JADT' 18

PROCEEDINGS OF THE
14TH INTERNATIONAL CONFERENCE
ON STATISTICAL ANALYSIS OF TEXTUAL DATA

JADT' 18

PROCEEDINGS OF THE
14TH INTERNATIONAL CONFERENCE
ON STATISTICAL ANALYSIS OF TEXTUAL DATA

(Rome, 12-15 June 2018)

Vol. I

UniversItalia
2018

PROPRIETÀ LETTERARIA RISERVATA
Copyright 2018 - UniversItalia - Roma
ISBN 978-88-3293-137-2

Program Committee

Ramón Álvarez Esteban: Univ. of León, E
Valérie Beaudouin: Telecom ParisTech, F
Mónica Bécue: Poly. Univ. of Catalunya, E
Sergio Bolasco: Sapienza Univ. of Rome, I
Isabella Chiari: Sapienza Univ. of Rome, I
François Daoust, UQÀM, Montreal, CDN
Anne Dister, FUSL, Bruxelles / UCL, Louvain, B
Jules Duchastel: UQÀM, Montreal, CDN
Serge Fleury: Univ. Paris 3, F
Cédrick Fairon: UCL, Louvain, B
Luca Giuliano: Sapienza Univ. of Rome, I
Serge Heiden, ENS, Lyon, F
Domenica Fioredistella Iezzi, Univ. of Tor Vergata, I
Margareta Kastberg, Univ. of Franche Comté, F
Ludovic Lebart: CNRS / ENST, Paris, F
Jean-Marc Leblanc: Univ. of Créteil, F
Alain Lelu: Univ. of Franche Comté, F
Dominique Longrée, Univ. of Liège, B
Véronique Magri: Univ. of Nice Sophia-Antipolis, F
Pascal Marchand: Univ. of Toulouse, F
William Martinez: Univ. of Lisboa, P
Damon Mayaffre: CNRS, Nice, F
Sylvie Mellet: CNRS, Nice, F
Michelangelo Misuraca: Univ. of Calabria, I
Denis Monière: Univ. of Montréal, CDN
Bénédicte Pincemin: CNRS, Lyon, F
Céline Poudat: Univ. of Nice Sophia-Antipolis, F
Pierre Retinaud: Univ. of Tolouse, F
André Salem: Univ. Paris 3, F
Monique Slodzian: Inalco, F
Arjuna Tuzzi: Univ. of Padua, I
Mathieu Valette: Inalco, F

Organising Committee

Domenica Fioredistella Iezzi: Univ. of Tor Vergata, I
Sergio Bolasco: Sapienza Univ. of Rome, I
Livia Celardo: Sapienza Univ. of Rome, I
Isabella Chiari: Sapienza Univ. of Rome, I
Francesca della Ratta: ISTAT, I
Fiorenza Deriu: Sapienza Univ. of Rome, I
Francesca Dolcetti: Sapienza Univ. of Rome, I
Andrea Fronzetti Colladon: Univ. of Tor Vergata, I
Francesca Greco: Sapienza Univ. of Rome, I
Isabella Mingo: Sapienza Univ. of Rome, I
Michelangelo Misuraca: Univ. of Calabria, I
Arjuna Tuzzi: Univ. of Padua, I
Maurizio Vichi: Sapienza Univ. of Rome, I
Francesco Zarelli: ISTAT, I

Local Organisation

Francesco Alò, Giulia Giacco,
Paolo Meoli, Vittorio Palermo, Viola Talucci

Table of contents

Abstracts

Brexit and Twitter: The voice of people

Francesca Greco, Leonardo Alaimo, Livia Celardo
Sapienza University of Rome – francesca.greco@uniroma1.it; leonardo.alaimo@uniroma1.it; livia.celardo@uniroma1.it

Abstract 1

There is an increase in Euroscepticism among EU citizens nowadays, as shown by the development of the ultra-nationalist parties among the European states. Regarding the European Union membership, public opinion is divided in two. British referendum in 2016, where citizens chose to "exit" shaking the public opinion, and the following general election in June 2017, where the British Europeanist parties won the election according to the 1975 British referendum where 72% of citizens chose to "Remain", are clear examples of this fracture. There are still few studies concerning the investigation of Brexit discourses within the social media and most of them focus on the 2016 British referendum. Due to that, this exploratory research aims to identify how Brexit and the EU are nowadays discussed on Twitter, through a text mining approach. We collected all the tweets containing the terms "Brexit" and "EU", for a period of 10 days. Data collection has been performed with TwitteR package, resulting in a large corpus to which we applied multivariate techniques in order to identify the contents and the sentiments behind the shared comments.

