Text Mining and Social Media: When Quantitative Meets Qualitative, and Software Meets Humans

Lawrence Ampofo, Simon Collister, Ben O'Loughlin, and Andrew Chadwick

Abstract

The ongoing production of staggeringly huge volumes of digital data is a ubiquitous part of life in the early twenty-first century. A large proportion of this data is text. This development has serious implications for almost all scholarly endeavour. It is now possible for researchers from a wide range of disciplines to use text mining techniques and software tools in their daily practice. In our own field of political communication, the prospect of cheap access to what, how, and to whom very large numbers of citizens communicate in social media environments provides opportunities that are often too good to miss as we seek to understand how and why citizens think and feel the way they do about policies, political organizations, and political events. But what are the methods and tools on offer, how should they best be used, and what sorts of ethical issues are raised by their use?

In this article we proceed as follows. First, we provide a basic definition of text mining. Second, we provide examples of how text mining has been used recently in a diverse range of analytical contexts, from business to media to politics. Third, we discuss the challenges of conducting text mining in online social media environments, focusing on issues such as the problem of gaining access to social media data, research ethics, and the integrity of the data corpuses that are available from social media companies. Fourth, we present a basic but comprehensive survey of the text mining tools that are currently available. Finally, we present two brief case studies of the application of text mining in the authors' field of political communication. We conclude with some observations about the proper place of text mining in social science research.

New Political Communication Unit Working Paper

This article appears in Peter Halfpenny and Rob Procter (eds) *Innovations in Digital Research Methods* (Sage, 2015).

'At no time in history has so much of the public's discussion… been so accessible to a wide audience and available for systematic analysis.'
—Scott Keeter, Director of Survey Research for the Pew Center for the People and the Press, *Wall Street Journal*, February 10, 2012.

The ongoing production of staggeringly huge volumes of digital data is a ubiquitous part of life in the early twenty-first century. A large proportion of this data is text. This development has serious implications for almost all scholarly endeavour. It is now possible for researchers from a wide range of disciplines to use text mining techniques and software tools in their daily practice. In our own field of political communication, the prospect of cheap access to what, how, and to whom very large numbers of citizens communicate in social media environments provides opportunities that are often too good to miss as we seek to understand how and why citizens think and feel the way they do about policies, political organizations, and political events. But what are the methods and tools on offer, how should they best be used, and what sorts of ethical issues are raised by their use?

In this chapter we proceed as follows. First, we provide a basic definition of text mining. Second, we provide examples of how text mining has been used recently in a diverse range of analytical contexts, from business to media to politics. Third, we discuss the challenges of conducting text mining in online social media environments, focusing on issues such as the problem of gaining access to social media data, research ethics, and the integrity of the data corpuses that are available from social media companies. Fourth, we present a basic but comprehensive survey of the text mining tools that are currently available. Finally, we present two brief case studies of the application of text mining in the authors' field of political communication: a research project that analysed political discussion on the popular social media service Twitter during the British general election of 2010, and a study of the early-2010 'Bullygate' crisis in British politics. We conclude with some observations about the proper place of text mining in social science research. Our overall argument is that text mining is at its most useful when it brings together quantitative and qualitative modes of enquiry. The technology can be powerful but it is often a blunt instrument. Human intervention is always necessary during the research process in order to refine the analysis. Indeed, rather than assuming that text mining software and big datasets will do the work, social science researchers would be wise to begin any project from the assumption that they will need to combine text mining tools with more traditional approaches to the study of social phenomena.

Defining Text Mining

Text mining is the term used to describe either a single process or a collection of processes in which software tools actively engage in 'the discovery of new, previously unknown information by automatically extracting information from different written [or text] sources' (Fan et al 2006). Such text sources can be defined as information that has been indexed in specific ways, such as, for example, patient records, web pages, and information contained within an organisation’s customer relationship management software. Such data is usually termed ‘structured data’ and has been defined as ‘any set of [text] values conforming to a common schema or type’ (Arasu, Garcia-Molina, 2003: 337). Some textual information that is amenable to analysis by text mining software does not necessarily conform to common schemas or types. This typically resides on the Web and is generally not housed within specific databases or other data storage structures. This data may include documents, emails, tweets, blog posts, to name but a few, and is typically defined as ‘unstructured data’. The central challenge of text mining is the accurate analysis of both structured and unstructured data in order to extract meaningful associations, trends and patterns in large corpuses of text. The increasing volume and availability of digital data online in social media environments like Twitter, Facebook, Flickr and collaborative online environments like Wikipedia, provides new opportunities for researchers to investigate social, cultural, economic and political behaviour.

Recent Applications of Text Mining in the Social Sciences

Text mining is emerging as an important tool for the natural sciences due to its usefulness in deriving value from the unprecedented volume of scientific studies that are now generated every year by the global scientific community (Ananiadou et al, 2006). Text mining allows for connections to be made between discrete scientific studies and research databases—connections that cannot be made manually due to the sheer scale of human effort that this would entail

(Hearst 1999). Text mining has also been used to reduce duplication in scientific research, to identify areas for potential collaboration across scientific fields (Rzhetzky et al., 2008), and to evaluate the consistency of research data over time (Rzhetsky et al., 2008).

While these scientific applications of text mining are important, rapid growth in the production of personal information and public sharing afforded by the rise of Web 2.0 and social media, as well as the proliferation of tools and techniques for real-time data collection and automated analysis, are now expanding text mining's potential range of applications. In some respects, the emphasis is shifting away from the retrospective analysis of static datasets and toward real-time analysis and 'predictive intelligence', particularly in the commercial sector. While these changes are made possible through the emergence of new technologies, communication practices, and research methods, they are also being shaped by many of the traditional concerns of longstanding disciplinary fields and research practices (Anstead and O'Loughlin 2011a). Bollier (2009) has identified the currently most fertile contexts for text mining as politics, public health, and business.

Measuring and managing public opinion through polling and the media have long been core components of liberal democracies. Both of these practices are now undergoing something of a transformation shaped by the availability of text mining and natural language processing (NLP) software. As Anstead and O'Loughlin have argued, '[i]t is not unrealistic … to imagine totally amalgamated real-time data' being used to inform political strategies in the very near future

(Anstead and O'Loughlin 2011a). Some recent U.S. studies have revealed a strong correlation between political opinions expressed via Twitter and official poll data, leading to the suggestion among some that real-time text mining might become 'a substitute and supplement for traditional polling' (Lindsay 2008; O'Connor 2010). However, studies conducted during the 2010 UK general election found that social media environments may offer poor data with little predictive value, primarily because social media samples can be highly unrepresentative of the wider public. There are also problems with the basic analyses. Commercial text mining companies very rarely publish their methods, so the mechanisms of accountability in this sphere are much less developed than those for traditional public opinion polling (Chadwick 2011a). In addition, the validity of text mining is impaired if automated text mining cannot detect irony and sarcasm, two linguistic techniques that feature prominently in online commentary of all kinds (Anstead and O'Loughlin, 2011b).

If the accurate prediction of public opinion remains elusive, text mining's usefulness for rapidly distilling the structures and meanings of large quantities of online political discourse is easier to grasp. For example, Wanner et al (2009) successfully applied real-time text mining of news coverage to gauge sentiment around specific topics and candidates during the fast-moving 2008 U.S. presidential campaign. These new forms of 'semantic polling' are coming to be seen by campaign managers in much the same vein as the focus groups that were first used by marketing companies in the post-war era to elicit more fine-grained interpretations of how consumers (and citizens) respond to particular aspects of a product or personality. In the field of politics, such intelligence is already evolving into a tool used by political actors seeking to strengthen their electoral strategies by adapting the content and delivery of their speeches and news announcements to public sentiment in real time (Anstead & O'Loughlin 2010; Chadwick 2011a; Chadwick 2011b; Chadwick 2013).

While such studies focus on text mining's application in broadly democratic contexts, scholars such as Leetaru (2011) have applied sentiment and geo-location analysis of broadcast, newspaper, and online sources to identify the precursors of unrest and to predict possible political uprisings or disruptive action by social movements. Adopting a similar technique, Papacharissi and de Fatima Oliveira (2011) have used text mining to explore the use of affective language and the geopolitical context in which it occurred to map the escalation and trajectory of revolutionary movements after the January 25, 2011 uprising in Egypt.

Business and economics scholars have also investigated text mining. Here, the emphasis has been on deriving commercial value from new types of market research and 'predictive sales intelligence'. In the entertainment industry, for example, sentiment and content analyses of social media data have been used to predict the likelihood of a film becoming a box-office hit (Asur 2010; Mishne 2006). Using a similar approach, Lee et al (2008) and Pang and Lee (2008) analysed customers' online reviews in order to distil this public feedback for providers eager to improve their products or services. Archak et al (2011) and Ghose and Ipeirotis (2007) have taken the text mining of product reviews still further, with an econometric analysis that seeks an '*objective, quantifiable,* and *context-sensitive* evaluation of opinions' (Ghose & Ipeirotis 2007: 416. Italics in original) and a set of techniques that can predict how differences in product reviews may predict levels of sales. Real-time quantitative and qualitative analysis are also now routinely performed on data from television audiences who share their opinions via Facebook and Twitter while watching a show live as it is broadcast (Wakamiya 2011a; Wakamiya 2011b).

Text mining studies of social media have also been applied to commercially-sensitive environments, such as financial markets, to try to develop predictive models. A number of studies have found that the sentiment and the volume of online 'buzz' correlates with stock movements. Some argue that it may be possible to predict likely market movements in close to real-time (Bollen 2011; Gilbert 2010; Lidman 2011). Others, however, are less bullish, and suggest that despite strong statistical correlations, the link between sentiment on social media and trading can be tenuous and difficult to predict (Antweiler and Frank 2004).

In the area of public health, quantitative and qualitative text mining analyses of social media have been used to identify and track the spread of natural disasters and epidemics. Culotta (2010) mined Twitter conversations in order to validate the service's role as a means of alerting public health officials to influenza outbreaks in the US, while Chunara et al (2012) were able to track the spread of disease during Haiti's 2011 cholera epidemic. Both studies point to the prospect of real-time predictive modeling in the field of epidemiology. Similarly, Chew and Eysenbach (2010) conducted content analysis of tweets relating to the 2009

H1N1 Swine Flu epidemic in order to assess the public's awareness of public health advice. Their conclusions revealed high levels of accurate knowledge among the public and they suggest that text mining of social media will offer a new way for health authorities to measure public awareness of their campaigns and respond to shifting concerns in real time. In a related field, public safety, the finding that there are strong correlations between discussions of earthquakes on Twitter and real earthquake events has revealed the potential of text mining as an addition to established early-warning systems (Sakaki et al 2010; Son Doan 2011). Indeed, Sakaki et al claim that Twitter may often be a more efficient early warning system than traditional systems.

