Social Indicators Research (2021) 156:669–688
https://doi.org/10.1007/s11205-020-02442-4

ORIGINAL RESEARCH

Emerging Topics in Brexit Debate on Twitter Around the Deadlines

A Probabilistic Topic Modelling Approach

Emiliano del Gobbo[1] · Sara Fontanella[2,3] · Annalina Sarra[4] · Lara Fontanella[4]

Accepted: 12 July 2020 / Published online: 22 July 2020

Abstract
The present study is focused on the online debate relating to the Brexit process, three years and half since the historical referendum that has sanctioned the divide of the United Kingdom from the European Union. In our analysis we consider a corpus of approximately 33 million Brexit related tweets, shared on Twitter for 58 weeks, spanning from 31 December 2019 to 9 February 2020. Due to its great accessibility to data, Twitter constitutes a convenient data source to monitor and evaluate a wide variety of topics. In addition, Twitter's marked orientation towards news and the dissemination of information makes this microblogging network more connected to politics compared to other platforms. Through static and dynamic topic modelling techniques, we were able to identify the topics that have attracted the most attention from Twitters users and to characterise their temporal evolution. The topics retrieved by the static model highlight the major events of the Brexit process while the dynamic analysis recovered the persistent themes of discussion and debate over the entire period.

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s11205-020-02442-4) contains supplementary material, which is available to authorized users.

✉ Sara Fontanella
s.fontanella@imperial.ac.uk

Emiliano del Gobbo
emiliano.delgobbo@unich.it

Annalina Sarra
asarra@unich.it

Lara Fontanella
lara.fontanella@unich.it

[1] Department of Neuroscience and Imaging, University "G. d'Annunzio" of Chieti-Pescara, Chieti, Italy

[2] National Heart and Lung Institute, Imperial College London, London, UK

[3] Department of Economics and Statistics, University of Torino, Turin, Italy

[4] Department of Legal and Social Sciences, University "G. d'Annunzio" of Chieti-Pescara, Chieti, Italy

Keywords Social media · Twitter data · Brexit · Topic models · Latent Dirichlet Allocation

1 Introduction

The development of Internet and mobile technologies has dramatically changed the way individuals communicate and acquire information. Consequently, social media have gained a central role for research in the field of computational social science that investigates questions using quantitative techniques and big data (Lazer et al. 2009; Cioffi-Revilla 2010).

Social media can be defined as any web-based and mobile-based Internet applications that provide real time communication and information, allowing the creation, access and exchange of user-generated content that is ubiquitously accessible. A taxonomy of social media can be found in Kaplan and Haenlein (2010) and a survey of techniques, tools and platforms exploited in social media analytics is provided by Batrinca and Treleaven (2015). The application of quantitative techniques to social media data enables a deeper comprehension of social, political, and economic phenomena. In particular, the impact of digital platforms on the production, distribution, and consumption of political information has been extensively analysed in literature. For instance, Ceron et al. (2017) provide a review of different techniques using social media to nowcast and forecast elections. A comprehensive overview of the main transformations induced by social media on the information process is presented in Casero-Ripollés (2018), while Jungherr and Theocharis (2017) discuss the opportunities along with the pitfalls of the continuously growing use of digital data in political science. A survey on political event researches on social media can be found in Korakakis et al. (2017). The authors focus mainly on Twitter and identify three main areas of interest: prediction of electoral results, sentiment analysis in political topics and opinion polls, and, finally, social analysis of human behaviour related to the interaction between politicians and citizens.

Social media have also acquired a central role during the most significant political event of the last fifty years in the UK as testified by the huge volume of information and discussion on Brexit on online platforms. The Brexit process, in spite of its recent date, has been the subject of several social media studies which concentrate mainly on the prediction of the results of the referendum and on the influence of social media in shaping the vote. Focusing on the information shared on Twitter, Khatua and Khatua (2016) investigated the 2016 Brexit referendum analysing an exhaustive set of hashtags, selected by considering the lack of ambiguity of their political leaning. Hänska-Ahy and Bauchowitz (2017) analysed 7.5 million tweets and found how the predominance of Euroscepticism on social media mirrored its dominance in the press. Howard and Kollanyi (2016) carried out a preliminary study on the use of political bots during the Brexit campaign. Grčar et al. (2017) addressed the stance of the Twitter users in relation to the referendum outcome polls and identified the influential users on both sides of the Brexit debate. An opinion analysis on the British EU membership referendum within Twitter is reported in the study of Llewellyn and Cram (2016). By adopting three different search strategies, the authors collected tweets from specific groups to explore how topics and language differ among groups and how those groups influence each other. Lansdall-Welfare et al. (2016), using simultaneous multiple change-point analysis, tried to capture changes in public mood in the days before and after the Brexit referendum. Moving along these lines of research, Hürlimann et al. (2016) present a dataset of sentiment-annotated tweets targeting the historical event of Brexit

to categorise the social and discourse dynamics behind this political event as well as the strength of the sentiment.

In this paper, we explore Twitter conversations collected in proximity of the UK's planned withdrawal from the EU. Specifically, we queried for the tweets containing the keyword Brexit and posted between the end of December 2018 and the first week of February 2020. By exploiting approximately 30 million Brexit related tweets, our overall goal is to explore the prominent themes discussed, and in particular to identify the topics that have attracted the most attention from users and how they evolve over time. We address these research questions through topic modelling techniques (Blei 2012). Probabilistic topic modelling consists of a collection of methods which specify a probabilistic generative model for the documents with the purpose of discovering and annotating large archives of textual documents with thematic information. In our study, we implement both standard (Blei et al. 2003) and dynamic (Blei and Lafferty 2006) Latent Dirichlet Allocation (LDA) models. The general idea beyond LDA is that documents with the same topic will use similar words and the key assumption is that documents are mixtures of topics, so that of central interest is how to discover a topic distribution over each document and a word distribution over each topic. In addition, the LDA dynamic version allows to analyse topic distributions over time and to gain insights on their changes and evolutions.

The remainder of this paper is organized as follows. Section 2 introduces probabilistic topic models, focusing in particular on LDA and its dynamic version, which relaxes the assumption that all documents are generated in the same time step. Section 3 describe the large collection of tweets related to Brexit, extracted from 31 December 2018 and 9 February 2020, and provides results of the exploratory analysis of the Twitter activity to uncover temporal patterns. Also results of hashtag analysis are presented. Section 4 discusses the main findings of the topic analyses. Finally, Sect. 5 concludes the paper and considers possible future works.

2 Probabilistic Topic Models

Topic modelling provides a powerful method for projecting text documents into topic space and it has been widely applied in many fields, ranging from information retrieval (Boyd-Graber et al. 2017; Wei and Croft 2006), to information visualization (Wang et al. 2016), and to recommendation systems (Wang and Blei 2011).

The application of automatic topic mining techniques to large electronic document archives, obtained from social media channels, constitutes an important tool in computational social science aiming at the detection of hidden topics in online discussions. Among several application fields, researchers have introduced the topic modelling approach also into political science studies focusing, in particular, on content shared on Twitter (see, for example, Karami et al. 2018; Fang et al. 2019 and references herein).

