


Geotagging Matters? The Interplay of Space and Place in Politicized Online Social Media Networks

Adrian Tear 

School of the Environment, Geography and Geosciences, University of Portsmouth, UK

Voting results in marginal constituencies often determine wider political outcomes. It is now apparent that key electorates in these areas have been geo-behaviourally targeted by elaborate operations intended to manipulate results through advertising, (mis)information, and/or ‘fake news’ disseminated via online social networks. Attempts to track the geographical diffusion of cyber politicking are hindered by incomplete geospatial referencing in social media (meta)data. Just about 1–2% of publicly posted Twitter tweets, and even fewer Facebook posts, are typically ‘geotagged’ with Latitude and Longitude coordinates. Many more records (about 25%) make toponymic mention of place. This paper examines about 8 million social media interactions, over 350,000 of which are geotagged, created during the 2012 US Presidential Election and the 2014 Scottish Independence Referendum campaigns, to assess the interplay of space and place in online communications. Results of text and data-mining show that coordinate-geotagging users of Twitter and Facebook, (a) make fewer references to place in their message text, (b) link to articles making fewer mentions of place in their content, and (c) make far fewer links to external content than their non-coordinate-geotagging peers. Despite providing some valuable geospatial information, coordinate-geotagged interactions offer only an inadequate proxy for tracking the spread of all places, linked content, or (mis)information shared online. As Twitter retires its tweet spatialization functionality, new regulatory and technical responses together with a better understanding of place will be required if electoral officials, platform operators, and researchers are to more easily and accurately identify nefarious content targeting specific areas as well as specific individuals during democratic elections.

Keywords: social media; geotagging; natural language processing; place detection

History: received on 7 October 2019; accepted on 27 November 2019; published on 27 January 2020

1 Introduction

Online social networks, politics, and Big Data are in the news. Alarming revelations surrounding Cambridge Analytica’s misuse of data for political marketing purposes prompted a US Congressional Committee to investigate data usage, sharing, and privacy policies at Facebook (U.S. House of Representatives, 2018), eventually leading to the imposition of a ‘record-breaking’ \$5 billion fine from the US Federal Trade Commission (Shepardson, 2019). Political campaigning using advanced behavioural and psychographic targeting, alongside geographical micro-marketing designed to bring out or win over key voters (Albright, 2017), may even have affected the outcome of the 2016 US Presidential Election, a contest which Cambridge Analytica claimed to have ‘won’ for Donald Trump (Lewis and Hilder, 2018).

Steiger et al. (2015, p. 816) note that coordinate-geotagged online social network (OSN) interactions, sourced primarily from Twitter, have demonstrated high degrees of utility in ‘research on event detection

A Tear (2020): *Geotagging Matters? The Interplay of Space and Place in Politicized Online Social Media Networks*. In: FB Mocnik and R Westerholt (eds.), *Proceedings of the 2nd International Symposium on Platial Information Science (PLATIAL'19)*, pp. 61–72

<https://doi.org/10.5281/zenodo.3628879>



Second International Symposium on Platial Information Science (PLATIAL'19)
Coventry, UK; 5–6 September 2019

Copyright © by the author(s). Licensed under Creative Commons Attribution 4.0 License.

[particularly in the] investigation of abnormal spatial, temporal and semantic tweet frequencies [surrounding] disaster and emergency [situations]'. The current research uses a mixture of coordinate-geotagged and non-coordinate-geotagged social media data from Facebook and Twitter, collected during the 2012 US Presidential Election (US2012) and the 2014 Scottish Independence Referendum (SCOT2014), to determine whether a similarly high level of analytical utility may be observed in political contexts. Understanding how different classes of social media users imprint their communications with place or, less frequently, space – or consume, link to, and share third party Uniform Resource Locator (URL) content imprinted with place – is essential when attempting to accurately track the downstream diffusion of deliberately geo-targeted political advertising.

2 Geographical Characteristics of Social Media Data

2.1 Space and Place

Senses of known-place(s), affirmed-place(s), and space(s), some of which may be accompanied by apparently accurate latitude and longitude coordinates, are often highly conflated in social media data. Users of Twitter, e.g., when registering, are asked 'Where in the world are you?' (Hecht et al., 2011) and may just as reasonably answer 'BRICK city bitch' or 'Somewhere, Overthere' as 'Concord, NC' or 'iPhone: 40.699490,-73.891556'. Difficulties inherent in identifying and parsing Potential Geographic Information (PGI) in free-form social media message text and associated (meta)data are amplified considerably when, as in this research, place-based geographical references must be detected computationally. Consequently, and as huge data volumes preclude individual human examination of over 8 million social media interactions, necessarily focused definitions of 'space' and 'place' are adopted in this research:

- **Space** - refers to geographically and explicitly *locational* data, i.e., to a point defined by a pair of latitude and longitude coordinates.
- **Place** – refers to computationally-identifiable geographical references in text, i.e., to *toponymic* place names, e.g., of towns, cities, counties, states, countries, etc.

Space, where it exists in social media interaction (meta)data, may generally be regarded unambiguously; the latitude and longitude coordinates of a user's location have been recorded alongside their message text by a Global Positioning System (GPS) equipped mobile device just at the moment of message creation (Li et al., 2016). Place, in social media data, retains many of the elements of ambiguity identified by Tuan (1977) and other geographical theorists but is referenced, on the admittedly narrower grounds adopted here, much more widely in message text, metadata, and linked/shared content than space.

