Chapter 10:

Discourse-Analytical Studies on Social Media Platforms:A Data-Driven Mixed-Methods Approach

Ehsan Dehghan, Axel Bruns, Peta Mitchell & Brenda Moon

As social media platforms have become ubiquitous in everyday life, calls for mixed-methods approaches to studying social media platforms have also increased. Especially following the so-called "computational turn" (Berry, 2011) in media and communication research, we have witnessed the development of novel quantitative approaches, tools, techniques, and methods to study social media platforms. Qualitative studies, in contrast, often continue to draw from traditional methods that are sometimes not well equipped to deal with the particularities of studying these spaces, including their affordances, dataset sizes, systematic sampling and downsizing of data, or simply put, finding the right datasets.

This becomes particularly important for studies employing a discourse-analytical perspective. Traditionally, selecting texts for discourse analysis has been straightforward, with researchers using contextual knowledge to select newspapers, articles, television programs, political speeches, and similar texts for qualitative analysis. However, when it comes to social media, especially if the researcher is interested in the broader discursive environment or the everyday communicative practices of social media users, data selection and downsizing becomes more complex.

We propose a mixed-methods approach to tackle some of the issues facing discourse analysts investigating social media platforms, especially Twitter. In doing so, we draw on social media analytics, network analysis, corpus linguistics, and discourse studies. The proposed methodological pipeline equips researchers with a series of metrics to identify social media

research objects (posts, tweets, number of followers, visibility, and so forth) with salient discursive power; observe the underlying network structures and investigate how the specific affordances of a platform influence the flow of information and shaping of the discourse; and select an objective, downsized, and representative sample of the data for further in-depth qualitative analysis.

We draw on examples from two different case studies to show how this methodology can provide a richer understanding of the context, discursive environment, and dynamics of communication on the platform. The case studies investigate recent and prominent issues in the Australian socio-political sphere: the controversial automated welfare-debt recovery scheme launched by the Australian Government in 2016, which spurred the nationally prominent #robodebt Twitter campaign; and the almost year-long discussions over proposed changes to Australia's Racial Discrimination Act in 2016–2017.

New Challenges for Discourse Analytics

Although the terms *discourse* and *discourse studies* encompass a wide array of theoretical and methodological approaches, a common characteristic of most, if not all, discourse-analytical studies is their emphasis on in-depth, rich, and context-aware analysis. This often requires a focus on relatively small datasets that are systematically selected and sampled to be representative of the discourses and topics of the study, yet manageable under the time and resource constraints of the researchers. In this regard, one key challenge for discourse-analytical studies on social media platforms arises when making critical and systematic decisions on where to look. This is particularly important for studies interested in the mundane, everyday discourses of social media users. Unlike studies focusing on the discourse of politicians, news organizations, activists, and other distinct actors, in which the data collection often follows a

straightforward process starting from the identified accounts of such individuals or organizations, the study of mundane social media discourses is often faced with large volumes of data, scattered across the platform under investigation. This necessitates the use of approaches that can account for such challenges in data collection, selection, and sampling.

Further complicating discourse-analytical studies of social media activities are the platform-specific characteristics of social media discourses. In any discursive environment, social actions are constrained by the material and the discursive limits of the spaces in which they occur. In the case of a social media platform, these constraints involve the logics of the platform itself (van Dijck & Poell, 2013), including its technological design, affordances, algorithmic curation and moderation practices, user demographics, perceived and expected communicative culture, and other factors. These technical and technological aspects of a platform constrain and constitute—and are constrained and constituted by—the discursive practices taking place on the platform. At the same time, such technical aspects of a platform are themselves rooted in the discourses of the designers, owners, content moderators, and, in Foucauldian terms, the orders of the discourses in which the platforms were created and within which they operate (Foucault, 1971). Social media analysis, therefore, has much to gain from a discourse-analytical approach, but, as noted above, long-established traditional approaches to discourse studies are not well equipped with the methodological and analytical toolkits required to account for such complexities.

Approaching these questions from a different disciplinary tradition, yet facing many of the same challenges, is research generally rooted in media and communication studies, and enhanced more recently with advanced computational methods. With the rise of the so-called participatory web, and with it the huge leap in the volume of available data, new approaches in

