How do politicians use Facebook? An applied Social Observatory

Simon Caton[1], Margeret Hall[2] and Christof Weinhardt[2]

Abstract

In the age of the digital generation, written public data is ubiquitous and acts as an outlet for today's society. Platforms like Facebook, Twitter, Google+ and LinkedIn have profoundly changed how we communicate and interact. They have enabled the establishment of and participation in digital communities as well as the representation, documentation and exploration of social behaviours, and had a disruptive effect on how we use the Internet. Such digital communications present scholars with a novel way to detect, observe, analyse and understand online communities over time. This article presents the formalization of a Social Observatory: a low latency method for the observation and measurement of social indicators within an online community. Our framework facilitates interdisciplinary research methodologies via tools for data acquisition and analysis in inductive and deductive settings. By focusing our Social Observatory on the public Facebook profiles of 187 federal German politicians we illustrate how we can analyse and measure sentiment, public opinion, and information discourse in advance of the federal elections. To this extent, we analysed 54,665 posts and 231,147 comments, creating a composite index of overall public sentiment and the underlying conceptual discussion themes. Our case study demonstrates the observation of communities at various resolutions: "zooming" in on specific subsets or communities as a whole. The results of the case study illustrate the ability to observe published sentiment and public dialogue as well as the difficulties associated with established methods within the field of sentiment analysis within short informal text.

Keywords

Social Observatory, Facebook, sentiment analysis, political discourse, text analytics, computational social science

Introduction

With social media, political parties can bring their message to the public faster, positing on recent events before the interaction and interpretation of local or national media (Stieglitz and Dang-Xuan, 2012). Putting issues onto the public stage they can directly interact with voters, supporters or residents of their election districts, thereby acting locally as well as nationwide. While this conversation is well-addressed on the micro-blog platform Twitter (Böcking et al., 2014; Housley et al., 2014; Mckelvey, 2013; Pak and Paroubek, 2010; Tumasjan et al., 2010), under-addressed are the characteristics of online political sentiment on Facebook. We argue that a significant entrance barrier to Facebook studies is a lack of easily employed, valid measurement systems and tools suitable for non-technically fluent users. Text from Facebook can be spliced for context and content, compared, and measured for sentiment and conceptual domains as a means of community assessment. Sentiment-based artefacts using publicly available data promise unprecedented access to the expectation of issues arising ex-ante, and the totality of effect of incidents ex-post, therefore, enabling researchers and decision makers to analyse, develop, implement and tune policies.

To address this, we present a *Social Observatory: an unobtrusive, low latency*[1] *multi-resolution framework for the observation, analysis and modelling of digital*

[1]National College of Ireland, Ireland
[2]Karlsruhe Service Research Institute, Karlsruhe Institute of Technology, Germany

Corresponding author:
Margeret Hall, Karlsruhe Institute of Technology (KIT), Englerstrasse, 11, Karlsruhe 76131, Germany.
Email: margeret.hall@kit.edu

societies in action (see 'Big Data challenges in the social sciences' section). With a Social Observatory, we aim to realize an automated framework that facilitates, reviews, and assesses specific aspects of online communities via Facebook using qualitative and quantitative methods (see 'Related work' section). Our research contribution is a framework that empowers interdisciplinary researchers with the tools to facilitate the understanding of phenomena within Facebook, as well as the communities they represent (see Appendix 1).

To validate our objective, we present a prototype implementation and case study analysing public political dialogue of German federal politicians (see 'Application of a Social Observatory: Political sentiment in Germany' and 'Zooming in and out of a social network' sections). Our dataset comprises all politicians with a Facebook presence from the five German federal parties:[2] the Christian Democratic Union (CDU/CSU), the Social Democrats (SPD), the Free Democrats (FDP), the Green Party (Grüne), and The Left Party (Die Linke). Using this data set, we evaluate the following research questions: Is Facebook a valid research medium for assessing political discourse online? If so, what are the characteristics of discourse and engagement with and of German politicians on Facebook (see 'Discussion' and 'Conclusion' sections)?

Big Data challenges in the social sciences

Our vision of a Social Observatory is a low latency method for the observation and measurement of social indicators. It is a computer-mediated research method at the intersection of computer science and the social sciences. The term Social Observatory is used in its original context (Hackenberg, 1970; Lasswell, 1967); our framework is the archetypal formalization of interdisciplinary approaches in computational social science. The essence of a Social Observatory is characterised by Lasswell (1967: 49) as follows:

> The computer revolution has suddenly removed age-old limitations on the processing of information [...] But the social sciences are data starved [...] One reason for it is reluctance to commit funds to long-term projects; another [...] is the hope for achieving quick success by 'new theoretical breakthroughs' [...] It is as though we were astronomers who were supposed to draw celestial designs and to neglect our telescopes. The social sciences have been denied social observatories and told to get on with dreams.

This is also in line with the approach of the American National Science Foundation's call for a network of Social Observatories.[3] Today, the notion of a Social Observatory lends itself to social media platforms, as digital mediators of social exchange, discourse and representation. This, as demonstrated by many researchers (Böcking et al., 2014; Burnap et al., 2014; Newman et al., 2003; Tumasjan et al., 2010; Xiang et al., 2010), becomes especially valuable for assessing, modelling, or predicting social phenomena. However, empowering social scientists to access data from Facebook is non-trivial (Burrows and Savage, 2014; Ruppert, 2013; Taylor et al., 2014; Tinati et al. 2014), and tends to be undertaken by scientists in cooperation with the Facebook Research team (e.g. Das and Kramer, 2013; Kramer, 2010; Kramer et al., 2014).

In Figure 1, we illustrate a general architecture of a modern Social Observatory entailing three processes; namely 1) Data Acquisition; 2) Data Analysis; and 3) Interpretation. Please see the Online Appendix for a discussion of the technical implementation. While it is apparent that a Social Observatory captures multiple streams of data, currently few scientific papers or services report this ability in a way easily replicable. This is despite prevalent availability of Application Programming Interfaces (APIs), and an almost endless supply of papers and studies that focus on specific platforms (Atefeh and Khreich, 2013; Burnap et al., 2014; Pak and Paroubek, 2010; Russell, 2013; Schwartz et al., 2013; Tanasescu et al., 2013). Though this article concentrates on Facebook, the architecture could be extended to other platforms.

Data Acquisition is well supported by most social media platforms via REST or streaming APIs, which are underpinned by lightweight data interchange formats like JSON, and authentication with technologies such as OAuth. The challenges instead lie in data volume, velocity, and variety, access rights, and cross-platform differences in curating data. The Big Data aspects of social media data are well known and do not need to be repeated here. With respect to access rights for data, however, we first need to distinguish between public data (like a Tweet or Facebook page) and personal data (like a Facebook profile). The authorisation rights for these types are significantly different. Although it has been shown that gamified settings can enable access to personal data (Hall et al., 2013a, 2013b), we expect a Social Observatory to rely mainly on public data, as opposed to studies like Kramer et al. (2014) and Schwartz et al. (2013).

Lastly, the method of data curation is not without its ambivalence. For example Twitter data curation tends to be forward-facing; accessing future Tweets that fulfil a specific set of attributes starting at a given time point. Facebook is retrospective; given a Facebook entity (e.g. a person, or page) researchers access current and historical posts, profiles, likes etc. From the perspective of analysing social data, this subtle difference significantly alters the effort and planning needed to curate a data
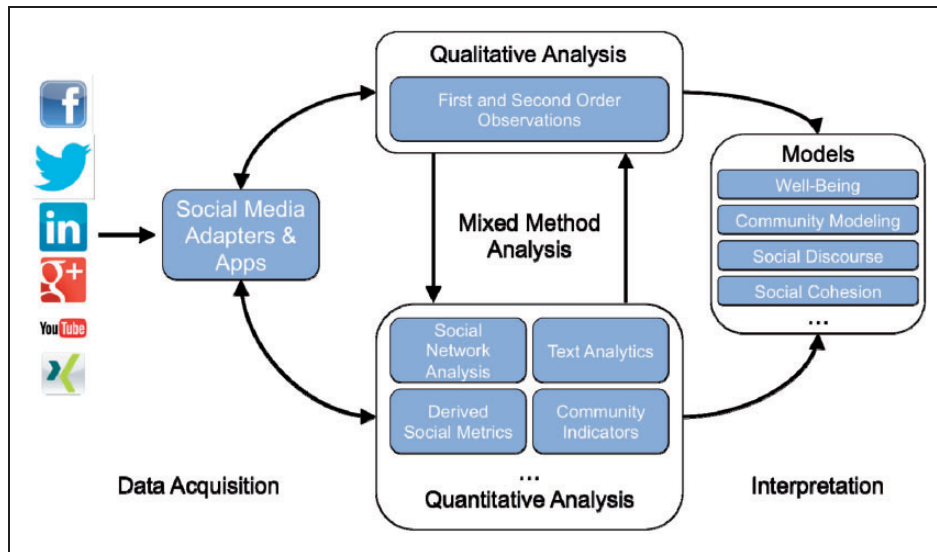
Figure 1. A general architecture for a Social Observatory.

set and the implicit biases associated with the method (González-Bailón et al., 2014; Ruths and Pfeffer, 2014). The technical challenges also differ significantly from receiving a continuous stream of data (i.e. tweets) vs. Facebook's paginated results. The latter incites large numbers of API calls, which are not limitless.

(Mixed Method) Analysis as illustrated in Figure 1 is inherently iterative and interdisciplinary. Foreseeable is repeated interaction with social media adapters and apps. While approaches from computer science and computational social science are becoming more prevalent, the question of research methodology is often a poignant discussion point and challenge that cannot be overlooked; computer and social scientists leverage diverse and often non-overlapping research methodologies. Therefore, a Social Observatory needs to accommodate a vast array of (interdisciplinary) methodological approaches.

Irrespective of methodology, an important feature of a Social Observatory is the ability to view a community at a variety of resolutions; starting from an individual micro layer, and progressively *zooming out* via ego-centric networks, social groups, communities, and demographic (sub)groups, up to the macro layer: community. This ability is of significant importance for understanding a community as a whole.

Interpretation is domain specific in nature, and should be decided according to the proposed research questions. Our architecture supports inductive and deductive research.