Abstract 2

Negli ultimi anni c'è stato un aumento dell'euroscetticismo tra i cittadini dell'UE, come testimoniato dallo sviluppo di partiti ultra nazionalisti in diversi stati europei. Sul tema "Europa", l'opinione pubblica è divisa fra europeisti e euroscettici. Un chiaro esempio di questa divisione è dato dalle recenti vicende britanniche: infatti, nel referendum del 2016 i cittadini britannici hanno scelto di "uscire" dall'UE scuotendo l'opinione pubblica, mentre le successive elezioni politiche di giugno 2017 hanno visto l'affermazione dei principali partiti filo-europeisti. Vi sono ancora pochi studi in letteratura che indagano come nei social media venga affrontato il tema della Brexit in relazione all'UE, dato che la maggior parte di essi si focalizza su cause e potenziali effetti del voto di giugno 2016. In tal senso, questa ricerca esplorativa ha lo scopo di identificare in che modo Brexit e l'Unione Europea vengano discusse su Twitter in questo momento storico attraverso l'analisi automatica del testo. A questo scopo sono stati raccolti tutti i messaggi contenenti i termini "Brexit" e "EU" per 10 giorni attraverso l'utilizzo del pacchetto TwitteR, ottenendo un corpus di grandi dimensioni a cui sono state applicate delle tecniche multivariate, al fine di individuare i contenuti e i sentimenti relativi al tema in esame.

Keywords: Brexit, Twitter, Emotional text mining.

1. 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 states. Regarding to European Union membership, public opinion is divided between Eurosceptics and pro-Europeans, as shown by the 2016 British referendum ("Brexit"), where 52% of citizens chose to "Leave". For further evidence of this division, the following general election of June 2017 saw the affirmation of the main Europeanist parties (especially the Labour Party) and the results led to a *hung Parliament*. Brexit has shaken the European public opinion as it revealed the relevance of the anti-Europeanist trend. During the 60th Anniversary of the Treaties of Rome in 2017, millions of citizens expressed their support to the EU participating to Europeanist demonstrations in many European cities.

One useful starting point for explaining the results of Brexit is to focus on the electoral issue: the relationship between the UK and Europe. This has always been a central and rather controversial issue in the British public debate. The media, public opinion and the political class have always been deeply critical and sceptical about the European integration. This position influences citizens' attitudes towards the Union, which is not only considered distant and inadequate to resolve everyday issues (immigration, unemployment, and so on), but it is often perceived as their major cause, by limiting the political and economic power of United Kingdom. The electoral outcome created disbelief all over the world. *Britain is the home of the term Euroscepticism* (Spiering 2004, p.127). But, while it is clear that a large proportion of UK residents are sceptical about Europe, it is not clear enough that this position coincides with the wish to leave the EU. However, Euroscepticism should not be confused with this wish. Szczerbiak and Taggart (2008) have distinguished two different types of Euroscepticism: the *Hard Euroscepticism* that is *a principled opposition to the EU and European integration* and *Soft Euroscepticism* that *concerns on one (or a number) of policy areas lead to the expression of qualified opposition to the EU.*

Although there are several studies exploring British Euroscepticism, only few of them investigate the Brexit discourses within the social media. Due to that, we decided to perform a quantitative study, where the online discourses regarding Brexit and EU are analysed using two different approaches, Content Analysis and Emotional Text Mining. The aim is to explore not only the contents but also the sentiments shared by users on Twitter. For this paper, we used one of the most important and known blog tools, Twitter. It is an online platform for sharing real-time, character limited communication with people partaking of similar interests that, in 2017, counted over than 300 million users and an average of about 500 million of tweets sent per day.