2.2 Geotagging Rates and Behaviour

Typically, and somewhat unfortunately for geographers, only small percentages of social media interactions are geotagged with latitude and longitude coordinates. Leetaru et al. (2013) report that just 1.6% of about 1.5 billion Twitter interactions analysed in their study contained 'Exact locations'. Slightly higher rates have been reported elsewhere (Croitoru et al., 2013) with variability attributed to event type (e.g., an elevated 16% following the Fukushima nuclear disaster in Japan), cultural practice (e.g., some nations use OSNs more frequently than others), and differing technological factors (e.g., smartphone adoption rates). While most users' mobile devices are perfectly capable of imprinting coordinates alongside their OSN posts, geotagging is an 'opt-in' feature which users must explicitly enable in their software application (Sui, 2017). Few users choose to deliberately activate geotagging facilities (Tasse et al., 2017) and, consequently, most mapping and geographical analyses of social media interactions are enabled not by the majority of OSN users, but by a *distinct minority* who choose to post with coordinates.

The current research questions whether there is an over-reliance on 'geosocial' data deposited by just about 1–2% of all social media users and whether expressions of 'place' in message text and linked/shared content are highly correlated with 'space' in coordinate-geotagged OSN interactions.

The work addresses a fundamental question: who makes, or links to external content containing, the most place-based references on social media networks, i.e., who is among the coordinate-geotagging or non-coordinate-geotagging users of these sites? Answering this question, to determine the toponymical representativeness of coordinate-geotagging users, helps determine whether or not this minority group may be used as ‘markers’ to accurately and spatially, through their latitude and longitude coordinates, trace the geographical diffusion of online opinion, and/or (mis)information.

3 Data Subjects

The subjects of this research are politically discursive social media messages; 8,196,380 OSN interactions created by 2,436,167 individual users of Twitter and Facebook in a roughly 90:10 ratio during two case study electoral events. Four sampled ‘streams’ of social media data were collected using the DataSift platform, a content aggregation system capable (at the time) of accessing Twitter’s full ‘Firehose’ of tweets, together with messages publicly-posted on Facebook. In each case, the source files in both Comma Separated Values (CSV) and JavaScript Object Notation (JSON) formats (ECMA International, 2017) were loaded into the Oracle 12c Relational Database Management System (RDBMS). The data collected for each event are further described below.

3.1 2012 US Presidential Election

During the two-month run up to the US Presidential Election of 6 November 2012, 1,661,402 Twitter tweets and 57,265 Facebook posts were sampled from contemporaneous OSN communications. Three sample sets from Twitter and Facebook were recorded, filtered on a range of identical text search terms, and controlled for explicit presence/absence of geographical coordinates, extent (country), and/or language. The interactions were filtered on any case-insensitive words or phrases matching those illustrated in Figure 1, which shows the contribution of each search term to the sample, noting that some terms (e.g., ‘US President’ and ‘Obama’) may have appeared multiple times within message text. Despite filtering on 15 search terms the top 3 terms account for 78.97% of the interactions sampled in the data set. Filtering selected for inclusion mainly on candidate surname, forename, or a combination of the two. In both political events (see also Figure 2) the top two terms usefully, and reasonably evenly, select interactions for the major protagonists in both contests. The three US2012 streams were recorded, stored, and downloaded from DataSift’s servers. The data set consists of 1,718,667 rows across three files each with up to 146 fields.

3.2 2014 Scottish Independence Referendum

The necessity for a second case study was prompted by early analysis of data from the first. The US2012 data set consisted of three sampled streams. Sampling was used to restrict data volumes and control costs, but also resulted in an ‘incomplete’ data set where the full network graph of tweeting, mentioning, and retweeting could not be examined. Consequently, a second real-time recording of OSN interactions during the much longer run up to the 2014 Scottish Independence Referendum was started on 18 September 2013, exactly one year before the vote was due to take place.

Scotland, with a much smaller population (about 5 million) than the US (about 320 million), was thought unlikely to create the sorts of OSN data volumes a 1:1 sampled US recording would have generated. Interactions were filtered on any case-insensitive words or phrases matching those illustrated in Figure 2, which also shows the contribution of each search term to the sample, again noting that some terms will have appeared multiple times within message text. Deliberate misspellings (‘independance’, etc.) were incorporated in the filter design as misspellings are a common feature of OSN communications (Deitrick and Hu, 2013).

Despite Scotland’s small population size, worldwide interest in the outcome of the referendum, coupled with the longer-running nature of the recording, eventually resulted in the collection of about 6.5 million OSN interactions. The top 3 of 27 search terms account for 63.25% of interactions sampled in the data set. Filtering has selected for inclusion on a mix of First Minister (and Vote Yes leader) Alex Salmond’s surname, the campaign slogan (‘Better Together’) of the Vote No (remain united) coalition, where no one political figure spearheaded the campaign, and the abbreviation ‘SNP’ (Scottish

