Policy and Internet

Mapping the Public Agenda with Topic Modeling: The Case of the Russian LiveJournal

Olessia Koltsova and Sergei Koltcov

This article describes agendas as "packages" of topics of varying salience, set by the Russian Internet users on Russia's leading blog platform LiveJournal. The research involved modeling LiveJournal's topic structure, viewed as an important component of what is termed here self-generated public opinion. Topic modeling was performed automatically with the LDA algorithm, and complemented with hand labeling of topics. Data were collected by software created by the authors to generate a relational database storing all posts by the top 2,000 LiveJournal users from three one-month periods: two during the Russian parliamentary and presidential elections 2011–2012, and one control period. We find that LiveJournal top users share their attention evenly between "social/political" and "private/recreational" issues, the proportion being very stable. However, the substitution of diverse public affairs issues by the topics related to national street protests in the politicized periods compared to the control period was found both automatically and manually. The group of topics centered around social issues demonstrates the biggest volatility in terms of its composition and may serve as the foundation for monitoring self-generated public opinion by further application of sentiment/opinion mining methods to these topics.

KEY WORDS: Russia, blogs, topic modeling, LDA, political protests, online public opinion

Introduction

In recent years, blogs and social networks have gained importance in many ways, but particularly in setting political agendas and forming collective opinions on a wide range of issues in many countries. The Russian parliamentary and presidential elections and protests of 2011–12 are a notable example of this. As diaries, usually maintained by ordinary individuals outside their professional activities and, therefore, a type of user-generated content, blogs may be regarded as "vox populi," or what we term here "self-generated public opinion." The long-term goal of the research presented in this article is to learn how to monitor this public opinion—consisting of both agendas and attitudes to them—in order to identify trends and derive sociological conclu-

1944-2866
Published by Wiley Periodicals, Inc., 350 Main Street, Malden, MA 02148, USA, and 9600 Garsington Road, Oxford, OX4 2DQ.

sions that, in turn, may be applied to the policymaking process. This article describes the topical structure of the Russian-language blogosphere as one of the two components of self-generated public opinion, in order to see what topics are salient, what social cleavages they indicate, and how this structure changes over time. No less important is to show how automatic text analysis, namely topic modeling, can be used to support this goal, and its advantages and limitations. The Russian political events of December 2011–March 2012 are the key test issues explored in this work.

These events started with the national Parliamentary elections held on December 4, 2011 and widely reported as fraudulent on the Internet, with an abundance of first-hand evidence. Researchers have assessed the scale of the fraud in these elections to be not less than 10 percent (Enikolopov, Korovkin, Petrova, Sonin, & Zakharov, 2013); most agree on the gradual growth of fraud in the last decade in Russia (for an overview, see Mebane & Kalinin, 2009). However, these were the first elections to provoke immediate and mass street protest actions, attracting tens of thousands of people in Moscow and about a hundred other cities and towns across the country. The protesters effectively organized a movement through extensive use of the Internet; they demanded cancellation of the parliamentary election results, the holding of new and fair parliamentary elections, and prevention of fraud and unfairness in the presidential elections held on March 4, 2012. After March 4, mass protests continued against the fraud in the presidential elections and against the inauguration of Vladimir Putin, until June 2012.

These events show that blogs may be especially important in countries like Russia—a society in which the Internet is not technically filtered, but where the leading traditional media are tightly controlled by the political elite. The Russian-language blogosphere contains about 65 million blogs, of which about nine million are stand-alone blogs, while others are hosted by approximately one hundred blog platforms. These data are available from Yandex,[1] the leading Russian search engine, which claims to index nearly the entire Russian blogosphere and maintains its popularity rating—an important reference point for Russia's educated public in its search of authoritative sources of information, independent from the mainstream media.

Since no single blog platform dominates in Russia, and the platforms are technically disconnected, the Russian blogosphere has become very clustered. For instance, most political and social discussion is hosted by LiveJournal (Alexanyan & Koltsova, 2009; Etling et al., 2010; Gorny, 2004), a major reason for our focus on this platform. LiveJournal blogs collect thousands of friends, thereby competing with the regular media: the top 30 LiveJournal users have 20,000+ friends each, a number that would be considered good circulation for an average Russian newspaper. LiveJournal is one of the four leaders in terms of the number of accounts (approximately 2.8 million) and is the leader in the number of daily posts (approximately 90,000). More importantly, as the Yandex rating suggests, most top bloggers come from LiveJournal.

