Conducting sentiment analysis on a broad topic: Analysing Brexit sentiment 18-months after the EU referendum

Introduction

Brexit has been a dominant topic of UK news, politics and public discussion since the result of the EU referendum in June 2016 decided that the UK would leave membership of the European Union. The issues voters were asked to consider were complex and confusing, and the Leave and Remain groups conducted highly negative campaigns which left the public feeling ill-informed and untrusting (BBC 2016), with a 51.89%/48.11% vote to leave the EU. During the 18 months since the referendum, campaign promises were withdrawn (Doré 2016), the vastly complicated process of leaving began (Rankin 2017) and the UK economy has been negatively affected (Partington 2017). Typical of modern politics, social media was a key platform for campaigning during the referendum and has continued to be a contemporary tool for analysing public opinion since the outcome. Sociologists have developed digital methods for social media analysis to understand feeling and meaning in social media data, such as sentiment analysis which works well in deriving meaning from varied text (Thelwall 2017, pp120).

In the following essay, I will suggest that it is legitimate to analyse social media with a view to understanding meaning across a broad topic, despite the limitation of reduced context affecting our ability to understand meaning (boyd and Crawford, pp670-671). Whilst we can apply existing digital techniques in scenarios where they are recognised to be most effective and multi-disciplinary techniques in others, I argue that we must find better techniques to be able to analyse social media data with varying degrees of context. Initially I will look at existing research, focussing especially on sentiment analysis (Liu 2013, Thelwall 2017) and the importance of context when analysing Big Data (boyd and

Crawford 2012, Seaver 2015). I will then describe the methods used to collect and analyse the data used for this essay, using Brexit as the broad topic to which I will apply sentiment analysis. Next I will draw findings from the data to demonstrate the shortcomings of sentiment analysis when analysing broad topics without a narrowly defined context. Finally, I will conclude by aligning my findings with existing academic research and considering how we might develop future techniques to enhance sentiment analysis of broad topics.

Literature review

In this section I will look at existing academic research about the challenges and shortcomings of analysing social media data, especially regarding sentiment analysis.

In this age of Big Data, sociologists are amongst the many academics and professionals “clamoring for access to the massive quantities of information produced by and about people, things, and their interactions” (boyd and Crawford 2012, pp663). However, whilst new techniques have been developed to further digital social research, it is recognised that potential limitations include sociologists’ limited technical skills, access to data, data collection constraints controlled by social media services or the tools used, and the complex issues of whether digital data represents wider society and how we determine the validity of the data collected (Lupton 2014, pp60-62). Furthermore, the analysis of data is usually to understand meaning and generate knowledge, yet boyd and Crawford challenge whether meaning can ever be retained with Big Data and argue that “taken out of context, Big Data loses its meaning” (2012, pp670). Moreover, Seaver expands the argument by drawing on broader disciplines such as anthropology, linguistics and philosophy and suggests that “taken out of context, *everything* loses its meaning” (2015, pp1104).

Quantitative techniques can work highly effectively with social media data, revealing networks and communities that might otherwise be invisible, but qualitative techniques are also required to understand how and why people do things and to find meaning from data (Marwick 2014, pp119). Sentiment analysis, or opinion mining (Liu 2012), is a digital technique developed to analyse the meaning of social media data as people "use sentiment to help convey meaning" (Thelwall 2017, pp119). Sentiment analysis programs use a dictionary or predefined understanding of text to calculate meaning from data and ideally provide a weighting and rationale to their sentiment calculation, making them ideal for use in social research (Thelwall 2014, pp91). However, to interpret meaning we must recognise "that meaning crucially depends on context" (Seaver 2015, pp1103).