these disciplines have attempted to accommodate the vast increase in what is called "big data" or, more specifically, "big *social* data" (Halavais, 2015). But this sudden enthusiasm for big data has at times led to an exaggerated focus on the data themselves, without taking into account the origins and contexts of the data sources (boyd & Crawford, 2012). The large volumes of data collected from social media platforms have frequently been treated as ends in themselves, leading to a substantial number of studies focusing on relatively shallow analyses of patterns of interaction, collective sentiments, clustering words and language, and the like. Interest in big social data has produced great advances in managing and processing the sheer volume of data in a limited time. However, it has also generated many studies that focus only on exceptional events and phenomena (identified on Twitter for instance by hashtags and other explicit discursive markers). Through their hashtag- and keyword-centric data gathering strategies, such studies usually treat their datasets as representative of isolated communicative spaces that are separate from their everyday discursive environments—the surrounding social media platforms, and the media environments beyond the platforms themselves. In extreme cases, this has led to media-centric and technologically determinist moral panics blaming social media platforms for everything that is wrong with society. One can see traces of such perspectives in the line of arguments—such as those refuted by Bruns (2019)—about how the use of social media platforms leads to polarization, filter bubbles, and echo chambers.

This argument, of course, should not be taken as completely dismissing the impact of platforms on societal developments. As we have noted, the materialities and embedded biases in the design, affordances, and governance of platforms clearly influence discursive practices. But although the role of these platforms in the everyday lives of users and non-users is undeniable, a persistent focus on atypical, extreme examples of social media discourse cannot provide a

comprehensive picture of discursive practices on social media platforms. Disregarding the socio-political contexts of discourses and solely focusing on broad patterns of communication can lead to misguided conclusions about discourses of and on social media.

As boyd and Crawford (2012) argued, regardless of the size of the data available from a social media platform, it is always important to remember that big data are created by humans, in human and temporal contexts, in the context of platforms and algorithms created by humans, and are eventually interpreted by humans. This intertwined nature of socio-political and platform contexts means that studies of social media discourses must account both for what happens on the platform and for the platform-external context that shapes the discourses themselves. KhosraviNik and Esposito (2018) refer to these as the horizontal and vertical contexts of social media discourses. On the horizontal scale, we face the technological and discursive aspects of the discourses that have shaped the platform, its affordances, and its operations. On the vertical, there are issues of power, hegemony, socio-political contexts, and other factors. At the intersection of the horizontal and vertical scales, one can witness what Carpentier (2017) theorized as the "discursive–material knot," where no hierarchies are present between the discursive and the material. Rather, the discursive (horizontal and vertical) and the material (technological design, affordances, and so forth) are co-present and co-constitutive on a social media platform.

In a study of communicative and discursive practices on social media platforms, therefore, there is always a need to account for the contextual peculiarities of the datasets under investigation. Although big-data approaches can provide valuable insights about the horizontal discursive environment, communities of users, patterns of social actions on platforms, sentiments, and the like, these insights need to be interpreted in light of their vertical context(s),

the discourses that shape them, and discourses that are shaped by them. Furthermore, any quantification of the material should be seen alongside the discursive. Any large-scale, big-data analysis of social practices on platforms has to incorporate a level of context-aware, cultural, and discursive analysis. As Manovich (2012) put it, "a human is still needed to make sense of these patterns" (p. 469). This calls for methodologies that can address large volumes of data, incorporate platform-specific (horizontal and material) aspects, and analyze discursive practices simultaneously.

Toward a Cyclical Mixed-Methods Social Media Discourse Analysis

In this light, we propose a methodological pipeline that draws from both quantitative, big-data approaches and qualitative, in-depth methods and show how each of these can inform and complement the other. This methodology draws from theories of discourse to exemplify how patterns of practice on social media platforms can be interpreted from a discourse-theoretical lens. We will also show how big-data approaches can help a discourse analyst identify where to look when faced with the sheer volume of datasets readily available from social media platforms. The iterative, circular flow of movement from big-data to small-data approaches, in which each step both feeds into and is informed by the others, forms the backbone of this methodological approach. Our approach cyclically moves between temporal and aggregate metrics, social network analysis, and textual investigation of the datasets.

Data Capture

The first challenge for any researcher working with data from social media platforms is selecting what data to gather. This is inherently shaped by the available data access points—usually, the platform's Application Programming Interface (API) or alternative approaches, such as Web scraping—and by the tools that connect to such access points. Although a wide variety of

such tools is available for common social media platforms, those tools often implement similar data gathering approaches, which are in turn directly influenced by the data structures that APIs and data access cost structures privilege. For Twitter, the most common data-gathering techniques capture tweets containing preselected keywords or hashtags. Approaches that select tweets based on user location, time zone, or other demographic factors, or whole-of-platform approaches that do not preselect themes and populations *a priori*, are considerably less common (Burgess & Bruns, 2015). It is imperative that researchers remain aware of how these methodological and political-economic factors shape the data they work with, and consider the impact of these biases on their analyses.