Computer Assisted Qualitative Data Analysis Software (CAQDAS)

Text mining software can be conceptualized as being within the wider group of technologies known as computer assisted qualitative data analysis software (CAQDAS). Indeed, the use of computers to assist with qualitative research 'are inextricably tied to the character of qualitative data…[as] Qualitative research often produces an "assemblage" of data' (Fielding & Lee 1998). CAQDAS applications have been in use since the 1980s and many have included text mining functionality to deal with qualitative data in all forms, from fieldnotes to interviews to social media content.

CAQDAS was initially used by researchers for effective data management, such as text retrieval and simple searching. Such features can now be found in common word processors. The second generation of software introduced facilities for coding, text, and manipulating, searching and reporting on the text to be used for analysis. Today, CAQDAS software builders emphasise tools to help with the analytic processes of qualitative research such as examining relationships within text and building theories and models.

Natural Language Processing

A prominent component of CAQDAS and text mining software is Natural Language Processing (NLP), a technology that allows researchers to conduct in-depth analyses of content and everyday linguistic expression. As such, this particular software is useful for the analysis of social media content.

The development of NLP from the 1960s onwards focused on the 'the need not only for an explicit, precise, and complete characterisation of language, but for a well-founded or formal characterisation and, even more importantly, the need for algorithms to apply this description' (Jones 1994: 3). Further NLP research in the 1980s revealed the difficulties of developing reliable programs. During this period, work was characterized by a focus on developing computational grammar theory so that software could handle the refinements of linguistic expression, such as indications of time and expressions of mood. The 1980s was also marked by a focus on the lexicon, in the first attempts to exploit commercial dictionaries in machine-readable form.

Since the 1990s, NLP development has focused on statistical language data processing and machine learning: in other words, means by which software may use algorithms and probability calculations to undertake discrete analytical tasks, such as summarising the meaning of very large texts, or connecting information on location, time, and behaviour.

Sentiment Analysis

Sentiment analysis has quickly evolved into one of the most popular applications of text mining, not least because it holds out the promise of automating the interpretation of the semantic tone of large corpuses. Mejova has argued that sentiment analysis is the study of 'subjective elements' in language. These are 'usually single words, phrases or sentences' because 'it is generally agreed that sentiment resides in smaller linguistic units' (Mejova 2009: 5; see also Eguchi & Lavrenko 2006; Pang & Lee 2008). Typically, sentiment analysis involves using software to run pre-compiled dictionaries of known positive and negative words against a corpus in order to identify the frequency with which these words appear and the contexts in which they are used. Leetaru (2011), for example, uses this approach in his analysis of a large historical news archive, including the entire 3.9 million-article database of the *Summary of World Broadcasts*, 5.9 million articles from the *New York Times* from 1945 to 2005 and data from a variety of online news crawls. Leetaru claims that increased negative sentiment in news articles is statistically related to major news events, such as political unrest and the outbreak of wars.

Sentiment analysis often has trouble dealing with the inherent complexity of even the most basic everyday communication (Kanayama et al 2004; Read et al

2007; O'Connor et al 2010). The linguistic idioms of online communication only add to these difficulties. Nevertheless, some interesting applications of sentiment analysis are now emerging. A good recent example is Twipolitico's analysis of Barack Obama's and Mitt Romney's tweets during the early stages of the 2012 U.S. presidential campaign (see http://cs.uc.edu/twipolitico). Twipolitico extracted tweets referring to each candidate from Twitter's public streaming application programming interface (API)[1] and calculated each tweet's sentiment using software provided by a company, AlchemyAPI. This tool uses natural language processing and machine learning algorithms to identify positive and negative sentiment. [2] Nevertheless, while Twitpolitico's approach to text mining contains useful elements, it is not without weaknesses. For instance, it cannot identify whether a sentiment is being expressed in the grammatical first or third person. Human analysis of social media content is excluded in this approach, which relies wholly on computational analysis. Over-reliance on software programs and algorithmic analyses of text can overlook valuable insights that pattern recognition by human coders is better equipped to detect, such as how the meanings of certain terms may differ according to the communicative contexts in which they are employed.

Challenges: Practicalities, Ethics, and Access

As socially useful as these applications of text mining are, as we have intimated, they are not without pitfalls. On the one hand, there are clear practical benefits to mining 'big data' generated through social media, and evangelists have argued that this should become the new standard for scientific inquiry (Anderson 2008; Bollier 2009: 4–5). But a growing skepticism is also emerging. For example, boyd and Crawford (2011) ask whether data mining will 'narrow the palette of research options' (boyd and Crawford 2011: 1). There is a risk that computational, technologically-determined, automated research practice may lead us to believe we can always identify and meaningfully know the complex reality in which we exist, solely by mapping patterns in purely digital data.

One potential way forward is to marry automated analysis with more adaptive, online ethnographic methods that use theoretical hunches and qualitative analysis of online text to explore the dynamics of socially produced online information flows. Studies such as those by Veinot (2007), Chadwick (2011a; 2011b), Awan et al (2011), Al-Lami et al (2012), Papacharissi and de Fatima Oliveira (2011) and Procter et al (2013) have developed such approaches. Karpf (2012) puts the issue most starkly: since the internet keeps changing, the unit of analysis keeps changing; all we can aspire to in the field of internet research are short term, flexible and adaptive studies. A study of social media use over a 12 month period can lose validity if users start accessing those social media via different interfaces and devices and in different contexts and if the social media itself adapts. Long-term studies using 'our best methods will yield research that is systematically behind-the-times', Karpf writes (2012: 647).

Ultimately, we need to recognise that in the social sciences (and, indeed the natural sciences) theory and empirics are always symbiotic. Mining textual datasets, however large those datasets may be, must always be preceded by research questions that derive from the classical concerns of social research. There is no one-size-fits-all method or tool and compelling research questions cannot be generated entirely by the data itself. We maintain that as the field develops, social science approaches to mining social media text ought to be pluralistic, adaptive, and grounded in modes of enquiry and styles of presentation that are both intuitive and developed in dialogue with the norms and traditions of individual social science disciplines.

Practical Problems

On a practical level, users of text mining tools should be aware that the software is not a panacea and is inherently limited in what it can achieve. Sample bias and self-selecting samples are well-established risks with online studies. We should also be wary of conflating the expression of sentiment with actual behaviour. One of the main practical challenges of effective text mining is, ironically, linguistic. Despite the fact that English is still the internet's most prevalent language, Chinese is now used almost as frequently, while other languages such as Spanish, Japanese, and Portuguese are also widely used (Internet World Stats 2010). Moreover, the prevalence of non-Latin scripts in the top ten languages used online, such as Chinese, Japanese, Arabic, and Korean, only amplifies these problems. Some of these difficulties can be mitigated by the use of native-speaking human coders, and many commercial providers of text mining now claim that their products are "language agnostic" (Crimson Hexagon, 2012a). However, social scientists must be alert to potential problems here. There is the obvious but important point that many concepts and meanings, not to mention devices such as humour and sarcasm, do not traverse linguistic divides.

These challenges are further compounded by the idioms and meta-languages that have developed over the last two decades, and in some cases much earlier, in computer-mediated settings. Text mining tools that are only equipped to process grammatically-correct English are acceptable for formal documents like legislation or applications in natural science research, where constructs like biological and chemical compounds remain stable. However, these tools will often struggle to effectively analyse online idioms such as 'LOL-speak'. Online, many individuals deliberately contract and alter grammatically-correct language to provide more responsive answers to others. Many examples of this can be seen in the use of instant message clients and microblogging platforms such as Identi.ca and Twitter, but these language forms are now widespread across all online settings. In addition, language use may be instantly detected by humans but difficult to code in software. Sarcasm, irony, and *double entendres* can only be understood with reference to extra contextual detail and, in social media environments like Twitter and Facebook, that detail may derive from very broad and often-fast-changing cultural references that are very difficult to integrate without human intervention to guide the automated analysis.

Interpreting potentially ambiguous online content is therefore a common problem for researchers operating in this new environment and manual analysis is often essential to account for the wider social context of online discourse. However, given the huge volumes of data available to researchers, manual coding may not always be practical. Compromises exist in the form of 'machine-learning' tools such as Crimson Hexagon and Netbase, which enable the researcher to identify and manually review ambiguous data that the technology cannot accurately code. The software can then be instructed to code the data according to the manual instructions. In this way, researcher and software work together to continuously identify and improve the quality and accuracy of the analysis, but this is a process far removed from the promise of fully-automated text mining.

Ethics

As Jirotka and Anderson (Chapter 14) explain, another set of challenges associated with text mining derives from the ethical questions raised by this form of social enquiry. Is it ethical to mine data that is generally comprised of conversations between subjects who did not consent to having their utterances used for research purposes? Do the usual ethical standards for gaining consent in human subject research fully apply in these contexts?

In online research more generally, since the 1990s a rough consensus has emerged that the effective study of computer-mediated communication may often require a number of modifications to the standard human subjects model of research ethics. In these fast-moving environments, where there is a general expectation of public exposure, gaining the consent of individuals would make much text mining research impossible (Sveningsson 2003). In 1995 Sheizaf Rafaelli argued that researchers should treat 'public discourse on Computer

Mediated Communication as just that: public.' He went on: 'Such study is more akin to the study of tombstone epitaphs, graffiti, or letters to the editor. Personal? Yes. Private? No…' (Sudweeks and Rafaeli, 1995: 121). This perspective may be convenient, but does it always work in today's social media environments, particularly Facebook, whose privacy settings constantly change and are notoriously difficult to understand for many users? Facebook is, of course, a commercial environment and is regulated indirectly via the agreement users read and digitally sign when they join the service. But if online communication is used in large-scale text mining studies of public opinion, does this require a new set of ethical guidelines? After all, traditionally-conducted public opinion polls always require the active consent of participants. Perhaps the same rule ought to apply to text mining. These are important questions that must be addressed as text mining becomes more embedded in political organisations and government.

Access

A related challenge is the absence of open and universal access to social media data. With the Web's transition from a broadly autonomous and fragmented network infrastructure to an increasingly centralised and controlled commercial space, the abundance of personal data produced through social media has a great deal of commercial—and political—value (World Economic Forum 2012; Lohr 2012). It is this transformation that has led Anderson and Wolff (2010) to declare that the Web is 'dead', at least in the context of its original conception as an open network. Although Anderson and Wolff's perspective is arguably an exaggeration (see Schonfeld 2010), their assessment is useful in drawing attention to the spread of proprietary 'portals', 'walled gardens', and 'applications' which restrict the free flow of online information for commercial reasons (*The Economist* 2010). User data is increasingly locked within proprietary platforms, out of the reach of scholarly researchers.