Related work

A new approach in the area of information-driven decision support is found in computational social science (Cioffi-Revilla, 2014, 2010), where the interaction of technology, online communities and individuals' perceptions are investigated at a previously unmanaged scale (Burrows and Savage, 2014; Savage & Burrows, 2007; Taylor et al., 2014; Tinati et al., 2014). In an exhaustive survey, Wilson et al. (2012) constructed five supra-categories for Facebook-based research: descriptive analysis of users, motivations for using Facebook, identity presentation, the role of Facebook in social interactions, and privacy and information disclosure. In terms of a Social Observatory, all five categories could be addressed, whereas this paper concentrates on descriptive user analysis and social interactions in an unobtrusive manner. Recognizable is that the usage of Facebook's API by non-Facebook staff or partners to support unobtrusive studies is low.

Many of the commonly applied methods in community analysis like judging communal sentiment, assessing tie strength, or participation and/or exchange in given contexts are often done qualitatively. Human-centric approaches have a long history and are well applied in varied domains (Hsieh and Shannon, 2005; Kassarjian, 1977), but lack scalability. When dealing with the volume required by Big Data analyses, either crowdwork (e.g. Hall and Caton, 2014; Paolacci et al., 2010) or automated programs are generally required. Crowdwork for the analysis of items like status updates and tweets however poses ethical issues (Markham and Buchanan, 2012), and can run afoul of platforms' terms and conditions.

The (social) scientist needs the necessary systems, and tools to leverage computational approaches. Text analysis, as a mechanism for measuring social impact, is becoming increasingly validated as a proxy for social phenomena (Böcking et al., 2014;

Chung and Pennebaker, 2014; Housley et al., 2014; Mckelvey, 2013). Twitter-based studies are common in the social media space and address a variety of social science-oriented research questions. It has, however, been established that sentiment and conversation styles differ across platforms (Davenport et al., 2014; Lin and Qiu, 2013), though the available tools do not match this research need. Facebook tools tend to rely either on crawling techniques, which cannot fully acquire paginated Facebook data and are disallowed in the terms and conditions, or data extraction via the Graph API but either focus on the logged-in user[4] or do not return data in full.[5]

Several authors have addressed the creation of frameworks for supporting Twitter studies (Burnap et al., 2014; Housley et al., 2014; Stieglitz and Dang-Xuan, 2012; Pak and Paroubek, 2010). These lack the corresponding technical infrastructure that allows researchers to create new, build on, or replicate the studies.[6] The closest in reach to a Social Observatory are those where the infrastructure is both open-source and requires minimal knowledge of computational infrastructure in order to be accessed (Housley et al., 2014), or the tools are of a plug and play nature (Kivelä and Lyytinen, 2004; McCallum, 2002).[7]

Key contribution differences are the observation viewpoint and elicitation of points of reference. Many studies observe the Twitter landscape at a macro level, whereas our interest is to facilitate micro, meso and macro observations. For example, Calvo and D'Mello (2010), Hampton et al. (2011) and O'Connor et al. (2010) demonstrated the predictive power of self-reported interests in social profiles and the observation of social practices. While the scientific value of such work is significant, their isolated investigations only give us insights into well-grounded research processes rather than assisting in the construction of a general approach. Similarly, Mitchell et al. (2013) investigate a macro-scale dataset of happiness, urbanization and obesity correlates, but do not create a generalizable model for wide-scale usage. Allen et al. (2014) and Jaho et al. (2011) investigated how content traversed social graphs, and explored opportunistic mechanisms for the dissemination of content via social structures. A focus of their work was mechanisms for community detection, and subsequent analysis of social structures for observing information paths through social networks. However, the emphasis is not on analysing the communities themselves.

Two mechanisms are widely used to support the automated recognition of written sentiment: corpus-based approaches and dictionary-based approaches (Turney and Pantel, 2010). The corpus-based approach is based on the co-occurrence of words, relying on the latent relation hypothesis, stating that words with similar meaning or sentiment co-occur more frequently (Turney and Pantel, 2010). Given a set of known and evaluated words, this methodology identifies words with similar orientation. This can be especially useful when searching for instances of sarcasm or irony, otherwise lost in the dictionary-based approach (Liu, 2010). Dictionary-based approaches use predefined word lists containing sentiment-loaded words. By scanning the considered text, sums of positive and negative affect can be derived, usually normalized regarding the length of the overall text. Kramer subtracts said sums to get a one-dimensional measure of sentiment (Kramer, 2010; Kramer et al., 2004), whereas Golder and Macy argue the independence of both dimensions by measuring them separately (Golder and Macy, 2012). The dictionary-based approach, however, is unable to find domain specific orientations and context oriented sentiment (Dodds et al., 2011; Thelwall et al., 2010).

Notable dictionary-based tools are *Linguistic Inquiry and Word Count* (LIWC) (Pennebaker et al., 2007; Tausczik and Pennebaker, 2010), *Text Analysis and Word Count* (TAWC) (built upon the LIWC 2007 dictionary), *SentiWordNet* (Baccianella et al., 2010) and *OpinionFinder* (Wilson et al., 2005). *SentiWordNet* sums up possible positive and negative sentiment and the third term of "neutrality" (Baccianella et al., 2010), *OpinionFinder* classifies subjectivity and objectivity within sentences (Wilson et al., 2005). To date, both lack linguistic localization, a feature making LIWC's 13 languages favourable.

Using short informal text as the foundation of sentiment measurement is challenging (Thelwall et al., 2010) due to length restrictions, the usage of abbreviations and emotional tokens, slang in various forms and styles and truncated sentences (Wang et al., 2014). Yet, the existence of items like emoticons can help to understand the intended sentiment. The use of positive and negative, or positive, negative, and neutral classifications of social media texts as opposed to more contextual sentiment is a common method (Burnap and Williams, 2014; Pak and Paroubek, 2010). A foundational paper from Go et al. (2009) looked at the classification of Twitter sentiment from the commercial perspective, identifying positive and negative tweets based on query terms of emoticons. Kouloumpis et al. (2011) found that intensifiers are most useful in the automated detection of sentiment in tweets. This study found that part-of-speech features are not necessarily useful in automated sentiment detection. A study by O'Connor et al. (2010) applied positive and negative sentiment scoring to the 2008 US presidential elections and found the method can be used to supplement consumer confidence polls.

Notable studies from Facebook Research also look at public sentiment. Kramer (2010) used status

updates based in the United States to create a composite well-being index. This has since been criticised in Wang et al. (2014), who state that Facebook status messages are not appropriate for well-being assessment, but rather mood regulation. Another series of studies by Kramer (2012; Kramer et al., 2014) reviews emotional contagion on Facebook. These findings support that short informal text like Facebook status updates can be used to measure sentiment online.

LIWC originally was not intended to be used on short informal text, but to analyse text of expressive and therapeutic writing (Pennebaker et al., 2007; Wang et al., 2014). However, its expansive psychometric dictionary offers a unique opportunity to reveal the latent emotional context of text-based data. LIWC has been shown to possess excellent precision and recall abilities with high but not overfitting correlations in the analysis of latent sentiment (Mahmud, 2014; Salas-Zárate et al., 2014), but machine learning approaches often perform better for prediction tasks (Balahur and Hermida, 2012; Komisin and Guinn, 2012). LIWC returns the percentage of words across the categories of social processes, affective processes, cognitive processes, perceptual processes, biological processes, work and achievement, as well as punctuation and structural details (Pennebaker et al., 2007; Tausczik and Pennebaker, 2010). Per cent based information shows the latent context and relative worth of categories in speech. This facilitates measuring change, looking for group-based patterns, monitoring individual trends and identifying psycholinguistic profiles.

We note the study of Tumasjan et al. (2010), which concentrates on the application of LIWC to text gained from German politicians' Twitter handles in advance of the 2009 elections. Our analysis has several distinct differences. We used the German dictionary (Wolf et al., 2008), rather than translating text content to English for analysis to retain the original intention of the writer as closely as possible. Whereas Tumasjan and colleagues review selected LIWC categories, we consider all German dictionary categories and established psycholinguistic profiles. Finally, the aim of our study is a descriptive analysis of political messaging on Facebook. It is not a prediction task.

Application of a Social Observatory: Political sentiment in Germany

Still unknown and an open challenge for social researchers is the actual impact and relationship between politicians on Facebook and their followers. Does some relationship exist, and if so, what are the important parameters thereof? This study reviews 54,655 posts and 231,147 comments by 257,305 unique users at three granularity levels (all posts and comments per party; monthly posts and comments per party; individuals' posts and comments per party) in the year preceding and one month after the 2013 federal elections. We establish macro trends, leading to discussions on the difference between politicians and constituents, then concentrate on discourse related to campaigning and individual discourse patterns. Each 'zoom' of the Social Observatory reveals telling yet sometimes-contradictory indicators.

On the 2013 German federal elections[8]

Germany's multi-cameral Parliament tends to be composed of five different political parties. Centre of right and right parties are the former governing coalition of the CDU/CSU and the FDP, and centre of left and left are the Grüne, SPD, and Linke. The CDU/CSU is the largest party of both the 17th and 18th federal Parliament, and SPD is the largest of the opposition parties from the 17th and 18th Parliaments. The FDP did not obtain enough votes in total to maintain their representation and is not a member of the 18th Parliament. While the CDU/CSU came near to an absolute majority (which in and of itself signifies a largely stable political culture and election series), they did not have enough votes to form a government without a partner. Without the FDP in Parliament, the CDU/CSU was forced to find a new coalition partner. This was the SPD, which obtained the second highest amount of votes in the election.

Descriptive aspects of German Parliament members on Facebook

From the 620 members of the 17th German Parliament we established a convenience sample considering whether they have a publicly available Facebook account or not, finding 187 politician presences[9] with an open profile or page on Facebook, representing approximately 30% of Parliament. Post refers to text pushed by politicians; comments refer to responses by constituents and politicians themselves. Users who only liked a politician's Facebook page are disregarded. Table 1 illustrates some representative aspects of our dataset.

Facebook is used mainly as a medium for promoting individual (political) agendas. Interactions between politicians are relatively low: 3,883 occurrences (0.23%) across all profiles. Figure 2 visualises interactions between politicians and their audience, capturing 85,679 bi-directional edges considering only text-based interactions, 345,704 considering only likes and 385,936 when considering both. On average, politicians and individual audience members interacted 2.70 times via comments, with a maximum of 1,503; 4.30 and 998 respectively for likes, and 4.45 and 1554 considering both.

Table 1. Descriptive attributes of dataset (numbers are rounded for representation purposes).