Generic Social Media Analytics

Once the data are gathered, it is crucial to develop a bird's-eye view of what the dataset comprises. This requires the researcher to address a series of questions that emerge both from the logics of the platform and the discourses shaping interactions on it. Especially central to studies collecting data for a period of time (from a few days to a few years) is the question of the temporal dynamics of interactions on the platform. Rises and falls in the number of posts, tweets, comments, likes, and so forth provide clues about the news, events, themes, and topics that have affected the patterns of participation. These peaks and troughs in social media activities can point researchers to the periods of time that are most interesting or relevant, whether researchers are interested in the discourses that are represented, amplified, or foregrounded, or those that are absent or backgrounded. Different patterns of activity over time can reveal valuable information about the discourses and events that lead to a higher level of reaction, or conversely about those that are ignored by social media users.

The logics of platforms in terms of visibility are similarly important to understanding activity levels. In general, the algorithmic design of platforms often rewards forms of practice that increase user engagement. These include having more followers, friends, or subscribers; posting, commenting, or replying; frequently using reaction, retweet, like, upvote, and downvote buttons; and generally making more use of the platform's affordances. Thus the platform's technological design and algorithmic moderation could affect the discursive practices of its users.

Therefore, at the early stages of a study's methodological conceptualization, researchers need to consider these platform logics, by identifying the objects that are more visible in one sense or another and reflecting on how such visibility should be interpreted. Variations in the visibility of different types of objects also reveal information about the discursive environment of the study, in terms of the horizontal context of the platform and the vertical context of the discursive environment. The affordances of platforms often enact different discursive functions. On Facebook, for instance, one could infer quite different conclusions about posts receiving high numbers of likes and loves or angry and sad reactions. The same is true for the number of retweets and @mentions received by different accounts on Twitter, the most (or least) upvoted or downvoted posts on Reddit, or liked and disliked videos on YouTube. Although these different affordances might be treated more or less equally in the algorithmic moderation of content on the platforms, in that they eventually create engagement, they reveal significantly different information about discourses on social media platforms.

Therefore, we propose an early reflection on and incorporation of the range of tools, metrics, and methods commonly referred to as Social Media Analytics (Zeng, Chen, Lusch, & Li, 2010) in the methodological framework of a study, to provide a bird's-eye view of the

communicative environment and to identify the patterns and objects that are more or less visible because of the social media logics embedded into the platform. Metrics differ by platform studied, and must be systematically and critically considered. For Twitter-centric studies, Bruns and Stieglitz (2012, 2014) have proposed a range of metrics and patterns that are commonly of interest. However, such metrics should not be treated merely as objective statistical measures of engagement, participation, or interaction. Rather, any identification of what is visible on a platform should be accompanied by follow-up reflections on how and why it is visible, the discursive functions this visibility achieves, and the materialities giving rise to its visibility and thus to its accumulation of discursive power. Finally, each research project has its own questions and interests, which require researchers to adapt their methodological choices to the challenges and limitations of studying a certain platform—such as issues of data quality, topic discovery and delineation, and data processing (Stieglitz, Mirbabaie, Ross, & Neuberger, 2018)—rather than apply a one-size-fits-all, off-the-shelf analytics solution.

Social Network Analytics

Following on from the initial exploratory insights drawn from the social media analytics phase, another important level of analytical possibilities created by social media data can be broadly conceptualized as "networkedness." Social actors, media objects, platform features, and interactions among them created as a result of platform affordances can be conceived as building blocks of a network of interactions and articulatory practices. Depending on the platform, this could be a network of accounts following a certain page, commenting on a subreddit, subscribing to a YouTube channel, @mentioning and retweeting other accounts, and many more possibilities. These collective discursive actions point to certain common features among actors and elements in a social environment (McPherson, Smith-Lovin, & Cook, 2001). Someone

interested in videogames, for instance, may be more likely to interact with actors and content that reflect this interest. The same is true for other discourses and discursive practices.

Therefore, any collective discursive practice can exhibit homophilic tendencies that lead to the clustering of related elements. The detection and interpretation of such clusters provide the researcher with crucial information about the discourses and practices present and prevalent in a large volume of data. The second phase in our methodology involves employing the range of tools, techniques, and methods known as Social Network Analysis (SNA). Although one can generate similar findings through a manual reading and coding of the dataset, this would be very time-consuming for large-scale datasets. Instead, as we show below, SNA can provide a faster way of identifying the various discursive positions in a dataset.

In the simplest of terms, community detection algorithms follow the assumption discussed above, and cluster different elements together if there is heightened interaction among them (Fortunato, 2010). These clusters can act as primary indicators for the investigation of the discourses and discursive strategies present in the data. However, we emphasize again that these tools are not a one-size-fits-all solution, because they rely only on mathematical and statistical techniques in grouping and clustering elements together. A cluster of elements identified by an algorithm may make complete statistical sense, but fail to reveal any meaningful underlying discursive structures; the clusters identified by an algorithm always also require a qualitative, context-aware examination and interpretation. This is why we argue for a discourse-oriented framework in understanding the results of a social network analysis. Below, we show how different network structures can reveal valuable information about the discursive structures and strategies present in a study.