These developments have a number of implications for research. Consider Facebook, currently the largest global social networking platform. Given its extensive socio-cultural dominance and 1.1 billion global users by 2013 the quantity of personal data shared within its walled garden is vast. This information, however, remains locked within Facebook's proprietary platform, with full access available only to its own and approved researchers and commercial partners, for example, advertisers and developers (Bakshy 2012; Deloitte 2012). Facebook is arguably creating its own gigantic proprietary data repository. Social researchers can gain limited access to this data using Facebook's Graph API, which provides datasets of Facebook 'objects'—certain content, such as photos, Facebook Events and Pages, and the 'connections' between them, such as friend relationships, shared content and tagged photos (Facebook 2012; Knguyen 2010; Russell 2011). But a great deal of Facebook users' content is off limits to researchers. And the range and volume of data available via Facebook's Graph API is now more tightly controlled than it was in the service's early days, when researchers were allowed greater scope (Golder 2007; Gross 2005; Lampe 2006; Lewisa 2008; Mayer 2008).

The same may be said of Twitter, which in the early 2010s moved to more tightly control access to its data in an attempt to enhance the company's profitability. At the time of writing, researchers have a variety of options to access Twitter data, ranging from Twitter's complete public data stream (the 'firehose') through to a 10% or a 1% sample of public tweets (the 'gardenhose' and 'spritzer' respectively) (boyd and Crawford 2011; Gannes 2010). This approach, however, favours commercial users over scholars. The cost of accessing Twitter's complete dataset prohibits what are often poorly-funded academic researchers. Although costs to access Twitter's firehose vary depending on the volume of tweets returned by search queries, Twitter's two official data resellers, Gnip and Datasift currently license access to a sample of tweets from between $1,000 and $15,000 per month (Datasift 2012). Moreover, an update to Twitter's terms of service in 2011 further compounded researchers' ability to access data by expressly forbidding users from 'resyndicating' or 'sharing' Twitter content, even if the data is collected legitimately (Twitter 2011 cited in Freelon, 2012). As a result, researchers, including the authors of this chapter, who previously benefited from access to datasets gathered by the research community in services like Twapperkeeper, are prevented from conducting further studies because that data cannot be made public (Judd 2011; Freelon 2012; Shulman 2011). More worryingly, there is some evidence that social media companies are increasingly keen to police research agendas. Citing a keynote talk by Twitter's internal researcher, Jimmy Lin, at the 2011 International Conference on Weblogs and Social Media, boyd and Crawford (2011) have argued that Lin 'discouraged researchers from pursuing lines of inquiry that internal Twitter researchers could do better given their preferential access to Twitter data' (boyd and Crawford 2011: n4).

Twitter's restrictions on data access, however, are not without work-arounds. While Twitter prevents the sharing of individual tweet content or 'follow relationships', it does allow the distribution of 'derivative data, such as the number of Tweets with a positive sentiment' and 'Twitter object IDs, like a Tweet ID or a user ID' which 'can be turned back into Twitter content using the statuses/show and users/lookup API methods, respectively' (Twitter cited in Freelon 2012). But although this may offer some very useful and interesting possibilities for research, these provisions prohibit independent large scale text mining, given the technological skills necessary to identify tweet content from object IDs or the likely timescales required to reverse engineer potentially very large datasets.[3]

Walled gardens pose a substantial challenge to the open flow of information across the web, but so, too, does another important recent trend: the growth of application-based platforms or 'apps'. The rise of apps is largely attributable to the late-2000s growth in smartphone and tablet computing and the platforms and protocols that govern how these devices operate. The challenge for researchers lies in the proprietary infrastructure of apps and mobile devices that 'use the Internet for transport but not the browser for display' (Hands 2011). As a result, data remains locked away and private companies become newly-important data 'gatekeepers'. Given the significant growth of smartphones and tablets and current industry predictions that tablet devices will outsell traditional personal computers by 2014 (Dediu 2012), the value of this data needs to be taken seriously by researchers interested in large-scale text mining. In future, will academic text mining studies be able to compete with those carried out by commercial organizations? Will scholars have access to meaningful data when so much important discourse is taking place inside these closed environments?

Text Mining Tools: A Brief Survey

We turn now to a brief survey of the main text mining tools and services that are currently available and which have a focus on the analysis of online text. Before we do so, some caveats are necessary. First, given the rapidly evolving nature of this field, any overview is inevitably provisional. Free or open-source tools are continually developed and shared among user communities and established commercial technologies are often acquired and bolted on to other products. Second, while text mining may appear to be a cohesive field, as this chapter demonstrates there is a diversity of approaches and applications. There are no perfect technological solutions and this section makes no hard recommendations as to the suitability of specific tools. However, we do wish to highlight the tools that we believe offer good starting points. These are outlined immediately below and summarised in the chapter'sAppendix, alongwith a selection of other tools and services.[4]

Sysomos MAP

Originally developed by researchers at the University of Toronto, Sysomos MAP is now arguably one of the better commercial text mining products for running basic analyses of social media content. MAP provides access to a database of 20 billion social media conversations, spanning platforms such as blogs, message boards, Twitter, and a sample of public Facebook pages. MAP's retrospective database consists of two years of historical data and claims to index eight million posts an hour in close to real time, which makes it useful for tracking live events. Sysomos has access to the full Twitter firehose, allowing researchers access to data from over 100 million Twitter users.

Sysomos MAP users can filter data geographically, potentially down to city level, provided the data is available, and demographically, according to the age, gender, and profession of individuals. Although MAP's automated sentiment analysis is useful, in practice the benefits can be limited. MAP provides a workaround by enabling researchers to override automated sentiment results and manually code more accurate sentiment scores for subsequent analysis. However, MAP's search algorithm does not automatically 'learn' from manual sentiment overrides. Search results are fully downloadable in a variety of formats, most usefully as CSV files. There are other services that compare favourably with MAP, such as Radian6 and Attensity, but Sysomos' strength lies in its relatively easy setup and relatively low cost.

NetBase

NetBase's 'Enterprise Social Intelligence Platform' takes the fundamental features found in commercial text mining tools such as Sysomos MAP or Radian6 and overlays a range of additional functionality. The service provides access to 100 million conversations from social media platforms and offers the ability to group common phrases and keywords in a dataset and automatically code these in the same way during future analyses. While NetBase does not permit access to the Twitter firehose, it claims that it indexes all of the public pages on Facebook. This gives NetBase an advantage over comparable tools that typically offer access to only a sample of public Facebook pages. A downside to NetBase is that some elements of basic functionality, for example, exporting data as CSV files, are not currently offered.

Crimson Hexagon Forsight

Crimson Hexagon Forsight offers functionality comparable with Sysomos MAP and NetBase, such as theme, sentiment, demographic, and influence analysis. Importantly, however, Forsight's analysis algorithm uses machine learning and is therefore 'trainable'. Researchers can manually code a data sample and instruct Forsight to 'learn' from this and apply it to future analyses. While not as accurate as data that is coded entirely by humans, this feature provides researchers with better options for gaining accurate results than many other tools. Crimson Hexagon's Social Research Grant Programme, which offers 'in-kind access' for the academic and non-profit community (Crimson Hexagon 2012), makes Forsight a relatively attractive tool for scholarly researchers.

DiscoverText

DiscoverText is comparable with Crimson Hexagon's Forsight in that it offers a number of unique features likely to be of particular use to social scientists. It allows researchers to perform manual text coding collaboratively through the creation of cloud-based data 'buckets' which can be shared online among a dispersed researcher network. Validation tools enable lead researchers to test for coding validity at the micro-level of the coder as well as at the project level. DiscoverText incorporates the ability to automate the inter-coder reliability tests that are essential for team-based content analysis. A significant feature of DiscoverText is its 'ActiveLearning Customized Classifiers' functionality. Although still in beta phase, this allows researchers to customise coding classifications and 'train' the DiscoverText algorithms to detect sentiment and themes using machine learning. DiscoverText can also provide access to the Twitter firehose, though this incurs an additional cost.

Linguamatics I2E

I2E from Linguamatics provides text mining analysis using natural language processing. Originally used within the life sciences, I2E has recently been deployed for the analysis of social media content, as we discuss in more detail below.

Applying Text Mining I: Social Media Monitoring During the 2010 British General Election

To illustrate the potential—and some of the pitfalls and work-arounds—of using digital methods for real-time analysis of political events, we now turn to a discussion of how I2E, developed by the Cambridge-based company Linguamatics, was used to analyse the opinions of Twitter users during and immediately after the live televised prime ministerial debates of the 2010 British general election. This was part of a larger collaborative project carried out in 2009 and 2010 to develop a real-time methodology for analysing public responses to emergent events.[5] The project consisted of several other experimental studies, including work on the autumn 2009 swine flu vaccination campaign in Britain, the December 2009 Copenhagen Climate Summit, the January 2010 Haiti earthquake, and the collapse of the Sony Playstation online network in March 2010.

The televised prime ministerial debates were the first events of their kind to have been held in the United Kingdom and there was great media interest. The allure of real-time results that could be delivered to the public at the end of each debate led a number of established polling companies to promise instant polls to broadcasters. For example, Comres delivered a poll result within six seconds of the end of one debate by using a telephone panel survey in which a representative sample of voters were given keypads and told to press a button to indicate who they thought had 'won' (Anstead & O'Loughlin 2012). But digital media also offered other sources of data. By spring of 2010 the 'two-screen' media event had become common in Britain. Many audience members now use laptops or mobile devices to offer their personal social media commentary on political or celebrity television broadcasts, in real time as they watch a show (Anstead & O'Loughlin 2011b; Chadwick 2011a; 2011b). In this context, the Linguamatics project aimed to see if there were patterned responses on Twitter to the party leaders' performances during each debate, with a particular focus on how each candidate was deemed to have performed in response to each question in the debates. The research team was also curious to see if the method could be used to predict the eventual 'winner' of the televised debates, though this proved entirely problematic, as we discuss below. Nevertheless, the project team was contacted by journalists who requested from them text-mining 'poll' results within hours of the end of each debate. This compelled the team to reflect on the ethics of how to present their research in meaningful ways.

The methodology and workflow for these studies depended upon a combination of human and automated analysis of social media content.

Setup Before the Event

- Human: decide key search terms and relevant queries based on expertise in the given field (for example, pharmaceuticals, climate change, financial markets, party politics).
- Technology: initial data search, aggregation, classification.
- Human: clean the data by refining search terms and vocabulary.

Real-Time Monitoring During the Event

- Human and Technology: continuous data stream from social media according to key search terms.
- Technology: process the data using I2E software.
- Human: interpret findings on an ongoing basis.