Party	Proportion of 17th German Bundestag	Proportion of Facebook dataset	Posts	Comments	Likes	Audience[10]
Grüne	11	11	6,586	41,744	194,528	38,665
CDU/CSU	38	40	20,006	68,667	493,891	119,212
FDP	15	11	4,835	26,703	118,215	21,046
Die Linke	12	13	8,886	26,471	178,816	24,986
SPD	23	25	14,342	67,562	501,483	80,300
Total	100	100	54,655	231,147	1,486,933	257,305
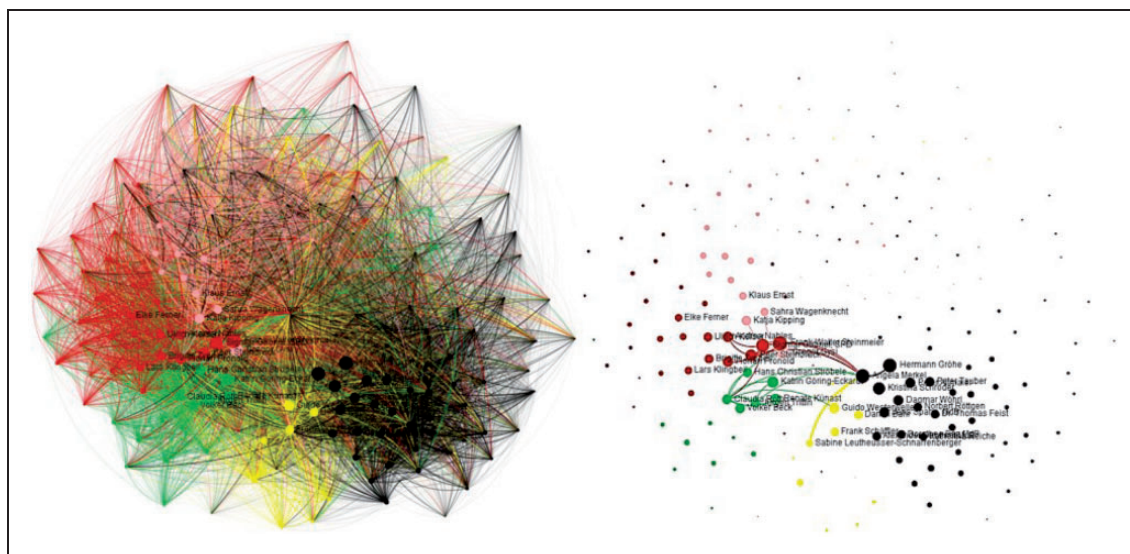
Figure 2. The extracted social interaction graph with all and weightiest edges.

Politicians posted on average 292 times (just over a post a day). The average profile contains 29,301 words, from which 25% were six letters or more (a measure of linguistic complexity). The average post length was 40.8 words, differing from the findings of Kramer (2010), who found that the average length of a Facebook post is nine words. This finding and its discrepancy compared to Kramer's results likely has its origin in the language of this sample.

Timing of posts and comments as well as daily patterns suggests that politicians see their positions as jobs, while constituents act as if their elected officials should be constantly available. Figure 3 depicts the continuum of hourly posting behaviour, with politicians posting in the morning and at lunchtime, and constituents responding in the afternoon. Politicians also tend to post on working days, whereas constituent volume shows no significant difference between weekdays and weekends (Figure 4).

Public conversation maps temporally and in interest-based ways to the realistic and real-time public events in Germany. The monthly distribution of posts and comments depicted in Figure 4 show an increase in activity leading to the elections with two exceptions: December 2012, also observed in 2009–2012, and July 2013 during Parliament's summer recess. December is also a "slow" period for comments. Posting activity significantly dropped in October 2013, directly after the elections. This drop is not reflected in the comments, nor is the recess drop in July. Comments show spikes in November 2012 and March 2013, corresponding to interest in the various public scandals of the former German President Christian Wulff (Figure 5).[11]

Negative emotions, anger and money discussions are positively related ($r_s(331) = .137$, $p < .0005$; $r_s(331) = .184$, $p < .0005$), reflecting on-going public sentiments at financial bailouts to neighbouring countries across the European Union. The most commonly repeated post was "STOPPT die Massentötung in Rumänien! STOPPT die Tatenlosigkeit aller Verantwortlichen in der EU! JETZT!" (*Stop the mass murders in Romania! Stop the inaction of EU stakeholders! Now!*), referring to wildly unpopular policy inaction over Romanian 'fur farming'. 117 unique users repeated this single post 234 times.
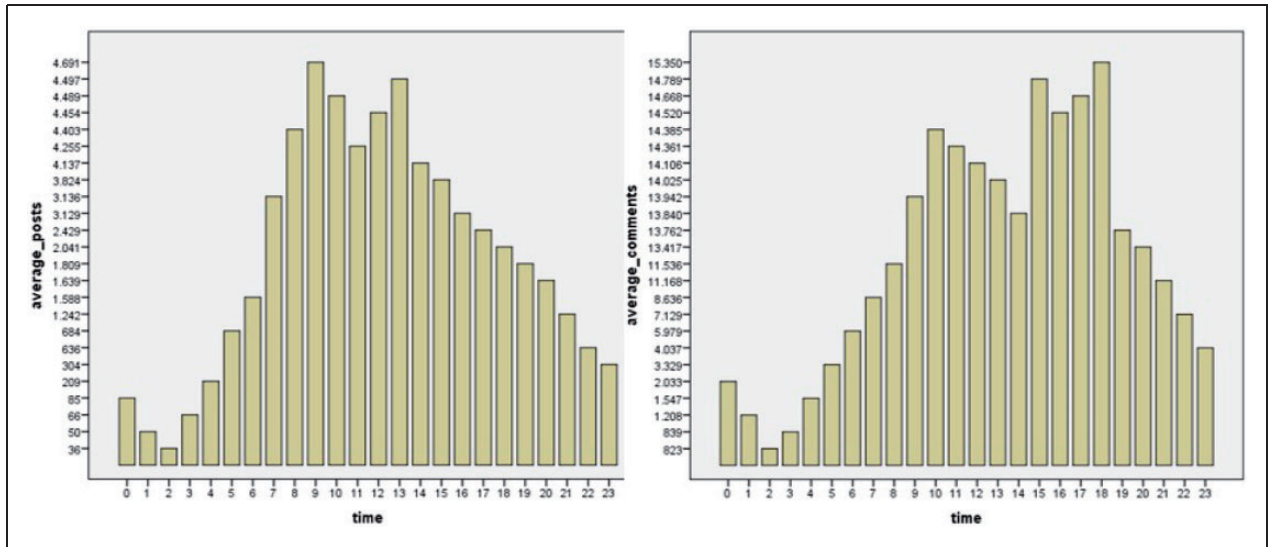
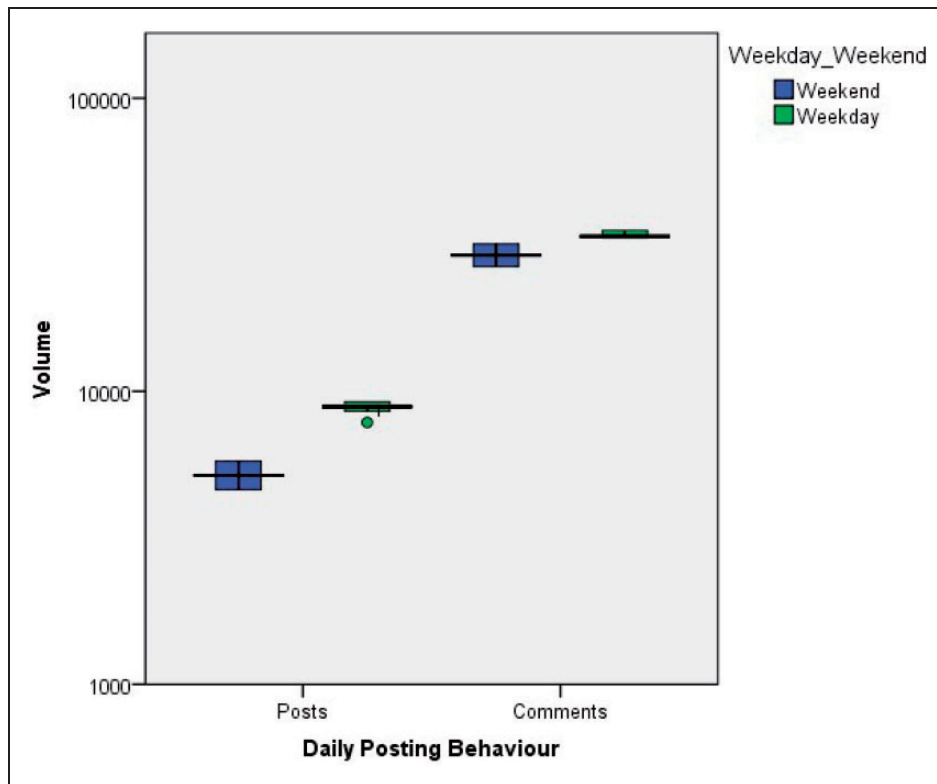**Figure 3.** Distributions of hourly posting behaviours, posts and comments.

Figure 4. Weekday and weekend post and comment activity (logarithmic scale).

Zooming in and out of a social network

Macro-level assessment

Table 2 is a rather unintuitive matrix of the nearest neighbours of each party's posts and comments, indicating that the platforms of the politicians are not responding well to the interests of their Facebook audiences – in fact, the audiences are in some cases nearest to the platforms of rival parties. Constituents and politicians are distinct groups: all comments are nearest to other comments and all posts are closest to other posts.
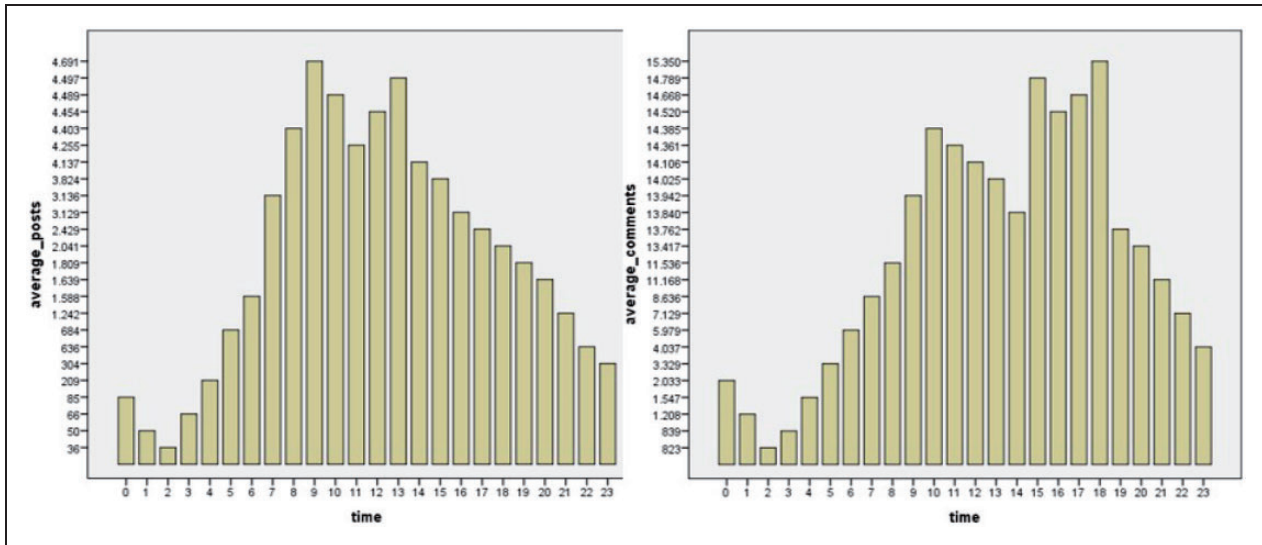
Figure 5. Monthly post and comment activity.