Depending on the problem at hand, there are many approaches and techniques one can use to extract and manage large volume of data. The two foundational probabilistic topic models are the probabilistic latent semantic analysis (pLSA, Hofmann 1999) and the Latent Dirichlet Allocation (Blei et al. 2003). The pLSA is a probabilistic variant of the Latent Semantic Analysis introduced by Deerwester et al. (1990) to capture the meaning or semantic information embedded in large textual corpora without human supervision. In the pLSA approach, each word in a document is modelled as a sample from a mixture model, where the mixture components are multinomial random variables that can

be viewed as representations of "topics". The pLSA model allows multiple topics in each document, and the possible topic proportions are learned from the document collection. Blei et al. (2003) introduced the LDA model which presents a higher modelling flexibility over pLSA by assuming fully complete probabilistic generative model where each document is represented as a random mixture over latent topics and each topic is characterized by a distribution over words. According to the generative model, each document in the corpus is generated in a two stage process (Blei 2012; Steyvers and Griffiths 2006). First, a distribution over topics is randomly chosen; then each word in the document is generated by first sampling a topic from this topic distribution, and choosing a word from the topic-word distribution over the vocabulary. A detailed derivation of the LDA can be found in Blei et al. (2003). Standard statistical techniques can be used to invert the generative process in order to infer the set of topics that were responsible for generating the collection of observed documents and a wide variety of approximate inference algorithms, such as sampling-based algorithms (see, for example Steyvers and Griffiths 2006, for a detailed derivation of Gibbs-sampler for LDA) and variational algorithms (Blei et al. 2003) can be considered.

The statistical assumptions behind standard LDA include that both words and documents are exchangeable (i.e. the order does not matter) and all documents are generated in the same time steps. This assumption is relaxed in dynamic topic models (Blei and Lafferty 2006; Wang and McCallum 2006; Wang et al. 2008; Iwata et al. 2010) which respect the ordering of the documents and give a richer posterior topical structure than LDA. In this case, rather than a single distribution over words, a topic is a sequence of distributions over word and it is possible to find an underlying theme of the collection and track how it has changed over time. In this work, we refer to the dynamic version of topic models (DTM) proposed by Blei and Lafferty (2006), which specifies a statistical model of topic evolution. In this approach, documents are divided into a set of sequential non-overlapping time slices and the basic assumption is that the topics associated with the temporal window t evolve from topics associated with slice $t-1$. Therefore, the documents of each slice are modelled through a K-component topic model, and both the natural parameters of the underlying topic distribution, and the natural parameters of the distributions for the document-specific topic proportions, associated with slice t, are chained in a state space model. Blei and Lafferty (2006) propose to use Variational Kalman Filtering or Variational Wavelet Regression to estimate the model parameters.

3 Data Collection and Twitter Traffic Temporal Evolution

For our study we use a Twitter dataset, collected for 58 weeks, spanning from 31/12/2018 to 09/02/2020. The data were extracted by using Twitter's Streaming Application Programming Interface (API) and we searched for tweets, written in English, containing the term Brexit. Our sample includes 135,607,216 tweets, of which 102,176,840 are retweets.

There are 6,811,652 tweets with at least one retweet. Of these, approximately 90% were retweeted less than 15 times. In our research, we focuses on pure tweets (i.e., tweets that are no retweets).

We developed a Python script to perform the screening and cleaning process (tokenization; lowercase conversion; special characters, URL, mentions and stop-words removal) of text documents in order to extract the relevant content and remove any unwanted nuisance terms. The nltk.word_tokenize function in the NLTK Python package (Bird et al. 2009) has

been employed for tokenization. Some tokens have been merged according to a multiword list, generated ad hoc for this dataset. The dataset contains 2.148.421 unique tokens; about 65% of these distinct tokens are hapax legomena (i.e. words that occur only once within the entire corpus) and around 90% are used less than 10 times in the entire corpus. For the analyses, the tokens with less than 4000 occurrences in the corpus have been pruned and the stop words removed. The stop words list contains a set of words that have no inherent useful information, either because they do not have any meaning (prepositions, coordinating conjunctions, determines, etc..) or because they are too frequent. Stop words create problems in identification of key concepts from textual sources when they are not removed due to their overwhelming presence. The stop word list for our analysis contains also the word Brexit, used as search term and thus presents in all the tweets. The final vocabulary consists of 6181 tokens and the most frequent words are visualised through a wordcloud in Fig. 1, where the word sizing is proportional to how frequently a term is used in the corpus. The words that appear more frequently are: "eu","party", "uk","vote", "people", "nodeal", "out", "deal", "labour", "may", "leave", "boris", "voted", "johnson", "country", "tory", "remain". These terms recorded more than one million occurrences. Political words such as "government", "parliament", party denominations and politician names are also evident in the cloud, highlighting that the actions taken by Government, Parliament and political actors are at the center of the discussion, whereas terms related to economy seem to be less recurrent.

Twitter activities can be tracked through the inspection of users posting activity (total number of tweets) and their ability to engage their followers for support (i.e., by retweeting their posts). The daily volume of tweets is represented in Fig. 2, along with some of the most relevant dates on the Brexit timeline. In Table 1, we indicate some specific dates in which we observe a relevant variation in Twitter activity, broken down by total number of tweets, pure tweets and retweets. We report also the ratio between retweets and pure tweets.

The peak dates are linked to important events in the Brexit process. The highest peaks in the Twitter traffic are clearly associated with 13 December 2019, the day of the General election results (when the Conservative party secured a huge majority) and with 27 May 2019, when the results of the UK's European Parliament elections were

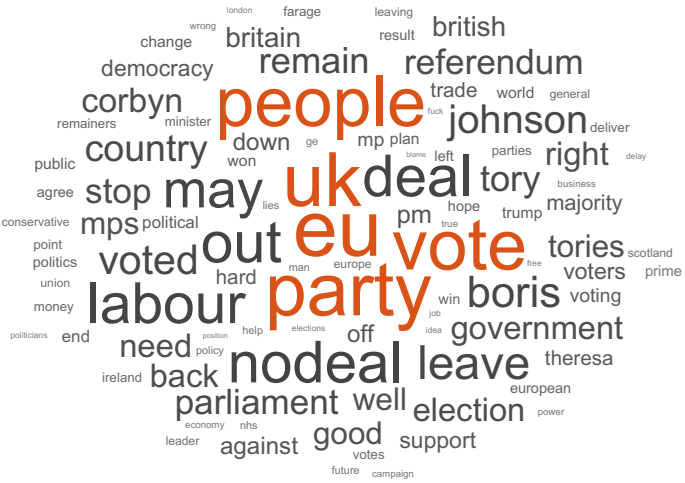

Fig. 1 Wordcloud of the most frequent words

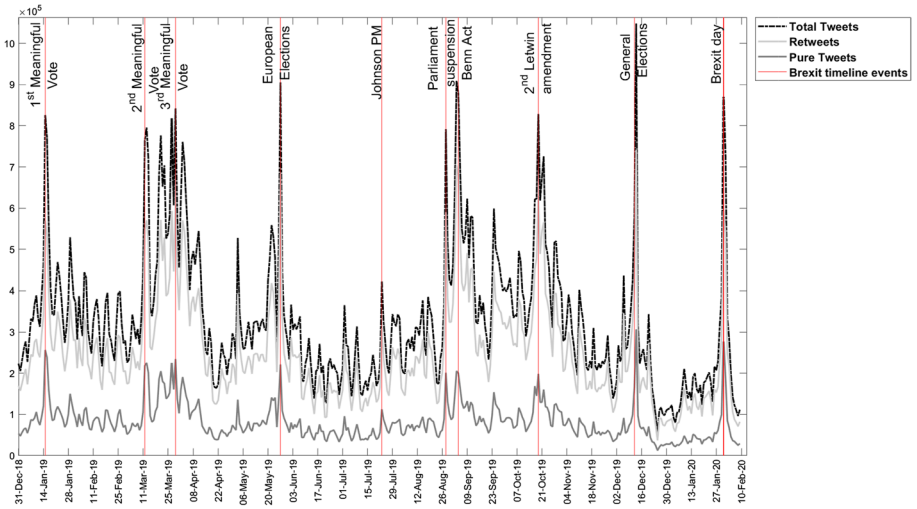

Fig. 2 Brexit-related tweet distributions over days, from 31/12/2018 to 09/02/2020