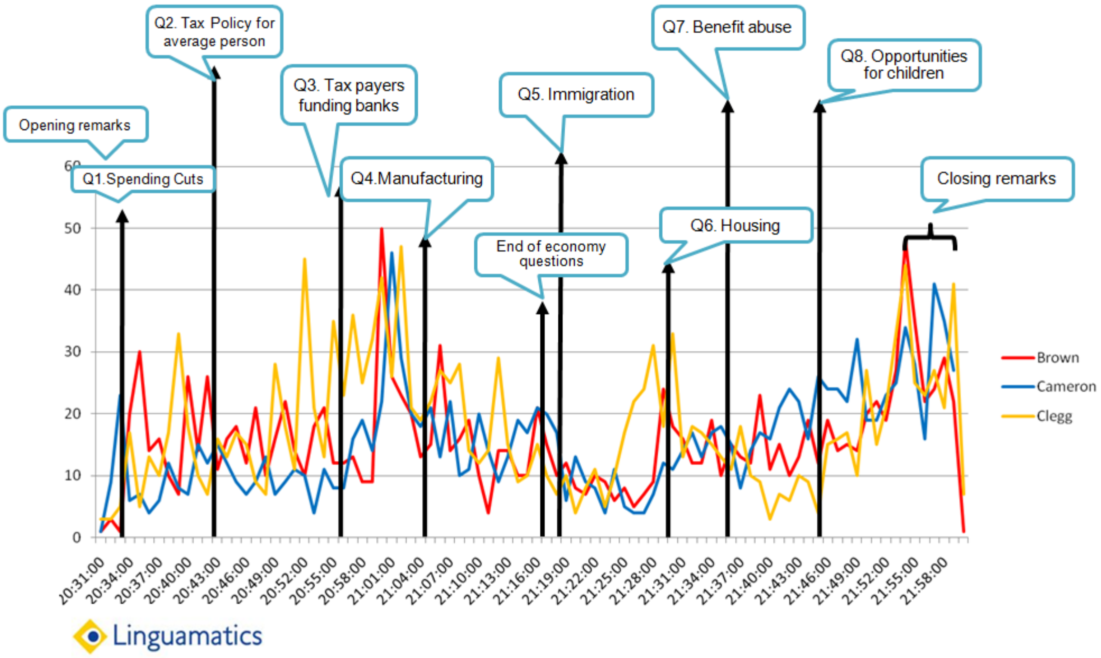
Integration and Presentation

- Human: integration of these results with other data, for example relationships between social media content and indicators such as share prices, sales of goods, or opinion surveys.
- Human and Technology: Visualisation tools to render key findings more quickly intelligible.

The three live televised debates offered a chance to test this methodology and workflow. The debates were held on April 15, 2010 (ITV), April 22 (Sky News), and April 29 (BBC). Approximately 567,000 tweets from 130,000 Twitter users were analysed during the three debates. The framework of the study went beyond traditional keyword searches of important terms in the data corpus. For each debate, a sentiment analysis of the content referring to each of the political leaders was examined using the NLP technology in Linguamatics' I2E text mining tool. I2E examines the grammatical structure of each tweet and uses a conceptual vocabulary that enables inferences about the intention of a person posting a message. It should be noted that, as a business, Linguamatics keeps the precise nature of I2E a closely guarded commercial secret. Scholarly researchers need to balance this drawback against having access to the expertise provided by private sector providers. Some of the services we listed above may offer the appearance of analytical power but if the process of research is not transparent and replicable, peer-reviewed scholarly journals may be wary of publishing the research. However, Linguamatics were granted access to Twitter's streaming API At the time of writing (March 2013), for independent scholarly researchers the API delivers one per cent of Tweets at no cost.

Figure 1 below presents the volume of tweets about each leader, organized by the questions asked during the third televised debate. The research team were able to conduct fine-grained analysis of the television coverage to identify precisely which statements or audience reactions correlated with these spikes in online commentary (see also Anstead & O'Loughlin 2011b). This chart also formed part of the coverage provided by the BBC's technology reporter Rory Cellan-Jones (2010).

Figure 1. Volume of Tweets About Each Leader's Response to a Question in the Third Debate



In Figure 1 the Y-axis refers to the number of tweets per minute that contained positive commentary on a leader's response to a question or issue. For example, the tweet 'Clegg strong on the "the outrageous abuse of bankers' bonuses"' was coded as for Liberal Democrat party leader Nick Clegg on debate question 3 (Q3) at 21:00 hours. Tweets were coded in 60-second chunks. Twitter's streaming API only provided ten per cent of all tweets, but still the numbers per minute seem relatively low (between zero and 50). However, Figure 1 excludes all tweets about leaders that were not connected to a particular issue or question. This relatively low number of tweets—4,082 in total for the third debate—allowed for human coding in the hours after the debate to check the validity of how each tweet was coded by I2E. The project team also analysed each leader's popularity by issue, and found that Clegg and Brown shared the lead on immigration, Clegg was ahead on banking and tax, while Brown clearly won on the economy. However, patterns of response were often uneven. In the second leaders' debate, for example, a question about whether the Pope should visit Britain while the Catholic church was confronting a sexual abuse scandal led to immediate responses on Twitter but also a later spike in interest, as Clegg briefly mentioned religion in response to another question. Discussion of issues and questions on Twitter did not map neatly onto the timelines.

In terms of overall sentiment towards each of the leaders across the three debates, Figure 2 below shows that Nick Clegg's share of positive sentiment dropped from 57 per cent in the first debate to 37 per cent by the end of the third and final debate. Gordon Brown's share stabilised at 32 per cent, while

David Cameron's rose from 18 per cent to 31 per cent. Intriguingly, these trends eventually converged roughly with the final vote share on election day: Cameron's Conservative Party won with 36.1 per cent of the popular vote, Labour came second with 29 per cent, and Clegg's Liberal Democrats fell to a final 23 per cent.

Figure 2. Share of Positive Sentiment for Party Leaders

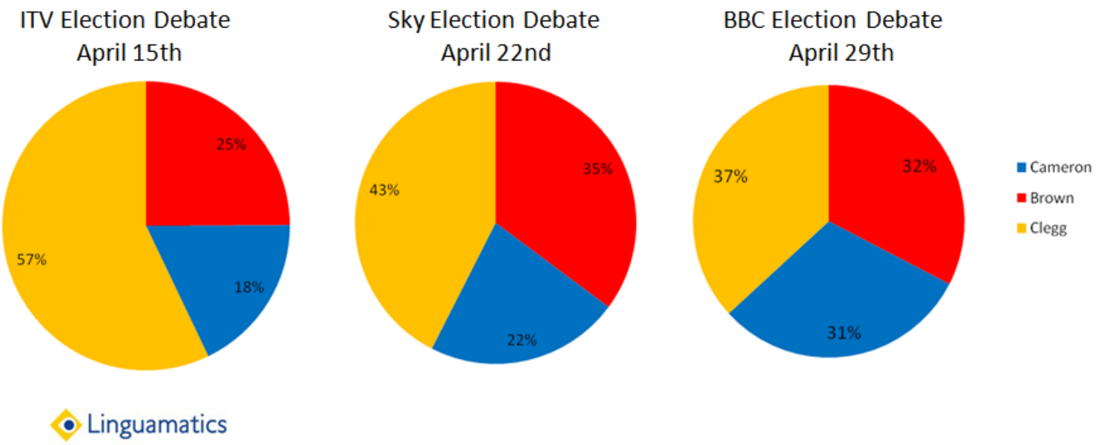


Figure 3. The Trend in Positive Sentiment for Cameron

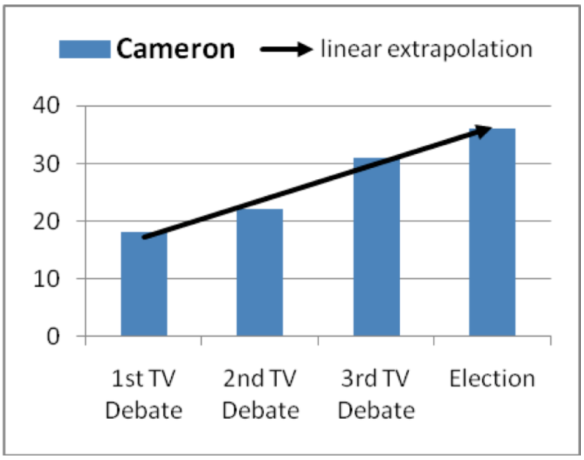
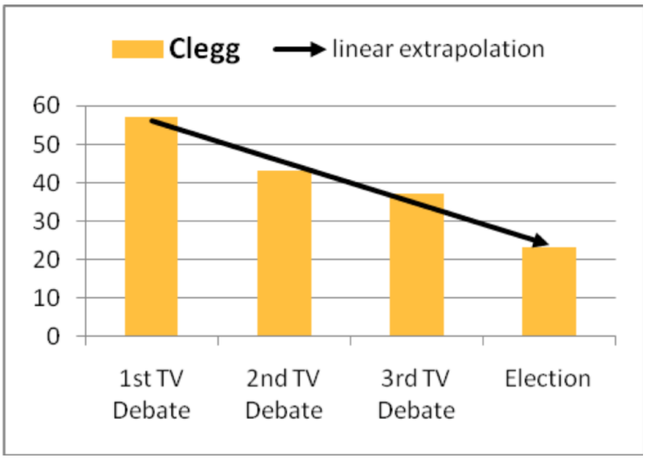


Figure 4. The Trend in Positive Sentiment for Clegg

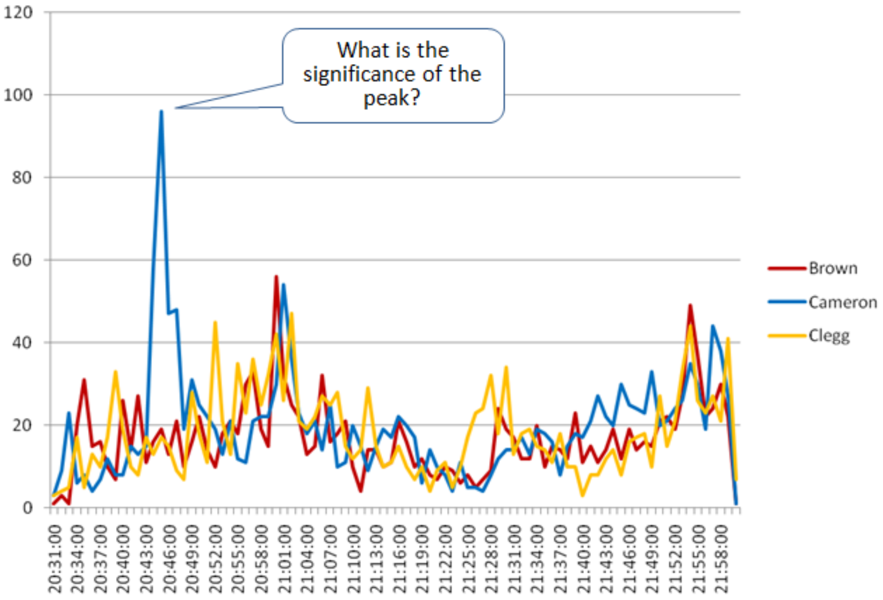


These trend lines make for striking visuals (Figures 3 and 4) that can be used to support simple media narratives about, for instance, the "Lib Dem surge" that saw Clegg unexpectedly "win" the first debate and then steadily lose support over the course of the general election campaign. However, using such analysis to predict election results is extremely problematic and should provoke instant caution. Twitter users are not representative of the whole electorate, a user's comments may not indicate how they will actually vote on election day, and of course events may occur between the final debate and election day that alter voting intentions. The same applies to any measurement of opinion around televised debates prior to elections, including telephone polls or devices such as "the worm" that monitors and visualises immediate audience responses to politicians as they speak. When the project team were contacted by journalists from national media organisations wishing for this social media analysis to be used in their reporting, it was not clear that journalists, the BBC's Rory Cellan-Jones aside, fully appreciated the differences between such analysis and traditional opinion polling (Anstead & O'Loughlin 2012).