Table 2. Nearest neighbours, politicians and constituents where $k = 5$.

	$k = 1$	$k = 2$	$k = 3$	$k = 4$	$k = 5$
CDU/CSU comments	Grüne (4.082)	SPD (4.209)	4.303 (FDP)	Linke (4.655)	Grüne$_p$ (10.487)
Linke comments	SPD (2.017)	Grüne (3.170)	FDP (3.413)	CDU/CSU (4.665)	FDP$_p$ (10.156)
FDP comments	Grüne (3.050)	Linke (3.413)	SPD (3.461)	CDU/CSU (4.303)	Grüne$_p$ (10.156)
Grüne comments	FDP (3.050)	Linke (3.170)	SPD (3.210)	CDU/CSU (4.082)	Grüne$_p$ (9.872)
SPD comments	Linke (2.017)	Grüne (3.210)	FDP (3.461)	CDU/CSU (4.209)	FDP$_p$ (9.982)
CDU/CSU posts	SPD (4.140)	Linke (5.201)	FDP (5.507)	Grüne (6.041)	SPD$_c$ (10.523)
Linke posts	SPD (4.386)	FDP (4.645)	CDU/CSU (5.201)	Grüne (6.089)	SPD$_c$ (10.523)
FDP posts	Linke (6.645)	SPD (4.730)	CDU/CSU (5.507)	Grüne (5.870)	SPD$_c$ (9.982)
Grüne posts	FDP (5.870)	SPD (5.898)	CDU/CSU (6.041)	Linke (6.089)	Grüne$_c$ (9.872)
SPD posts	CDU/CSU (4.140)	Linke (4.386)	FDP (4.730)	Grüne (5.898)	SPD$_c$ (10.184)

Comments are more similar to each other (2.017–4.665) than posts are similar to other posts (4.140–6.645). Distance is revealing: e.g. politicians from the governing party (CDU/CSU and FDP) and SPD are expected to be dissimilar but rather are one another's nearest neighbours, while governing and opposition block members largely do not occupy the same space. Only the SPD and Grüne have party and constituent closeness at $k = 5$, but this is not the case for the Linke, CDU/CSU, or FDP. In no case is a same party-constituent pairing closer than $k = 5$.

As the space is small but not equal with an absolute range from 2.017 (Linke comments and SPD comments), to 10.523 (Grüne posts and SPD comments) (Table 2), high dimensionality does not unexpectedly compress the data. As there are no "popular" hubs,

we can also reject that hubness is driving the results (Radovanovic et al., 2010).

As the above matrix occupies a relatively small space, a paired sample t test was employed (Table 3) to review if the patterns of speech are statistically the same. The results find that overall the five parties have significant differences in feed patterns as represented by their respective LIWC categorizations. A hyperplane of 64 LIWC sentiment categories are assessed for 45 unique party-constituent permutations. There are statistically significant differences in 35 political party and audience pairings out of the possible 45.

While some results are not unanticipated, other pairings are unusual. There is no significant difference between the posts or comments of the two centre-right parties CDU/CSU and FDP ($t(63)$ -1.788,

Table 3. Paired Sample t-Tests.

|			Paired Differences						
					95% Confidence Interval				
		Mean	Std. Deviation	Std. Error Mean	Lower	Upper	t	df	Sig. (2-tailed)
Pair 1	CDUCSU_comments - CDUCSU_posts	.37750	.87798	.10975	.15819	.59681	3.440	63	.001
Pair 2	CDUCSU_comments - DIE_Linke_comments	−.02328	.20852	.02606	−.07537	.02880	−.893	63	.375
Pair 3	CDUCSU_comments - DIE_Linke_posts	.33047	.86925	.10866	.11334	.54760	3.041	63	.003
Pair 4	CDUCSU_comments - FDP_comments	.01953	.18108	.02263	−.02570	.06476	.863	63	.391
Pair 5	CDUCSU_comments - FDP_posts	.31187	.83760	.10470	.10265	.52110	2.979	63	.004
Pair 6	CDUCSU_comments - Grüne_comments	.04047	.15789	.01974	.00103	.07991	2.051	63	.044
Pair 7	CDUCSU_comments - Grüne_posts	.40281	.82997	.10375	.19549	.61013	3.883	63	.000
Pair 8	CDUCSU_comments - SPD_comments	−.02422	.17064	.02133	−.06684	.01840	−1,135	63	.260
Pair 9	CDUCSU_comments - SPD_posts	.32328	.79619	.09952	.12440	.52216	3.248	63	.002
Pair 10	CDUCSU_posts - DIE_Linke_comments	−.40078	.86726	.10841	−.61742	−.18415	−3,697	63	.000
Pair 11	CDUCSU_posts - DIE_Linke_posts	−.04703	.27204	.03400	−.11498	.02092	−1,383	63	.172
Pair 12	CDUCSU_posts - FDP_comments	−.35797	.85170	.10646	−.57072	−.14522	−3,362	63	.001
Pair 13	CDUCSU_posts - FDP_posts	−.06563	.29366	.03671	−.13898	.00773	−1,788	63	.079
Pair 14	CDUCSU_posts - Grüne_comments	−.33703	.82788	.10348	−.54383	−.13023	−3,257	63	.002
Pair 15	CDUCSU_posts - Grüne_posts	.02531	.25991	.03249	−.03961	.09024	.779	63	.439
Pair 16	CDUCSU_posts - SPD_comments	−.40172	.88207	.11026	−.62205	−.18139	−3,643	63	.001
Pair 17	CDUCSU_posts - SPD_posts	−.05422	.15282	.01910	−.09239	−.01604	−2,838	63	.006
Pair 18	DIE_Linke_comments - DIE_Linke_posts	.35375	.82152	.10269	.14854	.55896	3.445	63	.001
Pair 19	DIE_Linke_comments - FDP_comments	.04281	.13607	.01701	.00882	.07680	2.517	63	.014
Pair 20	DIE_Linke_comments - FDP_posts	.33516	.79225	.09903	.13726	.53306	3.384	63	.001
Pair 21	DIE_Linke_comments - Grüne_comments	.06375	.15537	.01942	.02494	.10256	3.282	63	.002
Pair 22	DIE_Linke_comments - Grüne_posts	.42609	.82469	.10309	.22009	.63209	4.133	63	.000
Pair 23	DIE_Linke_comments - SPD_comments	−.00094	.10574	.01322	−.02735	.02547	−.071	63	.944
Pair 24	DIE_Linke_comments - SPD_posts	.34656	.77837	.09730	.15213	.54099	3.562	63	.001
Pair 25	DIE_Linke_posts - FDP_comments	−.31094	.80137	.10017	−.51111	−.11076	−3,104	63	.003
Pair 26	DIE_Linke_posts - FDP_posts	−.01859	.17153	.02144	−.06144	.02425	−.867	63	.389
Pair 27	DIE_Linke_posts - Grüne_comments	−.29000	.79408	.09926	−.48836	−.09164	−2,922	63	.005
Pair 28	DIE_Linke_posts - Grüne_posts	.07234	.28742	.03593	.00055	.14414	2.014	63	.048
Pair 29	DIE_Linke_posts - SPD_comments	−.35469	.84619	.10577	−.56606	−.14332	−3,353	63	.001
Pair 30	DIE_Linke_posts - SPD_posts	−.00719	.19851	.02481	−.05677	.04240	−.290	63	.773
Pair 31	FDP_comments - FDP_posts	.29234	.77422	.09678	.09895	.48574	3.021	63	.004
Pair 32	FDP_comments - Grüne_comments	.02094	.09772	.01221	−.00347	.04535	1.714	63	.091
Pair 33	FDP_comments - Grüne_posts	.38328	.79445	.09931	.18483	.58173	3.860	63	.000
Pair 34	FDP_comments - SPD_comments	−.04375	.11730	.01466	−.07305	−.01445	−2,984	63	.004
Pair 35	FDP_comments - SPD_posts	.30375	.75652	.09456	.11478	.49272	3.212	63	.002
Pair 36	FDP_posts - Grüne_comments	−.27141	.77293	.09662	−.46448	−.07833	−2,809	63	.007
Pair 37	FDP_posts - Grüne_posts	.09094	.29145	.03643	.01813	.16374	2.496	63	.015
Pair 38	FDP_posts - SPD_comments	−.33609	.81996	.10249	−.54091	−.13127	−3,279	63	.002
Pair 39	FDP_posts - SPD_posts	.01141	.22669	.02834	−.04522	.06803	.403	63	.689
Pair 40	Grüne_comments - Grüne_posts	.36234	.76808	.09601	.17048	.55420	3.774	63	.000
Pair 41	Grüne_comments - SPD_comments	−.06469	.12972	.01622	−.09709	−.03228	−3,989	63	.000
Pair 42	Grüne_comments - SPD_posts	.28281	.73739	.09217	.09862	.46701	3.068	63	.003
Pair 43	Grüne_posts - SPD_comments	−.42703	.84078	.10510	−.63705	−.21701	−4,063	63	.000
Pair 44	Grüne_posts - SPD_posts	−.07953	.24361	.03045	−.14038	−.01868	−2,612	63	.011
Pair 45	SPD_comments - SPD_posts	.34750	.79106	.09888	.14990	.54510	3.514	63	.001
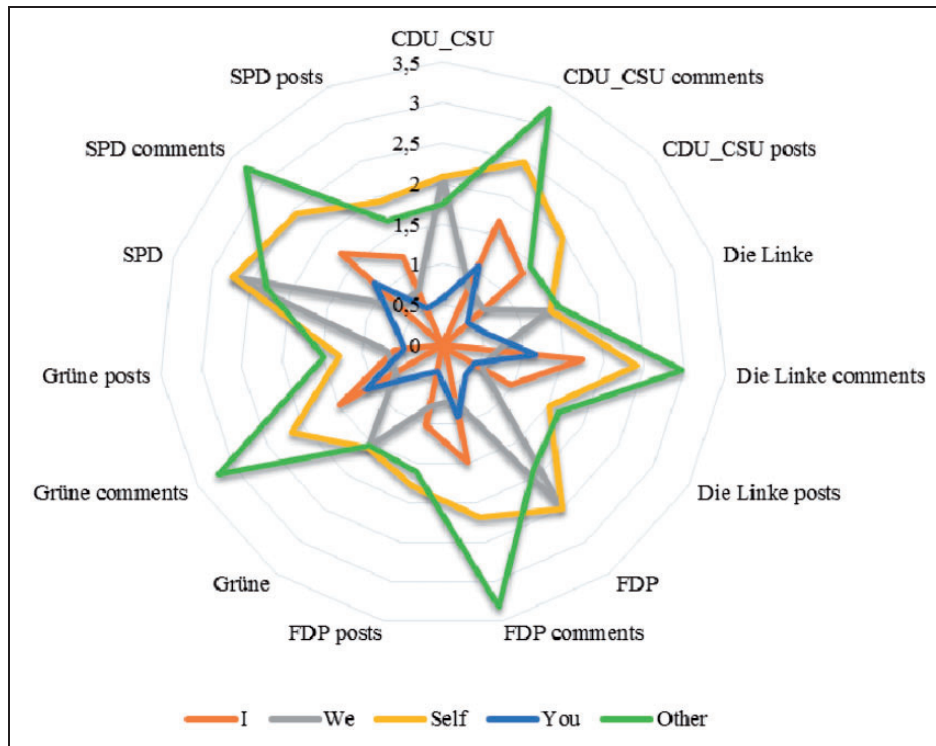
Figure 6. Social references in party manifestos, posts and comments.

