships to discover solutions that had not occurred to you before" (*ibidem*, p.259).

Thus, we have the 45-year journey of Crovitz's 42 relational words starting from Ogden's 1934 attempt to create the most basic system of learning English to a problem-solving organizational heuristic for Weick in 1979.

In this case, we are using the Crovitz relational words to investigate and help differentiae two different political elections around the search term "immigration" using Twitter as the event data. Rather than removing these "stop" or "drop" words, which is often the practice in automatic textual analysis, our goal is to reverse the situation and only include these words for the start of the analysis. It is the author's intent to show that a text analysis method can benefit from focusing in on relations and those few words that connect "one thing to another".

3. The case study: Immigration and Italian election campaigns

3.1. Data collection

For the Italian Election of 2018 Twitter messages in the Italian language were collected using the search term "immigration", "immigrant", and "immigrants" for ten days during the period of January 16th to January 25th, 2018, resulting in approximately 41,157 tweets or retweets.[4] For the European Election of 2019 Twitter tweets were collected likewise using the same search terms for 20 days during the period of March 19th to March 28th and from April 5th to April 14th, 2019, resulting in approximately 147, 236 tweets or retweets[5]. The package *rtweet* of R Statistcs (Gentry, 2016) was used to collect the Twitter data, which collects the tweets based on the Twitter API. No geographic restrictions were used, but the language was restricted to Italian. The European 2019 corpus is thus 3.5 times larger than the Italian corpus of 2018.

As language depends on the context, its use in twitter is quite different from blogs or expressive writing (Pennebaker et al., 2015). The tweet is similar to a slogan due to its character restriction (max 280) and the large amount of newly created words. As stated by Pennebaker and colleagues, « the LIWC2015 version captures, on average, over 86 percent of

[4] The **2018 Italian general election** was held on 4 March 2018 after the Italian Parliament was dissolved by President Sergio Mattarella on 28 December 2017. Source: Francesco Verderami (13 December 2017). "Elezioni 2018, si punta al 27 dicembre per lo scioglimento delle Camere: si vota il 4 marzo". *Corriere.it*. https://en.wikipedia.org/wiki/2018_Italian_general_election#cite_note-10, last accessed December 29, 2019.

[5] An election to the European Parliament was held in Italy on 26 May 2019. Source: https://en.wikipedia.org/wiki/2019_European_Parliament_election, last accessed December 29, 2019.

the words people use in writing and speech. Note that […], all means are expressed as percentage of total words used in any given language sample. Simple statistical tests indicate that nearly all language categories differ significantly between contexts » (*ibidem*, p. 10). The percentage of words in each category for the Twitter context was studied only for the English language. For this reason, the percentage of words in each category resulting from our analysis cannot be compared with a reference threshold. Nonetheless, we can assume that the word characteristics of the two corpora should be the same (same topic, same communication style) and that the language differences are the result of the way people talk about immigration in a given time. For this reason, we decide to compare the two corpora over time, matching the LIWC results of the 2018 General election to the 2019 European election.

3.2. Data analysis

The Twitter data for the general election of 2018 and for the European election of 2019 were analyzed using the two text software packages: Linguistic Inquiry and Word Count (LIWC) (Pennebaker et al., 2015) and WORDij (Danowski, 2013).

LIWC was originally designed by James Pennebaker to understand how some patients recover from traumatic experiences by writing about those experiences and the emotions associated with it then and afterwards. LIWC consists of a dictionary of words which assesses the percent that they occur in a given text. For this analysis an Italian LIWC dictionary was downloaded from the LIWC website for use in this analysis.[6] No modifications were made to the LIWC Italian dictionary.

There are 101 Italian dictionary categories of which for the purpose of this sentiment analysis we use just three categories: Negazio, Emotion Positive (Emo_Pos), Emotion Negative (Emo_Neg) and a calculation for the fourth, Positivity Index (the ratio of Emo_Pos divided by Emo_Neg).

For the WORDij analysis we preprocessed the two election files to remove any special characters and added a period at the end of every line of a tweet to create a “clean” file for analysis. A period was added to ensure the WORDij slide window would stop at the end of tweet.

A string replace file or recode file was created for a more complete translation of the Crovitz 42 words into Italian. An Italian include file was created for the translated Crovitz list of relational words along with three focus words: “immigrato” (immigrant), “immigrati” (immigrants), “immigrazione” (immigration). The WORDij Wordlink module was run to create a word frequency output for each file, then the WORDij Z-Word module was run to calculate a Z-Score and a Chi-Square to determine if there was any significance in the proportional frequency or counts in the occurrences of these words.

The Z-test is for proportions (relative frequencies) and the Chi-Square is a test of differences in counts. The Z-test cannot produce a value when one of the pairs has a frequency of zero, so we enter a very small constant to replace zero.