2. Data collection and analysis

In order to explore the sentiments and the contents on Brexit and EU in twitter communications during ten days, we scraped all the messengers in English language produced from September 22nd to October 2nd, 2017, containing together the words *Brexit* and *EU*. The data extraction was carried out with the TwitteR package of R Statistics (Gentry, 2016). We started collecting 221,069 messengers, including 83% of retweets, from which two samples of tweets were extracted. The first we used for the sentiment analysis is composed of 99,812 messengers, where the retweets were limited to the threshold of 31, resulting in a large corpus of 1,601,985 of tokens; the second one we used for content analysis, where we excluded all the retweets, resulted in a large corpus of 37,318 tweets and 618,255 tokens. In order to check whether it was possible to statistically process data, two lexical indicators were calculated: the type-token ratio and the hapax percentage ($TTR_{corpus\ 1} = 0.02$; $Hapax_{corpus\ 1} = 39.8\%$; $TTR_{corpus\ 2} = 0.04$; $Hapax_{corpus\ 2} = 52.31\%$). According to the large size of the corpus, both lexical indicators highlighted its richness and indicated the possibility to proceed with the analysis.

2.1. Emotional text mining

We know that people sentiments depend not only on their rational thinking but also, and sometimes most of all, on the emotional and social way of functioning of people's mind. If the conscious process set the manifest content of the narration, that is what is narrated, the unconscious process can be inferred through how it is narrated, that is, the words chosen to narrate and their association within the text. 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 (Greco, 2016). To this aim we perform a multivariate analysis based on a bisecting *k*-means algorithm to classify the text (Savaresi et Boley, 2004), and a correspondence analysis to detect the latent dimensions setting the cluster per keywords matrix (Lebart et Salem, 1994) by means of T-Lab software. The interpretation of the cluster analysis results allows to identify the elements characterizing the emotional representation of Brexit, while the results of correspondence analysis reflect its emotional symbolization. By the clusters interpretation, we classify the emotional representations in positive, neutral and negative sentiments, determining the percentage of messages for each sentiment modality. To this aim, first corpus was cleaned and pre-processed with the software T-Lab (T-Lab Plus version, 2017) and keywords selected. In particular, we used lemmas as keywords instead of types, filtering out the lemma *Brexit* and *EU* and those of the low rank of frequency (Greco, 2016). Then, on the tweets per keywords matrix, we performed a cluster analysis with a bisecting *k*-means algorithm limited to twenty partitions, excluding all the tweets that do not have at least two keywords co-occurrence. The percentage of explained variance (η) was used to evaluate and choose the optimal partition. To finalize the analysis, a correspondence analysis on the keywords per clusters matrix was made in order to explore the relationship between clusters and to identify the emotional categories setting Brexit representations.

2.2. Content analysis

Content analysis is a technique used to investigate the content of a text; in text mining, many methods exist to analyse it automatically. One of these is Text Clustering, where the corpus is splits in different subgroups based on words/documents similarities (Iezzi, 2012). In this paper, a text co-clustering approach (Celardo et al., 2016) 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 lemmatizing the texts, removing stop words and terms with frequency lower than 10. The weighted term-document matrix was then co-clustered through the double *k*-means algorithm (Vichi, 2001); the number of clusters for both rows and columns was fixed using the Calinski-Harabasz index.

3. Emotional text mining main results and discussion

The results of the cluster analysis for ETM show that the 655 keywords selected allow the classification of 88,6% of the tweets. The percentage of explained variance was calculated on partitions from 3 to 19, and it shows that the optimal solution is six clusters (η= 0.057). The correspondence analysis detected six latent dimensions. In table 1, we can appreciate the emotional map of Brexit and the EU emerging from the English tweets. It shows how the clusters are placed in the factorial space produced by five factors. The first factor represents the political and economic domain where Brexit seems to have its main impact; the second factor reproduces the possible solutions of Brexit: a separation or a new agreement; the third factor represents the national or European level of reaction to Brexit; the fourth factor is the blame, distinguishing the blame of politicians from the one of the willingness to be independent; and the fifth factor is the political leadership, differing old and new policies.