This raises important ethical questions about how text mining research is presented in the public domain. The Linguamatics team have tried to make clear that such research should be considered ‘qualitative’ insofar as it offers an understanding of how opinion forms and shifts.[6] Due to the unrepresentative sample, the statistical patterns identified in Twitter data lack the validity and generalisability of traditional polling and have little predictive value for the whole population. However, social media analysis does allow researchers to delve into the data to ask different questions: whose comments are creating what response and why? Why did an issue suddenly re-occur in a debate? Who has power and influence in this environment? The spontaneity of much social media commentary allows researchers to analyse individuals’ reasoning and their emotional responses to events, and on a large scale. Traditionally, this type of analysis has emerged only from research based on in-depth interviews or focus groups.

As we have argued, one of the challenges of mining text from real-time sources such as Twitter is establishing meanings through an awareness of linguistic idioms and broader cultural contexts. Consider Figure 5, which details the volume of positive sentiment tweets for each of the leaders during the third debate.

Figure 5. Volume of Tweets Expressing Positive Sentiment About Party Leaders in the Third Debate


What is the significance of the peak?

In Figure 5, the Y-axis shows tweets coded for positive sentiment towards each leader, measured in 60-second chunks. Notice the spike in the volume of commentary on David Cameron that occurred at 20:45 hours. The reasons for this were initially unclear. Twitter users often use irony and sarcasm, which can be frequently misinterpreted by text mining tools and even, of course, by humans. The increase in positive sentiment here was actually sparked by a well-known comedian and actor, Chris Addison, making a deeply sarcastic remark about David Cameron: '@mrchrisaddison: sky poll just in! David Cameron won the debate!...'. At that stage Addison had approximately 24,000 followers on Twitter (as of February 2013 he had 231,000). But, more importantly, his comment was retweeted by many others. On the night, Sky had published a poll in the middle of the debate that put Cameron in the lead. Many Twitter users felt that Sky were promoting Cameron to the extent of publishing polls that were biased in his favour, since Sky TV is part of the Rupert Murdoch-owned News International group, whose outlets historically tend to favour right-of-centre parties. This example shows the necessity of combining automated natural language processing with human analysis. I2E or any other software is unlikely to possess an understanding of who Chris Addison is or contextual knowledge of public opinion regarding media ownership and a media mogul's support for a party. The spike in data could have been taken at face value, leading to an erroneous finding. However, I2E directed the researchers to this spike and this led to a more detailed examination of how and why Addison's joke was worth retweeting, why some found it shocking that an opinion poll might become a political tool, and hence a broader exploration about prevailing conceptions of authority, credibility, and trust among the British electorate (Anstead and O'Loughlin 2011b).

This case study in 2010 was embryonic, and Linguamatics developers have been working on increasing the validity and reliability of their natural language processing, providing multi-lingual tools, and using social media analysis to segment and target social constituencies. Nevertheless, we contend that the most compelling research in this area will always involve an iterative workflow of human and automated analysis.

Applying Text Mining II: Analysing the Bullygate News Story of 2010

Our second example of applying text mining to online content is Chadwick's (2011a) study of 'Bullygate', a political crisis involving the British prime minister Gordon Brown during early 2010. Here we provide a summary of how some basic text mining and a great deal of manual work were used together in the qualitative analysis of a rapidly evolving political news story; one that revealed some important new aspects of news production.

Traditionally, the literature on news has been united by the fundamental assumption that the construction of political news is a tightly-controlled, even cosy game involving the interactions and interventions of a small number of elites: politicians, officials, communications staff, and journalists. While these elite-driven aspects of political communication are still much in evidence, the hybridisation of older and newer media practices in political communication requires a rejuvenated understanding of the power relations now shaping news.

During a weekend in February 2010, just a few weeks before the most closely-fought British general election campaign in living memory, Gordon Brown, then prime minister, became the subject of an extraordinary media spectacle. The crisis was sparked by revelations in a book about the Labour government by Andrew Rawnsley, one of Britain's foremost political journalists. Extended extracts from the book were printed in the paper edition of the *Observer*, one of Britain's oldest and most respected newspapers, as part of its 'relaunch' edition. The *Observer*'s extracts centered on the prime minister's alleged psychological and physical mistreatment of colleagues working inside his office in Number 10, Downing Street. Bullygate, as it became known, was potentially the most damaging political development of the entire Brown premiership, not only due to its timing—on the verge of a general election—but also its shocking and personalised nature. These were potentially some of the most damaging allegations ever to be made concerning the personal conduct of a sitting British prime minister. The Bullygate affair became a national and international news phenomenon.

But during the course of that weekend and into the early part of the following week, Bullygate took several momentous twists and turns. New players entered the fray, most notably an organization known as the National Bullying Helpline, whose director claimed that her organization had received phone calls from staff inside Number 10, Downing Street. This information created a powerful frame during the middle of the crisis. As the story evolved, events were decisively shaped by mediated interactions among politicians, not-for-profit organisation leaders, professional journalists, bloggers, and citizen activists organized on Twitter. Seemingly clear-cut revelations published in a national newspaper quickly became the subject of fierce contestation, involving competition, conflict, and partisanship, but also relations of interdependence, among a wide variety of actors operating in a wide variety of media settings. Over the course of a few days, following the introduction of largely citizen-discovered pieces of information, serious doubts about the veracity of the Bullygate revelations resulted in the story becoming discredited (Chadwick, 2011a.

Close, real-time, observation and logging, over a five-day period, of a wide range of press, broadcast, and online material, as the story broke, evolved, and faded, enabled a detailed narrative reconstruction of these interactions between politicians, broadcasters, newspaper journalists, and key online media actors. The aim of the analysis was to go beyond the accounts provided by traditional broadcast and newspaper media and to conduct a narrative reconstruction of the hybridised information flows surrounding the story. Chadwick was particularly interested in the roles played by non-elite actors, such as bloggers and influential Twitter users in the construction and contestation of the bullying allegations, and in how interactions between broadcast media and online media players came to shape the development of the story as part of what he termed a 'political information cycle.' Political information cycles, the study argued, are complex assemblages in which the personnel, practices, genres, technologies, and temporalities of online media are hybridized with those of broadcast and press media. This hybridization shapes power relations among actors and ultimately affects the flows and meanings of news.

Method and Setup

Studying political information cycles presents a significant challenge to researchers. Newspaper journalists now frequently post multiple updates to stories throughout the day and night and news sites have widely varying archive policies. The technological limitations of journalists' content management systems, as well as editorial policy, determine whether and how updates, additions, headline alterations, and picture replacements are signaled to readers. Most blogs and a minority of mainstream news outlets, such as the *Guardian* and the *Financial Times*, are transparent about an article's provenance. However, practices vary widely and it is common to see outdated time stamps, the incremental addition of paragraphs at the top or bottom of stories, and headline and URL changes to reflect new angles on developments as they emerge. Sometimes entire stories will simply be overwritten, even though the original hyperlink will be retained. All of these can occur without readers being explicitly notified.

Several 'forensic' strategies were used to overcome these problems. In addition to monitoring key political blogs and the main national news outlets' websites, the free and publicly-available Google Reader was used to monitor the RSS feeds and the timings of article releases from February 20 to February 25, 2010, for the following outlets: *BBC News* (Front Page feed), *Daily Express*, *Daily Mail*, *Daily Mirror*, *Daily Star*, *Daily Telegraph*, *Financial Times*, *Guardian*, *Independent*, *Independent on Sunday*, *Mail on Sunday*, *News of the World*, *Observer*, *Sun*, *Sunday Express*, *Sunday Mirror*, *Sunday Telegraph*, *Sunday Times* and the *Times*. Links were followed back to newspaper websites to check for article modifications, updates, and deletions. Google Reader consists of an effectively unlimited archive of every RSS feed dating back to when a single user first added it to Google's database. Evernote, free and publicly available software, was used to store selected news articles (see http://www.evernote.com).

The broadcast media archiving service, Box of Broadcasts, was used to store content from television, specifically Channel 4 News, BBC News at Ten, the BBC 24-Hour News Channel and ITV News. This enabled the qualitative analysis of pivotal moments during the flow of events on February 20, 21 and 22. This service is available to member institutions of the British Universities Film and Video Council (See http://bobnational.net). Where they existed, links to public transcripts of television and radio shows were also provided.

The Twitter search function (at http://search.twitter.com) was monitored in real-time using a number of queries, such as “national bullying helpline” and hashtags[7] such as “#rawnsleyrot” and “#bullygate” In the period between the introduction of the Twitter search engine and the time of the fieldwork, Twitter only made public the results from approximately three weeks prior to running a query and, at the time of the fieldwork, no robust and publicly-available means of automatically extracting and archiving individual Twitter updates existed. To circumvent these limitations, screen outputs of selected Twitter searches were captured in real-time and stored in Evernote. In April 2010, after the initial fieldwork was conducted, Google launched its Google Replay Search (this later became Google Real-Time Search but was withdrawn in July 2011). This enabled searches of the Twitter archive going back to early February 2010 and it presented the results in a timeline format, though it cannot automatically account for changes to the names of individual Twitter accounts, which had to be followed up manually. Where possible, the Google Replay Search service was used to track and present publicly available links to key Twitter updates.

