Brexit in Italy

Text Mining of Social Media

Francesca Greco, Livia Celardo, Leonardo Salvatore Alaimo

Abstract The aim of this study is to identify how Italian people talk about Brexit on Twitter, through a text mining approach. We collected all the tweets in Italian language containing the term "Brexit" for a period of 20 days, obtaining a large corpus on which we applied multivariate techniques in order to identify the contents and the sentiments within the shared comments.

Abstract *Questo studio ha lo scopo di identificare in che modo Brexit viene discussa su Twitter dagli Italiani attraverso l'analisi automatica del testo. A questo scopo sono stati raccolti tutti i messaggi in lingua italiana contenenti i termini "Brexit" per 20 giorni, ottenendo un corpus di grandi dimensioni su cui sono state applicate delle tecniche statistiche multivariate al fine di individuare i contenuti e i sentimenti relativi al tema in esame.*

Key words: Brexit, Twitter, Emotional text mining, Co-clustering

Introduction

There is a growing increase in Euroscepticism among EU citizens nowadays, as shown by the development of the ultra-nationalist parties among the European countries. Regarding to European Union membership, public opinion is divided between Eurosceptics and pro-Europeans, as clearly shown by the results of the 2016 British referendum. Many studies about Brexit are focused on the analysis of the electoral result, trying to highlight the effects of possible determinants - such as immigration, economic crisis, socio-economic and demographic characteristics of the voting population - on the electoral choices (Gietel-Bastel, 2016; Goodwin & Heath, 2016; Clark et al., 2017; Alaimo, 2018). Other studies analyse the possible consequences of Brexit, focusing on the economic consequences for United Kingdom (Dhingra et al., 2016; Dodds, 2016; Vikers, 2017). Although Brexit has shaken the European public opinion, there are few studies about how the referendum is actually

[1] Francesca Greco; Prisma S.r.l., Sapienza University of Rome; francesca.greco@uniroma1.it
Livia Celardo; Sapienza University of Rome; livia.celardo@uniroma1.it
Leonardo Salvatore Alaimo; Sapienza University of Rome; leonardo.alaimo@uniroma1.it

perceived by the citizens of European Member States. Moreover, not many analyses are available concerning how Brexit is discussed on the social media.

The wide diffusion of the internet increases the opportunity for millions of people to surf the web, create account profiles and search or share information every day. The constant rise in the number of users of social media platforms, such as Twitter, makes a large amount of data available; these data represent one of the primary sources for exploring people's opinions, sentiments, and emotions (Ceron et al., 2013; Pelagalli et al., 2017).

Due to that, we decided to perform a quantitative study where online discourses regarding Brexit are analysed using two text analysis techniques in parallel: Content Analysis and Emotional Text Mining. The aim is to explore not only the contents but also the sentiments shared by users on Twitter. In this paper we focused only on the analysis of the sentiments and contents published in Italian, in order to understand how Brexit is treated in Italy.

Methods

In order to explore the sentiments and the contents related to Brexit, we scraped from Twitter all the messages written in Italian containing the word *Brexit*, produced from January 23rd to February 18th, 2018. The data extraction was carried out with the TwitteR package of R Statistics. From the data we extracted, we decided to create two sub-corpora: the first one including the retweets and the other one excluding all the retweets. We chose to use two different corpora for the analyses because we were studying both sentiments and contents within our texts; for the analysis of sentiments and emotions it is important to consider also retweets, while for content analysis retweets are just the repetition of the same concepts. Then, the first corpus was composed of 13,662 messengers, including 76.4% of retweets, resulted in a large size corpus of 211,205 of tokens, which underwent the Emotional Text Mining (Greco et al., 2017). A second corpus was extracted excluding all the retweets, resulting in a large size corpus of 46,458 tokens, which underwent the content analysis. In order to check whether it was possible to statistically process data, two lexical indicators were calculated: the type-token ratio (TTR) and the hapax percentage ($TTR_{corpus\ 1} = 0.04$; $Hapax_{corpus\ 1} = 42.2\%$; $TTR_{corpus\ 2} = 0.16$; $Hapax_{corpus\ 2} = 61.3\%$). According to the large size of the corpora, the lexical indicators indicate the possibility to proceed with the analyses.