Sociologists have developed multi-disciplinary methods for decades to achieve a greater understanding of context, such as surveying accurately defined audiences, manually coding collected data or transcribing and annotating audio surveys (Marres 2017, pp29). When analysing social media data, to retain the context we can use manual methods for small samples. For larger samples using automated techniques we must attempt to maintain the context primarily through digital methods. This can include carefully defined data collection, programmatic content analysis or time specificity. For instance, for a known event, comparing sentiment before the event and after the event could be used to draw conclusions about the impact of the event on the audience's mood (Thelwall 2014, pp92).

Content analysis is the method of applying textual or non-textual rules and analyses to enhance a dataset (Einspänner et al 2014, pp97), such as to remove irrelevant data or enrich data through categorisation. Content analysis rules can be applied programmatically once defined but it is often a manual process to define rules that strengthen the meaning or identify unwanted terms that weaken it. This is a skilled iterative process which "requires manual checking and domain expertise to resolve" (Vis 2013, section 4.2.1), and hence is a labour and skill intensive task for social media research. For

Twitter data, hashtags often provide means of categorising or labelling data (Einspänner et al 2014, pp100), but also tools such as Mozdeh provide features such as co-word association to aid with content analysis.

In the following section I will explain in detail the specific techniques used for retrieving, processing and analysing the Twitter data for this essay.

Data and methods

In this section I explain the methods used to query Twitter to generate a dataset using Mozdeh. I go on to describe the processing and preparation of the data using Mozdeh and Excel, the sentiment analysis using SentiStrength and finally to visualise the network using Gephi. All computation was undertaken on a virtual Windows instance on Google Cloud for improved performance and to avoid interruption of the continual data-collection process.

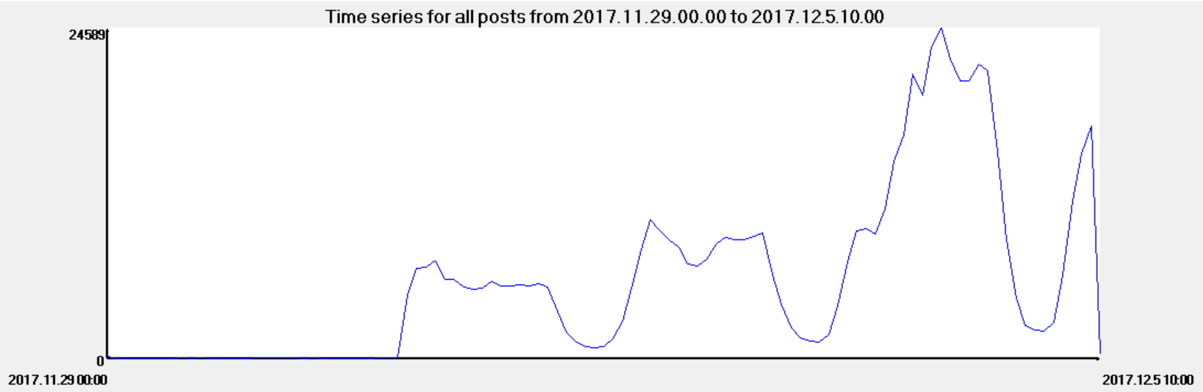
Mozdeh is a tool to retrieve and analyse Twitter data. Mozdeh submits a keyword query to the Twitter Search API which has a limit of 180 queries per 15 minutes and restricts results to tweets in the 7 days prior to the query. Mozdeh provides two data collection methods – a *one-off retrieval* returning up to 3200 tweets or a *continuous retrieval* returning potentially millions of tweets (depending how long the process is run). Since the EU Referendum, the Brexit topic on Twitter has become vast and unstructured, and initial tests demonstrated that one-off retrievals were relatively erratic and varied. The continuous retrieval method was therefore used to build a large dataset which could subsequently be analysed and refined in real-time within Mozdeh. A series of one-off retrievals were conducted to iteratively refine the keyword query, identifying that the additional terms ‘euref’ and ‘indyref’ were required. Furthermore, hashtags such as ‘#brexit’ were required in addition to ‘brexit’ and frequent misspellings were added (Table 1).