Textual Analysis

In the third phase of the methodology, we turn to the textual representations of discourses in the dataset. As argued above, an examination of a collective social practice cannot be considered satisfactory unless it accounts for the context and the broader discursive environment. This generally requires some form of in-depth, qualitative, context-aware analysis.

The first two stages of the methodological procedure discussed here point to a series of objects, discourses, and collective discursive formations that require further qualitative investigation. For instance, upon discovering a cluster of accounts with a high level of insider retweeting and a low level of outsider interactions, we are faced with a discursive formation created through the articulatory practice of retweets. A qualitative review of the key accounts in the cluster (where "key" may be defined by the metrics generated through generic social media analytics) often reveals information about the shared discursive interests of the accounts in the cluster. This can be an entry point to the data selection and sampling for a discourse-analytical study, enabling the researcher to select purposive samples of tweets from the key accounts in the cluster, or a stratified random selection from all the accounts in the cluster, depending on the research question. However, this might still be pose significant difficulties, especially for large datasets or studies interested in everyday practices and discourses.

In bridging the gap between large-scale datasets and the smaller samples required for an in-depth discourse-analytical study, scholars have noted the appropriateness of corpus linguistics approaches (Baker, 2012; Baker & Levon, 2015; Evans, 2014; Wiedemann, 2013). Techniques and approaches for keyword analysis have proven useful for (critical) discourse-analytical studies; they can be used to identify salient themes, topics, and discursive strategies in a corpus of systematically chosen texts (Baker, 2004). Going back to the example of a cluster of densely

connected accounts identified in the network analysis stage, all the posts/tweets/comments by the accounts in that cluster could be treated as a corpus of texts written around a shared discourse. Depending on the aims of the study, this corpus could be compared to a larger, generic reference corpus in order to identify its most distinct keywords, salient themes, and central topics. This positions the researcher much more comfortably for further in-depth discourse-analytical investigation.

Practical Applications for Cyclical Mixed-Methods Social Media Analysis

Having discussed the broader methodological considerations, we will now show how these methods can be used to provide context-aware, discourse-oriented insights into social media data. It is beyond the scope of this chapter to report the findings of all stages of these studies. Therefore, we focus only on elements of the two case studies that elucidate the preceding argument.

Case Studies: Political Discussions in the Australian Twittersphere

As noted, the methodological reflections we have introduced here can be adapted to account for different theoretical frameworks and theories of discourse. In our two case studies, we draw from the discourse theory of Laclau and Mouffe (2001) to show how the different metrics and analytical steps could be examined through the lens of discourse theory, and how a discourse-analytical study could benefit from the incorporation of these tools and metrics to enrich its findings, account for the materialities of the platform, and manage the sheer volume of data acquired from social media. Although we use Twitter data in these examples, the proposed methodology could be used for data collected from other platforms as well, with modifications as needed. Before presenting our findings, we briefly introduce the two case studies.

#RoboDebt: An algorithmic scandal

In 2016, the Australian government started using an automated data-matching algorithm to discover overpayments made to individuals receiving welfare assistance from Centrelink, the welfare program run by the Australian Department of Human Services. The algorithm compared an individual's annual employment records from the Australian Tax Office with the payments received from Centrelink; if discrepancies were found, the system automatically issued a debt notice letter, stating the amount owed to the government and giving a deadline to repay the money or prove that the person did not owe the debt identified. It did not take long before numerous reports of flaws in the system, incorrect debt notices, and problems in the process found their way into the media. On Twitter and elsewhere, activists started referring to the issue with the term (and hashtag) RoboDebt, and formed a campaign under the slogan "Not My Debt" to protest unfair treatment under this algorithmic data-matching regime. The scale of reports, campaigning activities, and reports of individuals receiving debt notices of many thousands of dollars led to two external investigations, both of which found flaws in how the algorithm was implemented and the complaints were handled.

We collected tweets discussing this campaign between December 2016 and May 2017, by tracking the keywords and hashtags Centrelink / Centerlink, Tudge (the Human Services Minister at the time), Robodebt / #robodebt, and NotMyDebt / not my debt / #notmydebt. This resulted in a dataset of about half a million tweets. The #RoboDebt case study serves as a useful indicator of the different voices, discourses, and actors that may be involved in a social media discussion about a government scandal. It shows how communities and other stakeholders strategically draw from different symbolic resources and interdiscursive references to resist a

social injustice. Further, it shows how the materialities of a social media platform and its affordances help in the creation of virtual discourse communities.