Methodological Approach and Related Work

Blogs as Sources of Agendas and Opinions

In the social science literature, blogs are often studied as a new kind of media that is able to set its own agenda or at least, able to influence the agenda of the regular media (Sayre, Bode, Shan, Wilcox, & Shah, 2010; Wallsten, 2007). Less frequently, blogs are studied as a new form of public opinion that produces both agendas (sets of topics of varying salience) and attitudes to them. Although this research deals first of all with agendas, it is important to separate our approach to them from agenda setting theory in media studies (McCombs & Shaw, 1972), and to instead place the notion of agenda into the context of a newly understood concept of public opinion. Traditionally, the latter is defined as "the distribution of individual preferences within a population" (Monroe, 1975, p. 6), while other authors would restrict it to the adult population (Erikson & Tedin, 2011, p. 8), or to the range of issues that are important to pollsters, to the community (Simon, 1974, p. 7), or to the government and political process.

When describing techniques that may be used to monitor public opinion, the relevant literature is, by default, centered on polling and rating instruments. This may seem natural for classic sources (e.g., Blumer, 1948; Droba, 1931), but the trend persists in the contemporary literature as well (e.g., Bardes & Oldendick, 2012; Erikson & Tedin, 2011), including those who criticize the very notion of public opinion (Bishop, 2005). In this view, public opinion is seen to exist in humans' minds, from where it has to be extracted by pollsters with special techniques that predetermine the structure of the output. Thus, the agenda for opinions is set by the researchers. This stream of literature contains no indication that public opinion may exist in the form of natural language texts that are not prestructured by external observers. However, unstructured texts are easily perceived to be a natural "filling" for media. The agenda setting theory (McCombs & Shaw, 1972) that stands behind much of the social research on blog texts views media as clearly separate from public opinion, which is understandable, given the time when it was developed. More specifically, it emphasizes the media's ability to tell the audience "what to think about," rather than "what to think" (Cohen, 1963, p. 120).

In their research, González-Bailón, Kaltenbrunner, and Banches (2012) point out that a fuller vision of public opinion may encompass both the selection of issues that are "thought about" *and* the state of public emotion about these issues. The second component can easily be broadened to include attitudes and preferences as more traditional "components" of public opinion. This new vision of public opinion can be termed "self-generated public opinion." It is free from the flaw for which polls have been consistently criticized, namely the imposition of an agenda on respondents, which makes them invent nonexisting opinions thus producing artifacts (for an overview of such a critique, see Bishop, 2005). On the other hand, this novel vision of public opinion is not free from some limitations. First, it is representative only of the self-selected population of those who have authored the texts under study, not of the whole population. Second,

like regular "polled" public opinion, the "online" public opinion still covers only those attitudes that bloggers are willing to share in public.

Pushing forward the argument that public agendas and public attitudes are components of the same phenomenon (i.e., public opinion), it can also be claimed that blogs are simultaneously a new form of public opinion *and* a form of media. As public opinion, the blogosphere is literally public, that is, it is publicly available and therefore able to function as a mass medium and to exercise influence, both within and outside itself. The closer to the top of the popularity list of bloggers, the more the blogosphere resembles the traditional media in all aspects of its functioning; the deeper into the "long tail" of the average bloggers, the more it functions as public opinion—although in principle it is available, it is largely anonymous and not easily accessed in an aggregated form. The "top list" effect may be especially important in societies, such as Russia, where popularity lists exert a visible influence on blogger's competitive behavior and on their audience's perception of a blogger's significance.

This also entails some reconsideration of the concept of an opinion leader. As first introduced by Lazarsfeld, Berelson, and Gaudet (1950), opinion leaders are included in "the public" and only influence their immediate environment, remaining invisible to the media and to the audiences of those media. Such local influentials can also be found in the "long tail" of the blogosphere. However, when their textual activity is aggregated, even they may have a visible public influence, for example on agenda setting through different kinds of "the most discussed news" ratings. As for the top bloggers, their reach their reach is disproportionately larger than that of offline opinion leaders, and is often close to that of media celebrities. No explicit qualitative boundary seems to be found in the blogosphere, either between its media and public opinion facets, or between local opinion leaders and global influentials. However, these two categories of bloggers—local and global—may have different agendas and attitudes, and to obtain an accurate picture, it is important to study both groups or to acknowledge the limitations of studying them separately.

Existing Methodological Approaches

These new research areas demand special methods of data extraction and analysis. So far, the traditional literature on public opinion does not consider (blog) texts as sources of public opinion, and the media studies literature is based mainly on traditional content analysis (for an agenda setting approach, see Sayre et al., 2010; Wallsten, 2007, for an online public opinion approach, see Zhou & Moy, 2007). The volume of data that can be processed by a human inevitably limits such studies, given that agendas may include hundreds of issues to track, as well as opinions on those issues. Papacharissi (2007) analyzed 260 randomly selected blogs and found them to be personalized, self-referential, and self-serving—without any public agenda—because A-list bloggers happened not to appear in the sample.

