



Geographies of Twitter debates

Detect public stances on Brexit at UK parliamentary constituencies' level

Emiliano del Gobbo¹ · Lara Fontanella² · Sara Fontanella³ · Annalina Sarra⁴

Received: 12 March 2021 / Accepted: 27 August 2021 / Published online: 12 September 2021
© The Author(s) 2021

Abstract

Over the last years, the prodigious success of online social media sites has marked a shift in the way people connect and share information. Coincident with this trend is the proliferation of location-aware devices and the consequent emergence of user-generated geospatial data. From a social scientific perspective, these location data are of incredible value as it can be mined to provide researchers with useful information about activities and opinions across time and space. However, the utilization of geo-located data is a challenging task, both in terms of data management and in terms of knowledge production, which requires a holistic approach. In this paper, we implement an integrated knowledge discovery in cyberspace framework for retrieving, processing and interpreting Twitter geolocated data for the discovery and classification of the latent opinion in user-generated debates on the internet. Text mining techniques, supervised machine learning algorithms and a cluster spatial detection technique are the building blocks of our research framework. As real-word example, we focus on Twitter conversations about Brexit, posted on Uk during the 13 months before the Brexit day. The experimental results, based on various analysis of Brexit-related tweets, demonstrate that different spatial patterns can be identified, clearly distinguishing pro- and anti-Brexit enclaves and delineating interesting Brexit geographies.

Keywords Social media · Knowledge discovery in cyberspace · Ensemble learning · Cluster detection and Brexit debate

Mathematics Subject Classification 62Pxx · 62-07

✉ Annalina Sarra
annalina.sarra@unich.it

Extended author information available on the last page of the article

Introduction

Nowadays, location-aware mobile devices are prevalent access points to social media services, and several social networking websites allow their users to share their location along with social media posts, either by explicitly specifying the geographical location or by embedding the spatial coordinates (i.e. latitude and longitude) in their posts. Therefore, location-based social media data are being increasingly used in research projects (see [16, 37], and references therein) and geo-tagged data can be useful to extract location-specific information, as well as to investigate the variation of information across different geographical regions [3]. To analyse the interdependent relationships among places, time, and contents shared on social media, Tsou and Leitner [55] introduced a new research framework, called knowledge discovery in cyberspace (KDC), which extracts information from geo-located social media data by using highly scalable mining and machine learning algorithms, computational linguistics, geographic information systems, visualisation tools, and spatial statistical methods. The spatial component is “the most important component in the KDC framework [...] because it is the glue to combine temporal information (time) and social media content (messages)” [55]. In social media research, the spatial dimension can also be of interest in mining people’s opinions, attitudes, and emotions towards entities, issues and events. Social media is becoming an alternative to polls for public opinion analysis [32], and geo-tagged social media data provides opportunities to explore the geographic locations of users expressing their ideas [12, 25]. In particular, geolocation information embedded in social media posts, along with the textual component, can be exploited to analyse the spatial distribution of opinions on a given topic. In this paper, we focus on stance classification, an opinion mining sub-domain that allows to automatically determine from the text whether the author is in favour of, against, or neutral towards a given target. Over the years, interest has rapidly grown around the classification of stances and most of the work has focused on debates in online forums (see [28, 53], and references therein). However, there is an increasing interest in stance classification on Twitter [2, 4, 42, 47], which, by providing a public application programming interface (API), represents the most accessible social networking application to researchers. As for opinion mining, Twitter offers a valid support to monitor and evaluate people’s views and belief on political phenomena and stands on referendums and elections [6, 18, 23, 56]. In this context, the analysis of the geographical dimension of the online political debate provides a picture of attitude and stance territorial heterogeneity and helps in exploring spatial patterns. In addition, spatial statistics techniques can be exploited to identify regions of space where the phenomenon shows distinct features and to detect territorial clusters.

In our research, we implement an integrated KDC approach whose building blocks are text mining techniques, supervised machine learning algorithms and a spatial cluster detection model. As real-world example, we consider the spatial analysis of the Twitter debate on the Brexit process during the year before the Brexit day (31 January 2020), when the UK officially left the EU. Specifically,

we sought to closely examine the relationship between communication on social media and the electoral geography that emerged from the 2016 United Kingdom European Union membership referendum [8, 17, 27, 31]. We are aware that information gathered from Twitter is biased due its non-random nature. In other words, Twitter data are not a representative sample of the population as a whole and thus they cannot be investigated from a broader perspective. For instance, [40], examining the distribution of demographic and attitudinal variables of social media users, found that Twitter platform has markedly different composition to the general UK population along most demographic and political variables. There is evidence that in UK Twitter global audience is younger than average, mainly composed by men and more likely to achieve better education attainment other than to be economically better (see results from [11] and [50]). Besides, studies using geotagged UK twitter data have found that Twitter users are more commonly found in urban areas and particularly wealthier areas, with younger populations [30]. In literature, different authors discuss methodologically, conceptually and technically the generalizability of social media data to broader populations (see, among others, [20, 44, 61]). As the focus on our analysis is more concerned with the online debate relating to the Brexit process rather than to making inference, we believe that the use of non probabilistic sampling does not weaken the design of our study. For the aims of our research, we first classify users with respect to their pro- or anti-Brexit attitude, on the basis of the contents they share on Twitter. This task is addressed by applying supervised machine learning algorithms onto the geo-tagged tweet corpus, and considering as units of analysis tweets' authors. Then, we map the obtained stance classification onto parliamentary constituencies in England, Wales, Scotland, and Northern Ireland. To investigate the spatial heterogeneity of the phenomenon under study and identify spatial clusters, we use the Bayesian scan statistics method [58]. This flexible method allows the identification of multiple clusters with arbitrary shapes, without any prior knowledge about their number and location.

The paper is organized as follows: “[Data collection and preprocessing](#)” details the experimental setup, including data collection, the construction of geo-tagged corpus and the related screening and cleaning operations. “[Methodology](#)” describes the methodological framework enabling the integrated analysis of tweets and location. “[Results](#)” evaluates the classification results and discusses the geography of stance occurrences in the geo-tagged corpus. Finally, “[Conclusion](#)” concludes and summarizes research perspectives.