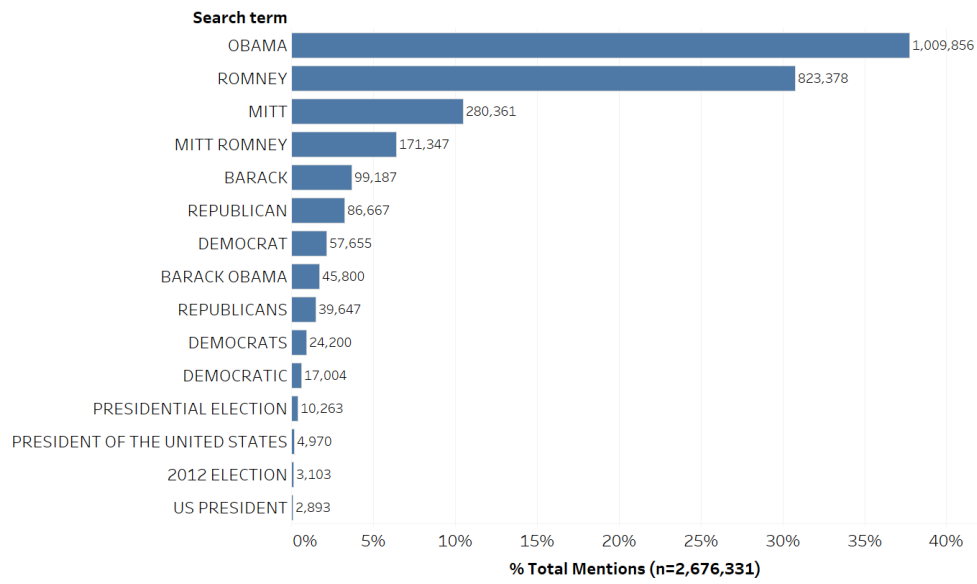


Figure 1: US2012: Search terms used. Numeric and percentage contribution to OSN interactions sampled (2,676,331 total mentions of search terms in 1,718,667 messages).

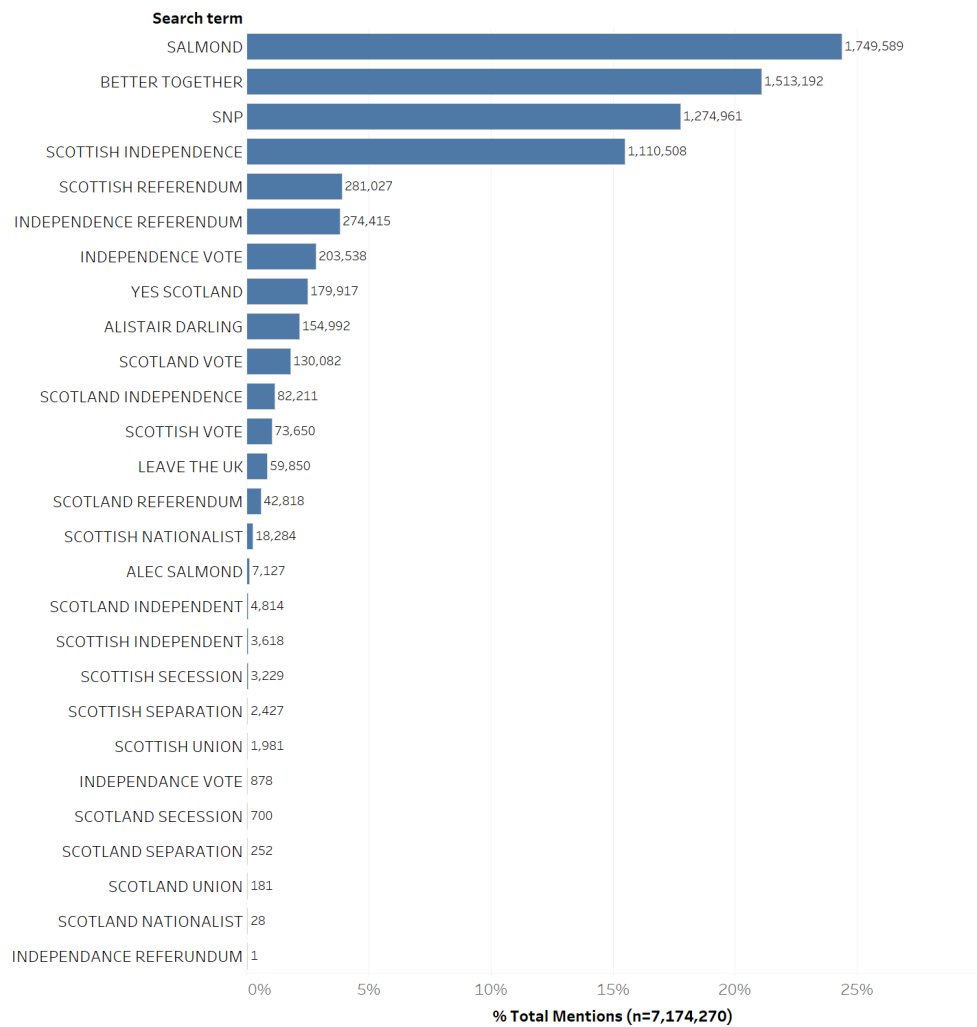


Figure 2: SCOT2014: Search terms used. Numeric and percentage contribution to OSN interactions sampled (7,174,270 total mentions of search terms in 6,477,713 messages).

Nationalist Party), the name of the pro-independence party in Scotland. As in the US2012 event, the top two terms usefully, and reasonably evenly, select interactions for the major protagonists in the 2014 Scottish Independence Referendum and are thought to offer a balance of messages for inclusion in the sample for both opposing sides of the political debate.

4 Data Analysis

Stock (2018, p. 209) has noted that ‘During the last ten years, a large body of research extracting and analysing geographic data from social media has developed’. Reviewing 690 papers accessing 20 social media platforms she states that ‘a wide array of [...] approaches have been developed, with methods that extract place names from message text providing the highest accuracy’. The three NLP packages used in this research to detect toponymic mentions in message text and, additionally, in linked/shared URL content are summarized below.