Section 18C of the Racial Discrimination Act

For our second case study, we focus on a long-standing and divisive political debate in the Australian public sphere. Over the past decade, increasing tensions related to freedom of speech in Australia have often revolved around Section 18C of the Racial Discrimination Act, which makes it unlawful to publicly discriminate against individuals or groups based on their ethnicity, race, religion, or other personal attributes if to do so is reasonably likely to "offend, insult, humiliate, or intimidate" them. In particular, the words "offend" and "insult" have been the focus of much controversy. Critics of the law argue that these words are subjective and open to interpretation. A number of Australian politicians have called for either a replacement of these terms with the stronger word "harass" or a complete repeal of 18C. In 2017, the Australian Senate voted down any proposals for change, and the section has remained intact. Our dataset of tweets discussing Section 18C, posted between August 2016 and May 2017, contains just under 200,000 tweets containing keywords and hashtags 18C, S18C, Section 18C, Racial Discrimination Act, or RDA.

Agonism and Antagonism: Distinguishing Discursive Patterns in Social Media Debates

A complete discussion of the history, context, and findings of these two case studies is not possible here. Instead, we review elements of the social network analysis stage, and demonstrate how discourse theory guides making sense of the interaction networks, informs data sampling, and underpins further qualitative investigation. In particular, building on Mouffe (2013), we focus on evaluating the agonistic and antagonistic qualities of the discourses we observe; in this context, *agonistic* represents "a struggle between adversaries," whereas

antagonistic is "a struggle between enemies" (Mouffe, 2013, p. 17). Laclau and Mouffe (2001) conceptualize discursive struggle as an ineradicable and necessary condition of democracy. In such an understanding, the goal of a democratic project is not to reach to consensus, but to channel antagonisms in a way that they are transformed from a struggle between enemies—the goal of which is to eliminate the enemy—to an acknowledgment of differences between adversaries, and working with the adversary toward a common goal (Mouffe, 1999).

In both cases, the retweet networks could be understood as networks of discursive amplification. Although a retweet does not necessarily mean an endorsement of a message, it is an indicator that a user deems something worthy of being circulated and seen by her or his followers. Thus, the network clusters of accounts created through retweeting as an articulatory practice point to a collective perception of the sorts of messages that these users believe are in need of circulation.

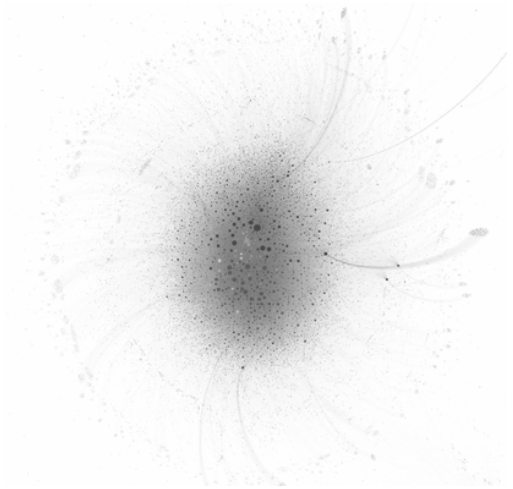

Figure 10.1: Retweet network of the #RoboDebt study

In our two examples, the retweet patterns (Figures 10.1 and 10.2) point to quite different network structures. The network of retweets in the #RoboDebt case study forms a star-shaped constellation of users retweeting a small number of key accounts in the discussion. In this core–

periphery structure, almost all users discussing the issue retweet certain accounts whom they presumably consider influencers in the debate. These activist accounts, which can be understood as crowd-sourced elites (Papacharissi, 2015), are those whose messages are perceived by Twitter users to be so significant that they should be seen by as many others as possible. Consequently, they are positioned in the center of the network. The lack of any other clusters in the network also shows that there might not be any competing discourses and actors regarding this issue. In other words, in curating and channeling the discussion about RoboDebt there appears to be a consensus on the key participants and whose voices need to be amplified to greater visibility. This network formation resembles the Broadcast Network in Smith et al.'s typology of communication patterns on Twitter (2014). As we discuss below, this could be an indicator of an agonistic space, created through retweets as one of the key affordances of Twitter, in which users with competing ideologies and discourses have formed a discursive alliance that at least temporarily transcends latent antagonisms for the sake of expressing a shared view or achieving a common objective.