Data collection and preprocessing

Data for this study were taken from Twitter conversations about Brexit, collected over 58 weeks, spanning from 31/12/2018 to 09/02/2020. Tweets were extracted in real-time through Socialgrabber, an application developed to exploit Twitter Streaming API. Data were gathered with a triple geographical redundancy in order to avoid any gaps or inconsistencies. Crucial to any form of data collection from Twitter is the provision of relevant search terms. In our work, we decided to focus on tweets,

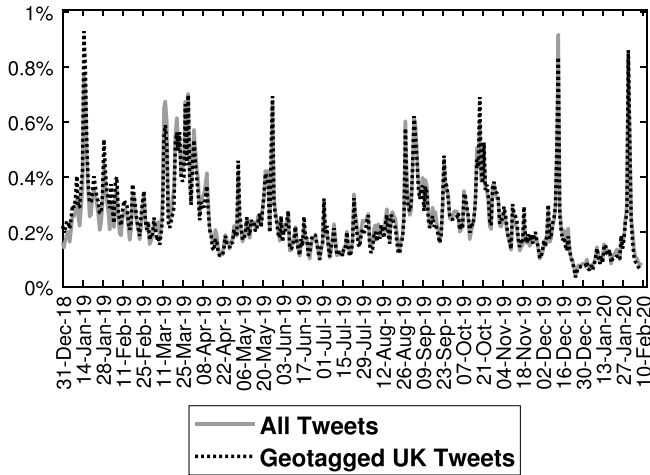


Fig. 1 Temporal dynamic of the UK geo-tagged tweets and the entire corpus, in percentage with respect to the corresponding total over all the observation period

written in English, containing the broad term “Brexit”. In total, we collected a dataset containing, along with the associated metadata, 135,607,216 tweets, of which 33,430,376 are pure tweets. A detailed description of their textual context and its dynamic can be found in [24]. Here, given the interest in analysing the geographical dimension of the online Brexit debate, we restrict the analysis to UK geo-tagged tweets.

Geo-localisation

We considered GPS coordinates provided by the users for selecting geo-tagged tweets, and we computed the haversine distance between the tweets’ coordinates and the latitudes and longitudes of 3173,958 cities included in the database provided by Maxmind.¹ To assign each tweet to the closest city in the database, we used the Python library Numba [36], whereas a custom GPU kernel has been built to accelerate the computation task with GPU high parallelization capability. In this way, we were able to identify the country for each geo-tagged tweet, out of which 957,356 (2.86% of the pure tweets) had been posted from the United Kingdom. To assign these tweets to the UK parliamentary constituencies, we relied on the shapefile provided by the UK Office for national statistics.²

The dynamic of the UK tweets shows the same temporal pattern of the entire corpus (Fig. 1). As for the vocabulary, after stopwords removal, there is a high correlation (0.764) between term occurrences in the total corpus and in the UK geo-tagged one.

¹ <https://www.kaggle.com/max-mind/world-cities-database>.

² https://geoportal.statistics.gov.uk/datasets/5ce27b980ffb43c39b012c2ebeb92c0_2.

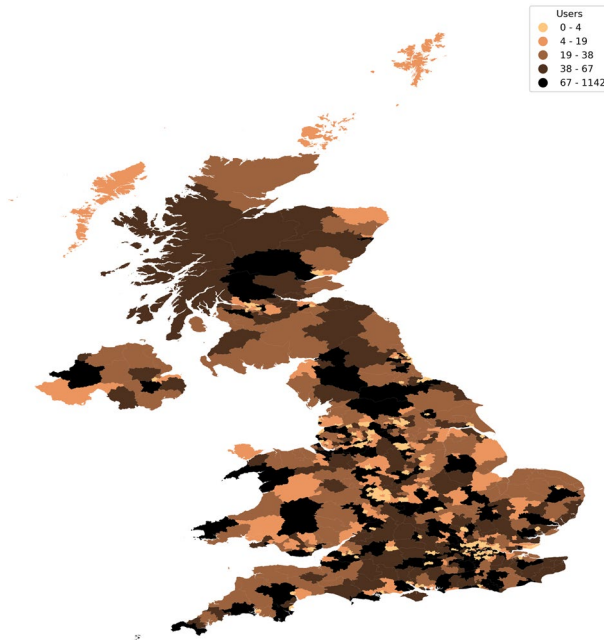


Fig. 2 Selected geo-tagged UK users distribution by constituency

In the investigation of the geography of Brexit stances, we consider as unit of analysis tweets’ authors. More in detail, we focus on users with at least one geo-tagged tweet in UK, and at least 10 tweets in the entire corpus. Users matching this condition are 40,563. Through this criterion, we avoid including users who took little part in the online Brexit conversations over 13 months. Users’ assignment to the UK constituencies is based on the majority rule applied to tweets’ locations. Figure 2 displays the number of users across the constituencies and gives evidence of a not uniform territorial distribution.

Text preprocessing

We developed a Python script to perform the screening and cleaning process (tokenization; lowercase conversion; special characters, URL, mentions and stop-words removal) of the textual context of the UK geo-tagged tweets to extract the relevant content and remove any unwanted nuisance terms. The `nlk.word_tokenize` function in the NLTK [38] Python package [10] has been employed for tokenization. We essentially relied on the English stopwords dictionary of the NLTK package, with some minor variations, including the tweet filtering keyword “brexit” and allowing the “no” term, as it is relevant to several patterns like “no deal”. The full geotagged UK corpus has 153,162 unique tokens; about 53% of these distinct tokens are hapax legomena (i.e. words that occur only once within the entire corpus) and around 81% are used less than 10 times in the complete textual dataset. After all

cleaning operations and the pruning of tokens occurring less than 150 times in the selected corpus, the final number of tokens is 7246. As a consequence of dictionary pruning, we needed to remove from the corpus 90 users left without any token.