Table 1 Peak dates in the temporal distribution of Brexit-related tweets

Date	Total tweets	Pure tweets	Retweets	Retweets/ pure tweets	Brexit timeline
13-Dec-19	1,048,141	306,539	741,602	2.42	General election results
03-Sep-19	908,868	205,302	703,566	3.43	Emergency debate motion on the Benn Act
27-May-19	905,231	221,442	683,789	3.09	European election results
04-Sep-19	880,212	200,266	679,946	3.40	MPs vote on the Benn Act
31-Jan-20	870,169	277,154	593,015	2.14	Brexit day
29-Mar-19	841,933	234,523	607,410	2.59	Third meaningful vote
19-Oct-19	828,414	198,644	629,770	3.17	Second Letwin amendment
15-Jan-19	825,812	256,526	569,286	2.22	First meaningful vote
27-Mar-19	818,065	225,097	592,968	2.63	First round of indicative votes
13-Mar-19	795,517	225,269	570,248	2.53	MPs' vote to reject no-deal Brexit
28-Aug-19	791,675	201,252	590,423	2.93	Parliament suspension
21-Mar-19	776,371	205,129	571,242	2.78	Article 50 extension to 30 June
16-Jan-19	771,807	238,839	532,968	2.23	First meaningful vote
12-Mar-19	764,284	218,019	546,265	2.51	Second meaningful vote
02-Apr-19	760,882	190,827	570,055	2.99	Second round of indicative votes

announced: the Brexit Party was the clear winner, the pro-EU Liberal Democrats came second, and the Conservative and Labour Parties suffered heavy losses. Starting from the beginning of 2019, important peaks coincide with the dates of the Brexit meaningful votes and indicative votes. The meaningful votes on the Withdrawal Agreement that the Conservative government had reached with the European Union took place in the House of Commons between January and March 2019: the bill was three times decisively

defeated, following a major revolt amongst Conservative backbenchers. The two rounds of indicative votes, on a series of non-binding resolutions on alternative Brexit options, were all rejected. A sudden increase in Brexit tweets is evident on 28 August, the date when the Queen granted Prime Minister Boris Johnson's request to suspend Parliament from 10 September until 14 October. Following this, a massive escalation in the number of Twitter users joining in the Brexit debate is apparent on 3 and 4 September. This coincides with the emergency debate on the so-called Benn Bill, or Benn Act, that was aimed at ruling out a unilateral no-deal Brexit by forcing the Government either to reach an Agreement or to get parliamentary approval for a no-deal Brexit, or else (if neither condition was fulfilled by 19 October) extend the deadline to 31 January 2020. A high Twitter traffic was recorded on 19 October 2019, when MPs, instead of backing Johnson's agreement in a meaningful vote, passed the second Letwin amendment, which forced the Government to request from the EU a delay to Brexit until 31 January 2020. An inspection of Fig. 2 reveals a final sharp increase in the number of tweets (more than 870,000 in total) on 31 January 2020, Brexit day.

An in-depth analysis can be obtained by tracking the top hashtags, which are a well-established means of categorising tweets by content and are included for emphasis. The number of pure tweets containing at least one hashtag is 7,994,833. Apart from #Brexit, which appears in 5,135,423 pure tweets, according to our data the hashtags contained in more than 100,000 pure tweets are: #peoplesvote, #eu, #uk, #stopbrexit, #remain, #brexitshambles, #borisjohnson, #revokearticle50, #fbpe, #labour, and #nodeal. In the contest between the two opposing positions of leavers and remainers, Fig. 3 shows that #remain (153,945 pure tweets) greatly exceeds #leave (79,109), with the biggest gap evident in May 2019. There was a reduction in the occurrences of both hashtags over the following months and the difference between their frequency fell sharply, starting from the second half of December 2019 when some key events related to Brexit occurred. In particular, on 19 December 2019 the New Withdrawal Agreement Bill was introduced

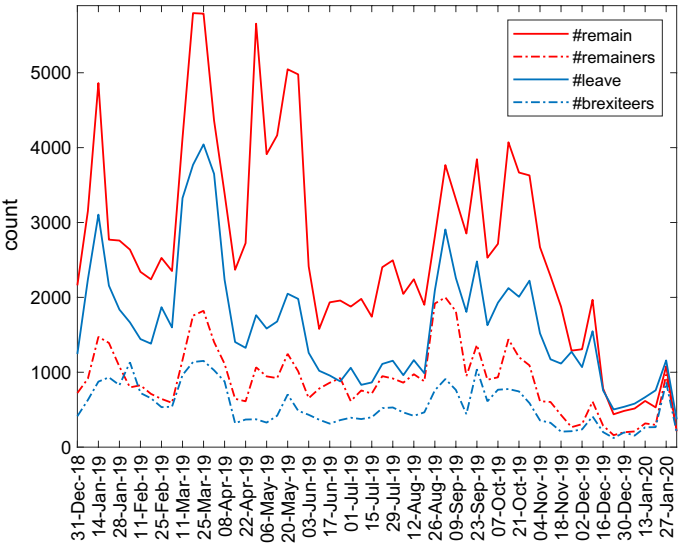

Fig. 3 Temporal distribution of "Leave" and "Remain" hashtags over weeks, from 31/12/2018 to 09/02/2020

Table 2 Occurences of the most popular hashtags for the period ranging from 31/12/2018 to 09/02/2020

Anti Brexit		Pro Brexit		Brexit criticism	
Hashtag	Pure tweets	Hashtag	Pure tweets	Hashtag	Pure tweets
peoplesvote	346,597	leavemeansleave	54,148	brexitshambles	144,669
stopbrexit	216,563	brexitbetrayal	45,925	brexitchaos	47,130
revokearticle50	117,622	britishindependence	34,858	brexshit	15,506
fbpe	116,048	getbrexitdone	30,512	bollockstobrexit	14,068
revokea50	98,422	standup4brexit	26,585	brexitcrisis	12,689
finalsay	60,445	letsgowto	11,481	brexitmayhem	10,174
indyref2	52,275	gowto	10,801		
stopbrexitsavebritain	37,039	nodealnoproblem	10,681		
peoplesvotemarch	25,680				
putittothepeople	19,895				
revokea50now	13,606				
peoplesvotenow	11,993				

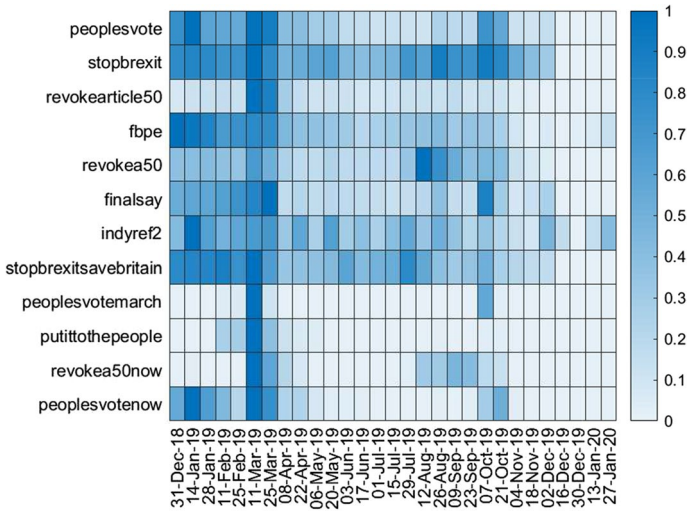

Fig. 4 Heatmap of anti-Brexit hashtags. The color represents biweekly occurrences from 31/12/2018 to 09/02/2020, normalised with respect to the total occurrences of the each hashtag. (Color figure online)

in Parliament. The hashtags #remainers and #brexiteers show analogous popularity and similar temporal patterns.