Table 1: Refined keyword list for retrieving Twitter data using Mozdeh

eu AND referendum	brexit	euref
eu AND referndum	#brexit	#euref
eu AND refrendum	bexit	indyref
eu AND referendm	brext	#indyref
eu AND rferendum		
eu AND referedum		

A continuous retrieval method was run for ~72 hours (2/12/17 8:30am – 5/12/17 10:00am), resulting in a dataset of 762,341 tweets between 29/11/17 – 05/12/17 (Figure 1).

Figure 1: Time Series graph visualising initial dataset by time

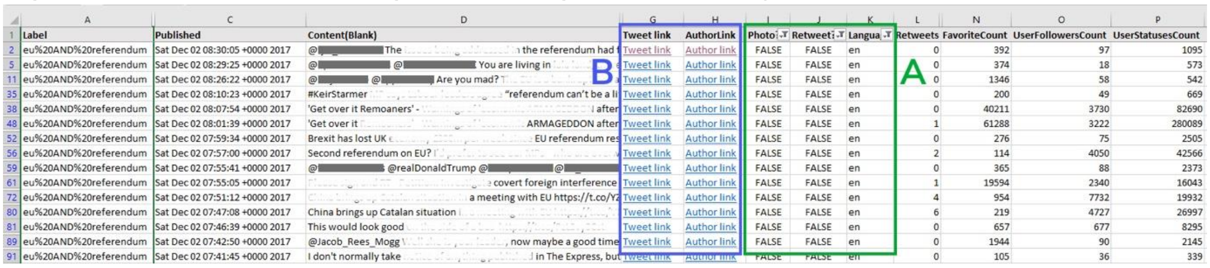


The Time Series graph reflects that Brexit is a vast topic on Twitter. The data collection was run for over 72 hours, yet the majority of the data retrieved remains within that timeframe.

After manual exploration, a small amount of legitimately collected data was discovered which was irrelevant to the dataset (e.g. tweets about the Catalan referendum using the #euref hashtag). The Mozdeh spam filtering functions were used to remove tweets by keyword search, creating a new Mozdeh project with a final dataset of 749,498 tweets. Subsequently word analysis was conducted to investigate high-frequency words within the dataset and correlations between words and to identify key themes within the topic. Coword analysis was conducted on the words ‘leave’ and ‘remain’ to investigate whether the opposing options of the referendum were still prominent themes.

Excel was used extensively to process the data. Functions were created which enabled data filtering – such as removing tweets which only contained a photo and couldn't be used for sentiment analysis (Figure 2:A). Further functions were created to provide additional columns which assist in manual data exploration and qualitative investigation such as one-click links to open a tweet in a browser (Figure 2:B).

Figure 2: Data in Excel showing data filtering and extra utility functions



SentiStrength uses the lexical approach to text analysis, utilising dictionaries with predefined weighting of understood terms to enable the calculation of an overall score for a text. Excel was used to organise and analyse the SentiStrength output. The Positive/Negative score columns created by SentiStrength were used to create a difference column (Figure 3:A). This was used to apply Conditional Formatting to the tweet data and visually differentiate the tweets based on their difference in the range -5 (Very negative) to +5 (Very positive). This enhances visual differentiation of the data when conducting qualitative investigation. This data was then used to generate a PivotTable to compare the scores proportionally.

Figure 3: Data in Excel showing Negative/Positive difference and Conditional Formatting

Excel was used to format the Mozdeh data for use in Gephi. Tweets that had been retweeted were extracted along with the original tweeter and retweeter to conduct network analysis.

Gephi is a tool for visually representing network data that would be verbose and overly complex to understand in tabular form. Gephi was used to investigate the actors (tweeters and retweeters) within the data and to understand who shapes the Brexit discourse on Twitter.