While this approach is obviously more time-intensive than using automated text mining, it offered several advantages for study of a political crisis that emerged and evolved very quickly ‘in the wild’ and which could not have been predicted in advance. While many text mining studies focus analysis on specific platforms like Twitter or Facebook, in this case it was essential to capture the Bullygate story as it emerged *across and between* media, in unforeseen locations and from the interventions of many previously unknown actors. Focusing on one medium alone would not have captured the story's spread and wider impact. As the episode developed, new information emerged, language use shifted and salient keywords evolved. Indeed, these shifts were part of the power relations in play. Twitter hashtags were particularly important here: they were created, adopted, and dropped with remarkable speed; and new hashtags were added to the information flows as political parties, journalists, and citizen activists sought to exercise power by steering developments. Creating automated and inflexible search queries at the beginning of the crisis would not have captured the story's evolving narratives.

In short, this research combined almost constant and real-time human intervention with a number of tools used for the efficient storage and analysis of digital text and audio-visual content. The application of basic text mining in this case enabled a more nuanced and detailed understanding of the power relations in contemporary networked news systems. It was useful in generating a complex picture of the twists and turns of political news and the increasing centrality of actors such as grassroots activists and citizen journalists who are able to intervene in the news making process for brief but often decisive moments using social media like Twitter, often in real time.

Conclusion

Text mining technologies are likely to become increasingly relevant for social science research, whether we like it or not. Text mining of social media data has already enabled the identification, analysis, and potential prediction of patterns of behaviour and opinion. It is clear, however, that when opening the Pandora's box of big data, researchers will increasingly encounter ontological, ethical, technical, and legal issues. While technology is now essential for the large-scale analysis of big data, the inherent irreducibility and complexity of the social remains. It is extremely unwise, and in any case almost certainly impossible, to leave text mining to software automation.

We can distinguish between discrete methods from methodology. We can use qualitative or quantitative methods, but the most appropriate response to big data for social scientists seeking to explain social, economic and political behaviour is to combine methods into a broader methodology, as in the two case studies we have presented in this chapter. Crawford (2013; see also Lewis et al., 2013) writes, 'new hybrid methods can ask questions about *why* people do things, beyond just tallying up *how often* something occurs. That means drawing on sociological analysis and deep ethnographic insight as well as information retrieval and machine learning'. Substitute hybrid methodologies for methods, and we agree.

These challenges also problematise some traditional distinctions between qualitative and quantitative research. Traditionally, qualitative research has used methods such as focus groups, interviews, and observation to elicit data that enable researchers to interpret sense making among social actors. In many respects, being able to monitor and analyse huge swathes of naturally-occurring online conversations is akin to eavesdropping on large-scale versions of these traditional contexts of qualitative research. Now, however, the sheer quantity of freely-available digital data may often require that qualitative researchers use *quantitative* methods to get a basic grip on their data before qualitative analysis can sensibly begin. In another context, Nigel Thrift has argued that a new style of knowing that he terms 'roving empiricism' is emerging, 'which is more controlled *and* also more open-ended' (2005: 223, italics in original). John Law, meanwhile, has written of the emergence of what he terms 'qualculation': the statistical sorting and ranking of objects, for example, through databases, in order to arrive at qualitative judgements about the justice or significance of situations (Callon and Law 2003: 3). In fields such as security, welfare, and public health, quantitative data is now being analysed on massive scales to help policy makers arrive at fine-grained decisions on whom to target—in these cases for interrogation, aid, or treatment. But such decisions will and should always depend on qualitative decisions and contextual understanding.

Text and data mining must be understood within the context of broader social and technological shifts that have been shaped by the emergence of computerisation and data analysis since the mid-twentieth century. Contemporary programmes of 'e-research' in the sciences, social sciences, and humanities constitute a 'vision' (Dutton 2010: 33) that the integration of disciplinary knowledge, network infrastructures, tools, services and data will allow complex social problems to be addressed (Jeffreys 2010: 51). Social media commentary, bank transactions and weather data, for example, all have different social meanings, and can be archived, analysed and visualised, often in real-time. But just because data can be gathered and stored does not make it valuable, though the current imperative appears to be to collect data now and hope that its usefulness may become clearer later (Wilks & Beston 2010). How commercial and scholarly researchers ought to treat this new mass of data will always be subject to debate. We hope that this chapter has illuminated this uncertain terrain and we invite readers to think imaginatively about how these methods can be combined with others to create compelling new forms of knowledge about the social world.

Appendix: An Overview of Text Mining Tools

Text Mining Tool	Brief Description
Attensity Analyze (Commercial) http://www.attensity.com	Specialises in social media and other unstructured data such as emails and text messages.
ClearForest OneCalais (Commercial) http://www.clearforest.com/	Analyses unstructured data through natural language processing.
COSMOS (Collaborative Online Social Media Observatory) (To be launched in 2014) http://www.cosmosproject.net/	One-stop 'social computational toolkit'. Real-time social media data gathering; various analysis and visualisation services.
Connotate (Commercial) http://www.connotate.com	Cloud-based solution monitors and analyses a wide variety of online content in real time.
Crimson Hexagon Forsight (Commercial) http://www.crimsonhexagon.com	Analyses and visualises social media content, users, and basic audience demographics, as well as proprietary internal enterprise data.
Diction (Commercial) http://www.dictionsoftware.com/	Uses dictionaries (word lists) to search texts for attributes like complexity, activity, optimism, realism, and commonality.

DiscoverText (Commercial) http://texifter.com/Solutions/DiscoverText	Enables collaborative manual and 'teachable' machine analyses of social media and other unstructured documents.
General Sentiment (Commercial) http://www.generalsentiment.com	Analyzes social media content in real time to determine sentiment.
I2E (Commercial) http://www.linguamatics.com/welcome/software/I2E.html	Enterprise-level. Mines unstructured text documents. Allows for building and refining queries.
Language Computer Corporation (Commercial) http://www.languagecomputer.com/	Uses natural language processing technologies, including named entity recognition, information extraction, and question answering.
Lexalytics (Commercial) http://www.lexalytics.com/	Comprises multiple tools, including sentiment analysis, named entity extraction, entity and theme sentiment, and summarisation.
Lextek (Commercial) http://www.lextek.com	Provides information retrieval and natural language processing technology.
Luxid (Commercial)	Searches and analyses

http://www.temis.com/?id=201&selt=1	information within structured databases.
MAXQDA (Commercial) http://www.maxqda.com/service	Content analysis and visualisation, with a module for quantitative text analysis.
Meltwater Buzz (Commercial) http://buzz.meltwater.com/products/buzz/	Monitoring dashboard for analysing content themes, influence, and sentiment.
Mindshare Text Analytics Suite (Commercial) http://www.mshare.net/solutions/mindshare-technologies-text-analytics.html	Analyses a range of online consumer conversations from social media.
Netbase (Commercial) http://www.netbase.com	Uses natural language processing to provide theme, sentiment, and influence analysis of social media content.
Netlytics (Commercial) http://www.netalytics.com/netalytics/	Web-based. Allows users to automate analysis and identify social networks in online communication.
NVIVO (Commercial) http://www.qsrinternational.com/products_nvivo.aspx	Content analysis. Now includes a web and social media module.
Philologic (Free)	A full-text search, analysis and

https://sites.google.com/site/philologic3/	retrieval tool for the analysis of large bodies of text.
Radian6 (Commercial) http://www.radian6.com/	Real-time monitoring dashboard to track and analyse social media content, map demographic and gender data, and gauge sentiment.
Rosette Linguistics Platform (Commercial) http://www.basistech.com/products/	Allows for analysis of unstructured text in Asian, European, and Middle Eastern languages.
Sysomos MAP (Commercial) http://www.sysomos.com/products/overview/sysomos-map/	Analyzes social media content, identifies influential participants, maps demographic and gender data, and gauges sentiment.
TextAnalyst (Commercial) http://www.megaputer.com/textanalyst.php	Summarizes, analyses, and clusters unstructured text documents.
Text Pair (Free) http://code.google.com/p/text-pair/	For identifying similar passages in large volumes of text.
Text Stat (Free) http://neon.niederlandistik.fu-berlin.de/en/textstat/	Produces word frequency lists from multiple languages and file formats.
Visible Intelligence (Commercial) http://www.visibletechnologies.com/produ	For the analysis of unstructured social media data to conduct

cts/visible-intelligence/	sentiment, theme, and influencer analysis.

About the Authors

Lawrence Ampofo earned his PhD in social media, security, and online behaviour at the New Political Communication Unit at Royal Holloway, University of London in 2012. He is founder and director of Semantica Research, a company that provides social media analysis for public, voluntary, and private sector organisations. Lawrence tweets as @lampofo.

Simon Collister is Senior Lecturer in Public Relations and Social Media at London College of Communication, University of the Arts, London. He is currently conducting PhD research at Royal Holloway, University of London's New Political Communication Unit on the mediation of power in networked communication environments. Before entering academia, Simon worked for a number of global communications consultancies, planning and implementing research-led campaigns for a range of public, voluntary, and private sector organisations. Simon tweets as @simoncollister.

Ben O'Loughlin is Professor of International Relations and Co-Director of the New Political Communication Unit at Royal Holloway, University of London. He is specialist advisor to the UK Parliament's soft power committee. He is co-editor of the Sage journal *Media, War & Conflict*. His last book was *Strategic Narratives: Communication Power and the New World Order* (Routledge, 2013). He has recently completed a study with the BBC on international audience responses to the 2012 London Olympics. Ben tweets as @Ben_OLoughlin.

Andrew Chadwick is Professor of Political Science in the Department of Politics and International Relations at Royal Holloway, University of London, where he founded the New Political Communication Unit in 2007. His books include the award-winning *Internet Politics: States, Citizens, and New Communication Technologies* (Oxford University Press); the *Handbook of Internet Politics* (Routledge), which he co-edited with Philip N. Howard, and the multiple award-winning *The Hybrid Media System: Politics and Power* (Oxford University Press). Andrew is the founding series editor of Oxford University Press's book series *Studies in Digital Politics*. He tweets as @andrew_chadwick.

References

Al-Lami, M., Hoskins, A. and O'Loughlin, B. (2012) 'Mobilisation and violence in the new media ecology: the Dua Khalil Aswad and Camilia Shehata cases'. *Critical Studies on Terrorism*, 5 (2), 237-256.

Anderson, C. (2008) 'The End of Theory: The Data Deluge Makes the Scientific Method Obsolete'. *Wired*, July 16th. New York, Conde Nast.

Anderson, C. and Wolff, M. (2010) 'The Web is Dead. Long Live the Internet.' *Wired*, August 17th. New York, Conde Nast.

Anstead, N. and O'Loughlin, B. (2010) 'The Emerging Viewertariat: Explaining Twitter Responses to Nick Griffin's Appearance on Question Time.' UEA School of Political, Social and International Studies Working Paper Series. Norwich, University of East Anglia.