The critical z value for two proportions are:

$$p < .05 \text{ is } + /- 1.64,$$

$$p < .01 \text{ is } +/- 2.389$$

[6] http://dictionaries.liwc.net/index.php/site/login, last accessed December 29, 2019.

p < .001 is +/- 3.5

The Chi-Square test may be preferred by some analysts because it is not an inferential statistic whereas Z-tests are. Nevertheless, if the number of occurrences in one or both of the files is less than 5 then Chi-Square statistics should not be used because the estimates are invalid. The value of Chi-Square that is statistically significant for degrees of freedom 1 and p < .05 (number of cells -1) is 3.841. Values higher than this are significant at higher levels. For example: p < .01 the critical value is 6.635, p < .005 is 7.879.

Table 2 - WORDij String Replace File:

grossomodo->circa	poiché->perché	nello->in	sopra->su	seppure->sebbene
incirca->circa	poichè->perché	nella->in	contrario->opposto	eppure->sebbene
tra->tra_fra	perchè->perché	spento->via	intorno->dietro	fintanto->fino
fra->tra_fra	siccome->perché	sul->su	allora->poi	finché->fino
allo->a	però->ma	sullo->su	nonostante->sebbene	finchè->fino
alla->a	entro->vicino	sulla->su	benché->sebbene	
agli->a	basso->giù	sulle->su	benchè->sebbene	
alle->a	nel->in	sugli->su	tuttavia->sebbene	

4. Results

The LIWC Sentiment Results are presented in Table 3. The Italian 2018 Election tweets were more negative than the European 2019 Election tweets as represented by the three LIWC categories of Negations, Emotion Positive (Emo_Pos), Emotion Negative (Emo-Neg) and the calculated Positivity Index. Negations declined by 37.0% from 2.57% from the Italian 2018 Election to 1.62% a year later in the European 2019 Election. Similarly, Emotion Positive (Emo_Pos) rose by 77.8% or .21 from .27 to .48 and Emotion Negative (Emo-Neg) rose 9.9% from 1.51 to 1.66 with an increase of .15. The Positivity Index, which is a ratio of Emo_Pos on Emo_Neg, rose by 61.7% or .11 from .18 to .29 from the Italian 2018 Election to the European 2019 Election. Even though the Emotion Negative decreased in the European elections, Anxiety rose 62.5% from .16 to .26 an increase of .10.

Table 3 - LIWC Sentiment Results

Filename	*Word Count*	*Negations*	*Emo_Pos*	*Emo_Neg*	*Positity Index*
Italian election 2018	793,558	2.57	0.27	1.51	0.18
European election 2019	2,040,530	1.62	0.48	1.66	0.29
Row Difference (Italian- European)	-1,246,972	0.95	0.21	0.15	0.11
Percent (Row Difference/Italian)		37.0%	77.8%	9.9%	61.7%

The WORDij Z-Word Results are presented in Table 4. The 37 words are ranked by their Z-Score from -26.92 to a + 71.51. There are seven columns in Table 4, labeled A to G:

- Column A is the English Crovitz relational word;
- Column B is the Italian translated Crovitz relational word;
- Column C is Group 1 Frequency Count (FC) for the General Election (FC-GE) (2018);

- Column D is Group 2 Frequency Count for the European Election (FC-EE) (2019);
- Column E is Group 1 Proportion (P) for the General Election (P-GE) (2018);
- Column F is Group 2 Proportion (P) for the European Election (P-EE) (2019);
- Column G is the Z-Score or standard deviation, significant values are +/- 1.64, p<.05;
- Column H is the Chi Square, where a minimum count of 5 is needed for each word, significant values are +3.841, p<.05.

Table 4 - WORDij Z-Word Results for Crovitz Relational Words Translated into Italian