From an ethical perspective, several issues were considered by applying an ethical framework for social media research (Townshend and Wallace 2016, pp8). At the end of the project the cloud-based virtual machine was destroyed and no retrieved data remained in the cloud. Regarding the retrieved data used for the essay, most of the analysis was conducted computationally. Where manual qualitative analysis was undertaken, the researcher predominantly used the retrieved data or sparingly accessed the original public Twitter data on the web. Any data shown in the diagrams is redacted so that usernames or text cannot be used to manually search the content online. Finally, when visualising the data using Gephi, the top 30 tweeters were individually manually checked and only included on the graph if they are public figures e.g. a politician, journalist or organisation specifically engaged in the Brexit topic.

In the following section, I will present the findings of the research carried out using the above methods.

Findings

The overall findings of the research are that Brexit is a vast and complicated topic on Twitter, which remains active and prominent on social media and which further reflects the negativity, uncertainty and unease which was evident during the referendum campaigns and immediately after the outcome.

The word association analysis (Figure 4) for the word ‘brexit’ demonstrates that the predominant theme in the discourse is now primarily about how Brexit will be conducted and what “deal” can be reached – represented by the words ‘may’ (the surname of the UK Prime Minister), ‘deal’, ‘talk’, ‘britain’ and ‘negotiation’. 37.2% of all tweets collected containing the word ‘brexit’ also had one of these 5 associated words in the tweet.

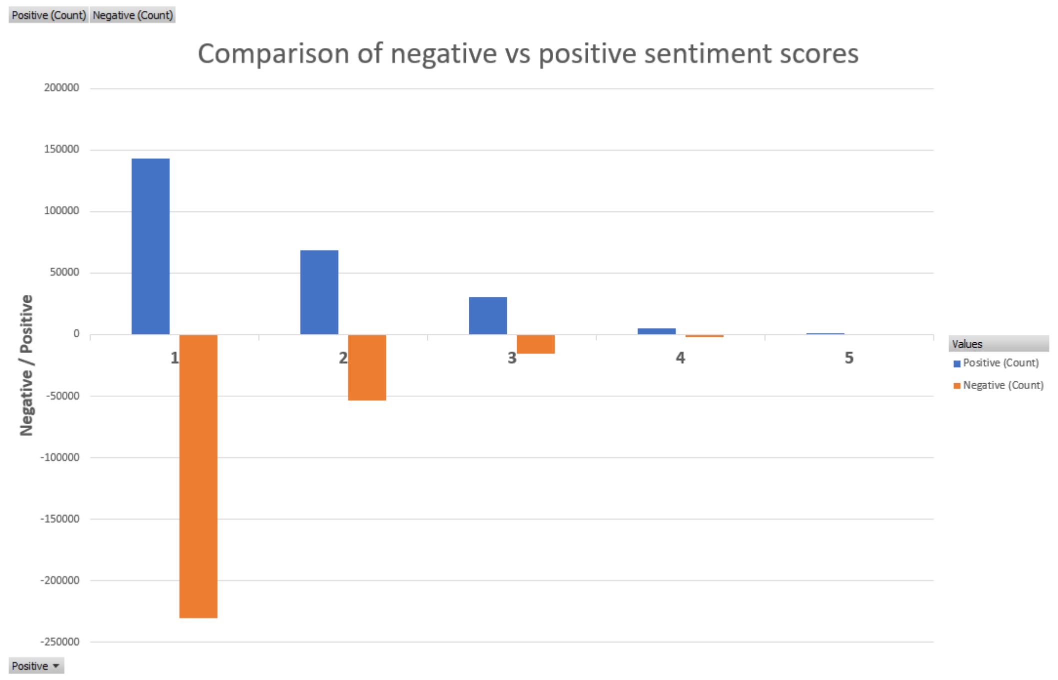
Figure 4: Word association for ‘brexit’ after frequent English words removed e.g. is, in, it, with, etc.