Then, on the first corpus was performed a sentiment analysis, which is a technique used to investigate the sentiments of a text. In text mining, many methods exist to analyse it automatically, which are supervised and unsupervised (e.g., Carli & Paniccia, 2002; Hopkins & King, 2010; Bolasco, 2013; Ceron et al., 2016). We performed the Emotional Text Mining (ETM) (Greco, 2016; Pelagalli et al., 2017; Greco et al., 2018), which is an unsupervised method derived from the Emotional Textual Analysis of Carli and Paniccia (Carli & Paniccia, 2002; Bolasco, 2013); it is based on the idea that people emotionally symbolize an event or an object, and socially share this symbolisation. The words they choose to talk about this event or object is the product of the socially-shared unconscious symbolization. According to this, it is possible to detect the associative links between the words to infer the symbolic matrix determining the coexistence of these terms in the text. In order to perform ETM, first corpus was cleaned and pre-processed with the software T-Lab (version T-Lab Plus 2018) and keywords were selected. In particular, we used lemmas as keywords instead of types, filtering out the lemma *Brexit* and those lemmas of the low rank of frequency (Greco, 2016). Then, on the tweets per keywords matrix, we performed a cluster analysis using the bisecting *k*-means algorithm, limited to twenty partitions (Savaresi & Boley, 2004) and excluding all the tweets having less than two keywords co-occurrence. The eta squared index was used to evaluate and choose the optimal solution in terms of number of clusters. To complete the analysis, a correspondence analysis (Lebart & Salem, 1994) on the keywords per clusters matrix was made, in order to explore the relationship between clusters and identify the emotional categories. The main advantage connected with this approach is the possibility in interpreting the factorial space according to words polarization, thus identifying the emotional categories that generate Brexit representations, facilitating the interpretation of clusters and exploring their relationships within the symbolic space.

On the other hand, content analysis is a technique used to investigate the subjects treated within a text; in text mining, many methods exist to analyse the contents automatically. One of these is the Text Clustering; it consists of splitting the corpus in different subgroups based on words/documents similarities (Iezzi, 2012). In this paper, a text co-clustering approach (Celardo et al., 2016) for contents analysis is used. The objective is to simultaneously classify rows and columns, in order to identify groups of texts characterized by specific contents. To do that, data were pre-processed with Iramuteq software: we lemmatized the texts and we removed stop words and terms with frequency lower than 2. The weighted terms-documents matrix was then co-clustered through the double *k*-means algorithm (Vichi, 2001); the number of clusters for both rows and columns was identified using the Calinski-Harabasz index.

Results

The results of the ETM show that the 240 keywords selected allowed us to classify the 79% of the tweets. The eta squared index was calculated on different partitions – from 3 to 9 clusters, and the values we found showed that the optimal solution is four clusters. The correspondence analysis detected three latent dimensions (Table 1). In figure 1, we can see the emotional map of Brexit emerging from the Italian tweets. It shows how the clusters are placed in the factorial space. The first factor represents the evaluation of the Brexit deal, considering the exit from the EU a bad deal or a good one; the second factor reflects the British political strategy, differing the partisan strategy – which is led by specific political and economic interests, from the public strategy, which is carried on by the government; finally, the third factor represents the post-Brexit effects evaluation, distinguishing the forthcoming impact from the future one. The four clusters are of different sizes (Table 2), and they reflect the Italian sentiments toward Brexit. The first cluster represents the Soros scandal as a conspiracy that disregards the democratic expression of Britons' choice; the second cluster reflects Italian satisfaction in considering Brexit as a bad deal for Britons, who are punished for their betrayal; the third cluster concerns the negative European economic impact of the British policy, which is perceived as an unfair British advantage causing a loss for EU citizens; and the fourth cluster highlights the hope for a British comeback through a new referendum. By clusters interpretation, no group highlights a positive sentiment in the direction of British exit from the EU. Nevertheless, we have considered as positive or satisfactory (51.5%) the British punishment and the hope for a British comeback, and negative the other two (48.5%), the unfair British advantage and the citizen right violation.