$p < .05$), or between the leftist parties SDP and Linke ($t(63) = -.290$, $p < .05$). Unexpectedly, no significant differences between the posts and comments of either the right-oriented parties CDU/CSU or FDP, and the socialist party Linke ($t(63) = -.893$, $p < .05$); ($t(63) = -.867$, $p < .05$) are found. Interestingly, the only non-significant difference of the Grüne was between that of the posts of the CDU ($t(63) = .799$, $p < .05$). All other pairings with the Grüne were significantly different. All post-comment combinations have significant differences, which is supported by the results of the nearest neighbour test.

These differences between relationships as found in the nearest neighbours and t-tests are interesting, as it suggests that politicians and their audiences on Facebook often concentrate on different points, giving importance to different topics across their general discussions. When considering only the posts, this finding supports the assumption that there is a diversity of political conversation amongst Facebook users. As the parties are platform based, this is a positive finding. The results defy the thesis of linguistic accommodation of Niederhoffer and Pennebaker (2002); a reason for the lack of coalescence here could be that conversation partners change too rapidly to adapt to one another. It is worth noting that the overall corpus follows the pattern of polite discussion put forth in Brown and Levinson (1987) and Pennebaker et al. (2003).

Meso-level assessment

There is a distinct propensity to discuss in present tense, which suggests that politicians on Facebook are not 'campaigning' in the traditional sense, but are rather discussing their daily activities. Considering the population, this is an unexpected finding. Whereas it may not be unusual for politicians and political discourse to focus on the present rather than the past, the absence of future references, especially in the face of national elections, is unanticipated (Figure 6). Manifestos have 3.19 times more references to the present than the past and 3.05 times more references to the present than the future, with the exception of the Grüne manifesto that has an inverse present–future relationship. Posts are slightly more balanced with present/past references having a 1.57 difference and present/future discrepancies at 2.73. Comments are the most present-focused, with audiences referring to the present 3.23 times more than the past and 4.46 times more than the future. The findings reported in Tausczik and Pennebaker (2010) of a political discourse study by Gunsch et al. (2000) state that this could also be related to positive campaigning rather than 'dirty' campaigning.

Reviewing this further, no significant correlation exists between positivity, negativity, use of first or third person, and tense and thereby does not replicate (Gunsch et al., 2000). The authors also state that first person references are related to positive campaigns and
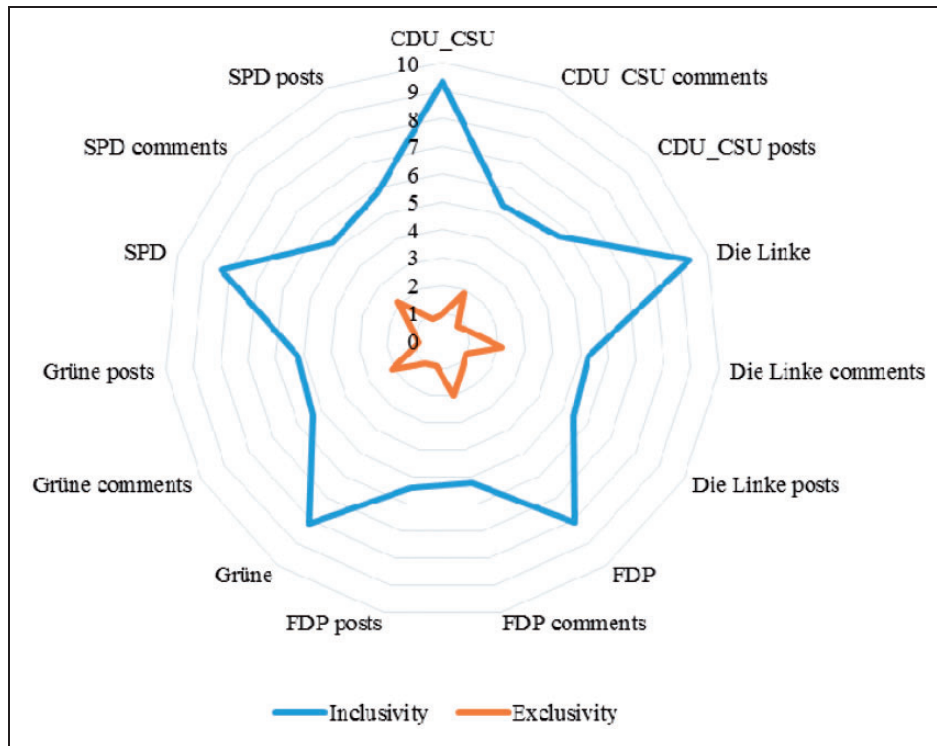
Figure 7. Inclusive and exclusive references in manifestos, posts and comments.

third person campaigns are related to negative campaigning. This is again a positive finding. Also rejected is that the social aspects reflect an "Us-Them" mentality, when taking the relative frequency of inclusivity and exclusivity into consideration (Figure 7). Especially manifestos and posts orient towards inclusive discourse. Comments, while having spikes of exclusionary sentiment, are also overarchingly inclusive.

Political discourse does seem to be communal discourse as displayed by the manifestos and Facebook activity. Social references rank well above references to the self; first person plural and the second person "you" come before first person singular (Figure 8). There is no cause to believe that the politicians or constituents are using the "Royal We," in which "we" is used to imply cohesion but indicates commands (Tausczik and Pennebaker, 2010).

Micro-level sentiment

While warning scholars to proceed with caution, Pennebaker et al. (2003) identified sentiment analysis as an area of future research in their 2003 article. As expected, emotion words are relatively low, accounting for 1.5–4% of the party's corpus (Figure 9). Gathering all posts and analysing for monthly changes, we cumulated all posts and comments, resulting in the graph depicted in Figure 9. A bump in positive sentiment

for both posts and comments is visible coinciding with the lead-up to the federal elections, along with a minor drop in negatively intoned posts. The rise in positive sentiment within the last month of 2012 is due to increased use of holiday wishes analogous to the finding of Dodds et al. (2011) and Kramer (2010).

As seen in Figure 10(a–d), the message that the parties would like to display is not necessarily being followed in day-to-day interactions of politicians and their constituencies. Overall, manifestos have nearly double the occurrence of positive emotion words as compared to posts and comments, and are more negatively intoned than posts in all cases. Positive sentiment within the posts and comments often concerns congratulations on birthdays, campaigning activities and self-promotion.