	A	B	C	D	E	F	G
1	**Word**	**Matches**	**NoMatch**	**Matches**	**Total**	**DiffPZ**	**Chisq**
2	brexit	100.00%	0.00%	316232	316232	785.1	616334
6	may	11.70%	10.30%	36967	67794	17.8	316.1
7	deal	11.20%	6.40%	35354	54460	66.5	4428.5
10	no	8.30%	7.20%	26312	48027	15.9	252.1
12	talk	6.30%	1.90%	20007	25659	87.3	7619.4
16	irish	5.10%	3.70%	16072	27287	25.7	658.6
18	referendum	4.50%	3.60%	14143	24998	17	289.4
20	britain	4.20%	3.20%	13147	22611	21	439
23	negotiation	3.80%	1.10%	12139	15564	67.5	4551.3
24	want	3.70%	2.70%	11837	20079	22	485.4
29	scotland	3.20%	2.60%	10095	17787	14.7	217.5
30	vote	3.00%	2.10%	9621	16059	22.1	488.3
31	reason	2.90%	1.20%	9160	12707	47.4	2242
38	trump	2.30%	1.00%	7195	10205	39.1	1530.6
39	blair	2.30%	1.00%	7137	10272	37.2	1380.7
40	election	2.30%	1.10%	7231	10519	36.1	1301.9

Looking beyond the top 50 words the word association suggests that the topic is very diverse and covers many words across politics, the economy, popular culture, regions of the UK, celebrities and brands. Not only does this reflect the diversity and broadness of the topic, but also the lack of focus and context in much of the discourse. Almost any part of UK politics and society may be mentioned in association with Brexit. A prominent story recently discussed how many leave voters regret their vote (Lynskey, 2017), yet the word ‘regret’ didn’t feature in the word association list and was evident in less than 0.02% of the data retrieved. Similarly, conducting coword analysis on the two main polemic words ‘leave’ and ‘remain’ did not produce a meaningful outcome, but rather suggested that the Brexit discourse has moved on from *whether* the UK should leave the EU to *how* the UK should leave. Furthermore, the proportion of data represented by retweets suggests this is not a discourse in which the public feel empowered (negatively or positively) or individually engaged. For example, there was relatively little original comment in the sample, with over 80% of the data collected being retweets.