Such studies suffer from a lack of sampling procedures, which makes their conclusions difficult to generalize. Sayre et al. (2010) and Zhou and Moy (2007)

restricted their studies to the extraction of relevant information through keyword searches, which tells little about the sample's completeness and representativeness, because of the nontransparent algorithms used by their search engines. While Papacharissi (2007) benefited from the random sampling service of blogger.com, Wallsten (2007) acknowledges multiple problems in making a random sample suitable for his goals.

Ironically, the computer science literature is flooded with methodological papers offering an abundance of algorithms that allow automatic analysis of big textual data. This often solves problems of insufficient and biased sampling, because, in some cases, entire populations become available for analysis. Methods of automatic analysis cannot entirely replace manual work with texts, but they can help reduce it to the most necessary stages of the research, or to the most meaningful and representative areas of the textual space that they help to map. In fact, they are the only means to monitor agendas and attitudes across multiple sources, and over long periods of time and at a large scale.

This literature can be divided into two streams. First, the sentiment analysis and opinion mining literature usually focuses on extracting opinions on the preset issues/agendas that are thought to already be present in the texts (for reviews, see Pang & Lee, 2008; Thelwall, Buckley, & Paltoglou, 2012). Most often, this is limited to dividing sentences or texts into negative and positive, although more nuanced classifications are developing rapidly. These techniques are based on thesauri of human-coded emotional words (Thelwall et al., 2012), on an algorithmic construction of sentiment dictionaries (Godbole et al., 2007), or on machine learning alone, when the "machine" classifies the texts into negative or positive, based on comparisons of texts that were labeled manually beforehand (Pang, Lee, & Vaithyanathan, 2002). A number of applications of these techniques exist for studying public opinion in blogs (Dey & Haque, 2009; Godbole, Srinivasaiah, & Skiena, 2007; Ku et al., 2006), although initially and most successfully, this approach was applied to product review analysis (Pang et al., 2002). In these papers, the possibility of extracting public opinion from natural texts—rather than from polls—is usually an implicit assumption, but is not discussed as a novelty.

The second stream of literature deals with revealing the agenda, that is a set of issues addressed in a sample of texts, which is done by topic modeling (for reviews, see Daud, Li, Zhou, & Muhammad, 2010; Steyvers & Griffiths, 2007) or using text clustering approaches (for reviews, see Andrews & Fox, 2007; Carpineto, Osiński, Romano, & Weiss, 2009). Both approaches "discover" issues (topics, agendas) in collections of texts without requiring a researcher to know beforehand what issues might exist in them. No keywords or "keytexts" are needed, so the algorithms do not depend on the researchers' preconceptions. Instead, they enable the division of texts into topically (or more precisely, lexically) similar groups that can later be interpreted and named (i.e., labeled) by researchers. While clustering techniques compare texts directly by means of a distance measure, based on the lexical similarity of the texts, topic modeling compares each text to a set of latent variables (called topics), whose distribution

over words and texts is simultaneously modeled. Topic modeling is a rapidly developing technique, and has been applied to blogs (Li, Xu, & Zhang, 2007; Liu, Niculescu-Mizil, & Gryc, 2009; Nallapati and Cohen, 2008). However, until recently, neither of these methods has measured attitudes to the agendas they find. Only the most recent papers have begun to develop algorithms that combine topic modeling and sentiment analysis of those topics (Li et al., 2010; Lin & He, 2009; Mei, Cai, Zhang, & Zhai, 2008).

Due to their recentness, and—often—immaturity, as well as their mathematical and computational complexity, the algorithms and approaches reviewed above are rarely applied by social scientists. A notable exception is the study by González-Bailón et al. (2012). However, this is not based on full-text analysis, but on headings of "discussions" (i.e., threads of posts). Etling et al. (2010) clustered the Russian-language blogosphere (and previously, the Persian and the Arabic blogospheres) based on hyperlinks, and then labeled clusters on the basis of hand coding randomly selected blogs. An important contribution to the development of automated text analysis in social science was made by Hopkins and King (2010), including an analysis of blogs by King, Pan, & Roberts (2013). However, this approach, known as supervised text classification, demands prior knowledge of the topics to be searched for, and what is most limiting, their operationalization through keywords. This approach usually works well for simple issues, such as named entities (personalities, trademarks) or events; it is much less efficient in extracting such topics as social problems or leisure domains, especially when they are not known beforehand. Diesner and Carley (2010) applied unsupervised topic modeling in the social sciences to classify research proposals. To our knowledge, topic modeling has not been applied for the extraction of agendas from blogs in any social science research.