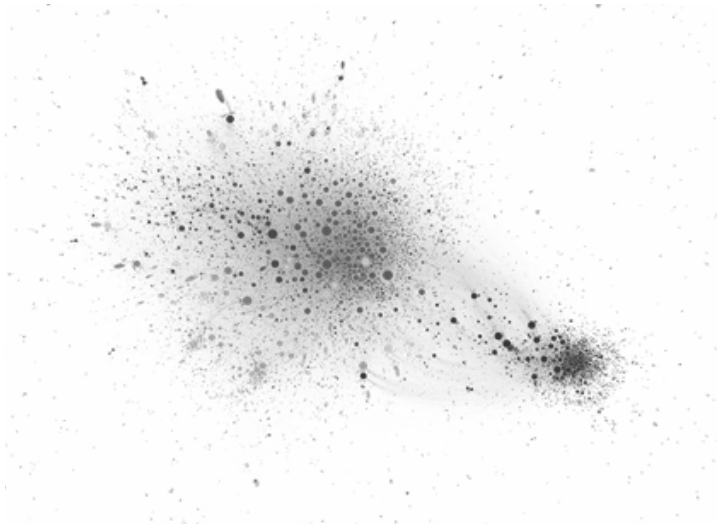

Figure 10.2: Retweet network of the 18C study

In contrast, the retweet network of the 18C case study forms quite a different structure. Three distinct clusters of users exhibit a higher propensity to retweet fellow members of their cluster than to retweet others outside of the cluster. Of these three clusters, one is comparatively less connected to the other two. This structure sits somewhere between Smith et al.'s Polarized Crowds and Community Clusters network types (2014), without entirely resembling either of these ideal types. Following the points made in the previous section, in such cases it is usually necessary to qualitatively examine the core accounts in each cluster and thereby contextualize these accounts in order to make inferences about the discourses shared and worldviews represented by the accounts in each cluster.

A qualitative examination of the three clusters in this network, by reviewing the profile information and tweets of key accounts, reveals three distinct groups of users: those generally opposing changes to Section 18C, Indigenous rights activists who have specific reasons to oppose changes, and those calling for a repeal of the section. The clusters of 18C supporters and Indigenous rights activists are tightly interconnected and therefore comparatively close to each other, whereas the repeal cluster has fewer connections to the other two. This can be explained by the close attitudinal alignment between the two pro-18C clusters: over the course of the debate, Indigenous activist organizations in Australia campaigned actively against making any changes to Section 18C, arguing that the proposed changes could put Aboriginal and Torres Strait Islander communities at an increased risk of racial vilification and racism.

The structure of retweet networks in this case reveals two interesting discursive practices. First, the polarized network structure between the pro- and anti-18C actors points to the presence of antagonistic discourses and discourse communities. This was predictable, but it is nonetheless important that the empirical evidence supports our working hypothesis on this point. Second, and

perhaps more interesting, the distinct yet closely related clusters of Indigenous rights activists and progressive users resisting changes to 18C indicate a discursive alliance and an agonistic discourse, formed around the notions of social justice, anti-racism, and equality as its nodal points, that envelops both communities; at the same time, however, the distinction between these two clusters on the pro-18C side shows that these discursive allies maintain their own identities and discourses through drawing from different symbolic resources, even as they form a discursive alliance against a common antagonistic opponent. In this case, the affordance of retweeting provided by Twitter helps in the formation of a shared agonistic discursive practice across two otherwise distinct communities.

Using the example of the retweet networks, we have shown how network analysis can lead a discourse analyst to form a working hypothesis about the range of discourses and discursive practices present in the data under investigation. We have also shown how awareness of context and discourse theory can guide a network analyst in connecting network structures to socio-political findings. However, the analysis should not necessarily conclude at this early interpretive stage; especially for a discourse analyst, the linguistic and textual representations of discourse remain central to the study. From this perspective, in addition to identifying discursive formations and communities, the findings of the network analysis stage should inform further data sampling considerations. For instance, in the case of the Section 18C debate, two major communities of pro- and anti-18C users were identified through network analysis. The discourse analyst can construct a purposeful sample of tweets posted in each cluster to serve as the basis for further qualitative investigation.

In approaching this sampling, although one may focus on tweets posted by the core accounts in each cluster (to study the discourse of key activists and influencers), our research

interest in the present project was the collective discourse among members of a discursive cluster. It is essentially impossible to conduct a qualitative discourse-analytical study of the hundreds of thousands of tweets posted by the members of each cluster. However, corpus linguistics tools can help identify the themes, topics, and keywords salient in the discourse of each cluster community in comparison with the broader discussions around Section 18C. We used the tweets posted by each community as the sample corpus, and all tweets in the dataset as the reference corpus. This technique assigns a rank to each keyword in the corpus, based on its frequency and statistical significance (that is, on how unusually frequent the keyword is in the sample corpus as compared to the reference corpus). This enables us to identify features that are salient in the discourse of a certain community (e.g. because they express pro- or anti-18C sentiment), including the themes, topics, and discursive strategies, the symbolic resources, and other features. (Such comparative analysis is impossible in a single-cluster dataset as encountered in the #RoboDebt case; still, the discourse of the anti-RoboDebt community on Twitter could perhaps be compared to the pro-RoboDebt rhetoric of government statements and interviews.)