Anstead, N. and O'Loughlin, B. (2011a) 'Semantic Polling and the 2010 UK General Election'. Paper presented at the ECPR General Conference, Reykjavik. Retrieved March 1st, 2012, from: http://www.ecprnet.eu/conferences/general_conference/reykjavik/paper_details.asp?paperid=2590

Anstead, N., and O'Loughlin, B. (2011b) 'The Emerging Viewertariat and BBC Question Time: Television Debate and Real-Time Commenting Online'. *The International Journal of Press/Politics*,16(4): 440-462.

Anstead, N. and O'Loughlin, B. (2012) 'Semantic Polling: The Ethics of Online Public Opinion.' LSE Media Policy Brief 5. Retrieved August 13th, 2013 from: http://www2.lse.ac.uk/media@lse/documents/MPP/Policy-Brief-5-Semantic-Polling_The-Ethics-of-Online-Public-Opinion.pdf

Arusu, A. and Garcia-Molina, H. (2003) 'Extracting Structured Data from Web Pages'. Proceedings of the 2003 ACM SIGMOD International Conference on Management of Data, New York: ACM, 337-348. Retrieved August 13th, 2013 from: http://dl.acm.org/citation.cfm?doid=872757.872799

Archak, N., Ghose, A. and Ipeirotis, P. G. (2011) 'Deriving the Pricing Power of Product Features'. *Management Science* 57(8): 1485–1509.

Asur, S. and Huberman, B.A. (2010). 'Predicting the Future With Social Media'. Paper presented at the International Conference on Web Intelligence and Intelligent Agent Technology, IEEE. Retrieved February 12th, 2012, from http://arxiv.org/pdf/1003.5699.

Awan, A.N., Hoskins, A. and O'Loughlin, B. (2011) *Radicalisation and Media: Terrorism and Connectivity in the New Media Ecology*. London: Routledge.

Bakshy, E., Rosenn, I. Marlow, C. and Adamic, L. (2012) 'The Role of Social Networks in Information Diffusion'. Paper presented at ACM WWW, Lyon, France. Retrieved March 13th, 2012, from http://arxiv.org/abs/1201.4145

Baym, N. K. (2009) 'A Call for Grounding in the Face of Blurred Boundaries'. *Journal of Computer-Mediated Communication* **14**: 720–723.

Bollen, J. (2011) 'Computational Economic and Finance Gauges: Polls, Search, & Twitter'. Paper presented at the Behavioral Economics Working Group, Behavioral Finance Meeting. Palo Alto, CA. Retrieved Februafy 12th, 2012, from http://www.nber.org/~confer/2011/BEf11/BEf11prg.html.

Bollier, D. (2009) 'The Promise and Peril of Big Data. Paper presented at Extreme Inference: Implications of Data Intensive Advanced Correlation Techniques', The Eighteenth Annual Aspen Institute Roundtable on Information Technology, Aspen, Colarado, The Aspen Institute. Retrieved February 12th, 2012, from http://bollier.org/sites/default/files/aspen_reports/InfoTech09_0.pdf

boyd, d. (2008) 'How Can Qualitative Internet Researchers Define the Boundaries of Their Projects: A Response to Christine Hine.' In Annette Markham and Nancy Baym (eds.), *Internet Inquiry: Conversations About Method*. Los Angeles, Sage**:** 26-32.

boyd, d. and Crawford, K. (2011) 'Six Provocations for Big Data. A Decade in Internet Time'. Paper presented at the Symposium on the Dynamics of the

Internet and Society. Oxford. Retrieved 12 February, 2012, from http://www.zephoria.org/thoughts/archives/2011/09/14/six-provocations-for-big-data.html

Callon, M. and Law, J. (2003) ‘On Qualculation, Agency and Otherness’. Centre for Science Studies, Lancaster University, Lancaster LA1 4YN, UK, at http://www.comp.lancs.ac.uk/sociology/papers/Callon-Law-Qualculation-Agency-Otherness.pdf

Castells, M. (2009) *Communication Power*. Oxford, Oxford University Press.

Cellan-Jones, R. (2010) ‘Online ‘sentiment’ around the prime-ministerial debates’. BBC News, April 30th. Retrieved on March 13th, 2013 from: http://www.bbc.co.uk/blogs/thereporters/rorycellanjones/2010/04/online_sentiment_around_the_pr.html

Chadwick, A. (2011a) ‘The Political Information Cycle in a Hybrid News System: The British Prime Minister and the ''Bullygate'' Affair.’ *The International Journal of Press/Politics*, 16(3): 3-29.

Chadwick, A. (2011b) ‘Britain’s First Live Televised Party Leaders’ Debate: From the News Cycle to the Political Information Cycle.’ *Parliamentary Affairs*, 64(1): 24-44.

Chadwick, A. (2013). *The Hybrid Media System: Politics and Power*. Oxford, Oxford University Press.

Chew, C. and Eysenbach, G. (2010) 'Pandemics in the Age of Twitter: Content Analysis of Tweets during the 2009 H1N1 Outbreak. ' *PLoS ONE* 5(11): e14118.

Chunara, R., Andrews, J.R. and Brownstein, J.S. (2012) 'Social and News Media Enable Estimation of Epidemiological Patterns Early in the 2010 Haitian Cholera Outbreak'. *American Journal of Tropical Medicine and Hygiene.* 86(1): 36–45.

Crawford, K. (2013) 'Think Again: Big Data'. *Foreign Policy*, May 9th. Retreived on August 13th, 2013 from http://www.foreignpolicy.com/articles/2013/05/09/think_again_big_data

Culotta, A. (2010) 'Towards detecting influenza epidemics by analyzing Twitter messages'. 1st Workshop on Social Media Analytics. Washington, DC, USA. Retrieved March 14th, 2012, from http://snap.stanford.edu/soma2010/papers/soma2010_16.pdf

Crimson Hexagon (2012a) 'Technical Specifications'. Retrieved May 24th, 2012, http://www.crimsonhexagon.com/technical-specifications/

Crimson Hexagon (2012b) 'Our Quantitative Analysis Methods'. Retrieved May 24th, 2012, from http://www.crimsonhexagon.com/quantitative-analysis/

Dahlberg, L. (2005) 'The Corporate Colonization of Online Attention and the Marginalization of Critical Communication?' *Journal of Communication Inquiry* 29(2): 160-180.

Datasift (2012) 'Pricing'. Retrieved March 13th, 2012, from http://datasift.com/pricing.

Dediu, Horace (2012) 'When will tablets outsell traditional PCs?' *Asymco*. Retrieved March 22nd, 2012, from http://www.asymco.com/2012/03/02/when-will-the-tablet-market-be-larger-than-the-pc-market/

Deloitte (2012) *Measuring Facebook's Economic Impact in Europe*. Retrieved, March 12th, 2012, from https://www.facebook.com/notes/facebook-public-policy-europe/measuring-facebooks-economic-impact-in-europe/309416962438169

Dredge, S. (2011) 'Smartphone and Tablet Stats: What's Really Going on in the Mobile Market?' *Guardian Apps Blog*. London, Guardian Media Group. Retrieved March 26th, 2012, from http://www.guardian.co.uk/technology/appsblog/2011/aug/01/smartphone-stats-2011

Dutton, W.H. (2010) 'Reconfiguring Access in Research: Information, Expertise, and Experience'. In W.H. Dutton and P.W, Jeffreys (eds.) *World Wide Research: Reshaping the Sciences and Humanities*, Boston, MA: The MIT Press.

The Economist (2010) 'The Web's New Walls: How the Threats to the Internet's Openness can be Averted.' London, The Economist Newspaper Limited.

Facebook (2012) 'Graph API'. Retrieved 13th March, 2012, from https://developers.facebook.com/docs/reference/api/.

Freelon, D. (2012) 'Arab Spring Twitter data now available (sort of)'. dfreelon.org. Retrieved March 1st, 2012, from http://dfreelon.org/2012/02/11/arab-spring-twitter-data-now-available-sort-of/

Gannes, L. (2010) 'Twitter Firehose Too Intense? Take a Sip From the Gardenhose or Sample the Spritzer'. *All Things D*. Retrieved March 13th 2012, from https://allthingsd.com/20101110/twitter-firehose-too-intense-take-a-sip-from-the-garden-hose-or-sample-the-spritzer/

Ghose, A., Ipeirotis, P.G. and Sundararajan, A. (2007) 'Opinion Mining Using Econometrics: A Case Study on Reputation Systems'. Paper presented at the 45th Annual Meeting of the Association of Computational Linguistics, Prague, Czech Republic, Association for Computational Linguistics. Retrieved March 1st, 2012, from http://pages.stern.nyu.edu/~aghose/acl2007.pdf

Gibbs, G.R., Friese, S., & Mangabeira, W. C., (2002) 'The Use of New Technology in Qualitative Research. Introduction to Issue 3(2) of FQS'. *Forum: Qualitative Social Research SozialForschung*. Volume 3, No.2, Art. 8, May 2002.

Gilbert, E. and Karahalios, K. (2010) 'Widespread Worry and the Stock Market'. Paper presented at the Fourth International AAAI Conference on Weblogs and Social Media, Washington, DC, AAAI. Retrieved February 22nd, 2012, from http://comp.social.gatech.edu/papers/icwsm10.worry.gilbert.pdf

Gluck, J. and C. Meador (no date) *Analyzing the Relationship Between Tweets, Box-Office Performance, and Stocks*. (Unpublished thesis) Swarthmore PA, Swathmore College. Retrieved March 1st, 2012, from http://www.sccs.swarthmore.edu/users/12/jgluck/resources/TwitterSentiment.pdf

Golder, S., Wilkinson, D. and Huberman, B. (2007) 'Rhythms of Social Interaction: Messaging within a Massive Online Network'. Paper presented at the Third International Conference on Communities and Technology, London. Retrieved March 1st, 2012, from *http://*www.hpl.hp.com/research/idl/papers/facebook/facebook.pdf

Gross, R. and Acquisti, A. (2005) 'Information Revelation and Privacy in Online Social Networks'. Paper presented at WPES'05. 12th ACM Conference on Computer and Communications Security Alexandria, VA. Retrieved March 1st, from, http://www.heinz.cmu.edu/~acquisti/papers/privacy-facebook-gross-acquisti.pdf

Hands, J. and Parikka, J. (2011) ‘Platform Politics’. Retrieved 23rd February, 2012, from http://www.networkpolitics.org/content/platform-politics.

Internet World Stats (2010) 'Internet World Users by Language.' Retrieved 18th May, 2012, from http://www.internetworldstats.com/stats7.htm

Jeffreys, P.W. (2010) ‘The Developing Conception of e- Research’. In W.H. Dutton and P.W, Jeffreys (eds.) *World Wide Research: Reshaping the Sciences and Humanities*, Boston, MA: The MIT Press.