Methodology

The methodological setup of the research comprises two fundamental steps for classification (ensemble learning) and spatial cluster detection. The single steps are described in the following sub-sections.

Remain versus leave: ensemble learning for choosing sides

We wish to predict the stands of Twitter users based on the content of their tweets related to the Brexit debate. Labelling large amount of data are a complex computationally expensive and time-consuming process. To overcome these limitations, given the large volume of tweets and considerable numbers of users, we proceeded in a stepwise fashion. First, we generated a training set consisting of human-classified accounts, and, second, we adopted supervised machine learning (ML) techniques to predict the stances of the remaining users based on the content of their tweets on the Brexit debate. The core training set has been constructed through a random sampling procedure, where the probability of inclusion was set proportional to the number of tweets posted by each user. We visited the Twitter accounts of the sampled users to check if there was evidence of their opinion on Brexit. In addition, we checked the contents of their Brexit related tweets and the hashtags used. In this way, 994 users were manually labelled as pro-Brexit (398 users) and anti-Brexit (596 users). We excluded from the training dataset users not showing a clear stance on Brexit. We then proceeded to classify the remaining users by adopting supervised ML techniques, using the carefully constructed training data. Specifically, we applied ensemble learning [21, 45, 49] to estimate the class membership of the remaining users. In ensemble learning, a number of base learners are trained, and their individual results are combined for prediction to achieve strong generalizability. The base learners act as a “committee” of decision-makers. The careful selection of the models and appropriate combination of the resulting individual predictions should lead, then, to an increased overall accuracy. In our context, ensemble learning should guarantee more robust and accurate predictions: as single learners are affected by small amounts of labelled data, these are expected to lead to weak classifiers unable to generalise to unseen cases. The ensemble learning becomes then appealing, as it enhance the weak learners performance, resulting in learners with high prediction accuracy. In this paper, we adopted a common approach in ensemble learning that leads to the definition and training of a collection of individual models, whose outputs are integrated into a final decision by means of a voting process. Here, the final classification was achieved using the majority voting scheme, which combines the individual outputs into a final classification by selecting the mode or, in other words, the most popular label prediction. For the ensemble, we

adopted five different base learners, namely: logistic regression (LR), Gaussian naive Bayes (GNB), decision tree (DT), support vector machines (SVM) and neural network (NNet). The selected models make different assumptions and adopt different strategies to explore the hypotheses space. This diversity is an important feature in ensemble learning, which guarantees the construction of a more performant classifier. The description of the individual methodologies is beyond the scope of this paper and we refer the readers to well-known texts on Statistical Machine Learning [29, 33, 43]. We evaluated the classifications achieved by the individual models using performance measures, such as *accuracy*, defined as the rate of correctly classified instances, *precision*, defined as the fraction of relevant instances among all retrieved instances, and *recall*, also known as *sensitivity*, which consider the fraction of retrieved instances among all relevant instances. Moreover, to meaningfully combine the classifiers, we considered also the diversity of the individual results. To this end, we computed four different pairwise diversity measures: the *Yule's Q statistics*, the *correlation coefficient* ρ , the *disagreement measure* (Dis) and the *double fault measure* (DF) as described in [35].

Spatial cluster detection

To examine the Brexit debate, as conducted on Twitter, in its geographical distribution, we carried out a spatial analysis aiming to identify statistically significant areas of Great Britain where the prevalence of pro or anti-Brexit viewpoints differs significantly from what under the assumption of uniform spatial distribution one would expect.

Researchers in many areas of science are interested in assessing the significance of the occurrence of clusters in space. Cluster detection methods clarify whether there is a pattern in the examined data and help in achieving a better understanding of a geographic phenomenon, by an automatic identification of regions of space that are “anomalous”, “unexpected” or otherwise “interesting”. Spatial cluster detection has been primarily applied in epidemiological research to identify disease hotspots. A comprehensive review on disease cluster detection methods can be found in [39]. As pointed out in this review, techniques for detecting local spatial clusters typically consist of two interrelated components: a geographical search method, aimed at identifying local concentrations of cases to be tested for clustering, and a hotspot detection method, including a statistical model for determining if the local concentration is significantly higher or lower than expected. Within cluster detection techniques, the circular spatial scan statistic [34] has been used extensively. In addition to epidemiological studies, this methodology has been applied in other research fields, including among others seismology [54], environment [52], criminology [41], and economic activity [9]. Scan statistic provides a likelihood ratio-based framework for hotspot evaluation and significance testing. Basic ingredients of this technique consist of the geometry of the area being scanned, the probability distribution generating responses under the null hypothesis of chance variation and the shape and sizes of the scanning window. Given the partition of the study region into a given number of subregions, and the availability, for each of these subregion,

of the number of cases, the total background population, and the centroid location (i.e. Gauss-Boaga coordinates), the scan statistic allows circular clusters to be centred on each area centroid. Under the null hypothesis, we expect that there is no cluster in the study region and the cases are uniformly distributed, such that the expected number of cases in each area is proportional to its population. A likelihood ratio statistic is then computed for each cluster, and the significance of the maximum of the likelihood ratio statistics over all circles is evaluated using a Monte Carlo p value. Several extensions and modifications of the scan statistics have been proposed in published works to make it applicable to hotspot-related issues encountered by scientists. In this paper, to detect spatial clusters of stances of Twitter users regarding Brexit, we rely on the Bayesian version of Kulldorff's scan statistic proposed by [58]. This cluster detection method leads to a partition of the study region such that each area is either within a cluster/anti-cluster or null: areas within a cluster are associated with increased relative risk, while areas within an anti-cluster are associated with decreased relative risk. To approximate the model parameters, inference is carried out adopting a sampling MCMC schema, tuned to the cluster method. A key feature of the method is the focus in computing the expected relative risk surface and the posterior probability that a specified area lies in a cluster/antcluster. To further have insights into the estimates of cluster structure the posterior summaries can be integrated into interpretable graphical displays, and eventually thresholded to be more informative about the areas of interest.