At this granularity level, there are almost no differences in the means of negative emotion usage, with posts tending to contain slightly less negative emotion words as compared to party manifestos and comments. The greater use of words bearing positive sentiment compared to words bearing negative sentiment is noticeable, especially in light of 60% more words within the LIWC dictionary being associated with negative sentiment (Pennebaker et al., 2007; Wolf et al., 2008). The highlights in negative sentiment typically detail posts about child abuse, angry discussion on night flight operations, as well as reflections on situations in the Middle East and Greece. While criticism
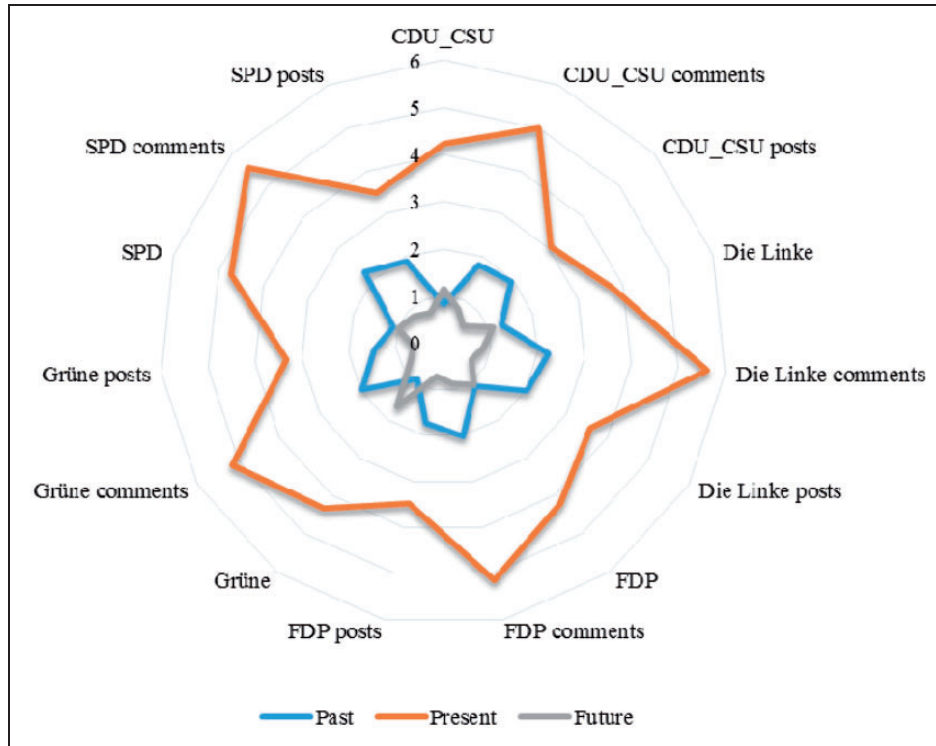
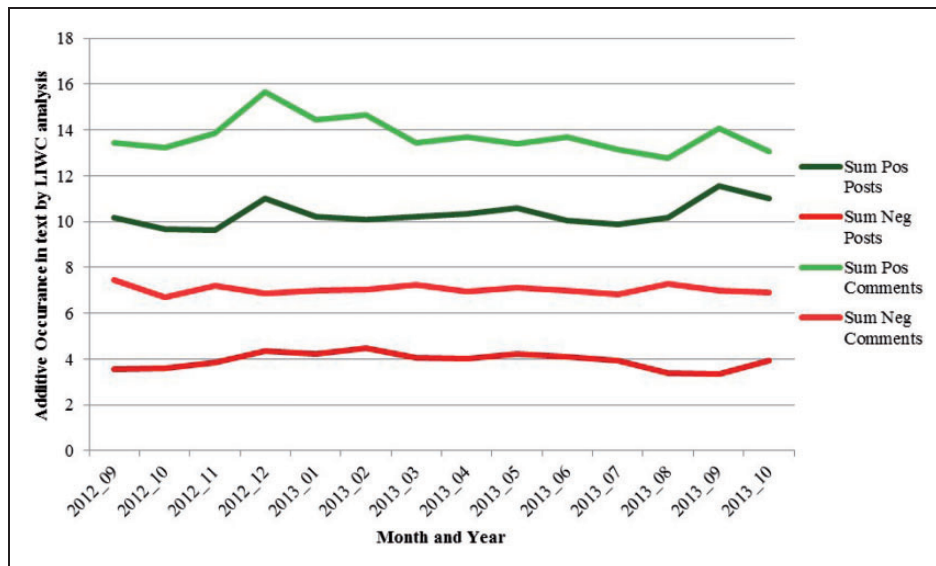**Figure 8.** Language tense patterns of party manifestos, posts and comments.

Figure 9. Average sentiment per month, posts and comments.

of opposing parties is present, the low negativity levels suggest that 'dirty' campaigning on Facebook is kept to a minimum, supporting our previous finding and diverging from Gunsch et al. (2000) for this user sample.

Positive emotion aligns with electability. Seven of the ten most positive commentaries are directed at CDU/CSU politicians, and the remaining three are directed at the SPD (Table 3). The incumbent CDU went on to nearly outright win the election, and the SPD is the second largest party, and went on to form the coalition of the 18th German Federal Parliament. Linke politicians, part of the opposition in the 17th and 18th Parliaments, have six of ten of the most negative politicians. Another notable feature is that while posts
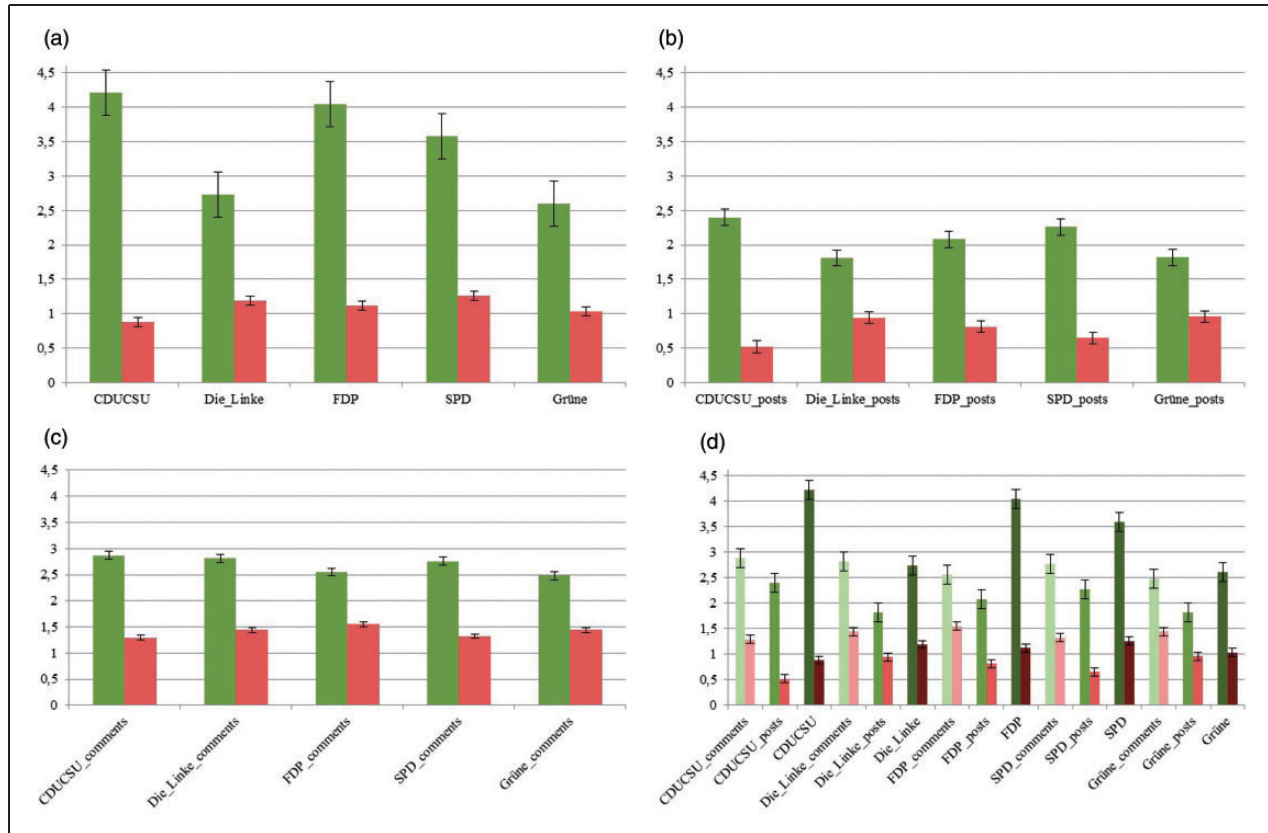
Figure 10. Sentiment by (a) manifesto, (b) politicians, (c) constituents and (d) overview of all (error bars at 95% confidence interval).

Table 3. Most positive and negative posts and commentator groups by relative per cent.

Name of politician	Party	Positive	Negative	party	Name of politician
Gero Storjohann comments	CDU/CSU	9.9	3.85	Linke	Andrej Hunko comments
Albert Rupprecht comments	CDU/CSU	8.78	2.75	Linke	Karin Binder comments
Peter Wichtel commenst	CDU/CSU	8.64	2.04	SPD	Sascha Raabe comments
Ewa Klamt comments	CDU/CSU	8.47	1.97	Linke	Dorothée Menzner comments
Sabine Weiss comments	CDU/CSU	8.31	1.88	Linke	Richard Pitterle comments
Günter Glose posts	SPD	6.17	1.66	Grüne	Marieluise Beck posts
Ingo Wellenreuther posts	CDU/CSU	3.62	1.65	CDU/CSU	Ernst-Reinhardt Beck posts
Hens Peter Friedrich posts	CDU/CSU	3.61	1.54	Linke	Ulla Jelpke posts
Peer Steinbrück posts	SPD	3.59	1.54	Grüne	Omid Nouripour posts
Franke Edgar posts	SPD	3.53	1.40	CDU/CSU	Guido Westerwelle posts

from Peer Steinbrück, the SPD contender for Chancellor in 2013, are amongst the most positive, Chancellor Angela Merkel appears neither in the most positive nor negative posts and comments. Without an existing benchmark in literature, this relationship between positive emotions, parties and politicians is left for future work.

A further look at social discourse between individual politicians to their constituents bears more interesting features. At the politician level, there are no significant differences in speech patterns based on gender, nor are there gender differences found in constituents' responses to politicians. There are no indications of the psycholinguistic indicators common to deception (more negative emotion, more motion words, fewer exclusion words, and less first-person singular) (Newman et al., 2003). Posts tend to be statements and comments tend to ask questions, which is indicative of an implicit hierarchy in

politician discourse according to the finding that higher status people ask less questions (Tausczik and Pennebaker, 2010). Anecdotally, Chancellor Merkel's posts did not contain a single question mark for the 13 months of this analysis.

Discussion

German political discourse is a rich, dense network. A major characteristic of German political discourse is that it occupies a close space, though distinct patterns appear at the correct resolution. Political discourse on Facebook is polite yet hierarchical. Especially the two largest parties (CDU/CSU and SPD) tend to use online speech in similar ways, while the three smaller parties have attributes onto themselves. Where the Grüne can be characterised as the least similar and most future-oriented party, the Linke has the highest concentration of negative commentators. Distinct in its nondescriptness, the FDP showed no discrete patterns. This lack of platform-based engagement is quoted as a major reason why the FDP did not meet the minimum criteria of to be re-elected into the 18th Parliament in favour of its larger, less conservative partner CDU/CSU.[12] Having established that a signal was available in the data that the FDP was losing constituent engagement, the next question is if and how this information could have been utilised by campaign managers and policy workers as a prediction tool.

Facebook offers an open, deliberative and participatory civil society forum for exchange. This was illustrated in the lack of gendered discourse and gender-directed responses in the face of a growing body of literature stating that Internet anonymity can increase sexist remarks.[13]

However, where politicians seek to be as inclusive as possible, constituents are careful to make distinctions in their viewpoints, thereby delimitating their own environments. Active campaigning is kept to a minimum, in favour of daily updates of how the politician is serving their community. One overarching fact of this study is that posts and comments are oftentimes intransitive, indicating that politicians and constituents are more often than not talking past one another.