Crovitz	***Italian***	***Group 1***	***Group 2***	***Group 1***	***Group 2***	***Z-***	***Chi***
Word	***Translation***	***FC-GE***	***FC-EE***	***P-GE***	***P-EE***	***Score***	***Square***
for to	per	7435	23052	0,049835	0,070412	-26,92	7999,83
in	in	11538	32955	0,077337	0,100661	-25,67	10309,21
about	circa	20	1081	0,000134	0,003302	-21,12	1022,45
at to	a	14557	38309	0,097572	0,117015	-19,82	10671,46
while	mentre	160	1518	0,001072	0,004637	-19,26	1099,03
but	ma	3756	11157	0,025176	0,034079	-16,37	3672,96
after	dopo	524	2390	0,003512	0,007300	-15,56	1194,91
now	di	22116	53245	0,148239	0,162637	-12,63	12858,30
when	quando	1447	4403	0,009699	0,013449	-10,90	1493,66
now	ora	628	2231	0,004209	0,006815	-10,80	898,78
through	sotto	45	435	0,000302	0,001329	-10,37	316,88
under	dove	477	1739	0,003197	0,005312	-9,95	718,70
from by	da	5805	14326	0,038910	0,043759	-7,72	3606,75
out	fuori	201	768	0,001347	0,002346	-7,10	331,77
immigrant	immigrato	2909	7340	0,019498	0,022420	-6,45	1915,68
opposite	opposto	58	307	0,000389	0,000938	-6,35	169,87
among between	tra_fra	927	2585	0,006213	0,007896	-6,30	782,73
though	sebbene	72	337	0,000483	0,001029	-5,98	171,70
because	perché	3	69	0,000020	0,000211	-4,97	NA
still	ancora	461	1307	0,003090	0,003992	-4,75	404,82
till	fino	112	384	0,000751	0,001173	-4,19	149,16
as	come	2202	5291	0,014760	0,016161	-3,61	1273,44
and	e	13719	31016	0,091955	0,094738	-3,05	6687,97
with	con	4653	10517	0,031188	0,032124	-1,71	2266,74
against	contro	1005	2266	0,006736	0,006921	-0,72	486,13
or	or	2	1	0,000013	0,000003	1,32	NA
on over up	su	4208	8970	0,028205	0,027399	1,57	1720,80
round	dietro	85	129	0,000570	0,000394	2,66	9,05
then	poi	825	1599	0,005530	0,004884	2,91	247,14
down	già	83	70	0,000556	0,000214	6,12	1,10
near by	vicino	180	188	0,001206	0,000574	7,29	0,17
off	via	617	917	0,004136	0,002801	7,54	58,67
if	se	4707	8325	0,031550	0,025429	12,02	1004,44
before	prima	1461	1567	0,009793	0,004786	20,17	3,71
not	non	13728	23275	0,092016	0,071093	25,03	2463,18
immigrants	immigrati	20111	27433	0,134799	0,083794	54,49	1127,62
immigration	immigrazione	8355	5884	0,056002	0,017973	71,51	428,81

Thirty-four of the thirty-seven words (92%) have significant Z-Scores of +/- 1.64. The three words that do not have significant Z-Score are: “contro” (against), “o” (or), and “su” (up). Two words do not meet the Chi Square minimum count of 5 in each file. They are: “perché” (why) and “o” (or) and they are designated with “NA” (not applicable). Three additional words fall below the Chi-Square significant value of + 3.841. They are: “giù” (down), “vicino” (near), and “prima” (first). Thus, thirty-two of the thirty-seven words (86%) have significant Chi-Square values.

Overall, these Italian translated Crovitz relational words are in this specific case very discriminating between the general election of 2018 and the European 2019 election in the Twitter sphere.

5. Discussion and conclusion

The results of the LIWC and the Crovitz relational words are in line with the results highlighted in previous studies performed on the same corpora with a different text mining approach, Emotional Text Mining (ETM) (Greco, 2016; Greco and Polli, 2019). ETM is an unsupervised methodology, which allows the profiling of people based on their communication; it is a bottom-up semiotic approach used to classify unstructured data according to word co-occurence. It allows the understanding of people’s symbolizations, representations and sentiment, about one or more discourse topics. ETM is a fast and relatively simple procedure, which is used to extract meaningful information from large text corpora.

It is interesting to note that the sentiment difference among the election campaigns of 2018 and 2019 are highlighted both by the semiotic approach (ETM) and the semantic approach (LIWC and Crovitz relational words). Each procedure highlights some specific characteristics of the communication in term of words choice (LIWC) and words relationship (Crovitz relational word and ETM). And, taken as whole, they allow for a deeper understanding of the role played by immigration in the political debate during the election campaigns.

According to ETM results, the sentiment in the Italian election of 2018 lacks positivity and only 12% of the classified messages are neutral (Greco, 2019). Tweets are mostly negative (88%) and negativity can be distinguished as negative for the community (33%), negative for immigrants (39%), and gender negativity (16%). Immigrants are represented as dangerous invaders (29%), violent against women (16%), dangerous regulars (4%), irregular workers (30%), new slaves (9%), and a social issue (11,8%). The negative sentiment seems to focus more on personal aspects (negative for immigrant and gender negativity = 55%) rather than community ones. This result seems to suggest that Italian culture is more sensitive to individual elements. Moreover, the different geopolitical conditions that characterize the two countries probably involve less positive sentiment regarding the need to manage the problem of immigration costs at the European level.

In line with the results of the Crovitz and the LIWC analysis, the sentiment on immigration during the European election measured with ETM was more positive (Greco et Polli, in press). The Italian reaction to immigration was: zero tolerance (22%), border closure (8%), selection of worthy immigrants (16%), countering crimes (24%), and dangerous aiding (30%). Positivity related to the possibility to welcome worthy legal immigrants in Italy (16%), while 84% of tweets were negative. The sentiment on immigration in the European elections in Italy confirmed that of the general election, with virtually the same negative sentiment percentage.