- **GATE** – The General Architecture for Text Engineering’s TwitIE processor, a Twitter Information Extraction engine, and ‘open-source NLP pipeline customized to microblog text at every stage’ (Bontcheva et al., 2013), running on the highly-scaleable GATEcloud.net system (Tablan et al., 2013), was used to identify various entities, e.g., toponyms, people, or organizations, mentioned in message text. GATE cannot, yet, append coordinates to detected platial references.
- **CLAVIN** – The Cartographic Location and Vicinity Indexer ‘extracts location names from unstructured text and resolves them against the GeoNames gazetteer to produce data-rich geographic entities’ (Berico-Technologies, 2017). As a specialist ‘geo-parser’, CLAVIN appends GeoNames-derived coordinate pairs to toponymic place names detected in interaction message text. CLAVIN cannot, however, identify other entity types (e.g., people, organizations, etc.) in text.
- **AlchemyAPI** – AlchemyAPI, now re-branded as Watson Natural Language Understanding (IBM, 2017), is Cloud-hosted, commercial software. Academic usage is restricted to 30,000 ‘daily transactions’, a processing restriction which prevented the timely processing of all about 8 million OSN interactions. Instead, a sample of 311,575 messages were processed, alongside 641,472 distinct URLs shared 3,485,840 times in the research data corpus, just 45,492 of which were shared by coordinate-geotagging users.

All about 8 million OSN messages were passed through TwitIE on GATEcloud and CLAVIN-rest. AlchemyAPI was used to detect toponyms in the linked/shared content of OSN users. The analytical pipeline, centred around an Oracle 12c database, relied upon both Cloud-hosted and locally virtualized computing resources running CentOS or Ubuntu Linux under Oracle VirtualBox on a Windows 10 host.

5 Results

Despite adopting technically different solutions to the ‘challenge’ of toponymic place detection in free form text (Li et al., 2016) both GATEcloud and CLAVIN-rest, the two systems successfully used to text-mine all about 8 million social media messages in the research data corpus, produced highly comparable results.

Table 1 shows the number of locations resolved by GATEcloud in interaction message text. The main entity type of interest returned in GATEcloud JSON following TwitIE processing (Bontcheva et al., 2013) is Location, in the locType key, which references indexed characters (i.e., the n-th characters in the message text) containing the detected location. These are coded as region, province, post, unknown, country, country_abbrev, city, airport, racecourse, or pre. These codings are mainly self-evident, except for pre and post which refer to first/last matches to parts of a location such as ‘Mount’, ‘East’, ‘Cape’, ‘Isle of’, etc., which co-occur with a proper noun. TwitIE on GATEcloud detects locType in the message text of 263,296 (15.32% of all) US2012 interactions, identifying a further 2,088,788 messages containing locations (32.25% of all) in the SCOT2014 data set. The ratio of resolved locations per interaction is higher in the larger and more recent SCOT2014 data set and is significantly higher for message text sourced from Facebook.

Table 1: GATEcloud place detection results. US2012/SCOT2014: Number of resolved locations detected by GATEcloud in Facebook (FB), Twitter tweet (TW), and Twitter retweet (RT) interactions.

	US2012	Ratio	SCOT2014	Ratio	Total
FB resolved locations	25,405	1.90	2,097,506	5.66	2,122,911
FB n interactions	13,341		370,774		384,115
TW resolved locations	153,085	1.23	1,087,698	1.31	1,240,783
TW n interactions	123,960		833,235		957,195
RT resolved locations	151,899	1.21	1,130,344	1.28	1,282,243
RT n interactions	125,995		884,779		1,010,774
Total Resolved	330,389	1.25	4,315,548	2.07	4,645,937
Total Interactions	263,296		2,088,788		2,352,084

Table 2: CLAVIN-rest place detection results. US2012/SCOT2014: Number of resolved locations detected by CLAVIN-rest in Facebook (FB), Twitter tweet (TW), and Twitter retweet (RT) interactions.

	US2012	Ratio	SCOT2014	Ratio	Total
FB resolved locations	22,019	1.80	1,573,145	3.98	1,595,166
FB n interactions	12,199		395,112		407,311
TW resolved locations	120,491	1.19	832,941	1.25	953,433
TW n interactions	101,664		667,263		768,927
RT resolved locations	119,979	1.18	856,383	1.22	976,363
RT n interactions	101,599		700,567		802,166
Total Resolved	262,489	1.22	3,262,469	1.85	3,524,959
Total Interactions	215,462		1,762,942		1,978,404

Table 2 shows the number of locations resolved by CLAVIN-rest in interaction message text. It is apparent, as with TwitIE on GATEcloud, that 45.25% of resolved locations ($n = 1,595,166$) detected by CLAVIN-rest stem from 407,311 geoparsed OSN interactions sourced from Facebook (i.e., 20.59% of all interactions processed), most of which ($n = 395,112$) were collected during the 2014 Scottish Independence Referendum. Larger numbers of Twitter tweet ($n = 768,927$) and retweet ($n = 802,166$) interactions (total $n = 1,571,093$) were geoparsed by CLAVIN-rest but yielded a total of only 1,929,794 ($n = 953,433$ and $n = 976,363$, respectively) resolved locations, 87.54% of which ($n = 1,689,324$) were found in the SCOT2014 data set which features a higher proportion of Twitter retweets. Most locations resolved in retweets will, of course, be duplicates of locations found in the originating tweet.