Results

This section illustrates the experimental results of applying the proposed-integrated approach. First, we display the users' stance in online debate about Brexit on Twitter achieved via ensemble learning algorithms and report the evaluation metrics findings. Hence, we account for spatial information of geo-tagged tweets and detect and map the areas where the prevalence of pro- and anti-Brexit viewpoints differs significantly from the assumption of uniform spatial distribution.

Classifying pro-Brexit and anti-Brexit accounts

In the classification task, the individual models were trained using 5-fold cross validation. Specifically, LR and SVM were trained setting the class weight as balanced to mitigate the unbalanced nature of the training dataset; DT and GNB were run with default parameters. NNet was trained using a solver L-BFGS [13] and with a small hidden layer of 5 perceptron. The analyses were run using Python and the Scikit-learn library [46] version 0.23.2.

On the training set, all the models showed very accurate classification, with SVM and NNet achieving the best performance (Table 1). We observed also a strong level of agreement between the classifiers (Table 2). This scenario does not generally prescribe the use of ensemble learning: to obtain higher performances, not only the individual classifiers should have good accuracy, but also

Table 1 Performance comparison of the single learners in stratified 5-folds cross validation

	SVM	LR	GNB	DT	NNet
Accuracy	0.97	0.96	0.95	0.88	0.95
Acc. Variance	0.00	0.00	0.01	0.01	0.00
Precision AB	0.96	0.95	0.95	0.90	0.96
Precision PB	0.98	0.96	0.93	0.84	0.95
Recall AB	0.99	0.98	0.96	0.89	0.97
Recall PB	0.94	0.93	0.92	0.84	0.93

Acronyms: *AB* anti Brexit, *PB* pro Brexit

Table 2 Pairwise diversity measures for the base learners

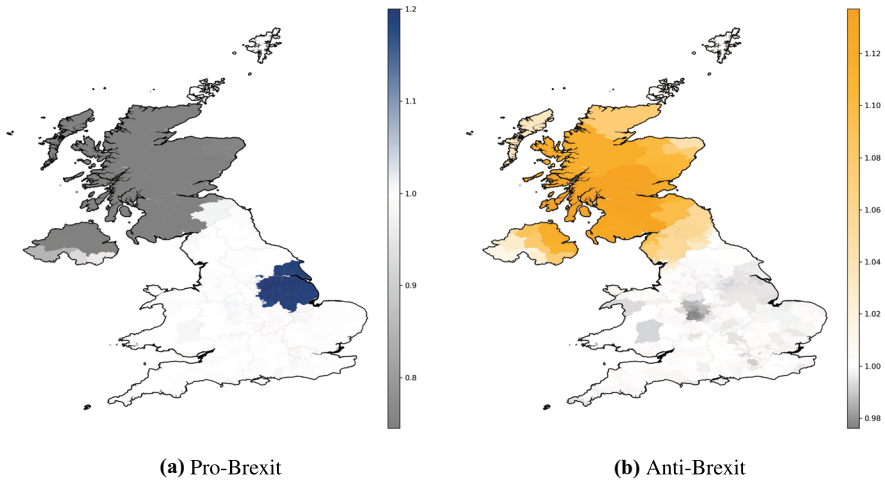
Learner 1	Learner 2	Training data				Complete data			
		Q statistics	ρ	Dis	DF	Q statistics	ρ	Dis	DF
SVM	LR	0.999	0.954	0.022	0.380	0.958	0.744	0.128	0.423
SVM	GNB	0.998	0.941	0.028	0.382	0.590	0.282	0.383	0.409
SVM	DT	1.000	0.975	0.012	0.391	0.399	0.204	0.396	0.243
SVM	NNet	1.000	0.975	0.012	0.391	0.925	0.565	0.234	0.271
LR	GNB	0.999	0.945	0.026	0.380	0.685	0.342	0.349	0.433
LR	DT	0.998	0.929	0.034	0.377	0.421	0.216	0.393	0.252
LR	NNet	0.998	0.929	0.034	0.377	0.867	0.494	0.272	0.259
GNB	DT	0.997	0.920	0.038	0.380	0.429	0.188	0.460	0.332
GNB	NNet	0.997	0.920	0.038	0.380	0.713	0.279	0.476	0.270
DT	NNet	1.000	1.000	0.000	0.400	0.443	0.215	0.364	0.167

show significant prediction diversity. Even though this is not reflected in the training set, the diversity measures obtained considering the whole dataset—labelled and unlabelled users—demonstrate that the individual models lead to different classifications (Table 2). This discrepancy is suggestive of overfitting on the training data: the training set is composed by users with a distinctive profile and a clear stance regarding the Brexit debate; the two classes are thereby highly separable, enabling the individual models to learn perfect classifications. On the premise that assigning the classification to a committee of experts is sometimes more reasonable than just putting all the stake on one expert [48], to reduce the overfitting and ensure a reliable classification of the unlabelled data, we, then, employ ensemble learning. Accordingly, we combined the predictions of the single learners, following the majority voting scheme. To account for uncertainty in the final classification, we adopted a simple modification. Specifically, the stance of each user was assessed considering the degree of agreement of the classifiers in assigning the users to one of the two classes. Table 3 shows how the classification varies according to different thresholds. In this study, we consider the high disagreement between single learners, as a proxy of the uncertainty of user

Table 3 Distribution of class assignment for different levels of agreement

Min. Accordance	PB count	%	AB count	%	Neutral count	%
5 of 5	6199	15.3	5661	14.0	28,613	70.7
4 of 5	14,942	36.9	12,829	31.7	12,702	31.4
3 of 5	21,900	54.1	18,573	45.9	0	0.0

Acronyms: *AB* anti Brexit, *PB* pro Brexit

**Fig. 3** Smoothed posterior estimates of the relative risk

opinion and, for each user, we attached the label “anti” or “pro” if all or at least four out of five machine learning algorithms were in agreement. The remaining users were classified as “Neutral”. The results show higher prevalence of anti-Brexiters (36.9%) compared to proBrexiters (31.7%); the remaining users (31.4%) were classified as Neutrals (Table 3).