Attitudes critical of the Brexit process are confirmed by several hashtags which reflect the viewpoint of Twitter users: as shown in Table 2, among the most widely-used hashtags are #peoplesvote, #revokearticle50, #fbpe (i.e. #FollowBackProEU), which have a significant larger promotion than #leavemeansleave, #brexitbetrayal and #britishindependence. Figures 4 and 5 reveal how most of the pro and anti Brexit hashtags had a wider diffusion during the first months of 2019; some of them (e.g. #revokearticle50, #peoplesvotemarch,

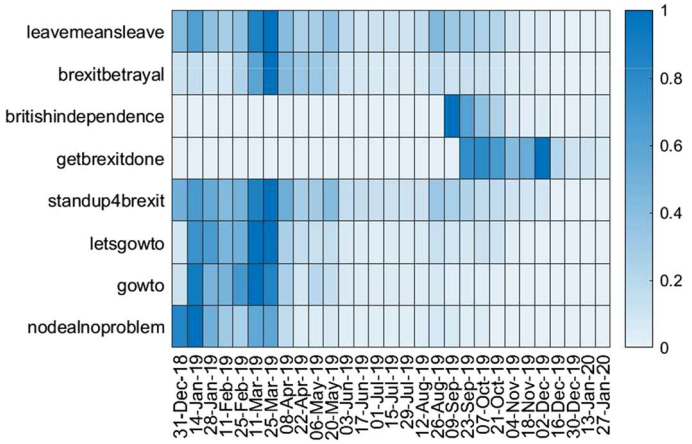

Fig. 5 Heatmap of pro-Brexit hashtags. The color represents biweekly occurrences from 31/12/2018 to 09/02/2020, normalised with respect to the total occurrences of the each hashtag. (Color figure online)

#peoplesvotenow, #putittothepeople, #leavemeansleave, #brexitbetrayal) reached their highest levels of popularity during March 2019. Hashtags calling for a stop to the Brexit process (#stopbrexit, #stopbrexitsavebritain, #indyref2), show a longer persistence, beginning to tail off in the second half of October 2019. September 2019 saw the rise of #britishindependence, followed by #getbrexitdone, which persisted until December 2019.

A number of hashtags indicate the difficulties encountered during the Brexit process (e.g. #brexitshambles, #brexitchaos, #brexitmayhem) and their temporal dynamic shows that they were mostly in use just before the first Brexit withdrawal deadline (29 March 2019) and after the suspension of Parliament on 10 September 2019 (Fig. 6).

Details of the most widely used hashtags relating to politicians and political parties, and of their temporal dynamic, can be found in the online Supplementary Material.

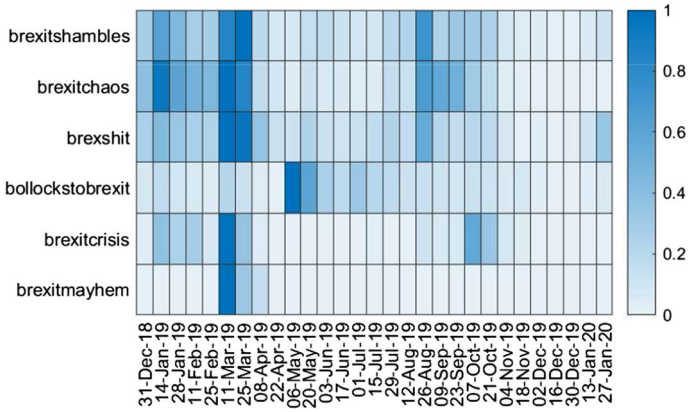

Fig. 6 Heatmap of critical hashtags of the Brexit process. The color represents biweekly occurrences from 31/12/2018 to 09/02/2020, normalised with respect to the total occurrences of the each hashtag. (Color figure online)

4 Results

In order to investigate recurrent themes emerging from the Brexit debate, probabilistic topic models were used which allow the extraction of coherent topics hidden within a huge volume of text. For this purpose both the standard LDA and its dynamic version were applied; the results are provided in the following sections.

4.1 Discovering Topics Associated with Brexit Tweets Through LDA

To perform LDA, we consider a corpus where each document ($N = 9744$) consists of the bunch of tweets posted in a hour time span and we set the input parameter related to the number of desired topics (K), in turn, equal to 10, 15, 20, 25, 30. The analysis was performed through the fitlda Matlab routine available in the Text Analytics Toolbox (MATLAB 2018). In particular, we used collapsed variational Bayesian algorithm (Teh et al. 2006).

To select the number of topics we considered UMASS coherence measure (Mimno et al. 2011) that, for each topic k, is defined as

$$C(k, W^{(k)}) = \sum_{m=2}^{M} \sum_{l=1}^{m-1} log \frac{D\left(w_m^{(k)}\right), D\left(w_l^{(k)}\right) + 1}{D\left(w_l^{(k)}\right)}$$

where $W^{(k)} = (w_1^{(k)}, \ldots, w_M^{(k)})$ is the list of the M-top words for topic k, $D(w)$ is the document frequency of word type w (i.e., the number of documents with at least one token of type w) and $D(w, w')$ is the co-document frequency of word types w and w'. The coherence measure, computed considering 25-top words for each topic, suggested a twenty-topic solution, as shown in Table 3.

The sensitivity analysis performed varying the number of top words in the range 10–30 confirmed the choice of 20 topics, as shown in the Supplementary Material. From an interpretative point of view, the model estimated with $K = 20$, guarantees the right trade-off between having enough words to disclose relevant information without making the topics cluttered.

In LDA, the topics are assumed to be latent variables, which need to be intuitively interpreted and, as point out by Steyvers and Griffiths (2006), the topics are individually interpretable, providing a probability distribution over words that picks out coherent clusters of correlated terms. All the estimated topics, along with the top ten relevant terms, are listed in Table 4. Words are ordered according to their relevance, obtained by normalising the posterior word probabilities per topic by the geometric mean of the posterior probabilities for the word across all topics.

In Fig. 7, for each topic, we represent the document-topic probabilities. As previously stated, each document consists of tweets posted over a time span of an hour, and, therefore, this representation allows us to follow topics' temporal evolution. A more detailed representation of each topic, coupling wordclouds with temporal dynamic, is provided in the Supplementary Material.

Table 3 Coherence measures for the choice of the number of topics in the LDA

Top words	Number of topics				
	10	15	20	25	30
25	− 269.33	− 253.54	− 247.3	− 279.37	− 268.27

Table 4 LDA analysis: top ten terms within the 20 topics sorted according to their relevance scores