Welch Two Sample T-tests, calculated using R (The R Foundation, 2018), compare distributions of numbers of NLP-detected toponymic mentions per interaction, or per user, for non-coordinate-geotagged/ing and coordinate-geotagged/ing interactions/users (Table 3). During the US2012 event, in 9 out of 20 cases like-for-like comparisons of geoparser, OSN source and level for non-coordinate-geotagged message text or linked/shared URL content against coordinate-geotagged corollaries are statistically significant with >95% confidence. The null hypothesis, that there is no difference in the distribution of numbers of toponymic mentions detected in OSN message text or linked/shared URL content by OSN source for the given NLP/geoparsers at interaction and user levels, can be rejected. In most ($n = 6$) of these statistically significant comparisons non-coordinate-geotagged interactions or non-coordinate-geotagging users make more NLP/geoparser-detectable toponymic mentions in message text, or link to and share URLs having more detectable toponymic mentions in content, than their coordinate-geotagged or geotagging corollaries. Statistics for 6 cases (marked 'N/A' in Table 3) cannot be calculated in R for the US2012 data set as no coordinate-geotagged Facebook interactions are present. During the SCOT2014 event, statistical significance with >95% confidence is found in 18 of 20 like-for-like cases comparing numbers of toponymic detections by NLP/geoparser in message text and linked/shared URLs for Facebook, Twitter tweet, and retweet data at interaction and user levels. In over half ($n = 11$) of these statistically significant comparisons non-coordinate-geotagged interactions or non-coordinate-geotagging users make more NLP/geoparser-detectable toponymic mentions in message text, or link to and share URLs having more detectable toponymic mentions in content, than their

Table 3: Statistical results. Welch Two Sample T-tests, at interaction and user levels, comparing the number of resolved locations detected by the three NLP systems in coordinate-geotagged and non-coordinate-geotagged Facebook (FB), Twitter tweet (TW), and Twitter retweet (RT) message text and linked/shared content.

System	Level	Source	US2012		SCOT2014	
			<i>t</i>	<i>t</i> > ±2	<i>t</i>	<i>t</i> > ±2
GATEcloud (Messages)	Interaction	FB	N/A	N/A	1.03	✗
GATEcloud (Messages)	Interaction	TW	-32.54***	✓	3.37***	✓
GATEcloud (Messages)	Interaction	RT	0.44	✗	7.01***	✓
AlchemyAPI (Messages)	Interaction	TW	-1.04	✗	-3.55***	✓
CLAVIN-rest (Messages)	Interaction	FB	N/A	N/A	4.91***	✓
CLAVIN-rest (Messages)	Interaction	TW	3.92***	✓	2.99**	✓
CLAVIN-rest (Messages)	Interaction	RT	2.63**	✓	4.68***	✓
AlchemyAPI (Links)	Interaction	FB	N/A	N/A	1.10	✗
AlchemyAPI (Links)	Interaction	TW	1.29	✗	-29.85***	✓
AlchemyAPI (Links)	Interaction	RT	2.21*	✓	3.44***	✓
GATEcloud (Messages)	User	FB	N/A	N/A	83.69***	✓
GATEcloud (Messages)	User	TW	-34.48***	✓	69.32***	✓
GATEcloud (Messages)	User	RT	14.01***	✓	-13.52***	✓
AlchemyAPI (Messages)	User	TW	-20.24***	✓	-7.17***	✓
CLAVIN-rest (Messages)	User	FB	N/A	N/A	97.36***	✓
CLAVIN-rest (Messages)	User	TW	37.18***	✓	47.42***	✓
CLAVIN-rest (Messages)	User	RT	1.25	✗	-12.41***	✓
AlchemyAPI (Links)	User	FB	N/A	N/A	52.12***	✓
AlchemyAPI (Links)	User	TW	2.92**	✓	-33.01***	✓
AlchemyAPI (Links)	User	RT	0.25	✗	-7.20***	✓

* $P \leq 0.05$; ** $P \leq 0.01$; *** $P \leq 0.001$

coordinate-geotagged or coordinate-geotagging corollaries.

These findings are at odds with the research hypothesis that coordinate-geotagging users are the most geographically expressive of all OSN users. Although they do actively (or accidentally) coordinate-geotag their Twitter tweets or Facebook posts, this small group of social media users are not, in three important respects, representative of all OSN users. Of course, OSN users in general (Diaz et al., 2016), and geotagging users in particular (Sloan and Morgan, 2015), are not thought to be representative of the general population. During elections, they are likely to be even less so, probably being younger and living in urban areas and often, according to Barberá and Rivero (2015), exhibiting ‘extreme ideological preferences’. The results of a comprehensive analysis and cross-comparison of toponymic mentions detected in the message text and URL link shares of coordinate-geotagging and non-coordinate-geotagging users interacting online during two data-rich political case study events show that

1. coordinate-geotagging users make fewer toponymic mentions in message text than non-coordinate-geotagging users of two popular OSN platforms;
2. coordinate-geotagging users make far fewer URL link shares than non-coordinate-geotagging users, and
3. the content of URLs shared by coordinate-geotagging users makes fewer mentions of place than content shared by non-coordinate-geotagging users.

The research findings presented here imply that geographical outputs (such as point maps and counts or aggregations to larger areal units such as constituencies or states) based on searches for specific words, toponyms, #hashtags or @mentions in message text or URL link shares, which may readily be mapped using the interaction latitude and longitude coordinates of Twitter tweets or Facebook posts deposited by coordinate-geotagging users (or aggregated to wider areas using a GIS), are unlikely to be representative of the spread of all such content within online social networks.

6 Discussion

Just as Massey and Allen's *Geography Matters!* (1984) reaffirmed the relevance of geography in socio-spatial, environmental, political, and economic spheres, a conception of place clearly matters when individuals interact online using social media platforms. In the 2012 US Presidential Election and the 2014 Scottish Independence Referendum case studies examined here, about 3.5–4.5 million toponymic mentions have been identified in around one quarter of the about 8 million interactions in the research data corpus. Around one quarter of the about 7 million entities identified in about 650,000 distinct URLs – posted, tweeted, or retweeted about 3.5 million times – also contained toponymically identifiable content. Elections are peculiarly *geographic*, as well as *political*, events. It is, therefore, both unsurprising and reassuring to find that electorates and commentators make frequent geographical references online during electoral campaigns, and that many of these mentions refer to the 'swing' states or constituencies the ballot results of which typically shape wider political outcomes.