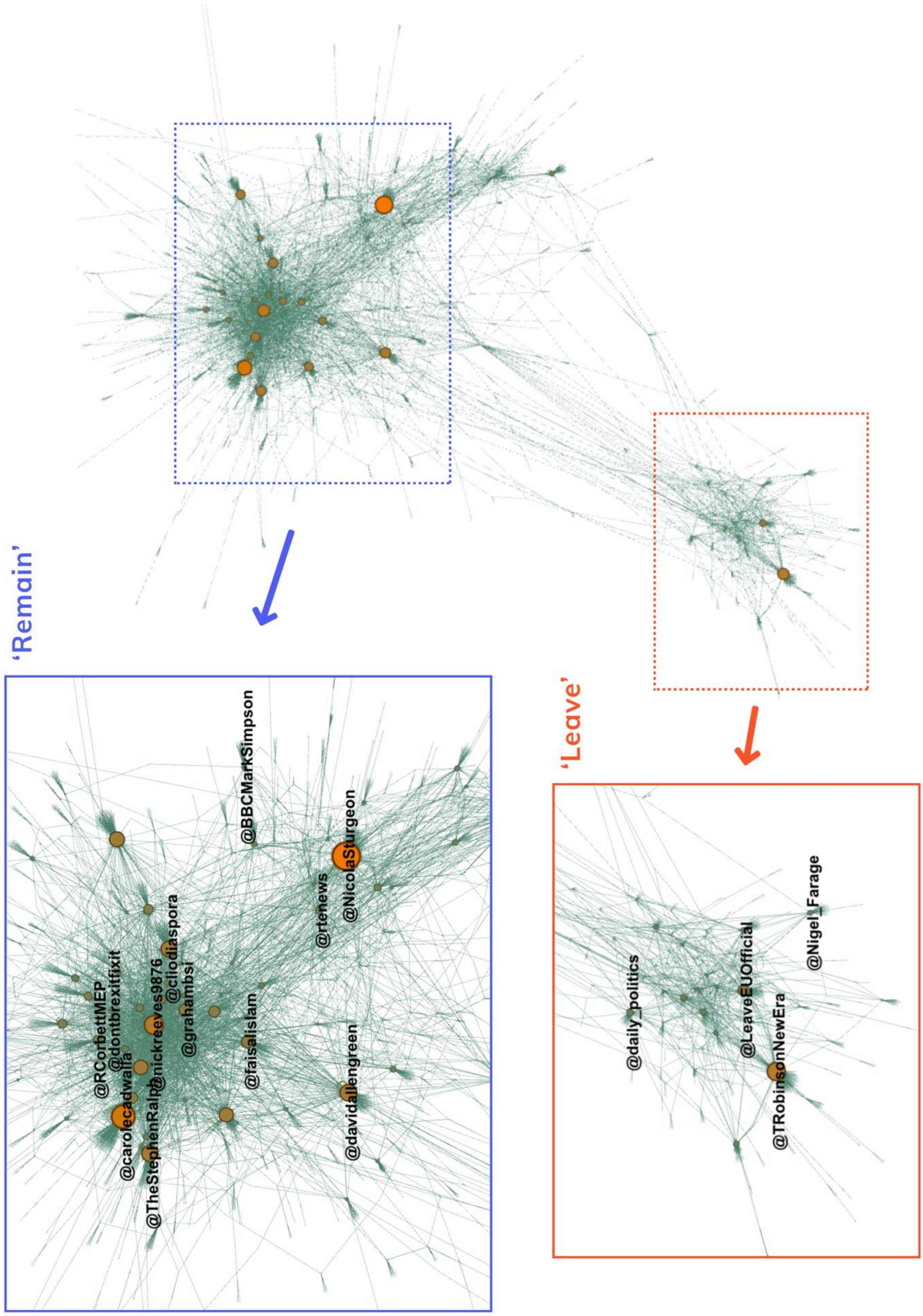
The Brexit topic presents a significant challenge for sentiment analysis; the breadth in which the topic is discussed involves a large and diverse vocabulary, the time-period is unspecific, the hashtag #brexit has become very generic and people rarely share personal opinions or feelings about the topic. Analysing the sentiment of the tweets overall confirmed the broadly unfocussed and neutral engagement of the tweet content. Comparing the overall negative and positive counts proportionally, the majority of the sentiment can be observed as neutral (a score of 1 or -1) with relatively equal 'slightly negative' and 'slightly positive' sentiment. This could be deduced as a lack of original opinion and commentary in the data due to the proportion of retweets, or indeed a growing fatigue in public engagement with the topic. However, the largely neutral sentiment analysis most likely demonstrates that without a specific context to the main topic, the ability to deduce meaning from sentiment analysis is very limited.

Figure 5: Graph of sentiment proportions



In contrast to the ineffectiveness of sentiment analysis, the more quantitative technique used for the network visualisation potentially reveals some interesting insights. Looking at the network visualisation (Figure 6), several “opinion leaders” (Stieglitz et al 2014, pp92) dominate the data. When viewing a diverse and active topic such as Brexit it might be expected to see a far more even distribution of low-to-medium volume tweeters, but at this resolution it is clear that a majority of the users involved in the discourse are only retweeting a few (e.g. less than 100) opinion leaders. Upon investigation, these opinion leaders are predominantly reporters, politicians or organisations involved in Brexit campaigning. Furthermore, two distinct communities are visible in the network visualisation – the larger network towards the top of the graph and the smaller network towards the bottom. Upon manual investigation of the Twitter profiles in each community, the larger community can be seen to represent ’Remain’ and the smaller one ‘Leave’. As already discovered, the Brexit discussion is no longer about leave or remain, and hence the ‘Leave’ voice could be deduced as no longer needing to be as active whereas the ‘Remain’ voice continues to try and drive the so-called ‘soft Brexit’.