TOPIC 1		TOPIC 2		TOPIC 3		TOPIC 4		TOPIC 5	
Word	Score	Word	Score	Word	Score	Word	Score	Word	Score
nodeal	0.054	nodeal	0.338	honda	0.076	mps	0.143	extension	0.228
ferry	0.021	backstop	0.053	nissan	0.055	vote	0.138	delay	0.206
grayling	0.016	ireland	0.023	japan	0.031	nodeal	0.138	march	0.106
dyson	0.012	irish	0.021	diesel	0.025	deal	0.102	petition	0.071
wto	0.009	food	0.020	ford	0.020	amendment	0.084	article50	0.036
singapore	0.008	shortages	0.015	plant	0.017	defeat	0.082	revoke	0.033
warns	0.007	border	0.012	production	0.017	parliament	0.078	people	0.021
ferries	0.006	varadkar	0.010	sunderland	0.017	bill	0.058	october	0.021
contract	0.006	recession	0.009	jobs	0.016	reject	0.031	halloween	0.020
ships	0.005	trade	0.008	industry	0.016	meaningful-vote	0.012	protest	0.017
TOPIC 6		TOPIC 7		TOPIC 8		TOPIC 9		TOPIC 10	
Word	Score	Word	Score	Word	Score	Word	Score	Word	Score
bercow	0.065	party	0.120	labour	0.166	hunt	0.067	parliament	0.092
april	0.037	elections	0.083	libdems	0.079	farage	0.045	johnson	0.072
speaker	0.037	vote	0.069	elections	0.067	peterbor-ough	0.039	queen	0.049
mps	0.035	farage	0.066	ukip	0.045	leadership	0.026	democracy	0.032
indicative	0.029	ukip	0.048	ge	0.022	rory	0.024	stop	0.030
revoke	0.017	519	0.041	local	0.019	deliver	0.022	prorogation	0.019
options	0.016	deselect	0.028	tories	0.017	widde-combe	0.017	block	0.018
parliament	0.015	milkshake	0.014	victory	0.013	byelection	0.014	suspension	0.015
deadlock	0.012	democracy	0.009	general	0.012	raab	0.011	coup	0.015
extension	0.008	respect	0.008	green	0.012	hammond	0.010	bercow	0.011
TOPIC 11		TOPIC 12		TOPIC 13		TOPIC 14		TOPIC 15	
Word	Score	Word	Score	Word	Score	Word	Score	Word	Score
court	0.072	dup	0.060	boris	0.143	corbyn	0.128	corbyn	0.087
supreme	0.046	ireland	0.027	vote	0.093	johnson	0.088	labour	0.080
proroga-tion	0.045	extension	0.021	election	0.082	nhs	0.052	election	0.038
extension	0.039	letter	0.019	libdems	0.060	swinson	0.025	nhs	0.028
johnson	0.036	juncker	0.019	labour	0.055	deal	0.024	antisem-itism	0.014
rebel	0.026	backstop	0.017	candidate	0.042	pm	0.022	seats	0.012
libdems	0.026	northern	0.016	seats	0.041	nodeal	0.021	working-class	0.011
unleash	0.025	irish	0.016	farage	0.037	labour	0.021	racism	0.008
judges	0.024	border	0.012	tories	0.035	trump	0.016	defeat	0.008
unlawful	0.017	treaty	0.011	conservative	0.016	debate	0.015	blame	0.007

Table 4 (continued)

TOPIC 16		TOPIC 17		TOPIC 18		TOPIC 19		TOPIC 20	
Word	Score	Word	Score	Word	Score	Word	Score	Word	Score
january	0.055	happy	0.027	trump	0.158	theresa	0.356	boris	1.213
50p	0.036	eu	0.021	tusk	0.104	may	0.276	johnson	1.097
bong	0.027	celebrating	0.016	trade	0.065	pm	0.098	october	0.095
ben	0.026	fuck	0.015	uk	0.046	prime	0.069	pm	0.087
trade	0.025	celebrate	0.014	eu	0.029	minister	0.053	prime	0.072
postbrexit	0.024	fireworks	0.012	postbrexit	0.023	deal	0.049	election	0.054
coin	0.014	britain	0.011	america	0.014	nodeal	0.042	minister	0.050
transition	0.012	farage	0.010	hell	0.013	resign	0.024	deliver	0.030
sajid	0.008	congratulations	0.008	ireland	0.012	deliver	0.023	bojo	0.025
cummings	0.007	celebration	0.007	healthcare	0.007	confidence	0.016	ge	0.023

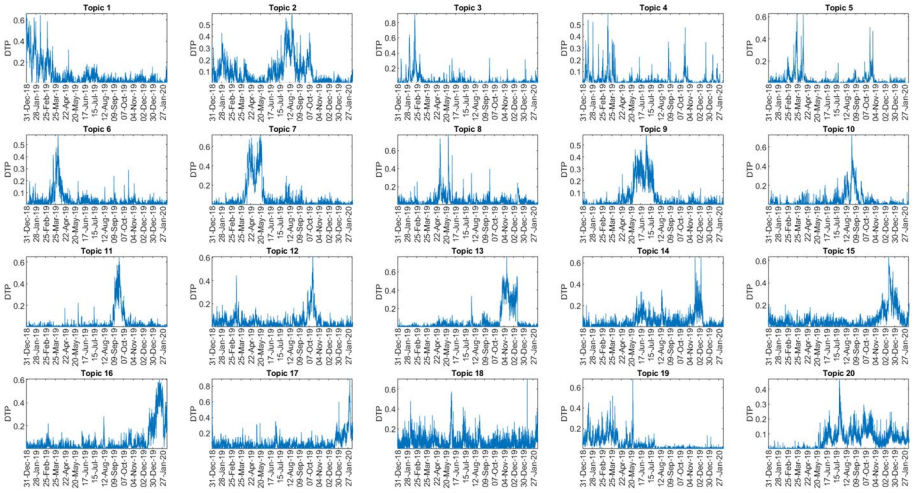

Fig. 7 Trajectories of topic-document probabilities in the LDA analysis

As can be seen from Table 4, Topics 1 and 2 relate to fears for the economic and social consequences of a "no-deal" Brexit scenario. Specifically, Topic 1 contains references to contracts awarded to three ferry companies to handle the potential additional need for roll-on roll-off lorry freight capacity in the event of a no-deal Brexit. It was later discovered that one of these companies, Seaborne Freight, did not, in fact, have any ferries. The Transport Secretary, Chris Grayling, came under considerable pressure and later decided to cancel the ferry contracts. Topic 2 focuses on alarm over possible shortages of food and medical supplies, and on the debate over the compromise on the Irish border backstop included in Theresa May's Brexit deal and aimed at ensuring that no "hard border" (physical checks and infrastructure) should be reintroduced in Ireland. It is evident from Fig. 7 that Topics 1 and 2 were very prominent in the Brexit debate from December 2018 to March 2019,

while Topic 2 also dominated the discussions during the months preceding the vote held in October 2019. The words retrieved under Topic 3 clearly refer to the car industry crisis of February 2019, which concerned Japanese automobile and insurance companies. On 19 February, Honda announced its intention to close its Swindon manufacturing plant in 2021. At the same time, Nissan decided to withdraw investment from its Sunderland plant, while Jaguar Land Rover and Ford announced job cuts. Discussions concerning the meaningful votes and the defeat of the government are captured in Topic 4. This topic appears to be highly localised in time and the peaks correspond to the days in which the votes were held. Topic 5 captures Article 50 extensions and the protest movements organised especially around the first Brexit withdrawal deadline (29 March 2019). In March 2019, the Speaker of the House of Commons, John Bercow, ruled out a third meaningful vote on Theresa May's Brexit deal. This event is found in Topic 6, along with discussion of the indicative votes. Topic 7 focuses on the pro-Brexit politician Nigel Farage, and especially refers to an incident which saw Farage being assaulted by an opponent of Brexit during an electoral campaign event held in Newcastle before the European elections. Topic 8 characterises the Brexit debate during the month preceding the May 2019 European Parliament election. It reflects the electoral debate featuring the UK's major political parties. Topic 9 was evidently highly relevant during June and July 2019. The words retrieved relate to concerns expressed by the most important members of the Conservative and Brexit parties after the Labour party victory in the Peterborough by-election of 6 June 2019. The top scoring words for Topic 10 clearly refer to the attempt by Prime Minister Boris Johnson to suspend Parliament's activities. Chaos broke out in Parliament after the suspension, and impromptu protests were held in major cities across the country to "stop the coup". Indignation was also expressed by John Bercow, the House of Commons Speaker, who described the initiative as a "constitutional outrage". This topic was prominent during September 2019. Figure 7 shows that, soon after this event, the Brexit debate focused on the Supreme Court judgement over the attempt to prorogue Parliament: the Court ruled that the prorogation of Parliament was unlawful and in breach of Britain's constitution. These events feature in Topic 11. Discussions over Brexit intensified as the October deadline drew closer, and concerned the UK and EU positions over the agreed deal (Topic 12). After he had reached an agreement with the EU leaders, Boris Johnson sought Parliamentary approval. While European Commission president Jean-Claude Juncker appeared to rule out an extension to Brexit, Northern Ireland's Democratic Unionist party rejected the deal. The UK general election held on December 2019 characterises Topic 13. Topic 14 captures discussions of the negotiations involving the National Health Service as part of the USA-UK deal. The most relevant terms highlight the strong position of the Labour leader, Jeremy Corbyn, who raised concerns over the implications of giving US companies access to the British health service. Reflections on the defeat of the Labour party in the UK general election, which are captured in Topic 15, dominated the discussion in mid to late December. There was widespread speculation that responsibility for the defeat could be ascribed only to Jeremy Corbyn, and in particular to his reputation for anti-Semitism. He appeared to be widely mistrusted by the British electorate. Words in Topic 16 are linked to discussions trending in the days preceding the Brexit deadline on 31 January 2020: there were references to a crowdfunding campaign, run by the StandUp4Brexit group, to raise money to pay for making the bell of Big Ben 'bong for Brexit' (which never in fact happened); to issue of a commemorative 50p Brexit coin; and to plans for, and concerns about, the post-Brexit transition period.