Our analysis of political sentiment mining indicates that modern assessments of public opinion are largely improperly scaled. Individual sentiment scoring is an especially revealing method for community modelling. Positive and negative sentiment display interesting characteristics but show only limited potential as public opinion gauges, in agreement with Chung and Mustafaraj (2011), Jungherr et al. (2011) and Pennebaker et al. (2003). Much more revealing is the meso-analysis, as aggregating sentiment levels of users at the macro level leads to an averaging value without distinct significance, causing a blurred view. Accordingly, it is striking that

when observing at different levels, i.e. all, a party, or an individual, we uncover subtleties otherwise lost in the aggregation method.

Our Facebook-based Social Observatory facilitates interdisciplinary mixed method research on aspects of online communities. It adds to the toolbox of social media researchers that is today predominantly occupied by Twitter applications. Using a point and click style interface (social science) researchers can avoid the technical challenges of extracting social media data. We have also provided basic analytical capabilities that we will extend as required by future use cases. Although we present a case study in political science, our Social Observatory can be leveraged for any case study requiring Facebook data. Here, we envisage case studies in areas such as business as well as competitive intelligence, marketing and campaign management, and community detection and monitoring. Using our approach, researchers can mitigate the research biases common to social media research (see González-Bailón et al., 2014; Ruths and Pfeffer, 2014) as we can extract complete timelines, not samples thereof. We do, however, note that just because Facebook data is made available in this manner, not all research facilitated by a Social Observatory is ethical. Researchers need to be aware of the ethical boundaries of Facebook-based studies as the recent Facebook contagion study (Kramer et al., 2014) painfully demonstrated. Anecdotally, most users are completely unaware that Facebook pages are publically accessible and consequently do not provide informed consent to studies conducted by third parties. Simply by clicking like on a page, they conceivably become an entity in a dataset that a Social Observatory can curate for a researcher to analyse.

Conclusion

The continuing integration of the offline and digital self creates new requirements for social researchers and stakeholders. Missing has been a generalisable, open-source tool for accessing and analysing these phenomena that is specific to Facebook. We have presented the vision and architecture of a Social Observatory: a low latency method for the observation and measurement of social indicators within an online community. To explore the usefulness and possibilities of a Social Observatory for policy and decision makers, we implemented a Facebook adapter that allowed us to focus the Observatory on 187 German federal politicians and 257,305 lay constituents, as proxies to public opinion. We were able to observe how users interacted, with whom and at what volume. In addition, by leveraging the LIWC text analysis toolkit, we were able to identify different facets of communication processes and observe significant differences in sentiment between the politicians and their followers.

The implications of this work are threefold; firstly, we offer a framework to automatically extract public data troves (even from Facebook profiles) for use in studies related to online communities. Secondly, that by providing a few generalizable tools quite complex interdisciplinary research processes can be undertaken. Finally, using only a small number of points of reference, i.e. the 187 politicians, our approach can discover and analyse the actions of an entire (sub)community. By employing similar techniques and extending the analysis stages, we would be able to undertake the same study on any online social community, shedding light on specific social dynamics, and identifying key or influential actors unobtrusively. This ability is of key strategic use for public figures that wish to assess, for example, their public standing, or the reactions to specific actions.

While we believe the results of our case study are encouraging, the methods are not without fault. Within our quality control of selected users we found posts with incorrectly labelled sentiment scores. A misinterpretation by the word/word stem approach is likely, as these methods are notoriously hard to apply to cases of irony and sarcasm (Tsur and Rappoport, 2010). We will also revisit our post filtering approach; we included only status updates without photos, videos or links. Some politician profiles heavily use media content and are consequently largely omitted from our analysis. Politicians have PR teams that may post on their behalf; as such, we will extend our feature extraction and filtering methods to enable differentiated author studies. We will also automate the text analytics functionality currently provided by LIWC, making it an invokable tool in the future iterations of the Social Observatory workflow.

Notes

1. Latency here means visibility, not timespan, as commonly understood in Computer Science.
2. Thus excluding marginal, extremist and localized parties.
3. http://www.socialobservatories.org/vision
4. See for example: https://apps.facebook.com/namegen-web/ (accessed May 2014).
5. See for example: Wandora, http://wandora.org/ (accessed May 2014).
6. We note that Truthy (Mckelvey, 2013) states they aim to open their API for queries but have to date not done so.
7. The Java-based tool Wandora (Kivelä and Lyytinen, 2004) has a Facebook plug-in, but a limited ability to parse entire graph, leaving it with limited research functionality.
8. This is a summary of information available on the federal government's website: http://www.bundestag.de/htdocs_e/
9. We note that 25 officials' Facebook accounts were deleted following the elections and thus not considered.
10. Audience relates to the number of unique Facebook IDs that interacted with one or more politicians. Note: the total audience is not the sum for each party, indicating that Facebook users interact with more than one party.
11. For more information on this politician's public mishaps and scandals, please see: http://www.spiegel.de/international/topic/christian_wulff/archiv-2011235.html
12. http://wahl.tagesschau.de/wahlen/2013-09-22-BT-DE/analyse-wanderung.shtml#11_Wanderung_UNION (accessed November 2013, infographic in German).
13. For examples, please review: http://www.nationinstitute.org/blog/nationbooks/3014/challenging_online_sexism/; http://www.theguardian.com/commentisfree/2011/nov/07/abusive-sexist-comments-online (accessed May 2014) or review the conversation surrounding #Aufschrei.

References

Allen SM, Chorley M, Colombo G, et al. (2014) Exploiting user interest similarity and social links for micro-blog forwarding in mobile opportunistic networks. *Pervasive and Mobile Computing* 11: 106–131.

Atefeh F and Khreich W (2015) A survey of techniques for event detection in twitter. *Computational Intelligence* 31(1): 132–164.

Baccianella S, Esuli A and Sebastiani F (2010) SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In: *Proceedings of the seventh international conference on language resources and evaluation* (LREC'10). Paris, France: European Language Resources Association, pp.2200–2204.

Balahur A and Hermida JM (2012) Extending the EmotiNet knowledge base to improve the automatic detection of implicitly expressed emotions from text. In: *Proceedings of the 8th international conference on language resources and evaluation*. Paris, France: European Language Resources Association, pp.1207–1214.

Beck S (1953) The science of personality: Nomothetic or idiographic? *Psychological Review* 60(6): 353–359.

Böcking B, Hall M and Schneider J (2015) Event prediction with learning algorithms—A study of events surrounding the egyptian revolution of 2011 on the basis of micro blog data. *Policy & Internet* 7(2): 159–184.

Braga-Neto UM and Dougherty ER (2004) Is cross-validation valid for small-sample microarray classification? *Bioinformatics* 20(3): 374–380.

Brown P and Levinson SC (1987) *Politeness: Some Universals in Language Usage*. Cambridge: Cambridge University Press.

Burnap P, et al. (2014) Towards an integrated. COSMOS and scalable service for analysing social media on demand. *International Journal of Parallel, Emergent and Distributed Computing* 37–41.

Burnap P, et al. (2014) Towards real-time probabilistic risk assessment by sensing disruptive events from streamed news feeds. In: *2014 Eighth international conference on complex, intelligent and software intensive systems.* Birmingham, UK: IEEE, pp. 608–613. Available at: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6915582 (accessed 20 November 2015).

Burnap P and Williams ML (2014) Hate speech, machine classification and statistical modelling of information and flows on Twitter: Interpretation and communication for policy decision making, pp. 1–18.

Burnap P and Williams M (2015) Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making. *Policy & Internet* 7(2): 223–242.

Burrows R and Savage M (2014) After the crisis? Big Data and the methodological challenges of empirical sociology. *Big Data & Society* 1(1): p.2053951714540280. Available at: http://bds.sagepub.com/lookup/doi/10.1177/2053951714540280 (accessed 14 October 2014).

Calvo RA and D'Mello SK (2010) Affect detection: An interdisciplinary review of models, methods, and their applications. *IEEE Transactions on Affective Computing* 1(1): 18–37.

Chung CK and Pennebaker JW (2014) Counting little words in Big Data: The psychology of communities, culture, and history. In: Forgas J, Vincze O and Laszlo J (eds) *Social Cognition and Communication*. New York, NY: Psychology Press, pp. 25–42.

Chung J and Mustafaraj E (2011) Can collective sentiment expressed on twitter predict political elections? In: *Proceedings of the twenty-fifth AAAI conference on artificial intelligence*. Palo Alto, CA: AAAI Press, pp.1770–1771. Available at: http://www.aaai.org/ocs/index.php/AAAI/AAAI11/paper/viewPDFInterstitial/3549%26lt%3B/4126 (accessed 20 November 2015).

Chung J and Mustafaraj E (2010) Can collective sentiment expressed on Twitter predict political elections? In: *Proceedings of the twenty-fifth AAAI conference on artificial intelligence*, pp. 1770–1771.

Cioffi-Revilla C (2010) Computational social science. *Computational Statistics* 2(3): 259–271.

Cioffi-Revilla C (2014) *Introduction to Computational Social Science*. Berlin: Springer Texts in Computer Science.

Creswell JW (2003) *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches*. Thousand Oaks: SAGE Publications.

Das S and Kramer A (2013) Self-censorship on Facebook. In: *Seventh international AAAI conference on weblogs and social media*. Palo Alto, CA: AAAI Press, pp. 120–127.

Davenport SW, et al. (2014) Twitter versus Facebook: Exploring the role of narcissism in the motives and usage of different social media platforms. *Computers in Human Behavior* 32: 212–220.

Dodds PS, et al. (2011) Temporal patterns of happiness and information in a global social network: Hedonometrics and twitter. *PLoS ONE* 6(12).

Fowler RL (1987) Power and robustness in product-moment correlation. *Applied Psychological Measurement* 11(4): 419–428.

Go A, Bhayani R and Huang L (2009) *Twitter Sentiment Classification using Distant Supervision*. Stanford, CA.

Golder SA and Macy MW (2012) Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures. *Science* 333: 1878–1881.

González-Bailón S, et al. (2014) Assessing the bias in samples of large online networks. *Social Networks* 38: 16–27.

Gunsch MA, et al. (2000) Differential forms linguistic content of various of political advertising. *Journal of Broadcasting and Electronic Media* 44(1): 27–42.

Hackenberg R (1970) The social observatory series data for health and behavioral research. *Social Science and Medicine* 4: 343–357.