Pro-Brexit and anti-Brexit enclaves

To identify significant spatial clusters in the geographical distribution of users’ stances on Brexit, we applied the methodology described in Sect. 3.2 through the R SpatialEpi package [19]. For each parliamentary constituency, we consider the number of users as total background population, while the number of cases is given, in turn, by the aggregate count of “anti” and “pro” Brexit users. In the Bayesian estimation procedure, we retain the same assumptions found in [58].

The posterior estimates of the relative risk, represented in Fig. 3, provide a measure of the incidence rate of users on each constituency being pro-Brexit or anti-Brexit, compared to the users in the remaining region. The maps clearly shows a

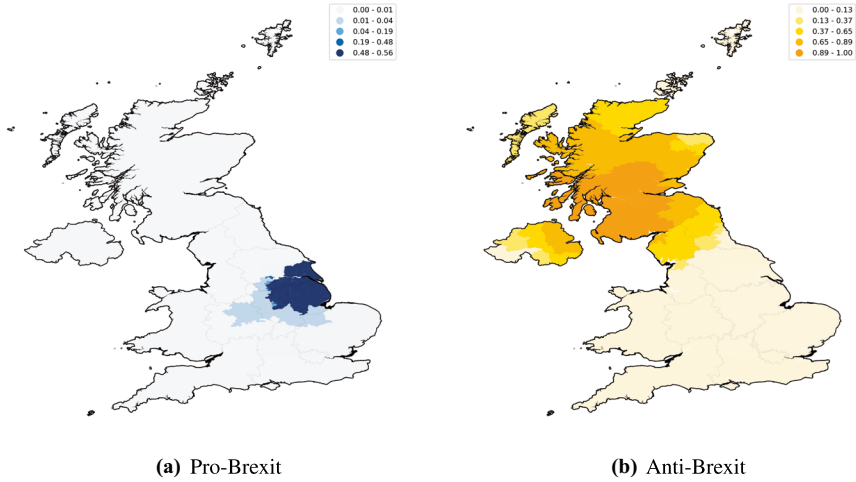


Fig. 4 Posterior probability of cluster membership

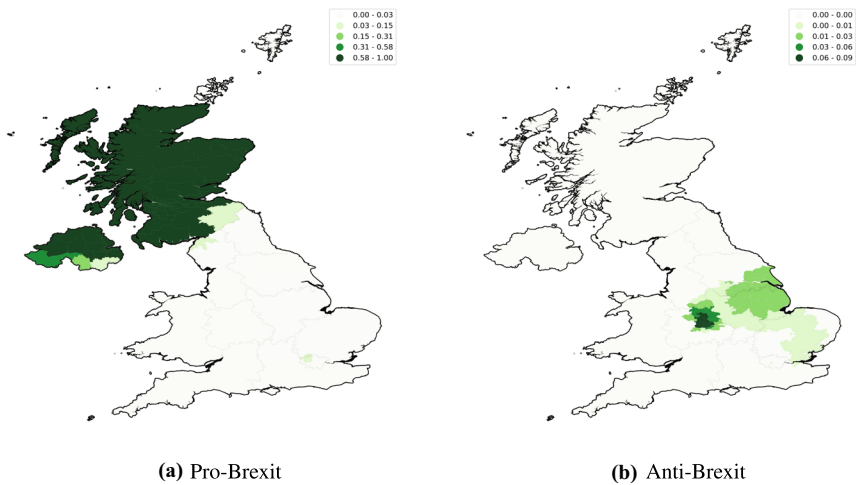


Fig. 5 Posterior probability of anti-cluster membership

higher association of North Ireland and Scotland with anti Brexit stances, while users in the East Midlands are more likely to have a pro-Brexit attitude.

Another useful graphical summary is produced by computing the posterior probability of cluster and anti-cluster membership for each constituency, which give a clear indication to detect areas of clustering related to a significantly higher (Fig. 4) or lower prevalence (Fig. 5) of the corresponding opinion on the Brexit process. The cluster of constituencies where it is more likely to find Twitter’ users against Brexit and less likely to find opinion in favour of Brexit are clearly in line with the findings

derived above from the analysis of the posterior estimates of the relative risk, with the except of London for which it can be noted a significant probability to be in an anti-cluster. On the other hand, the cluster related to Pro-Brexit position in addition to Midlands extends to parliamentary constituencies of West Midlands. In addition, it is worth noting that, considering also the anti-cluster detected for stances against Brexit, the cluster extends also to the Kent constituencies.

Conclusion

Today, with the proliferation of location aware devices, a large proportion of user-generated content contributed through social media sites is geolocated, thus fostering the emergence of geosocial media. The volume and richness of user-generated geospatial information open new frontiers and avenues in geosocial knowledge discovery and offer enormous opportunities to acquire new scientific insights. Allowing the access to fine-grained social media footprints at individual levels, location-based social media data provide a set of new lens to examine complex social dynamics. It should be noted that, through geolocated social media, researchers are able to observe human activities in scales and resolutions that were so far unavailable. In this respect, extensive studies have been already carried out to extract knowledge from geo-located data that convey information about significant events or communicate users opinion and views. However, peculiar characteristics of location-based social media data pose challenges in terms of representativeness and modeling, other than computational issues. In this research, we put forward an integrated KDC model with the aim of handling and analyzing very large human messages collected from cyberspace and social media. The framework is implemented based upon a public data stream of Twitter feeds around Brexit debate, posted on UK over 13 months. Specifically, to exploit the unstructured content of Brexit related geo-tagged tweets, a mixed approach involving text mining techniques, supervised machine learning algorithms and a spatial cluster detection method has been designed. In such way, we were able to gain valuable information on public stances on the divisive topic concerning the United Kingdom European Union membership referendum. Findings from the analysis carried out through the Bayesian scan statistics on a corpus of Brexit-related tweets classified for user location and Brexit stance, have revealed that Twitter general view regarding the remain and leave sides differs by constituencies in a way that resembles the referendum results. Twitter users from Northern Ireland and Scotland predominantly express anti-Brexit attitudes. A different picture emerges from England, where Midlands and West Midlands constituencies are more likely to be in the pro-Brexit enclave. By “zooming in” to a finer spatial resolution, an opposite direction emerges from Twitter activity of London’s users who clearly support the Remain stance. Our experimental results have shown that the sheer volume of geosocial data could lead to results of similar predictive power as traditional election polls. Several studies have shown how to roughly predict public opinion, as in political elections, by analysing user activities in blogging platforms and online social networks (see, among others, [1, 5, 57, 60]). Besides, existing literature has provided specific insights regarding the role of Twitter in reaching political partisans