Figure 6: Network visualisation showing opinion leaders and ‘Remain’/’Leave’ communities



Discussion and conclusions

The volume of data collected from a 72-hour continuous query relating to the Brexit topic demonstrates that Brexit continues to be a large, active and broad topic on social media. Applying social media research techniques to the Twitter data, the findings of the word association content analysis showed that the key focus of the topic now concerns the process of leaving the EU and the terms on which the UK will leave. Furthermore, the quantitative-based findings of the network analysis revealed low levels of original content creation and points towards the notion that a majority of the data focussed on the Brexit topic on Twitter consists of retweets.

The aim of the essay was to consider how effective sentiment analysis techniques can be for a broad topic. Without focussing on a specific time or event within the topic, the diverse language used in discussions about Brexit and the lack of easily identifiable polemic positions makes it very difficult to use sentiment analysis to effectively understand the meaning of the data. The lexical approach of sentiment analysis provides a large dictionary of understood words and how to weight them, but without context or meaning the technique can be ineffective. This may be because of the lack of context, that the techniques are in their infancy or that social media text is hard to analyse due to the use of emoticons, slang, sarcasm and media content (Stieglitz et al 2014, pp91). A further concern regarding sentiment analysis is the potential for ambiguity about the cause of a neutral outcome. This raises several questions that warrant further analysis, such as: ‘Does a neutral outcome reflect a thorough and well-deduced neutral analysis or is it the result of poor data or lack of context?’ This is a complicated differentiation but one which might be considered in future evolutions of the sentiment scoring technique.

As the term ‘Brexit’ has come to represent the whole topic, it is more difficult to identify individual context within the data. Furthermore, hashtagging is not significantly used to sub-divide the discussion or to associate a position, and hence identifying sub-themes is difficult. Regarding chronological context, studies have demonstrated how effective social media analysis can be during or immediately after an event (Procter et al 2011). However, as the Brexit topic has moved on from the referendum the time-context is ambiguous. Specific moments within the topic may create opportunity for specific research, but that was not the aim of this essay. Future research are likely to be required to expand our understanding about how we effectively analyse long-lifecycle topics on social media. Furthermore, we may understand patterns across a topic lifecycle such that we can appreciate different techniques are effective at different stages. Considering the type of data available from Twitter, we might consider that Twitter's appeal to social researchers is also one of its challenges. The benefit of the short, text-based format with inherent relational network metadata is unstructured and often context-bound to the moment it was created, making for complex interpretation (Gaffney and Puschmann 2014, pp65). To improve the context of data covering a broad topic we can adopt mixed-methods approaches that can maximise the opportunities and effectiveness of quantitative and qualitative techniques whilst also reducing the shortcomings of each (Einspänner et al 2014). A mixed-method approach to the Brexit topic would allow for greater qualitative investigation to try and understand meaning in the data. Furthermore, in doing so, digital methods could positively embrace the often opposing position of qualitative and quantitative methodologies (Marres 2017, pp35).

In analysing the broad topic of Brexit using social media methods, this essay has considered whether the use of sentiment analysis can be effective to deduce meaning from a broad topic – Brexit - in social data. The evidence presented within this essay suggests that without an ability to maintain strong context, sentiment analysis cannot currently be effective for broad topics, such as Brexit, with social media data. Potential future research questions has been suggested and I have sought to bring attention to the important role that mixed-method approaches represent as a way to bring a more balanced way of conducting social media analysis of broad topics. Yet the opportunities presented to sociologists by social media research and the future of digital methods remain hugely exciting, and as Vis points out, “[whilst] this is difficult to deal with analytically, that does not mean researchers should not try” (2013, section 2.1).