What then of geotagging; does it *matter*? Geotagging is a relatively recent socio-technological phenomenon, primarily enabled by the worldwide proliferation and usage of GPS-equipped mobile, or smartphone, devices. The increasingly large volumes of Ambient (and/or Volunteered) Geospatial Information now available (Goodchild, 2007; Stefanidis et al., 2013b) offer new research opportunities for scholars in geography and related social science disciplines. Increased scrutiny of 'Geo-social Networks' (Bahir and Peled, 2013), and the possibilities they afford for wider geographical analysis, are demonstrated by the growing number of academic articles and specialist journals published in the last decade or so, many of them cited in this paper. Geotagged photographic images publicly posted on Flickr have been used to combat wildlife poaching in protected areas and in criminological research (Lemieux, 2015). Geotagged social media and other 'Big Data' have been used to monitor natural disaster situations (Burns, 2018; Goodchild and Glennon, 2010). OpenStreetMap has been used in the study of the production and 'prosumption' of user-generated geographic Big Data (Cockayne, 2016). Human interaction data sourced from Twitter and, to a lesser extent, Facebook have been used, seemingly, to 'do everything'; from monitoring earthquakes (Crooks et al., 2013) to tracking riots (Bonilla and Rosa, 2015; Crampton et al., 2013), helping to demarcate urban areas (Yin et al., 2017), and much else besides (see Kapoor et al., 2018, for a recent and comprehensive summary of application areas). This proliferation of research activity, e.g., '[delineating] city cores, [gaining] insights into travel plans and tourism, [characterizing] urban landscapes, [studying] global migrations or [identifying] mobility patterns', has also been identified by Rzeszewski and Beluch (2017), who go on to note that 'much less attention' has been devoted to studies investigating the 'subgroup of users that produce (or rather contribute since they may not be aware of it) [...] ambient geospatial information' on social media networks.

7 Conclusion

Bespoke geographical targeting campaigns, as developed by Cambridge Analytica (Albright, 2017), may exploit toponymic references found in users' self-reported 'Location' fields (Hecht et al., 2011), toponymic references found in users' publicly-posted message text (Stock, 2018), and/or latitude and longitude coordinates deposited in OSN metadata when users optionally choose to 'geotag' their social media posts (Kumar et al., 2014). On Facebook (2018), advertisements can be displayed to all users in (or within a radius of) selected locations, users who live in those locations (also 'validated' by Internet Protocol, IP, address), users currently in those locations ('as determined only by mobile device') or people just passing through ('as determined by mobile device [when it is] greater than 100 miles from their stated home location from their Facebook profile'). Other major social media websites operated by Google, Instagram, Snapchat, Twitter, and YouTube offer broadly similar facilities to advertisers – or political campaigners – using their services. All also offer targeting on age, basic demographics (e.g., gender), and, in several cases, on more advanced behavioural or similarity traits (e.g., interests and 'Lookalike Audiences' in Facebook's case). It is currently unclear whether recent attempts to distort the outcome of democratic elections through geo-behavioural targeting have shown clear 'monolithic effects [but] the impact of social media in political campaigning around the world is undeniable' (Dimitrova and Matthes, 2018). It is also, unfortunately for concerned citizens, electoral regulators, and others, much easier to set up a geo-targeted online political advertising campaign than it is for third parties to

monitor the downstream geographical diffusion or effectiveness of such communications.

Coordinate-geotagging users are much less widely followed than others on OSNs and, somewhat counter-intuitively, express themselves less geographically than others in their message text and through the URLs they choose to link to and share. The comprehensive analysis of social media interactions presented here reveals important differences in the posting behaviour of coordinate-geotagging and non-coordinate-geotagging users during two political case study events. The more fundamental question, whether geotagging matters, is not so easily answered. To a professional geographer, the vast number of coordinate pairs now deposited online by social media users appears highly propitious. On a massive scale, arguably for the first time in human history, it is possible to know who is saying what, when, and where. However, as detailed earlier, and remarked upon by Paraskevopoulos and Palpanas (2016, p. 1), 'only a very small percentage of [OSN] posts are geotagged, which significantly restricts the applicability and utility of [many] applications'. Low rates of coordinate-geotagging in OSN data, and the unrepresentativeness of coordinate-geotagging users, definitely limit the 'applicability' of any analyses based solely upon geotagging users' message text, metadata, or spatial location in political contexts. When examining politicized communications made on social media networks, or determining how political opinion or (mis)information may be geographically tracked on these platforms, it appears that *geo* matters, but *tagging* matters much less. As Twitter, the most widely studied OSN (Stock, 2018; Tufekci, 2014), retires its tweet geotagging functionality (Leetaru, 2019) and Facebook, one of the world's key outlets for targeted political advertising (Lilleker et al., 2015), restricts data access in an operational environment ripe for regulation (McKinnon and Seetharaman, 2018), the need to efficiently and accurately detect place, in potentially dwindling volumes of social media message text and metadata, seems certain to increase.


Supplementary Information

This paper is an abridged synopsis of Adrian Tear's doctoral thesis (2018), supervised by Professors Richard Healey and Humphrey Southall in the Department of Geography at the University of Portsmouth. The full text of the thesis is available at [https://researchportal.port.ac.uk/portal/en/theses/geotagging-matters\(90524d65-f62c-4bfd-b201-642a90368f96\).html](https://researchportal.port.ac.uk/portal/en/theses/geotagging-matters(90524d65-f62c-4bfd-b201-642a90368f96).html).