Table 1 – Correspondence analysis results (the percentage of explained inertia is reported between brackets beside each factor).

Factor 1 (27.5%)		Factor 2 (24.3%)		Factor 3 (19.8%)		Factor 4 (15.6%)		Factor 5 (12.9%)	
NP	PP	NP	PP	NP	PP	NP	PP	NP	PP
future	negotiation	bill	try	Macron	Florence	blame	referendum	leader	people
war	Briton	Barnier	pro	European	Delay	march	Johnson	remain	Tory
support	chance	Brussel	deliver	good	withdrawal	stay	Verhofstadt	walk	hard
remainer	zero	divorce	brexiteers	miracle	blast	speech	independent	urge	voter
concern	better off	progress	help	market	states	conservative	destroy	May T.	party
save	laureate	negotiator	debate	union	finger	anti	migrant	hope	happen
proposal	leaving	pay	allow	single	finish	Blair	vow	call	Catalonia
fight	economist	chief	event	Merkel	row	reverse	adopt	time	die
0.07-0.02 ac	6.49-4.40 ac	4.72-1.50 ac	0.35-0.12 ac	1.5-1.61 ac	0.35-0.05 ac	0.55-0.29 ac	5.22-0.94 ac	3.65-1.24 ac	10.28-1.49 ac

NP =negative pole; PP = positive pole; ac = absolute contribution (10^{-3})

The six clusters are of different sizes and reflect the representations of Brexit (table 2), that correspond to three different sentiments: positive, negative for domestic reasons, and negative for foreign ones (table 1). The first cluster represents the choice to leave EU as a good option, underlining the need to proceed; the second cluster focuses on the EU political reaction fixing divorce conditions, perceiving EU political representatives as unfavourable and therefore threatening; the third cluster represents Britons' hope to improve their economic condition leaving EU as naive; the fourth cluster represents the old British political leadership as incompetent, being unable to protect and adequately inform Britons in order to support them in remaining in the EU; the fifth cluster reflects the negotiation of the divorce conditions, perceiving the negotiation as unfair and the costs of leaving EU as a punishment; and the sixth cluster represents Brexit as a Britons informed choice, highlighting that its consequences belong to the policy domain who should respect the citizens' choice.

Table 2 – Clusters (the percentage of context units classified in the cluster is reported between brackets).

Cluster 1 (10.0% CU) *Good Choice*	Cluster 2 (14.9% CU) *EU Reaction*	Cluster 3 (20.9% CU) *Uncertain Future*	Cluster 4 (13.4% CU) *British Leadership*	Cluster 5 (19.2% CU) *Divorce Conditions*	Cluster 6 (21.7% CU) *Informed Choice*
leader	Macron	negotiation	people	bill	referendum
remain	European	leaving	Tory	Barnier	Corbyn
time	good	Briton	hard	brussel	Johnson
Theresa May	market	chance	voter	progress	think
urge	warn	zero	party	divorce	independent
call	single	better off	happen	negotiator	Boris
walk	business	Nobel	Florence	pay	Verhofstadt
UKIP	minister	economist	stay	chief	Florence
government	Europe	laureate	Catalonia	demand	destroy
hope	move	tell	believe	national	try
look	Merkel	rating	Spain	Davis	policy
mean	miracle	law	Rees Mogg	offer	issue
From 1611 to 620 CU	From 2004 to 951	From 1844 to 668 CU	From 2506 to 461 CU	From 2705 to 843 CU	From 2098 to 512 CU

CU = context units classified in the cluster.

By the clusters interpretation, we detected six different representations of Brexit that correspond to three different sentiments (table 1). We have considered as positive (21,7%) the representation of Brexit as a Good Choice or an Informed Choice, and negatives all the other representations (78,3%). Among the negative clusters, we distinguished negativity according to the origin of the problem: Uncertain Future and British Leadership are negative for domestic reasons (34,2%), that is, the lack of UK political leadership's competences; and EU Reaction and Divorce Condition are negatives due to foreign factors (34,1%) as the EU after Brexit seems to be perceived as vindictive and, therefore, threatening.