The keyword analysis of the corpus of tweets from each community (Table 10.1) points us to the presence of two competing, antagonistic discourses. The pro-18C community of users frames the discussion by using interdiscursive references to the discourse of hate speech, White supremacy, and racism, arguing that the proposed changes pushed by liberal politicians and supported by right-wing groups in Australia will eventually lead to an increase in hate speech. Such references are almost absent from the discourse of the anti-18C community, which foregrounds discourses of free speech and criticisms of Islam and political correctness, arguing that Section 18C must either be reformed or repealed in order to ensure freedom of speech in

Australia, including the freedom to criticize Islam. Additionally, as Reisigl and Wodak (2005) argued, discursive strategies of nomination (how different actors, events, and so forth are named and referred to) and predication (the qualities attributed to different actors, events, and so forth) are often present in antagonistic discourses, which focus on an us/them dichotomy. Such strategies are frequently present in the discourses of the clusters we identified in the retweet network. In the discourse of both communities, there are frequent uses of diminutive pejoratives in reference to the Other: anti-18C discourses refer to the Other with terms such as "SJW" ("Social Justice Warrior") or "lefty snowflake" whereas pro-18C users employ pejoratives such as "free speech warrior", "RWNJ" ("Right Wing Nut Job"), or "old white male."

Table 10.1: Keywords in the discourse of pro- and anti 18C communities.

Anti-18C Discourse Keywords		**Pro-18C Discourse Keywords**	
Rank	**Keyword**	**Rank**	**Keyword**
5	reform	6	changes
6	petition	7	right
7	must	17	libs
8	islam	18	white
14	sign	25	liberals
15	should	32	rwnjs
16	repeal	33	rw
31	left	37	hate
35	removed	38	liberal
41	pc	41	agenda
52	abolished	45	rwnj
53	rid	49	racists
66	dismiss	50	bigots

Ranking of distinct keywords in each discourse community. Shared and generic terms (e.g. prepositions, articles, etc.) excluded from each list.

It is beyond the scope of this chapter to provide a full discourse-analytical account of the findings of these two case studies; detailed results of this research will be presented in other contexts. Instead, our intent is to outline the broad methodological considerations in undertaking such mixed-methods investigations. Our examples show how methods such as social media

analytics and social network analysis provide researchers with the tools to move from the large, noisy datasets obtained from social media platforms to a small, objectively selected sample of texts that is ready for further qualitative analysis. They also show how well-established discourse theories can inform working hypotheses about the bird's-eye participation and interaction patterns obtained through quantitative approaches such as social media analytics and social network analysis—hypotheses that can be further evaluated through purposeful qualitative analysis.

Enlisting Computational Methods in Doing Discourse Analysis

Not least because of the potential size of the big social data collections they make available, social media continue to present a substantial challenge to researchers in the humanities and social sciences, and especially to scholars from traditionally more qualitatively focused disciplines. Recent years have seen a "computational turn" (Berry, 2011) toward new research practices that are sometimes categorized under umbrella terms such as "digital humanities" or "computational communication studies," but it is critical that such developments are not misunderstood as implying a wholesale move from qualitative to quantitative approaches; purely quantitative research methods are valuable for particular research questions, but tend to produce aggregate and bird's-eye perspectives that identify broad patterns yet lack equally important insights into the finer detail. Conversely, purely qualitative investigations reveal details, but are often unable to contextualize the findings against the backdrop of larger communicative spaces.

Rather, as we argue, the combination of large-scale quantitative and fine-grained qualitative methods in genuine mixed-methods research designs presents a particularly fruitful path toward methodological advancement. In the abstract, this is not a new observation; what we

have provided is a structured overview of the concrete tools and practical steps involved in working through a research agenda that interweaves both quantitative and qualitative elements. Our brief discussion of how we have applied this research agenda to two political debates on Twitter in Australia, and of how we have operationalized this research in pursuit of questions emerging from discourse theory, demonstrates the use and utility of this mixed-methods approach in everyday research practice.

The methodological frameworks and tools for social media analytics, social network analysis, and textual analysis presented here are not the only such methods that could have been applied to the two case studies, nor necessarily the most sophisticated; methodological innovations since the emergence of big social data have produced a vast array of approaches and tools that can be brought to bear on datasets representing social media and other forms of communication, to the point that any attempt to comprehensively catalog them would be a fool's errand. Researchers should always seek to develop and maintain their methodological literacies, but happily even the fairly basic methods we have employed produce valuable results when they are sensibly combined and sequenced. It is less important to employ the very latest and most complex methods than it is to draw on methodological approaches that are relevant to the research questions being asked.