The increase of positivity in Twitter's communication in 2019, detected by both the ETM and the LIWC, could be explained by the decision of the *Lega-M5S* government to strengthen border control, which led to a drastic drop in the flow of illegal migrants. As a matter of fact, in the first four months of 2018, 9,467 landings of illegal immigrants were registered, compared to the 746 illegal landings (-92%) recorded in the same period of 2019.

Also, the Crovitz's relational words results seems to confirm this hypothesis. The communication during the general election campaign was characterized by words reflecting a refusal: *away* (away), *behind* (hidden), *not* (negation), *if* (condition), and *after* (consequences), highlighting a threatening and persecutory perception of immigrants. While in 2019 communication, the tone is more «objective» with the presence of Crovitz's relational words reflecting time and space specification (e.g., *of*, *to*, *for*, etc.).

The results of the LIWC showed that although negative sentiment eased in the European elections, there was an increase in anxious words among negative ones. However, this contradiction is only apparent. Indeed, Pennebaker notes that by its nature the human mind constantly tries to understand the world around it. One of the reasons why we are obsessed with a given negative feeling is the constant attempt to understand it. An effective way to find an answer is to talk about it or put it into words.

Anxiety is the apprehensive anticipation of future dangers or negative events. The fact that both positive sentiment and an increase in anxious words emerge in the tweets during the 2019 European election campaign can be referenced back to the Pennebaker model: while writing about painful or negative things, you are feeling bad. Nevertheless, expressing the negative feelings makes you feel better and can lead you to express positive feelings also. The Pennebaker model explains the positive sentiment on immigration observed in 2019 despite the increase in words expressing anxiety. This result is confirmed by Crovitz's relational words, which reflect an attitude less persecutory and more oriented to explain, compare, and specify.

The combination of the three methodologies (ETM, LIWC and Crovitz's relational words) has some significant advantages. First, the combination of the three methodologies allows increasing the depth of the analysis, linking the interpretation of the results to consolidated theoretical frameworks. Furthermore, their joint use allows for a much more detailed study since each of them provides valuable insights and suggestions for the interpretation of the results as a whole.

In fact, the ETM allows extracting meaningful information from large corpora, clustering the texts, identifying the latent dimensions that organize the discourse about a topic, and the sentiment that characterizes each cluster. The LIWC enables deepening the sentiment analysis, making it more specific, and linking it to a well-founded psychosocial theoretical frame. Finally, the methodology based on Crovitz's relational words facilitates understanding how the various parts of the text connect to the topic and making comparisons between different corpora.

In summary, the text mining procedure described in this study helps answer three fundamental questions about a collection of texts related to a topic: what, how, and above all why. We aim to direct future developments of our research towards two main objectives. The first objective is to improve the potential of this text mining procedure, refining the final stage of interpretation of the results. Second, to extend its use to other areas of application, where it is crucial to have tools for fast and relatively inexpensive extraction of structured information from large text corpora.

References

Crovitz H.F. (1967). The form of logical solutions. *American Journal of Psychology*, 80: 461-462.

Crovitz H.F. (1970). *Galton's Walk*. New York: Harper and Row.

Danowski J.A. (2013). WORDij version 3.0: Semantic network analysis software. Chicago: University of Illinois at Chicago.

Gentry J. (2016). R Based Twitter Client. R package version 1.1.9.

Greco F. (2016). *Integrare la disabilità. Una metodologia interdisciplinare per leggere il cambiamento culturale*. Milano: Franco Angeli.

Greco F. (2019). Il dibattito sulla migrazione in campagna elettorale: Confronto tra il caso francese e italiano. *Culture e Studi nel Sociale,* 4*(2):* 205-13.

Greco F., Maschietti D. and Polli A. (2017). Emotional text mining of social networks: The French pre-electoral sentiment on migration. *RIEDS*, 71(2): 125:36.

Greco F. and Polli A. (in press). The political debate on immigration in the election campaigns in Europe. In P. Gloor, F. Grippa and A. Przegalinska (Eds.), *Proceedings of COINS 2019*. Pearson.

Greco F. and Polli, A. (2019). Emotional Text Mining: Customer profiling in brand management. *International Journal of Information Management.* DOI: 10.1016/j.ijinfomgt.2019.04.007

Ogden C.K. (1934). *The System of Basic English*. New York: Harcourt, Brace and Company.

Pennebaker J.W., Boyd R.L., Jordan K. and Blackburn K. (2015). *The development and psychometric properties of LIWC2015*. Austin, TX: University of Texas at Austin.

Weick K. (1979). *The Social Psychology of Organizing*. Philippines, Addison-Wesley Publishing Company.