Jones, K.S., (1994) ‘Natural Language Processing: A Historical Review’. *Current Issues in Computational Linguistics: in Honour of Don Walker,* ed. Antonio Zampoli, Nocoletta Calzolari, Martha Palmer (Linguistica Computazionale, vol. 9-10); Pisa, Dodrect, [1994].

Judd, N. (2011) Who Controls 'Twistory?'. *TechPresident.* Retrieved 13th March, 2012, from http://techpresident.com/short-post/who-controls-twistory

Karpf, D. (2012) ‘Social Science Research in Internet Time’. *Information, Communication & Society*, 15(5): 639-661.

Knguyen (2010) 'Facebook Crawler?' Retrieved 13th March, 2012, from http://stackoverflow.com/questions/2022929/facebook-crawler

Lampe, C., Ellison, N. and Steinfield, C. (2006) 'A Face(book) in the Crowd: Social Searching vs. Social Browsing'. Paper presented at CSCW-2006, ACM, New York. Retrieved March 13th, 2012, from www.msu.edu/~nellison/lampe_et_al_2006.pdf

Lee, Dongjoo, J., Ok-Ran and Lee, S. (2008) 'Opinion Mining of Customer Feedback Data on the Web'. Paper presented at the 2nd International Conference on Ubiquitous Information Management and Communication. New York, USA. Retireved March 13th, 2012, from http://ids.snu.ac.kr/w/images/7/7e/IC-2008-01.pdf

Leetaru, K. H. (2011) 'Culturomics 2.0: Forecasting large–scale human behavior using global news media tone in time and space.' *First Monday* 16(9). Retrieved January 12th, 2012, from http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/viewArticle/3663/3040

Lefler, J (2011) *I Can Has Thesis?: A Linguistic Analysis of Lolspeak*. Unpublished Masters Thesis. University of Louisiana at Lafayette, December 2011. Retrieved May 22nd, 2012, from http://etd.lsu.edu/docs/available/etd-11112011-100404/.../Lefler_thesis.pdf

Lewis, K., Kaufman, J., Gonzalez, M., Wimmer, A., and Christakis, N. (2008) ‘Tastes, ties, and time: A new social network dataset using Facebook.com’. *Social networks*, 30(4): 330-342.

Lewis, S.C., Zamith, R. and Hermida, A. (2013) ‘Content Analysis in an Era of Big Data: A Hybrid Approach to Computational and Manual Methods’. *Journal of Broadcasting & Electronic Media*, 57(1): 34-52.

Lidman, M. (2011) 'Social Media as a Leading Indicator of Markets and Predictor of Voting Patterns. Computing Science. (Unpublished Masters Thesis). Umea, Umea University. Retrieved March 13th, 2012, from www.christopia.net/data/school/2011/Fall/social-media-mining/project_proposal/sources/lidman-2011.pdf

Lindsay, R. (2008) ‘Predicting polls with Lexicon’. *Language Wrong*. Retrieved February 9th, 2012, from http://languagewrong.tumblr.com/post/55722687/predicting-polls-with-lexicon

Lohr, S. (2012) ‘The Age of Big Data’. *New York Times*, February 11. Retrieved August 13th, 2013, from http://www.nytimes.com/2012/02/12/sunday-review/big-datas-impact-in-the-world.html?pagewanted=all&_r=0

Manning, C.D., & Schütze, H., (1999) *Foundations of Statistical Natural Language Processing*. Cambridge, MA: The MIT Press.

Mayer, A. and S. L. Puller (2008) ‘The Old Boy (and Girl) Network: Social Network Formation on University Campuses.’ *Journal of Public Economics* 92(1): 329–347.

Mishne, G. and Glance, N. (2006) ‘Predicting Movie Sales from Blogger Sentiment’. Paper presented at the Spring Symposium on Computational Approaches to Analysing Weblogs AAAI. Retrieved March 13th, 2012, from www.nielsen-online.com/downloads/us/buzz/wp_MovieSalesBlogSntmnt_Glance_2005.pdf

O' Connor, B., Balasubramanyan, R,, Routledge, B.R. and Smith, N.A. (2010) ‘From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series’. Paper presented at the International AAAI Conference on Weblogs and Social Media, Washington, DC, May 2010. Retrieved March 13th, 2012, from www.cs.cmu.edu/~nasmith/papers/oconnor%2Bbalasubramanyan%2Broutledge%2Bsmith.icwsm10.pdf

Pang, B. and L. Lee (2008) ‘Opinion Mining and Sentiment Analysis.’ *Foundations and Trends in Information Retrieval* 2(1–2): 1–135.

Papacharissi, Z. and de Fatima Oliveira, M. (2011) ‘The Rhythms of News Storytelling on Twitter: Coverage of the January 25th Egyptian uprising on Twitter’. Paper presented at the World Association for Public Opinion Research

Conference. Amsterdam. Retrieved March 1st, 2012, from *http://*tigger.uic.edu/.../RhythmsNewsStorytellingTwitterWAPORZPMO.pdf

Procter, R., Vis, F., & Voss, A. (2013). Reading the riots on Twitter: methodological innovation for the analysis of big data. International Journal of Social Research Methodology, 16(3), 197-214.

Russell, M.A. (2011) *Mining the Social Web: Analyzing Data from Facebook, Twitter, LinkedIn, and Other Social Media Sites*. Sebastopol, CA, O'Reilly Media.

Rzhetsky, A., Seringhaus, M., & Gerstein, M. (2008) ‘Seeking a new biology through text mining’, *Cell*, 134(1): 9-13.

Sakaki, T., M. Okazaki, et al. (2010) ‘Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors’. Paper presented at WWW2010 - 19th International World Wide Web Conference, Raleigh, North Carolina. Retrieved February 9th, 2012, from http://ymatsuo.com/papers/www2010.pdf

Schonfeld, E. (2010) ‘Wired Declares The Web Is Dead—Don't Pull Out The Coffin Just Yet’. *Techcrunch*. Retrieved February 3rd, 2012, from http://techcrunch.com/2010/08/17/wired-web-dead

Shulman, S. (2011) ‘Twitter Asks DiscoverText to Stop Sharing Tweet Data’. *texifter*. Available at: http://blog.texifter.com/index.php/2011/05/04/twitter-cites-terms-of-service-violation/

Doan, Son, H. Vo, B.-K. and Collier, N. (2011) ‘An analysis of Twitter messages in the 2011 Tohoku Earthquake’. Paper presented at the 4th ICST International Conference on eHealth. Malaga, Spain. Retrieved March 13th, 2012, from http://arxiv.org/pdf/1109.1618

Sudweeks, F., and Rafaeli, S. (1995) ‘How do you get a hundred strangers to agree? Computer-mediated communication and collaboration’. In Harrison, T.M., & Stephen, T.D. (Eds.) *Computer networking and scholarship in the 21st century university*. New York : SUNY Press, 115–136.

Thrift, N. (2012) ‘The insubstantial pageant: producing an untoward land’. *cultural geographies*, 19(2): 141-168.

Thrift, N. (2005) *Knowing Capitalism*. London: Sage.

Veinot, T. (2007) ‘The Eyes of the Power Company: Workplace Information Practices of a Vault Inspector’. *The Library Quarterly* 77(2): 157-180.

Visible Technologies (2012) 'Visible Technologies Acquires Cymfony to Expand Global Product Offering'. Retrieved May 24th, 2012, from http://www.visibletechnologies.com/visible-acquires-cymfony/

Wakamiya, S., Lee, R. and Sumiya, K. (2011a) ‘Towards better TV viewing rates: exploiting crowd's media life logs over Twitter for TV ratings’. Proceedings of the

5th International Conference on Ubiquitous Information Management and Communication, New York. Retrieved March 13th, 2012, from http://dl.acm.org/ft_gateway.cfm?id=1968661&type=pdf

Wakamiya, S., Lee, R. and Sumiya, K. (2011b) ‘Crowd-Powered TV Viewing Rates: Measuring Relevancy between Tweets and TV Programs’. Paper presented at Proceedings of the 16th International Conference on Database Systems for Advanced Applications. Retrieved March 13th, 2012, from

Wanner, F., Rohrdantz, C., Mansmann, F., Oelke, D. and Keim, D.A. (2009) ‘Visual Sentiment Analysis of RSS News Feeds Featuring the US Presidential Election in 2008’. Paper presented at Visual Interfaces to the Social and the Semantic Web. Sanibel Island, Florida, USA. Retrieved March 13th, 2012, from http://data.semanticweb.org/workshop/VISSW/2009/paper/main/6/html

Wilks, Y.A. and den Besten, M. (2010) ‘Key Digital Technologies to Deal with Data’. In W.H. Dutton and P.W, Jeffreys (eds.) *World Wide Research: Reshaping the Sciences and Humanities*, Boston, MA: The MIT Press.

World Economic Forum (2012) *Big Data, Big Impact: New Possibilities for International Development*. Geneva, World Economic Forum. Retrieved, March 26th, 2012 from http://www.weforum.org/reports/big-data-big-impact-new-possibilities-international-development

Notes

[1] Twitter's Public Streaming API is part of the collection of Twitter's Streaming APIs which provide developers with access to the company's stream of Tweet data. 2013, For further detail see 'The Streaming APIs': https://dev.twitter.com/docs/streaming-apis.

[2] Machine learning algorithms can be defined as artificial intelligence that allows computers to 'learn' when exposed to new data, without the need for additional programming.

[3] Freelon suggests object ID datasets 'will be all but useless to anyone without at least a basic understanding of all of the following: APIs and how to retrieve data from them, a programming language like PHP or Python, and a relational database system such as MySQL.' Recreating a full dataset of tweets covering the Arab Spring political unrest using Twitter's permissible methods would take 'months of 24/7 automated querying given Twitter's API limits' (Freelon 2012).

[4] This overview of text mining tools is based on our personal experiences and invaluable consultation with colleagues. Particular thanks go to Peter Fontana and Carolin Gerlitz. Any errors or shortcomings are our own.

[5] 'Monitoring of Complex Information Infrastructure by Mining External Signals', Technology Strategy Board, Award Reference TP: BK067C: David Milward (Linguamatics Ltd) and Ben O'Loughlin (Royal Holloway, University of London).

[6] 'Semantic Polling: The Ethics of Public Opinion', Hansard Society and the LSE Media Policy Project, 5 July 2012, Millbank, London.

[7] A hashtag is a tag embedded in a message posted on the Twitter microblogging service, consisting of a word within the message prefixed with a hash sign (See en.wiktionary.org/wiki/hashtag).