[15, 59], who may then go on to influence their own friends and social connections [7, 14, 62]. Thus, experimental evidence suggests that Twitter may be useful in activating network of core supporters. The main methodological contribution of this work is a comprehensive framework, able to collect social media feeds, analyse them and unveil geosocial knowledge of users' opinions and trending topics, transferable to many other domains. However, along with the benefits of our approach, some limitations in analysing geo-tagged Twitter data have to be highlighted. First, we acknowledge that publicly available tweets are only a sample of all tweets and only a part of tweet traffic is geocoded [51] because a tweet geocode can only be generated from GPS-enabled devices [26]. Additionally, the population of social media users may be demographically skewed. For example, a recent survey in Great Britain [40] showed that Twitter is especially appealing to highly educated adults whose mean age was 34. In the same research it has been detected that Twitter users differ from the general population in terms of demographics, political attitudes and political behaviour. Since biases in these relevant dimensions can be significant, a possible solution for using social media data for studying public opinion and forecasting elections requires the data to be appropriately weighted using demographics, as suggested by [22]. Future research should seek to overcome some of these limitations and integrate the temporal component into unified framework proposed.

Funding Open access funding provided by Università degli Studi G. D'Annunzio Chieti Pescara within the CRUI-CARE Agreement.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

1. Adamic, L.A., Glance, N. (2005). The political blogosphere and the 2004 US election: Divided they blog. In *Proceedings of the 3rd International Workshop on Link Discovery* (pp. 36–43). Chicago, IL.
2. Addawood, A., Schneider, J., Bashir, M. (2017). Stance classification of twitter debates: The encryption debate as a use case. In: *Proceedings of the 8th International Conference on Social Media and Society* (pp. 1–10). Association for Computing Machinery <https://doi.org/10.1145/3097286.3097288>
3. Ahuja, A., Wei, W., Lu, W., Carley, K.M., Reddy, C.K. (2017) A probabilistic geographical aspect-opinion model for geo-tagged microblogs. In *2017 IEEE International Conference on Data Mining (ICDM)* (pp. 721–726). <https://doi.org/10.1109/ICDM.2017.82>
4. Al-Ghadir, A., Azmi, A., & Hussain, A. (2021). A novel approach to stance detection in social media tweets by fusing ranked lists and sentiments. *Information Fusion*, 67, 29–40. <https://doi.org/10.1016/j.inffus.2020.10.003>.

5. Albrecht, S., Lübcke, M., & Hartig-Perschke, R. (2007). Weblog campaigning in the German bundestag election 2005. *Social Science Computer Review*, 25(4), 504–520. <https://doi.org/10.1177/0894439307305628>.
6. Barberá, P., & Rivero, G. (2015). Understanding the political representativeness of twitter users. *Social Science Computer Review*, 33(6), 712–729. <https://doi.org/10.1177/0894439314558836>.
7. Barnidge, M., Ardévol-Abreu, A., & Gil de Zúñiga, H. (2018). Content-expressive behavior and ideological extremity: An examination of the roles of emotional intelligence and information network heterogeneity. *New Media & Society*, 20, 815–834. <https://doi.org/10.1177/1461444816675183>.
8. Bastos, M., & Mercea, D. (2018). Parametrizing Brexit: mapping Twitter political space to parliamentary constituencies. *Information, Communication & Society*, 21(7), 921–939. <https://doi.org/10.1080/1369118X.2018.1433224>.
9. Bersimis, S., Chalkias, C., & Anthopoulou, T. (2014). Detecting and interpreting clusters of economic activity in rural areas using scan statistic and LISA under a unified framework. *Applied Stochastic Models in Business and Industry*, 30(5), 573–587. <https://doi.org/10.1002/asmb.2003>.
10. Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python*. O'Reilly Media, Inc.
11. Blank, G. (2017). The digital divide among twitter users and its implications for social research. *Social Science Computer Review*, 35(6), 679–697. <https://doi.org/10.1177/0894439316671698>.
12. Bohannon, J. (2017). The pulse of the people. *Science*, 355(6324), 470–472. <https://doi.org/10.1126/science.355.6324.470>.
13. Bollapragada, R., Mudigere, D., Nocedal, J., Shi, H.M., Tang, P.T.P. (2018). *A progressive batching L-BFGS method for machine learning*. [arXiv:1802.05374](https://arxiv.org/abs/1802.05374)
14. Borge Bravo, R., & Esteve Del Valle, M. (2017). Opinion leadership in parliamentary Twitter networks: A matter of layers of interaction? *Journal of Information Technology & Politics*, 14, 263–276. <https://doi.org/10.1080/19331681.2017.1337602>.
15. Bright, J. (2018). Explaining the emergence of political fragmentation on social media: The role of ideology and extremism. *Journal of Computer-Mediated Communication*, 23, 17–33. <https://doi.org/10.1093/jcmc/zmx002>.
16. Cao, G. C., Wang, S., Hwang, M., Padmanabhan, A., Zhang, Z., & Soltani, K. (2015). A scalable framework for spatiotemporal analysis of location-based social media data. *Computers, Environment and Urban Systems*, 51, 70–82. <https://doi.org/10.1016/j.compenvurbsys.2015.01.002>.
17. Celli, F., Stepanov, E., Poesio, M., Riccardi, G. (2016). Predicting Brexit: Classifying agreement is better than sentiment and pollsters. In: *Proceedings of the Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media (PEOPLES)* (pp. 110–118). The COLING 2016 Organizing Committee.
18. Chauhan, P., Sharma, N., & Sikka, G. (2020). The emergence of social media data and sentiment analysis in election prediction. *Journal of Ambient Intelligence and Humanized Computing*. <https://doi.org/10.1007/s12652-020-02423-y>.
19. Chen, C., Kim, A., Ross, M., Wakefield, J. (2018). *Methods and data for spatial epidemiology, R Package Version 1.2.3*. <https://CRAN.R-project.org/package=SpatialEpi>
20. Davis, J., & Love, T. (2018). Generalizing from social media data: A formal theory approach. *Information, Communication & Society*, 22(5), 637–647. <https://doi.org/10.1080/1369118X.2018.1555610>.
21. Dietterich, T.G. (2000) Ensemble methods in machine learning. In *Multiple classifier systems, Lecture notes in computer science* (pp. 1–15). Springer. https://doi.org/10.1007/3-540-45014-9_1.
22. Filho, R., Almeida, J., Pappa, G. (2015). Twitter population sample bias and its impact on predictive outcomes. In *Proceedings of the 2015 IEEE/ACM international conference on advances in social networks analysis and mining* (pp. 1254–11264). ACM Press
23. Gaumont, N., Panahi, M., & Chavalarias, D. (2018). Reconstruction of the socio-semantic dynamics of political activist Twitter networks—Method and application to the 2017 French presidential election. *PLoS One*, 13(9), e0201879. <https://doi.org/10.1371/journal.pone.0201879>.
24. del Gobbo, E., Fontanella, S., Sarra, A., & Fontanella, L. (2020). Emerging topics in Brexit debate on Twitter around the deadlines. *Social Indicator Research*. <https://doi.org/10.1007/s11205-020-02442-4>.
25. Gong, Z., Cai, T., Thill, J., Hale, S., & Graham, M. (2020). Measuring relative opinion from location-based social media: A case study of the 2016 U.S. presidential election. *PLoS One*, 15(5), e0233660. <https://doi.org/10.1371/journal.pone.0233660>.