Table 1 - *Explained inertia for each factor*

Factor	*Eigenvalues*	*%*	*Cumul. %*
1	0,705	38,3	38,3
2	0,621	33,7	72,0
3	0,515	28,0	100,0

Figure 1 - *Factorial space set by three factors*

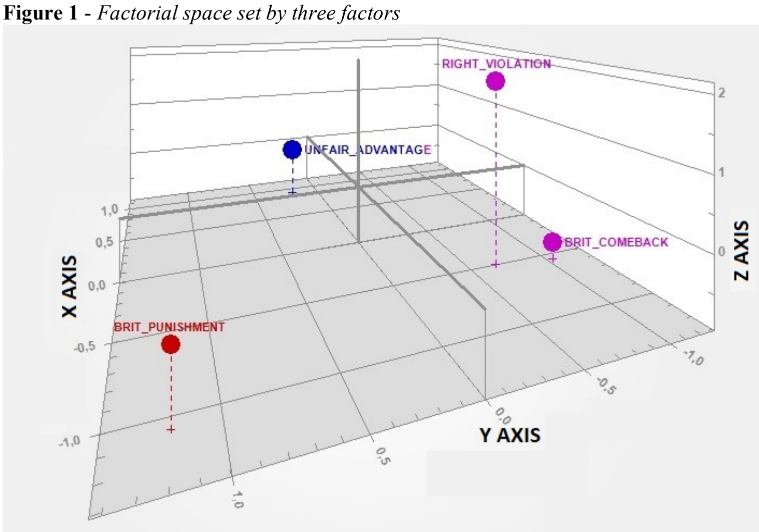

Table 2 – *Brexit representations and sentiments*

Clusters	*No. tweets classified*	*Size*	*Label*	*Keywords*	*CU*	*Sentiment*
1	1355	13.0%	Citizen right violation	Soros	792	Negative
				media	496	
				documento	425	
				telegraph	383	
				prova	372	
				cercare	359	
2	2517	24.0%	British punishment	ammettere	867	Positive
				Regno Unito	809	
				governo	789	
				peggio	692	
				britannico	606	
				Europa	436	
3	3733	35.5%	Unfair British advantage	europeo	784	Negative
				UK	717	
				Italia	618	
				occupazione	545	
				sterlina	469	
				anti-Brexit	461	
4	2894	27.5%	British comeback	UE	912	Positive
				Soros	769	
				Londra	675	
				May	378	
				voto	352	
				segreto	344	

The first six keywords of the clusters are ordered by the number of context units (CU) classified in each cluster

For the content analysis, we firstly pre-processed the corpus, removing all the noise – i.e. stop-words – cutting the messages contents by 83%. Then, on the terms-documents matrix (where documents are represented by the tweets) we calculated the Calinski-Harabasz index, in order to find the number of groups, for both the rows and the columns. The index indicated five groups for words and five for tweets; the results of the co-clustering procedure are shown in the Table 3. The first group of words identifies the common language, representative of all the messages; it is about the general conditions and aspects of Brexit. The 35% of the tweets has, in addition to this, a specific language. Almost the 30% of messages is about the impacts and the effects of Brexit on the local economies and on the political systems. A smaller share of tweets deals with the Soros scandal (5%), while just few messages are about what is going to happen in the near future in Italy.

Table 3: *Centroids matrix resulting from the co-clustering procedure*

Cluster - Label	***1***	***2***	***3***	***4***	***5***
	(65%)	(14%)	(13%)	(5%)	(3%)
1- Brexit	0.005	0.001	0.002	-0.001	0.001
2 - EU Impact	0.002	0.000	**0.049**	-0.002	-0.002
3 - Economic Effects	0.005	**0.066**	0.004	-0.003	0.002
4 - Soros Scandal	0.005	0.004	0.006	-0.007	**0.668**
5 - Forthcoming	0.009	0.004	0.001	**0.624**	0.028

Discussion and conclusion

The results of our analyses showed how "Brexit" is represented by Italians both in terms of sentiments and contents. The Emotional Text Mining revealed the presence of positive and negative sentiments in respect to the consequences of Brexit, but not directly toward the UK exit. A positive sentiment is connected to the punishment of the British choice and the hope of a rethink concerning the UK exit, which is basically perceived as a betrayal. In fact, negative opinions rely on the unfair British advantage and the disregards of the Britons' referendum choice.