Hall M and Caton S (2014) A crowdsourcing approach to identify common method bias and self-representation. *IPP2014: Crowdsourcing for Politics and Policy*. Oxford, England.

Hall M, Caton S and Weinhardt C (2013a) Well-being's predictive value. In: *Proceedings of the 15th international conference on human–computer interaction (HCII)* (eds AA Ozok and P Zaphiris), pp. 13–22. Berlin: LNCS, Springer Verlag.

Hall M, Glanz S, et al. (2013b) Measuring your best you: A gamification framework for well-being measurement. In: *Third international conference on social computing and its applications*, pp. 277–282. Karlsruhe, Germany: IEEE.

Hampton KN, et al. (2011) *Social networking sites and our lives*. Washington, DC: Pew Research Center Internet and American Life Project.

Hechenbichler K and Schliep K (2004) Weighted k-nearest-neighbor techniques and ordinal classification. *Molecular Ecology* 399: 17.

Housley W, et al. (2014) Big and broad social data and the sociological imagination: A collaborative response. *Big Data & Society* 1(2): p.2053951714545135. Available at: http://bds.sagepub.com/lookup/doi/10.1177/2053951714545135 (accessed 14 September 2014).

Hsieh H-F and Shannon SE (2005) Three approaches to qualitative content analysis. *Qualitative Health Research* 15(9): 1277–1288.

Jaho E, Karaliopoulos M and Stavrakakis I (2011) ISCoDe: A framework for interest similarity-based community detection in social networks. In: *2011 IEEE conference on computer communications workshops*. New York, NY: IEEE, pp.912–917. Available at: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber = 5928942 (accessed 20 November 2015).

Jungherr A, Jurgens P and Schoen H (2011) Why the Pirate Party won the German Election of 2009 or the trouble with predictions: A response to Tumasjan, A., Sprenger, T. O., Sander, P. G., & Welpe, I. M. "Predicting elections with Twitter: What 140 characters reveal about political sentiment.". *Social Science Computer Review* 30(2): 229–234.

Kassarjian HH (1977) Content analysis in consumer research. *Journal of Consumer Research* 4(1): 8–18.

Kivelä A and Lyytinen O (2004) Topic map aided publishing – A case study of assembly media archive. In: *STeP 2004 – The 11th Finnish artificial intelligence conference proceedings*, 1–3 September, Vantaa, Finland.

Komisin M and Guinn C (2012) Identifying personality types using document classification methods. In: *Proceedings of the twenty-fifth International Florida artificial intelligence research society conference*. Palo Alto, CA: AAAI Press, pp.232–237.

Kouloumpis E, Wilson T and Moore J (2011) Twitter sentiment analysis: The good the bad and the omg! In: *Proceedings of the fifth international AAAI conference on weblogs and social media (ICWSM 11)*. Palo Alto, CA: AAAI Press, pp. 538–541. Available at: http://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/download/2857/3251?iframe=true&width=90%&height=90% (accessed 20 November 2015).

Kramer A (2010) An unobtrusive behavioral model of "gross national happiness." In: *Proceedings of the SIGCHI conference on human factors in computing systems*. New York, NY: ACM, pp.287–290.

Kramer A (2012) The spread of emotion via Facebook. In: *Proceedings of the 2012 ACM annual conference on human factors in computing systems – CHI '12*. New York, NY: ACM Press, pp. 767–770. Available at http://dl.acm.org/citation.cfm?doid=2207676.2207787 (accessed 20 November 15).

Kramer A, Fussell SR and Setlock LD (2004) Text analysis as a tool for analyzing conversation in online support groups. In: *CHI2004*, Vienna, Austria, pp. 1485–1488.

Kramer A, Guillory JE and Hancock J (2014) Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences* 111(24): 8788–8790. Available at: http://www.pnas.org/cgi/doi/10.1073/pnas.1320040111 (accessed 3 June 2014).

Lasswell HD (1967) Do we need social observatories? *The Saturday Review* 49–52.

Lin H and Qiu L (2013) Two sites, two voices: Linguistic differences between Facebook status updates and tweets. In: Rau PLP (ed.) *Lecture notes in computer science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 8024 LNCS. Berlin: Springer, pp. 432–440.

Liu B (2010) Sentiment analysis and subjectivity. In: Indurkhya N and Damerau FJ (eds) *Handbook of Natural Language Processing*. Boca Raton, FL: CRC Press, pp. 1–38. Available at: http://www.cs.uic.edu/~liub/FBS/NLP-handbook-sentiment-analysis.pdf\nhttp://people.sabanciuniv.edu/berrin/proj102/1-BLiu-Sentiment (accessed 20 November 2015).

Mahmud J (2014) Why do you write this? Prediction of influencers from word use psycholinguistic analysis from text. In: *ICWSM*. Palo Alto, CA: AAAI Press, pp. 603–606.

Markham A and Buchanan E (2012) Ethical decision-making and internet research recommendations from the AoIR Ethics Working Committee, Association of Internet Researchers. Available at: http://www.aoir.org/documents/ethics-guide (accessed 20 November 2015).

McCallum AK (2002) MALLET: A machine learning for language toolkit. Available at: http://mallet.cs.umass.edu (accessed 14 October 2014).

Mckelvey K (2013) Truthy: Enabling the study of online social networks. In: *CSCW'13*. San Antonio, TX: ACM Press, pp. 23–25.

Mckelvey K (2013) Enabling the study of online social networks. In: *CSCW'13*, pp. 23–25. San Antonio, TX: ACM Press.

Niederhoffer K and Pennebaker J (2002) Linguistic Style Matching in Social Interaction. *Journal of Language and Social Psychology* 21(4): 337–360.

Newman ML, et al. (2003) Lying words: Predicting deception from linguistic styles. *Personality and Social Psychology Bulletin* 29: 665–675.

O'Connor B et al. (2010) From tweets to polls: Linking text sentiment to public opinion time series. In: *Proceedings of the fourth international AAAI conference on weblogs and social media*. Palo Alto, CA: AAAI Press, pp. 122–129.

Pak A and Paroubek P (2010) Twitter as a corpus for sentiment analysis and opinion mining. In: *Proceedings of the seventh conference on international language resources and evaluation*. Palo Alto, CA: AAAI Press, pp. 1320–1326. Available at: http://incc-tps.googlecode.com/svn/trunk/TPFinal/bibliografia/ (accessed 20 November 2015).

Paolacci G, Chandler J and Ipeirotis P (2010) Running experiments on Amazon mechanical Turk. *Judgment and Decision Making* 5(5): 411–419.

Pennebaker JW, et al. (2007) *The Development and Psychometric Properties of LIWC2007*. Austin, TX: University of Texas.

Pennebaker JW, Mehl MR and Niederhoffer KG (2003) Psychological aspects of natural language use: our words, our selves. *Annual Review of Psychology* 54: 547–577.

Radovanovic M, Nanopoulos A and Ivanovic M (2010) Hubs in space: Popular nearest neighbors in high-dimensional data. *Journal of Machine Learning Research* 11: 2487–2531.

Ruppert E (2013) Rethinking empirical social sciences. *Dialogues in Human Geography* 3(3): 268–273.

Russell M (2013) *Mining the Social Web*, 2nd ed. Sebastopol, CA: O'Reily Media.

Ruths D and Pfeffer J (2014) Social media for large studies of behavior. *Science* 346(6213): 1063–1064.

Salas-Zárate MP, et al. (2014) A study on LIWC categories for opinion mining in Spanish reviews. *Journal of Information Science* 1(13): 1–13.

Savage M and Burrows R (2007) The coming crisis of empirical sociology. *Sociology* 41(5): 885–899.

Schwartz HA, et al. (2013) Personality, gender, and age in the language of social media: the open-vocabulary approach. *PloS One* 8(9): e73791.

Stieglitz S and Dang-Xuan L (2012) Social media and political communication: a social media analytics framework. *Social Network Analysis and Mining* 3(4): 1277–1291.

Tanasescu V, et al. (2013) The personality of venues: Places and the five-factors ('big five') model of personality. In: *2013 fourth international conference on computing for geospatial research and application*, pp. 76–81.

Tausczik YR and Pennebaker JW (2010) The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology* 29(1): 24–54.

Taylor L, Schroeder R and Meyer E (2014) Emerging practices and perspectives on Big Data analysis in economics: Bigger and better or more of the same? *Big Data & Society* 1(2). Available at: http://bds.sagepub.com/lookup/doi/10.1177/2053951714536877 (accessed 7 October 2014).

Thelwall M, et al. (2010) Sentiment strength detection in short informal text. *Journal of the American Society for Information Science and Technology* 61(12): 2544–2558.

Tinati R, et al. (2014) Big Data: Methodological challenges and approaches for sociological analysis. *Sociology* 48(4): 663–681.

Tsur O, Davidov D and Rappoport A (2010) ICWSM – A great catchy name: Semi-supervised recognition of sarcastic sentences in online product reviews. In: *ICWSM*. Palo Alto, CA: AAAI Press, pp. 1–9. Available at: http://www.aaai.org/ocs/index.php/ICWSM/ICWSM10/paper/viewFile/1495/1851papers3://publication/uuid/9470D66E-1C15-4E9F-8B02-18A0AF76D8E3 (accessed 20 November 2015).

Tumasjan A, et al. (2010) Predicting elections with Twitter: What 140 characters reveal about political sentiment. In: *Proceedings of the fourth international AAAI conference on weblogs and social media*. Palo Alto, CA: AAAI Press, pp.178–185. Available at: http://www.aaai.org/ocs/index.php/ICWSM/ICWSM10/paper/viewFile/1441/1852 (accessed 20 November 2015).

Turney PD and Pantel P (2010) From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research* 37: 141–188.

Wang N, et al. (2014) Can well-being be measured? Using Facebook status updates validation of Facebook's gross national. *Social Indicators Research* 115(1): 483–491.

Wilson RE, Gosling SD and Graham LT (2012) A review of Facebook research in the social sciences. *Perspectives on Psychological Science* 7(3): 203–220.

Wilson T, Hoffmann P, Somasundaran S, et al. (2005) OpinionFinder: A system for subjectivity analysis. In: *Proceedings of HLT/EMNLP on interactive demonstrations*. Association for Computational Linguistics, pp. 34–35.

Wolf M, et al. (2008) Computergestützte quantitative textanalyse. *Diagnostica* 54(2): 85–98.

Xiang R, Neville J and Rogati M (2010) Modeling relationship strength in online social networks. In: *WWW2012*, pp. 981–990. Raleigh, NC: ACM Press.