26. Han, B., Cook, P., & Baldwin, T. (2014). Text-based twitter user geolocation prediction. *Journal of Artificial Intelligence Research*, 49, 451–500. <https://doi.org/10.1613/jair.4200>.
27. Hanretty, C. (2017). Areal interpolation and the UK's referendum on EU membership. *Journal of Elections, Public Opinion and Parties*, 27(4), 466–483. <https://doi.org/10.1080/17457289.2017.1287081>.
28. Hasan, K.S., Ng, V. (2013). Stance classification of ideological debates: Data, models, features, and constraints. In *Proceedings of the Sixth international joint conference on natural language processing* (pp. 1348–135). Asian Federation of Natural Language Processing. <https://www.aclweb.org/anthology/I13-1191>
29. Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The elements of statistical learning*. Springer series in statistics. Springer New York Inc.
30. Hecht, B., Stephens, M. (2014). A tale of cities: Urban biases in volunteered geographic information. In *Proceedings of the International Conference on Web and Social Media (ICWSM)* (pp. 197–205). AAAI Publications.
31. Johnston, R., Manley, D., Pattie, C., & Jones, K. (2018). Geographies of Brexit and its aftermath: Voting in England at the 2016 referendum and the 2017 general election. *Space and Polity*, 22(2), 162–187. <https://doi.org/10.1080/13562576.2018.1486349>.
32. Klačnja, M., Barberá, P., Beauchamp, N., Nagler, J., & Tucker, J. A. (2015). Measuring public opinion with social media data. In L. R. Atkeson & R. M. Alvarez (Eds.), *The Oxford handbook of polling and polling methods* (pp. 555–582). Oxford University Press. <https://doi.org/10.1093/oxfordhb/9780190213299.013.3>.
33. Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling*. Springer-Verlag. <https://doi.org/10.1007/978-1-4614-6849-3>.
34. Kulldorff, M. (1997). A spatial scan statistic. *Communications in Statistics - Theory and Methods*, 26(6), 1481–1496. <https://doi.org/10.1080/03610929708831995>.
35. Kuncheva, L. I., & Whitaker, C. J. (2003). Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Machine Learning*, 51(2), 181–207.
36. Lam, S.K., Pitrou, A., Seibert, S. (2015). Numba: A LLVM-based python JIT compiler. In *Proceedings of the Second Workshop on the LLVM Compiler Infrastructure in HPC, LLVM '15*. Association for Computing Machinery. <https://doi.org/10.1145/2833157.2833162>
37. Liu, P., & De Sabbata, S. (2021). A graph-based semi-supervised approach to classification learning in digital geographies. *Computers, Environment and Urban Systems*, 86, 101583. <https://doi.org/10.1016/j.compenvurbsys.2020.101583>.
38. Loper, E., Bird, S. (2002). NLTK: the Natural language Toolkit. In *Proceedings of the ACL-02 Workshop on Effective tools and methodologies for teaching natural language processing and computational linguistics* (vol. 1, pp. 63–70). Association for Computational Linguistics. <https://doi.org/10.3115/1118108.1118117>
39. McLafferty, S. (2015). Disease cluster detection methods: Recent developments and public health implications. *Annals of GIS*, 21(2), 127–133. <https://doi.org/10.1080/19475683.2015.1008572>.
40. Mellon, J., & Prosser, C. (2017). Twitter and Facebook are not representative of the general population: Political attitudes and demographics of British social media users. *Research & Politics*, 4(3), 1–9. <https://doi.org/10.1177/2053168017720008>.
41. Minamisava, R., Nouer, S. S., De Moraes Neto, O. L., Melo, L. K., & Andrade, A. (2009). Spatial clusters of violent deaths in a newly urbanized region of Brazil: Highlighting the social disparities. *International Journal of Health Geographics*, 8(1), 66. <https://doi.org/10.1186/1476-072X-8-66>.
42. Mohammad, S., Kiritchenko, S., Sobhani, P., Zhu, X., Cherry, C. (2016). SemEval-2016 task 6: detecting stance in Tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)* (pp. 31–41). Association for Computational Linguistics. <https://doi.org/10.18653/v1/s16-1003>
43. Murphy, K. P. (2012). *Machine learning: A probabilistic perspective*. The MIT Press.
44. Olteanu, A., Castillo, C., Diaz, F., & Kiciman, E. (2019). Social data: Biases, methodological pitfalls, and ethical boundaries. *Frontiers in Big Data*, 2, 13. <https://doi.org/10.3389/fdata.2019.00013>.
45. Opitz, D., & Maclin, R. (1999). Popular ensemble methods: An empirical study. *Journal of Artificial Intelligence Research*, 11(1), 169–198. <https://doi.org/10.1613/jair.614>.
46. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Müller, A., Nothman, J., Louppe, G., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, É.: Scikit-learn: machine learning in Python.