References

BBC (2016), 'Glaring deficiencies' in EU debate, Electoral Reform Society says, BBC [online]. Available at: http://www.bbc.co.uk/news/uk-politics-37238641 (Accessed 04/12/17)

boyd d. and Crawford K. (2012) ‘Critical Questions for Big Data’, Information, Communication & Society, Volume 15 (Issue 5), pp662-679

Doré L. (2016), 7 Brexit promises that have already been abandoned, The Independent [online]. Available at: https://www.indy100.com/article/7-brexit-promises-that-have-already-been-abandoned--b1Ne7rGnZDZ (Accessed 14/11/17)

Einspänner J., Dang-Anh M. and Thimm C. (2014) Computer-Assisted Content Analysis of Twitter Data, in Weller K. (Ed) Twitter and Society, New York, Peter Lang, pp97-108

Gaffney D and Puschmann C. (2014) Data Collection on Twitter, in Weller K. (Ed) Twitter and Society, New York, Peter Lang, pp55-68

Liu B. (2012) Sentiment Analysis and Opinion Mining, Synthesis Lectures on Human Language Technologies, Toronto, Morgan & Claypool

Lupton D. (2014) Digital sociology, London, Routledge

Lynskey D. (2017) ‘I thought I’d put in a protest vote’: the people who regret voting leave. The Guardian [online]. Available from: https://www.theguardian.com/politics/2017/nov/25/protest-vote-regret-voting-leave-brexit (Accessed 30/11/17)

Marres N. (2017) Digital sociology, Cambridge, Polity Press

Marwick A. (2014) Ethnographic and Qualitative Research on Twitter, in Weller K. (Ed) Twitter and Society, New York, Peter Lang, pp109-122

Partinton R. (2017) How has Brexit vote affected UK economy? October verdict, The Guardian [online]. Available from: https://www.theguardian.com/business/2017/oct/24/how-has-the-brexit-vote-affected-the-uk-economy-october-verdict (Accessed 11/12/17)

Procter R., Vis F. & Voss A (2011) How riot rumours spread on Twitter, The Gurdian [online]. Available from: https://www.theguardian.com/uk/interactive/2011/dec/07/london-riots-twitter (Accessed 10/12/17)

Rankin J. (2017), Painstaking detail of Brexit process revealed in EU documents, The Guardian [online]. Available from: https://www.theguardian.com/politics/2017/may/30/painstaking-detail-brexit-process-revealed-eu-documents (Accessed 11/12/17)

Seaver N. (2015) 'The nice thing about context is that everyone has it', Media, Culture & Society, Volume 37 (Issue 7), pp1101-1109

Stieglitz S., Dang-Xuan L., Bruns A. and Neuberge C. (2014), 'Social Media Analytics: An Interdisciplinary Approach and Its Implications for Information Systems', Business & Information Systems Engineering, Volume 6 (Issue 2), pp89-96

Thelwall M. (2014) Sentiment Analysis and Time Series with Twitter, in Weller K. (Ed) Twitter and Society, New York, Peter Lang, pp83-96

Thelwall, M. (2017) The Heart and Soul of the Web? Sentiment Strength Detection in the Social Web with SentiStrength, in Holyst, J. (Ed) *Cyberemotions: Collective Emotions in Cyberspace,* London, Springer International Publishing, pp119-134.

Thelwall, M., Buckley, K., and Paltoglou, G. (2012). Sentiment strength detection for the social Web, *Journal of the American Society for Information Science and Technology*, 63(1), 163-173

Townshend L. and Wallace C. (2016) Social Media Research: A Guide to Ethics. Available at: http://www.dotrural.ac.uk/socialmediaresearchethics.pdf

Vis F. (2013) 'A critical reflection on Big Data: Considering APIs, researchers and tools as data makers', First Monday, Volume 18 (Number 10), doi: http://dx.doi.org/10.5210/fm.v18i10.4878