Acknowledgements

The results reported here rely upon computational analysis of over 8 million social media messages publicly posted by just under 2.5 million users of Twitter and Facebook. I thank these users, the two online social network platforms which enable(d) data access and DataSift, without whom it would have been impossible to download and analyse so many individual, albeit terse and often ungrammatical, thoughts and opinions. Professor Kalina Bontcheva and colleagues at the University of Sheffield provided invaluable assistance in the use of GATE and GATEcloud natural language processing software. Further technical assistance was provided by staff at IBM Watson, Oracle Corporation, and the University of Portsmouth.

ORCID

Adrian Tear  <https://orcid.org/0000-0001-5028-2622>

References

- Albright, Jonathan: *Cambridge Analytica: the geotargeting and emotional data mining scripts*. <https://medium.com/tow-center/cambridge-analytica-the-geotargeting-and-emotional-data-mining-scripts-bcc3c428d77f>, 2017. Retrieved 27 November 2019
- Bahir, Eitan and Peled, Ammatzia: *Identifying and tracking major events using geo-social networks*. *Social Science Computer Review*, 31(4), 2013, 458–470. doi: 10.1177/0894439313483689
- Barberá, Pablo and Rivero, Gonzalo: *Understanding the political representativeness of Twitter users*. *Social Science Computer Review*, 33(6), 2015, 712–729. doi: 10.1177/0894439314558836

- Berico-Technologies: *About clavin*. <https://clavin.bericotechnologies.com/about-clavin/>, 2017. Retrieved 27 November 2019
- Bonilla, Yarimar and Rosa, Jonathan: *#Ferguson: digital protest, hashtag ethnography, and the racial politics of Social Media in the United States*. *American Ethnologist*, 42(1), 2015, 4–17. doi: 10.1111/amet.12112
- Bontcheva, Kalina; Derczynski, Leon; Funk, Adam; et al.: *TwitIE: an open-source information extraction pipeline for microblog text*. Proceedings of the International Conference on Recent Advances in Natural Language Processing, 2013, 83–90
- Burns, Ryan: *Datafying disaster: institutional framings of data production following super-storm sandy*. *Annals of the American Association of Geographers*, 108(2), 2018, 569–578. doi: 10.1080/24694452.2017.1402673
- Cockayne, Daniel G: *Affect and value in critical examinations of the production and 'prosumption of Big Data*. *Big Data and Society*, 3(2), 2016, 1–11. doi: 10.1177/2053951716640566
- Crampton, Jeremy W; Graham, Mark; Poorthuis, Ate; et al.: *Beyond the geotag: situating 'big data' and leveraging the potential of the geoweb*. *Cartography and Geographic Information Science*, 40(2), 2013, 130–139. doi: 10.1080/15230406.2013.777137
- Croitoru, Arie; Crooks, Andrew; Radzikowski, Jacek; and Stefanidis, Anthony: *Geosocial gauge: a system prototype for knowledge discovery from Social Media*. *International Journal of Geographical Information Science*, 27(12), 2013, 2483–2508. doi: 10.1080/13658816.2013.825724
- Crooks, Andrew; Croitoru, Arie; Stefanidis, Anthony; and Radzikowski, Jacek: *#Earthquake: Twitter as a distributed sensor system*. *Transactions in GIS*, 17(1), 2013, 124–147. doi: 10.1111/j.1467-9671.2012.01359.x
- Deitrick, William and Hu, Wei: *Mutually enhancing community detection and sentiment analysis on Twitter networks*. *Journal of Data Analysis and Information Processing*, 1(3), 2013, 19–29. doi: 10.4236/jdaip.2013.13004
- Diaz, Fernando; Gamon, Michael; Hofman, Jake M; Kiciman, Emre; and Rothschild, David: *Online and social media data as an imperfect continuous panel survey*. *PLOS ONE*, 11(1), 2016, e0145406. doi: 10.1371/journal.pone.0145406
- Dimitrova, Daniela V and Matthes, Jörg: *Social Media in political campaigning around the world: theoretical and methodological challenges*. *Journalism and Mass Communication Quarterly*, 95(2), 2018, 333–342. doi: 10.1177/1077699018770437
- ECMA International: *The JSON data interchange syntax*. Tech. rep., 2017. International Standard
- Facebook: *Location targeting*. <https://en-gb.facebook.com/business/a/location-targeting>, 2018. Retrieved 27 November 2019
- Goodchild, Michael F: *Citizens as sensors: the world of volunteered geography*. *GeoJournal*, 69(4), 2007, 211–221. doi: 10.1007/s10708-007-9111-y
- Goodchild, Michael F and Glennon, J Alan: *Crowdsourcing geographic information for disaster response: a research frontier*. *International Journal of Digital Earth*, 3(3), 2010, 231–241. doi: 10.1080/17538941003759255
- Hecht, Brent; Hong, Lichan; Suh, Bongwon; and Chi, Ed H: *Tweets from Justin Bieber's heart*. Proceedings of the 2011 Annual Conference on Human Factors in Computing Systems, 2011, 237–246. doi: 10.1145/1978942.1978976
- IBM: *Natural language understanding*. <https://console.stage1.bluemix.net/catalog/services/natural-language-understanding>, 2017. Retrieved 27 November 2019