- Journal of Machine Learning Research*, 12(85), 2825–2830 (2012). <http://jmlr.org/papers/v12/pedregosal1a.html>
47. Rajadesingan, A., & Liu, H. (2014). Identifying users with opposing opinions in Twitter debates. Lecture notes in computer science. In W. Kennedy, N. Agarwal, & Y. SJ (Eds.), *Social computing, behavioral-cultural modeling and prediction* (pp. 153–160). Springer. https://doi.org/10.1007/978-3-319-05579-4_19.
 48. Re, M., & Valentini, G. (2012). Ensemble methods: A review. In M. J. Way, J. D. Scargle, K. M. Ali, & A. N. Srivastava (Eds.), *Advances in machine learning and data mining for astronomy* (pp. 563–594). Chapman and Hall/CRC.
 49. Rokach, L. (2010). Ensemble-based classifiers. *Artificial Intelligence Review*, 33(1–2), 1–39. <https://doi.org/10.1007/s10462-009-9124-7>.
 50. Sloan, L. (2017). Who tweets in the United Kingdom? Profiling the Twitter population using the British Social Attitudes Survey 2015. *Social Media & Society*, 3(1), 1–11. <https://doi.org/10.1177/2056305117698981>.
 51. Sloan, L., & Morgan, J. (2015). Who tweets with their location? Understanding the relationship between demographic characteristics and the use of geoservices and geotagging on twitter. *PLoS One*, 10(11), 1–5. <https://doi.org/10.1371/journal.pone.014220>.
 52. So, H. C., Pearl, D. L., von Königsłow, T., Louie, M., Chui, L., & Svenson, L. W. (2013). Spatio-temporal scan statistics for the detection of outbreaks involving common molecular subtypes: Using human cases of Escherichia coli O157: H7 provincial pfge pattern 8 (National Designation Excai.0001) in alberta as an example. *Zoonoses and Public Health*, 60(5), 341–348. <https://doi.org/10.1111/zph.12012>.
 53. Sridhar, D., Getoor, L., Walker, M. (2015). Collective stance classification of posts in online debate forums. In *Proceedings of the Joint Workshop on Social Dynamics and Personal Attributes in Social Media* (pp. 109–117). Association for Computational Linguistics. <https://doi.org/10.3115/v1/w14-2715>
 54. Taylor, S. R., Arrowsmith, S. J., & Anderson, D. N. (2010). Detection of short time transients from spectrograms using scan statistics. *Bulletin of the Seismological Society of America*, 100(5A), 1940–1951. <https://doi.org/10.1785/0120100017>.
 55. Tsou, M., & Leitner, M. (2013). Visualization of social media: Seeing a mirage or a message? *Cartography and Geographic Information Science*, 40(2), 55–60. <https://doi.org/10.1080/15230406.2013.776754>.
 56. Tumasjan, A., Sprenger, T. O., Sandner, P. G., & Welpe, I. M. (2011). Election forecasts with Twitter: How 140 characters reflect the political landscape. *Social Science Computer Review*, 29(4), 402–418. <https://doi.org/10.1177/0894439310386557>.
 57. Tumasjan, A., Sprenger, T., Sandner, P., Welpe, I. (2010). What 140 characters reveal about political sentiment. In *Proceedings of the International AAAI Conference on Web and Social Media*. AAAI Publications
 58. Wakefield, J., & Kim, A. (2013). A Bayesian model for cluster detection. *Biostatistics*, 14(4), 752–765. <https://doi.org/10.1093/biostatistics/kxt001>.
 59. Weaver, I., Williams, H., Cioroianu, I., Williams, M., Coan, T., & Banducci, S. (2018). Dynamic social media affiliations among UK politicians. *Social Networks*, 54, 132–144. <https://doi.org/10.1016/j.socnet.2018.01.008>.
 60. Williams, C., Gulati, G. (2008). What is a social network worth? Facebook and vote share in the 2008 presidential primaries. In *Annual Meeting of the American Political Science Association* (pp. 1–17).
 61. Yıldız, D., Munson, J., Vitali, A., Tinati, R., & Holland, J. A. (2017). Using Twitter data for demographic research. *Demographic Research*, 37, 1447–1514. <https://doi.org/10.4054/DemRes.2017.37.46>.
 62. Gil de Zúñiga, H., Barnidge, M., & Diehl, T. (2018). Political persuasion on social media: A moderated moderation model of political discussion disagreement and civil reasoning. *The Information Society*, 34, 302–315. <https://doi.org/10.1080/01972243.2018.1497743>.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Emiliano del Gobbo¹ · Lara Fontanella² · Sara Fontanella³ · Annalina Sarra⁴

Emiliano del Gobbo
emiliano.delgobbo@unich.it

Lara Fontanella
lara.fontanella@unich.it

Sara Fontanella
s.fontanella@imperial.ac.uk

- ¹ Department of Neuroscience, Imaging and Clinical Sciences, G. d'Annunzio University Chieti-Pescara, Chieti, Italy
- ² Department of Legal and Social Science, G. d'Annunzio University Chieti-Pescara, Chieti, Italy
- ³ Faculty of Medicine, National Heart and Lung Institute, Imperial College of London, London, UK
- ⁴ Department of Philosophical, Pedagogical and Economic-Quantitative Sciences, G. d'Annunzio University Chieti-Pescara, Chieti, Italy