Although applicable to a broader range of domains, our observations are directed most immediately to discourse analysts, who in the context of social media data continue to struggle especially with the purposeful selection of appropriate samples. As in the case of the RoboDebt and 18C examples, the pathway from large-scale datasets through the application of standard quantitative analytical methods to the selection of meaningful samples for further discourse

analysis should help address this challenge, and in doing so support and enable new discourse-analytical studies of communicative phenomena in social media environments.

References

Baker, P. (2004). Querying keywords: Questions of difference, frequency, and sense in keywords analysis. *Journal of English Linguistics*, *32*(4), 346–359.

Baker, P. (2012). Acceptable bias? Using corpus linguistics methods with critical discourse analysis. *Critical Discourse Studies*, *9*(3), 247–256.

Baker, P., & Levon, E. (2015). Picking the right cherries? A comparison of corpus-based and qualitative analyses of news articles about masculinity. *Discourse & Communication*, *9*(2), 221–236.

Berry, D.M. (2011). The computational turn: Thinking about the digital humanities. *Culture Machine*, 12. Retrieved from http://www.culturemachine.net/index.php/cm/article/view/440

boyd, d, & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society*, *15*(5), 662–679.

Bruns, A. (2019) *Are filter bubbles real?* Cambridge, UK: Polity.

Bruns, A., & Stieglitz, S. (2012). Quantitative approaches to comparing communication patterns on Twitter. *Journal of Technology in Human Services*, *30*(3–4), 160–185.

Bruns, A., & Stieglitz, S. (2014). Twitter data: What do they represent? *it - Information Technology*, *56*(5), 240–245.

Burgess, J., & Bruns, A. (2015). Easy data, hard data: The politics and pragmatics of Twitter research after the computational turn. In G. Langlois, J. Redden, & G. Elmer (Eds.),

Compromised data: From social media to big data. (pp. 93-111). London, UK: Bloomsbury.

Carpentier, N. (2017). *The discursive-material knot: Cyprus in conflict and community media participation*. New York, NY: Peter Lang.

Evans, M. S. (2014). A computational approach to qualitative analysis in large textual datasets. *PLoS ONE*, *9*(2), e87908.

Fortunato, S. (2010). Community detection in graphs. *Physics Reports*, *486*(3–5), 75–174.

Foucault, M. (1971). Orders of discourse. *Social Science Information*, *10*(2), 7–30.

Halavais, A. (2015). Bigger sociological imaginations: Framing big social data theory and methods. *Information, Communication & Society*, 18(5), 583–594.

KhosraviNik, M., & Esposito, E. (2018). Online hate, digital discourse and critique: Exploring digitally-mediated discursive practices of gender-based hostility. *Lodz Papers in Pragmatics*, *14*(1), 45–68.

Laclau, E., & Mouffe, C. (2001). *Hegemony and socialist strategy: Towards a radical democratic politics* (2nd ed). London, UK: Verso.

Manovich, L. (2012). Trending: The promises and the challenges of big social data. In *Debates in the Digital Humanities*. Minneapolis, MN: University of Minnesota Press.

McPherson, M., Smith-Lovin, L., & Cook, J. M. (2001). Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, *27*(1), 415–444.

Mouffe, C. (1999). Deliberative democracy or agonistic pluralism?. *Social Research,* 66(3), 745–758.

Mouffe, C. (2013). *Agonistics: Thinking the World Politically*. London, UK: Verso.

Papacharissi, Z. (2015). *Affective publics: Sentiment, technology, and politics*. New York, NY: Oxford University Press.

Reisigl, M., & Wodak, R. (2005). *Discourse and discrimination: Rhetorics of racism and antisemitism*. London, UK: Routledge.

Smith, M.A., Rainie, L., Himelboim, I., & Shneiderman, B. (2014). Mapping Twitter topic networks: From polarized crowds to community clusters. *Pew Research Center*, 20, 1-56.

Stieglitz, S., Mirbabaie, M., Ross, B., & Neuberger, C. (2018). Social media analytics – Challenges in topic discovery, data collection, and data preparation. *International Journal of Information Management*, *39*, 156–168.

van Dijck, J., & Poell, T. (2013). Understanding Social Media Logic. *Media and Communication*, 1(1), 2–14.

Wiedemann, G. (2013). Opening up to big data: Computer-assisted analysis of textual data in social sciences. *Historical Social Research/Historische Sozialforschung*, 332–357.

Zeng, D., Chen, H., Lusch, R., & Li, S.-H. (2010). Social media analytics and intelligence. *IEEE Intelligent Systems*, *25*(6), 13–16.