Regarding the content analysis, it identified in the first cluster the common language shared within Twitter, used to describe what happened and is continuing to happen in Europe after Brexit. The other four clusters underline the presence of a specific language. In particular, the second and third clusters focus on the current political and economic effects of Brexit in Europe.

The results of the two analyses showed that Brexit is a theme with a strong emotional charge in the Italian context. Italian people seem to focus their attention mainly toward the economic and political effects of the British choice. This subject is treated negatively from the users, probably because of the worries for the consequences of the British vote in Europe.

References

1. Alaimo, L. S.: Demographic and socio-economic factors influencing the Brexit vote. Rivista Italiana di Economia, Demografia e Statistica (RIEDS), 72(1), 17–28 (2018).
2. Bolasco S.: L'analisi automatica dei testi: fare ricerca con il text mining. Carocci, Roma (2013).
3. Carli R., Paniccia R. M.: Analisi Emozionale del Testo. Franco Angeli, Milano (2002).
4. Celardo, L., Iezzi, D. F., Vichi, M.: Multi-mode partitioning for text clustering to reduce dimensionality and noises. In: Mayaffe, D., Poudat, C., Vanni, L., Magri, V., Follette, P. (eds) JADT 2016: Statistical Analysis of Textual Data. Les Press de Fac Imprimeur, Nizza (2016).
5. Ceron, A., Curini, L., Iacus, S.M.: Social Media e Sentiment Analysis. L'evoluzione dei fenomeni sociali attraverso la Rete. Springer, Milano (2013).
6. Ceron A., Curini L., Iacus S. M.: iSA: a fast, scalable and accurate algorithm for sentiment analysis of social media content. Information Sciences, 367, 105-124 (2016).
7. Clark, H.D., Goodwin, M., Whiteley, P.: Brexit: Why People Voted to Leave the European Union. Cambridge University Press, Cambridge (2017).
8. Dhingra, S., Ottaviano, G., Sampson, T., Van Reenen, J.: The consequences of Brexit for UK trade and living standards. Centre for Economic Performance (CEP), London School of Economics and Political Science (LSE) (2016).
9. Gietel-Bastel, S.: Why Brexit? The Toxic Mix of Immigration and Austerity. Population and Development Review, 42(4), 673–680 (2016)
10. Goodwin, M. J., Heath, O.: The 2016 Referendum, Brexit and the Left Behind: An Aggregate-level Analysis of the Result. The Political Quarterly, 87(3), 323-332 (2016)
11. Greco, F.: Integrare la disabilità. Una metodologia interdisciplinare per leggere il cambiamento culturale. Franco Angeli, Milano (2016).
12. Greco, F., Mascietti, D., Polli, A.: Emotional text mining of social networks: the French pre-electoral sentiment on migration. RIEDS (2018) Available from http://www.sieds.it/index.php?option=com_content&view=article&id=17:rivista-rieds&catid=26:pubblicazioni&Itemid=136
13. Hopkins D., King G.: A method of automated nonparametric content analysis for social science, American J. Pol. Sci., 54(1), 229-247 (2010).
14. Iezzi, D. F.: Centrality measures for text clustering. Communications in Statistics-Theory and Methods, 41(16-17), 3179-3197 (2012).
15. Lebart, L., Salem, A.: Statistique Textuelle. Dunod, Paris (1994).
16. Pelagalli, F., Greco, F., De Santis, E.: Social emotional data analysis. The map of Europe. In: Petrucci A., Verde R. (eds) SIS 2017. Statistics and Data Science: new challenges, new generations. Proceedings of the Conference of the Italian Statistical Society, Florence 28-30 June 2017.: Firenze University Press (2017).
17. Savaresi, S.M., Boley, D.L.: A comparative analysis on the bisecting K-means and the PDDP clustering algorithms. Intelligent Data Analysis, 8(4), 345-362 (2004).
18. Vichi, M.: Double k-means clustering for simultaneous classification of objects and variables. Advances in classification and data analysis, 43-52 (2001).
19. Vickers, J. Consequences of Brexit for Competition Law and Policy. Oxford Review of Economic Policy, 33 (2017).