- Kapoor, Kawaljeet Kaur; Tamilmani, Kuttimani; Rana, Nripendra P; et al.: *Advances in social media research: past, present and future*. Information Systems Frontiers, 20(3), 2018, 531–558. doi: 10.1007/s10796-017-9810-y
- Kumar, Shamanth; Morstatter, Fred; and Liu, Huan: *Twitter data analytics*. New York, NY, USA: Springer, 2014. doi: 10.1007/978-1-4614-9372-3
- Leetaru, Kalev: *Twitter's shift away from GPS-tagged tweets suggests it will be harder to fight 'deep fakes'*. Newspaper article in Forbes Magazine, 2019
- Leetaru, Kalev; Wang, Shaowen; Cao, Guofeng; Padmanabhan, Anand; and Shook, Eric: *Mapping the global Twitter heartbeat: the geography of Twitter*. First Monday, 18(5), 2013
- Lemieux, Andrew M: *Geotagged photos: a useful tool for criminological research?* Crime Science, 4, 2015, 3. doi: 10.1186/s40163-015-0017-6
- Lewis, Paul and Hilder, Paul: *Leaked: Cambridge Analytica's blueprint for Trump victory*. Newspaper article in The Guardian, 2018
- Li, Songnian; Dragicevic, Suzana; Castro, Francesc Anton; et al.: *Geospatial big data handling theory and methods: a review and research challenges*. ISPRS Journal of Photogrammetry and Remote Sensing, 115, 2016, 119–133. doi: 10.1016/j.isprsjprs.2015.10.012
- Lilleker, Darren G; Tenscher, Jens; and Štětka, Václav: *Towards hypermedia campaigning? Perceptions of new media's importance for campaigning by party strategists in comparative perspective*. Information, Communication and Society, 18(7), 2015, 747–765. doi: 10.1080/1369118X.2014.993679
- Massey, Doreen and Allen, John: *Geography matters!* Cambridge, UK: Cambridge University Press, 1984
- McKinnon, John D and Seetharaman, Deepa: *In Facebook hearings, lawmakers ramp up talk of regulation*. Newspaper article in The Wall Street Journal, 2018
- Paraskevopoulos, Pavlos and Palpanas, Themis: *Where has this tweet come from? fast and fine-grained geolocalization of non-geotagged tweets*. Social Network Analysis and Mining, 6, 2016, 89. doi: 10.1007/s13278-016-0400-7
- Rzeszewski, Michal and Beluch, Lukasz: *Spatial characteristics of Twitter users—toward the understanding of geosocial media production*. ISPRS International Journal of Geo-Information, 6(8), 2017, 236. doi: 10.3390/ijgi6080236
- Shepardson, David: *Facebook to pay record \$5 billion U.S. fine over privacy; faces antitrust probe*. <https://www.reuters.com/article/us-facebook-ftc/facebook-to-pay-record-5-billion-us-fine-over-privacy-faces-antitrust-probe-idUSKCN1UJ1L9>, 2019. Retrieved 27 November 2019
- Sloan, Luke and Morgan, Jeffrey: *Who tweets with their location? Understanding the relationship between demographic characteristics and the use of geoservices and geotagging on Twitter*. PLoS ONE, 10(11), 2015, e0142209. doi: 10.1371/journal.pone.0142209
- Stefanidis, Anthony; Crooks, Andrew; and Radzikowski, Jacek: *Harvesting ambient geospatial information from social media feeds*. GeoJournal, 78(2), 2013b, 319–338. doi: 10.1007/s10708-011-9438-2
- Steiger, Enrico; de Albuquerque, João Porto; and Zipf, Alexander: *An advanced systematic literature review on spatiotemporal analyses of Twitter data*. Transactions in GIS, 19(6), 2015, 809–834. doi: 10.1111/tgis.12132
- Stock, Kristin: *Mining location from social media: a systematic review*. Computers, Environment and Urban Systems, 71, 2018, 209–240. doi: 10.1016/j.compenvurbsys.2018.05.007

Sui, Daniel Z: *Understanding locational-based services: core technologies, key applications and major concerns*. In: Warf, Barney (ed.), *Handbook on Geographies of Technology*, Cheltenham, UK: Edward Elgar Publishing, 2017. 85–95

Tablan, Valentin; Roberts, Ian; Cunningham, Hamish; and Bontcheva, Kalina: *GATECloud.net: a platform for large-scale, open-source text processing on the cloud*. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 371(1983), 2013, 20120071. doi: 10.1098/rsta.2012.0071

Tasse, Dan; Liu, Zichen; Sciuto, Alex; and Hong, Jason I: *State of the geotags: Motivations and recent changes*. *Proceedings of the 11th International AAAI Conference on Web and Social Media*, 2017, 250–259

Tear, Adrian: *Geotagging matters? The role of space and place in politicised social media discourse*. Ph.D. thesis, University of Portsmouth, 2018

The R Foundation: *R: the R project for statistical computing*. <https://www.r-project.org/>, 2018. Retrieved 27 November 2019

Tuan, Yi-Fu: *Space and place. The perspective of experience*. Minneapolis, MN, USA: University of Minnesota Press, 1977

Tufekci, Zeynep: *Big questions for Social Media Big Data: representativeness, validity and other methodological pitfalls*. *Proceedings of the 8th International AAAI Conference on Weblogs and Social Media*, 2014, 505–514

U.S. House of Representatives: *Facebook: transparency and use of consumer data*. <https://energycommerce.house.gov/committee-activity/hearings/hearing-on-facebook-transparency-and-use-of-consumer-data-full-committee>, 2018. Retrieved 27 November 2019

Yin, Junjun; Soliman, Aiman; Yin, Dandong; and Wang, Shaowen: *Depicting urban boundaries from a mobility network of spatial interactions: a case study of Great Britain with geo-located Twitter data*. *International Journal of Geographical Information Science*, 31(7), 2017, 1293–1313. doi: 10.1080/13658816.2017.1282615