Topic 17 clearly refers to celebrations by Brexit supporters. On 31 January, after three and a half years of negotiations, the UK became the first country to leave the EU, and celebrations were held all over the whole country.

The discussion in Topic 18 is not clearly localised in time, and the principle terms do not help to identify a clear theme. Finally, the words in the last two topics refer to the two UK Prime Ministers who were protagonists of the Brexit negotiation, and Fig. 7 clearly highlights the transition between Theresa May's and Boris Johnson's leadership. On 24 May PM Theresa May bowed to intense pressure from her own party and named 7 June as the day she would resign as Conservative leader. At the end of July 2019 Boris Johnson won the Conservative leadership contest and took over as the UK's prime minister.

4.2 Dynamic Structure of Topics Associated with Brexit Tweets

To detect topics showing a stability over time, we applied dynamic topic modelling, as described in Sect. 2. After distinguishing topics in time periods, DTM applies a state space model that handles the transition of topics from one period to another.

In this analysis, we consider 58 temporal slices, spanning the period from 31 December 2018 to 09 February 2020, so that each slice corresponds to a weekly window. The corpus consists of 33,430,376 pure tweets. The analysis was performed using the Gensim Python library (https://radimrehurek.com/gensim; Rehurek and Sojka 2011).

To explore the resulting corpus and its themes, we estimated a 20-component dynamic topic model. In this analysis, in order to address word relevance, we take into account their topic-specific probability, and in particular, the top terms for each topic were selected using functional boxplots. Sun and Genton (2011) propose an extension of the classic boxplot to the functional data analysis framework by defining the descriptive statistics of a functional boxplot as: the envelope of the 50% central region, the median curve, and the maximum non-outlying envelope. They further develop the model to allow for detection of outliers. In this analysis, we make use of this model to identify superior outliers in the relevance trends. We are implicitly assuming that the superior outliers correspond to the terms having the highest relevance over time, thereby representing good candidates to characterise the topic dynamics. The functional boxplots are provided in the Supplementary Material and their visual inspection helps in detecting those topics whose dynamic shows a more stable structure over time.

Tracking theme temporal trends, we see that a stable topic is related to the juxtaposition, in the online debate, of the Leave and Remain stances (Topic 5). It is worth noticing how the most characteristic words for this topic (Fig. 8) are strictly linked to the leavers' argument that the result of the public vote in the referendum held in June 2016, when 17.4 million people opted for Brexit, must be respected: this gave the Leave side 51.9%, compared with 48.1% for Remain. This argument is also discussed in Topic 12 (Fig. 9), whose narrative concerns the negotiations for the UK's exit from the EU and the subsequent relationship. It is apparent from Fig. 10 how the Leavers' argument is stressed, in particular starting from the first extension of Article 50 until the European Election and then again from the beginning of August until 9 September, when the Benn Bill was approved. In the early months of 2019 and after the October 2019 general election there is a major discussion on freedom of movement and citizenship rights.

Persistent themes, widely debated, are also those linked to the major social, economic, and political consequences of Brexit (see wordcloud in Fig. 11). In particular, all keywords of Topic 1 are related to the implications for health, social care and education,

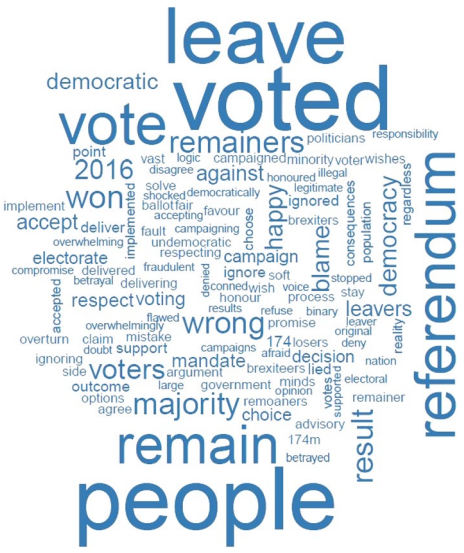

Fig. 8 DTM analysis: wordcloud of the most relevant words of Topic 5

Fig. 9 DTM analysis: wordcloud of the most relevant words of Topic 12

and principally reflect concerns about possible shortages in the supply of vital medicines, price increases for some medications, and greater difficulty in recruiting international staff. The discussion associated with Topic 3 undoubtedly refers to the negative impact that the exit of the UK from the EU may have on the economy in general and on trade in particular. The top scoring words reported for Topic 9 are highly coherent upon inspection and clearly suggest that the underlying discussion is linked to relations between England, Wales, and Scotland, and specifically to the debate over Scottish independence. Another steadily disputed theme, retrieved under Topic 18, refers to the impact of Brexit on the Irish border, which was due to become the only land border between Great Britain and the European Union. The Irish border issue is clearly one

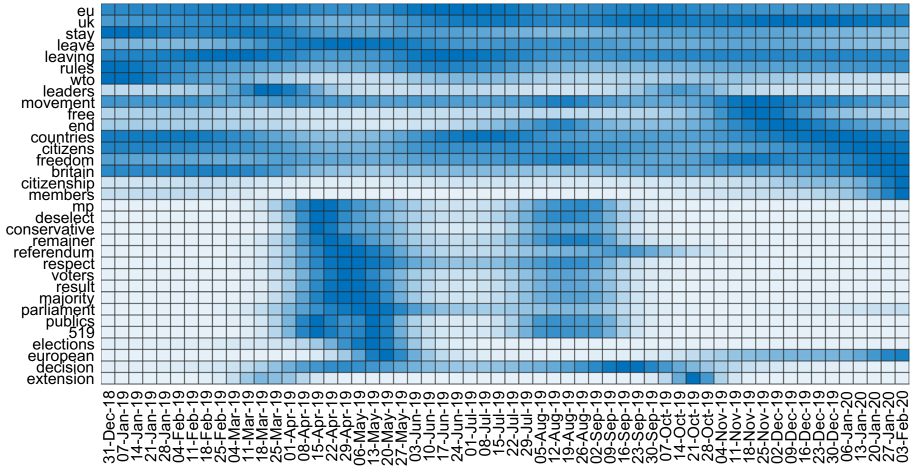

Fig. 10 DTM analysis: heatmap of the most relevant words of Topic 12. The color represents word-document probabilities, normalised with respect to the maximum for each word. (Color figure online)