4. Content analysis main results and discussion

The pre-processing phase, implemented on the second corpus, allowed us to identify a set of 1.957 keywords, representing the 97% of the tweets; so, on the term-document matrix of dimension (1.957 × 36.383) we calculated the Calinski-Harabasz Index in order to define the number of clusters for rows and columns. After calculating the index values for partitions from 2 to 10 for each dimension, the Calinski-Harabasz Index suggested to classify the words in three groups and the tweets in five groups. In table 3, the centroids of the clusters are exposed.

Table 3 – Centroids matrix (Terms × Documents).

	Cluster 1 (55%)	*Cluster 2* (20%)	*Cluster 3* (12%)	*Cluster 4* (11%)	*Cluster 5* (2%)
Cluster 1	**0,005**	0,003	0,004	0,000	0,000
Cluster 2	0,002	**0,063**	0,003	**0,149**	0,012
Cluster 3	-0,002	0,000	**0,090**	-0,003	**0,309**

Table 4 – Words groups (first 10 words listed below by frequency of occurrence).

Cluster 1 *Negotiation*	*Cluster 2* *Economic Transformation*	*Cluster 3* *British Identity*
stay	leave	home
Junker	move	sound
ambassador	transition	cake
cry	late	plan
track	deal	datum
surge	trade	live
peer	retain	finish
shape	post	Id
turmoil	Macron	idea
survive	urge	national

As shown in the table 3, the algorithm has identified five blocks of specificities; in fact; the first cluster of words is connected to the first group of tweets; the second is specific of the second and the fourth cluster of tweets and the third is related to the third and the fifth group of tweets. In table 4, the groups of words are presented.

The first group of words is related to the need of defining new rules and settlements within the negotiation and it represents more than half of the tweets; it has no strong specificities related to the texts, but in comparison to all the documents clusters, it seems to be more connected to those words. On the other hand, for the other two groups of words, there are more effective specificities; the second cluster of words is about the definition of new economic agreements, and it is connected to the 31% of the tweets, while the third one, related to the requirement in specifying a new identity after Brexit, is representative of the 14% of the corpus documents.

5. Conclusions

The results of the two analyses showed a strong relationship between the terms "Brexit" and "EU", not only in terms of sentiment, but also in terms of contents. According to the literature, the sentiment analysis revealed the presence of both positive and negative opinions in respect to the exit of United Kingdom from the EU. On the other hand, starting from the analysis of the contents we found that the Twitter communications on Brexit focuses primarily on the concept of *negotiation*. The remaining part of the messages take into account both the Brexit economic features and the need of the national identity redefinition. To conclude, the results of the two analyses revealed that Brexit is a theme with a strong emotional charge, mostly negative. British people seem to focus their attention basically toward three issues: the new asset, the economic consequences, and the national identity. These subjects are treated positively and negatively from the users, probably because of the lack of cohesion within the country.

References

Celardo L., Iezzi D.F. and Vichi M. (2016). Multi-mode partitioning for text clustering to reduce dimensionality and noises. In *Proceedings of the 13th International Conference on Statistical Analysis of Textual Data*.

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*. Franco Angeli.

Hobolt S. (2016). The Brexit vote: a divided nation, a divided continent. *Journal of European public policy*, 23(9): 1259–1277.

Iezzi D. F. (2012). Centrality measures for text clustering. *Communications in Statistics-Theory and Methods*, *41*(16-17), 3179-3197.

Lebart L. and Salem A. (1994). *Statistique Textuelle*. Dunod

Savaresi S. M. and Boley D. L. (2004). A comparative analysis on the bisecting K-means and the PDDP clustering algorithms. *Intelligent Data Analysis*, 8(4): 345-362.

Spiering M. (2004). British Euroscepticism. In Harmsen R. and Spiering M., editors, *Euroscepticism: Party Politics, National Identity and European Integration*. Editions Rodopi B.V.

Szczerbiak A. and Taggart P. (2008). *Opposing Europe? The Comparative Party Politics of Euroscepticism. Volume 1: Case Studies and Country Surveys*. Oxford University Press.

Vichi M. (2001). Double k-means clustering for simultaneous classification of objects and variables. *Advances in classification and data analysis*, 43-52.