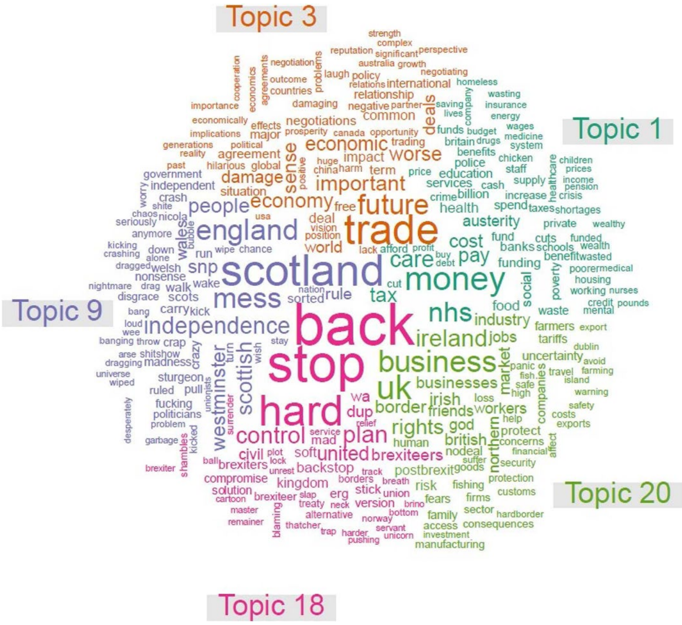

Fig. 11 Wordcloud for Topics related to socio-economic issues

of key arguments also captured by Topic 20, along with the challenges that Northern Ireland has to face, being more exposed to the impact of any trade barriers that might emerge as a consequence of Brexit. Exploring the most frequent terms associated with this topic, we find that the Twitter public expresses uncertainties about changes to trade, customs, investments, local economies, services, and other matters as a result of Brexit.

Finally, it is worth noting that DTM was able to identify a topic that melds together offensive and pejorative terms (see Topic 7 in Fig. 12) and which has remained stable over time.

5 Discussion and Future Research

Social media have emerged as a critical factor in information dissemination and, potentially, an important tool for mobilizing people. Social scientists have long recognized the importance of social networks in the spread of information (Granovetter 1973) and modern communications technologies have enhanced this role making social networks ubiquitous. Therefore, social media are an important channel for people to share information, simplifying and facilitating at the same time news sharing for media organizations.

This role is remarkably carried out by Twitter, which is used as a hybrid between a communication media and an online social network and hosts real-time discussion of current topics of popular interest (Kwak et al. 2010).

Focusing on political communication, Casero-Ripollés (2018) describes how social media, in general, and Twitter, in particular, widen the number and type of actors that interact and how the communication is, consequently, moving from a scenario characterised by relations between journalists and politicians, to a more scattered panorama in which a greater number of actors participate in the exchanges that contribute to define the public sphere, thanks to digital platforms. Besides the spreading of information, social influence can play a pivotal role also on the adoption of political opinions. Studies of political conversations have shown that both mass media and individuals can deliver political information that citizens employ in their voting decisions (see Vaccari et al. 2013, and references herein).

In this study, we have provided a comprehensive analysis of the Brexit-debate on the online Twitter platform. Britain's split from the European Union is one of the most important political happening of recent times as well as being one of the most debated events

Fig. 12 DTM analysis: word-cloud for Topic 7

almost from the moment the idea was broached, and this is testified by the large volume of tweets containing the term Brexit. Compared to previous studies, in our work we explored an exhaustive and updated Brexit-related Twitter activity, occurred between December 2018 and February 2020 and our investigation has been aimed at building up an understanding of the online debate around Brexit and its dynamic across time.

To map Twitter info-sphere, we collected more than 135 million Brexit-related tweets. The temporal analysis of Twitter traffic allows to clearly identify the key dates in the UK's divorce from the EU, supporting the role of Twitter as a communication channel exploited by political and social actors as well individuals to spread information and news. Also the hashtag dynamic follows closely the Brexit timeline and reveals two opposite viewpoints: hard "brexiteers" who would exit the EU with no deal, and, on the other side, remainers who support people's vote campaign and call for a second referendum on the final Brexit deal. Apart from some hashtags that are popular only over a limited time frame, the hashtag evolution shows a high permanence and stability, confirming the results of Romero et al. (2011) on the persistence of hashtags on politically controversial topics. According to the authors, this is an example of the "complex contagion" principle which states that repeated exposures to an idea are particularly crucial when the idea is in some way controversial or contentious.

The clear temporal evolution of the debate and its links to the most relevant events in the Brexit process is registered also in the latent topics retrieved by the probabilistic topic models which testify Twitter's role inside the information dissemination process. The use of LDA models enabled us to gain valuable insights into various aspects of the debate. Use of the static model revealed transient topics (having a significant localisation in time) inspired by the general sequence of events, while the main underlying topics that have characterised the debate from start to finish were captured by the dynamic model: these are not localised in time, but they represent the fundamental elements of the Brexit debate.

An interesting aspect to further investigate, and a possible dimension for our future research on intermedia agenda setting in the spread of information on Brexit, is the mechanism of content transfers between traditional mass media and social media. Questions exist over the extent and the direction of the interaction between those two categories of media (Rogstad 2016; Harder et al. 2017). Su and Borah (2019), comparing the agendas on Twitter and newspapers through rank-order and cross-lagged correlations between both platforms, have found out that Twitter is more likely to influence newspapers agenda in terms of breaking news, whereas newspapers are more likely to lead Twitters agenda in terms of ongoing discussions during non-breaking news periods. The analysis of the intermedia agenda setting can be accomplished by using time series models to identify influence in media networks as well as convergence behaviour in the topics being discussed across source (Meraz 2011).

References

Batrinca, B., & Treleaven, P. C. (2015). Social media analytics: A survey of techniques, tools and platforms. *Ai and Society*, *30*(1), 89–116. https://doi.org/10.1007/s00146-014-0549-4.

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with python*. Newton: O'Reilly Media Inc.

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, *55*(4), 77–84. https://doi.org/10.1145/2133806.2133826.

Blei, D. M., & Lafferty, D. J. (2006). Dynamic topic models. In *ICML 2006—Proceedings of the 23rd international conference on machine learning* (pp. 113–120). https://doi.org/10.1145/1143844.1143859.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, *3*(1), 993–1022. https://doi.org/10.1162/jmlr.2003.3.4-5.993.

Boyd-Graber, J., Hu, Y., & Mimmo, D. (2017). Applications of topic models. *Foundations and Trends in Information Retrieval*, *11*(2–3), 143–296. https://doi.org/10.1561/1500000030.

Casero-Ripollés, A. (2018). Research on political information and social media: Key points and challenges for the future. *El profesional de la informaciòn*, *27*(5), 964–974.

Ceron, A., Curini, L., & Iacus, S. M. (2017). *Politics and big data*. London: Routledge. https://doi.org/10.4324/9781315582733.

Cioffi-Revilla, C. (2010). Computational social science. *Wiley Interdisciplinary Reviews: Computational Statistics*, *2*(3), 259–271. https://doi.org/10.1002/wics.95.

Deerwester, S., Dumais, G., Furnas, S., Landauer, T., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, *41*(6), 391–407.

Fang, A., Habel, P., Ounis, I., & MacDonald, C. (2019). Votes on Twitter: Assessing candidate preferences and topics of discussion during the 2016 U.S. presidential election. *SAGE Open*, *9*(1), 2158244018791653. https://doi.org/10.1177/2158244018791653.

Granovetter, M. S. (1973). The strength of weak ties. *American Journal of Sociology*, *78*(6), 1360–1380.

Grčar, M., Cherepnalkoski, D., Mozetič, I., & Kralj Novak, P. (2017). Stance and influence of Twitter users regarding the Brexit referendum. *Computational Social Networks*, *4*(1), 1–25. https://doi.org/10.1186/s40649-017-0042-6.

Hänska-Ahy, M., & Bauchowitz, S. (2017). Tweeting for Brexit: How social media influenced the referendum. In J. Mair, T. Clark, N. Fowler, R. Snoddy, & R. Tait (Eds.), *Brexit, trump and the media* (pp. 31–35). Bury St Edmunds: Abramis Academic Publishing.

Harder, R. A., Sevenans, J., & Van Aelst, P. (2017). Intermedia agenda setting in the social media age: How traditional players dominate the news agenda in election times. *The International Journal of Press/Politics*, *22*(3), 275–293. https://doi.org/10.1177/1940161217704969.

Hofmann, T. (1999). Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 50–57). https://doi.org/10.1145/312624.312649.

Howard, P. N., & Kollanyi, B. (2016). Bots, #StrongerIn, and #rexit: Computational propaganda during the UK-EU referendum. arXiv:1606.06356 [physics]. https://doi.org/10.2139/ssrn.2798311.

Hürlimann, M., Davis, B., Cortis, K., Freitas, A., Handschuh, S., & Fernández, S. A. (2016). A Twitter sentiment gold standard for the Brexit referendum. In *Proceedings of the 12th international conference on semantic systems, Leipzig, Germany* (pp. 193–196). https://doi.org/10.1145/2993318.2993350.

Iwata, T., Yamada, T., Sakurai, Y., & Ueda, N. (2010). Online multiscale dynamic topic models. In *Proceedings of the 16th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 663–672). New York, NY, USA: ACM. https://doi.org/10.1145/1835804.1835889.

Jungherr, A., & Theocharis, Y. (2017). The empiricist's challenge: Asking meaningful questions in political science in the age of big data. *Journal of Information Technology and Politics*, *14*(2), 97–109. https://doi.org/10.1080/19331681.2017.1312187.

Kaplan, A. M., & Haenlein, M. (2010). Users of the world, unite! The challenges and opportunities of social media. *Business Horizons*, *53*(1), 59–68. https://doi.org/10.1016/j.bushor.2009.09.003.

Karami, A., Bennett, L. S., & He, X. (2018). Mining public opinion about economic issues: Twitter and the U.S. presidential election. *International Journal of Strategic Decision Sciences*, *9*(1), 18–28. https://doi.org/10.4018/IJSDS.2018010102.

Khatua, A., & Khatua, A. (2016). Leave or remain? Deciphering Brexit deliberations on Twitter. In *16th international conference on data mining workshops (ICDMW), IEEE* (pp. 428–433). https://doi.org/10.1109/ICDMW.2016.0067.

Korakakis, M., Spyrou, E., & Mylonas, P. (2017). A survey on political event analysis in Twitter. In *2017 12th international workshop on semantic and social media adaptation and personalization (SMAP)* (pp. 14–19). https://doi.org/10.1109/SMAP.2017.8022660.

Kwak, H., Lee, C., Park, H., & Moon, S. (2010). What is Twitter, a social network or a news media? In *Proceedings of the 19th international conference on world wide web* (pp. 591–600). New York, NY, USA: ACM. https://doi.org/10.1145/1772690.1772751.

Lansdall-Welfare, T., Dzogang, F., & Cristianini, N. (2016). Change-point analysis of the public mood in UK Twitter during the Brexit referendum. In *IEEE international conference on data mining in politics workshop (DMIP)* (pp. 434–439). https://doi.org/10.1109/ICDMW.2016.0068.

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabasi, A. L., Brewer, D., et al. (2009). Computational social science. *Science*, *323*(5915), 721–723. https://doi.org/10.1126/science.1167742.

Llewellyn, C., & Cram, L. (2016). Brexit? Analyzing opinion on the UK-EU referendum within Twitter. In *Proceedings of the tenth international AAAI conference on web and social media* (pp. 1760–1761).

MATLAB. (2018). *version 9.5.0.944444 (R2018b)*. The MathWorks Inc., Natick, Massachusetts.

Meraz, S. (2011). Using time series analysis to measure intermedia agenda-setting influence in traditional media and political blog networks. *Journalism and Mass Communication Quarterly*, *88*(1), 176–194. https://doi.org/10.1177/107769901108800110.

Mimno, D., Wallach, H. M., Talley, E., Leenders, M., & McCallum, A. (2011). Optimizing semantic coherence in topic models. In *Proceedings of the conference on empirical methods in natural language processing, EMNLP '11* (pp. 262–272). USA.

Rehurek, R., & Sojka, P. (2011). Gensim-statistical semantics in Python. EuroScipy 2011, Paris, 25–28/8/2011.

Rogstad, I. (2016). Is Twitter just rehashing? Intermedia agenda setting between Twitter and mainstream media. *Journal of Information Technology and Politics*, *13*(05), 1–17.

Romero, D. M., Meeder, B., & Kleinberg, J. (2011). Differences in the mechanics of information diffusion across topics: Idioms, political hashtags, and complex contagion on Twitter. In *Proceedings of the 20th international conference on world wide web* (pp. 695–704). New York, NY, USA: ACM. https://doi.org/10.1145/1963405.1963503.

Steyvers, M., & Griffiths, T. (2006). Probabilistic topic models. In T. Landauer, D. McNamara, S. Dennis, & W. Kintsch (Eds.), *Latent semantic analysis: A road to meaning*. Hillsdale: Lawrence Erlbaum.

Su, Y., & Borah, P. (2019). Who is the agenda setter? Examining the intermedia agenda-setting effect between Twitter and newspapers. *Journal of Information Technology and Politics*. https://doi.org/10.1080/19331681.2019.1641451

Sun, Y., & Genton, M. G. (2011). Functional boxplots. *Journal of Computational and Graphical Statistics*, *20*(2), 316–334. https://doi.org/10.1198/jcgs.2011.09224.

Teh, Y. W., Newman, D., & Welling, M. (2006). A collapsed variational Bayesian inference algorithm for latent Dirichlet allocation. In *Proceedings of the 19th international conference on neural information processing systems* (pp. 1353–1360). Cambridge, MA, USA: MIT Press.

Vaccari, C., Valeriani, A., Barberà, P., Bonneau, R., Jost, J., Nagler, J., et al. (2013). Social media and political communication: A survey of Twitter users during the 2013 Italian general election. *Rivista Italiana di Scienza Politica*, *43*(12), 381–410. https://doi.org/10.1426/75245.

Wang, C., & Blei, D. M. (2011). Collaborative topic modeling for recommending scientific articles. In *Proceedings of the 17th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 448–456). ACM. https://doi.org/10.1145/2020408.2020480.

Wang, C., Blei, D., Heckerman, D. (2008). Continuous time dynamic topic models. In *Proceedings of the twenty-fourth conference on uncertainty in artificial intelligence, UAI'08* (pp. 579–586). Arlington, Virginia, United States: AUAI Press.

Wang, X., Liu, S., Chen, J., Zhu, H., & Guo, B. (2016). Topicpanorama: A full picture of relevant topics. *TVCG*, *22*(12), 2508–2521. https://doi.org/10.1109/TVCG.2016.2515592.

Wang, X., & McCallum, A. (2006). Topics over time: A non-Markov continuous-time model of topical trends. In *Proceedings of the 12th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 424–433). New York, NY, USA: ACM. https://doi.org/10.1145/1150402.1150450.

Wei, X., & Croft, W. B. (2006). LDA-based document models for ad-hoc retrieval. In *Proceedings of the 29th annual international ACM SIGIR conference on research and development in information retrieval* (pp. 178–185). New York, NY, USA: ACM. https://doi.org/10.1145/1148170.1148204.