Both topic modeling/clustering *and* sentiment analysis/opinion mining are needed to effectively monitor self-generated public opinion. However, the development of the latter in Russia is hindered by the absence of elaborated, publicly available thesauri, which are much better developed in the English-language domain. Although the long-term goal of this research effort is to combine these two approaches, this paper presents the results of the first stage, in which we map the topic structure of the Russian blogosphere and tackle its most salient topics.

The Concept of Topic

For this purpose, the concepts or notions of "agenda" and "topic" should be clarified. Agenda can be defined as a set of issues or topics addressed in a text or in a sample of texts. Intuitively clear, the concept of topic can be generally defined as the main subject matter of a text. However, on a more detailed level, topic is much more difficult to capture. In people's minds, topics are centered on events, social problems, unproblematic issues, constant or transient aspects of life, discourse types, or value systems. Topics of different scales can be found, with the varying "power of a microscope" through which a text (document) collection

(corpus, sample) is viewed. Some of these topics are organized in hierarchies, and others stand alone. Often topics meet or overlap within the same text, while other texts are mono-topical. In some discursive spaces, such as the blogosphere, topics change fast, not only emerging and disappearing, but also breaking, uniting, and evolving. Finally, some portions of texts have no topic or legible meaning and cannot be understood or classified at all. Such texts can be considered as noise. Thus, these everyday topical classifications have neither clear grounds, nor shared ad hoc reasons, and they are very difficult to unambiguously and comprehensively define for research purposes.

While analytical definition is difficult, there are two ways to detect topics empirically: the first is to ask people (coders) to determine them, while trying to reach a desirable level of agreement between them on the same texts, and the other is automatic. Since, for large text collections, only the second approach is feasible, the only available material for topic detection is words or word sequences in texts. Topic then appears to be something like a list of words. Although this approach has obvious limitations in tackling subtle meanings and links in texts, experiments with checking its results against human coding (i.e., with topics assigned to texts by human coders) show over 90 percent accuracy under some conditions (see, e.g., Blei, Ng, Jordan, & Lafferty, 2003, p. 1013). With this approach, a topic is a result of the work of an algorithm; this means that it is defined algorithmically and no "textual" definition is possible. A brief description of the algorithm used in this present research is provided below.

Data, Algorithms, and Software

Data

To create a representative sample of the entire Russian blogosphere, one would have to develop a software application similar to that of the leading search engines. Each blog platform has its unique data format that demands special processing if one wishes to go beyond the first available page and to store the entire relational structure of a blog. This was one of the reasons that we limited our current research to the leading political blog platform, LiveJournal. We have developed software that downloads blog accounts from the LiveJournal list of bloggers and stores their full text content for a given period of time in a relational MS SQL database. This database links bloggers' nicknames, texts, and the URLs of their posts with dates, the texts of related comments with dates, and commentors' nicknames; it also supports different kinds of sampling and exports data to a number of analytical tools. For the purposes of agenda detection, only the texts of the posts were used. The top 2,000 bloggers, rated by their numbers of followers, are taken into this analysis to detect the agenda that potentially may tell the maximal number of readers what topics are salient.

Three datasets were created for this study, with 10E4 posts in each set (Table 1). The length of the time period was determined based on previous studies of news lifecycles (Koltsova, 2011; Wu, Hofman, Mason, & Watts, 2011).

Table 1. Time Period of the Three Data Samples

Period	Political Context	Sample
August 15–September 15, 2011	Politically "calm" period	Top 2,000 bloggers, no more than 50 posts per blogger
November 27–December 27, 2011	Around Russian parliamentary election of December 4, which covers both preelection discussion and all protest actions before the Russian Christmas break in early January	Top 2,000 bloggers, no more than 50 posts per blogger
March 4–April 4, 2012	After the presidential elections with some protest actions	Top 2,000 bloggers

Algorithm Properties Important for Social Science

The demand for social analysis of large text data is currently greater than the development of both data collection instruments, and of analytical instruments that social scientists can use—that is, of thoroughly tested, methodologically transparent, and user-friendly software. Both clustering and topic modeling algorithms and software are multiplying rapidly, while their comparative and independent testing is lagging. Seven software-related areas of concern that relate to our goals are: the computational complexity of the algorithms, the quality of the algorithms, the correctness of the software, the ability to detect the optimal number of clusters (topics), the functionality for cluster/topic labeling, the functionality for text preprocessing, and applicability to certain types of texts, for example, Russian language blogs.

The computational complexity of an algorithm is—roughly—how fast it can work, how much data it can digest, and how powerful a computer it requires. We have not found any comprehensive, comparative assessment studies for such algorithms in research literature. Computational complexity is usually derived, if at all, from the mathematical properties of algorithms, without reference to real experiments (e.g., Zhong, 2005). The quality of an algorithm refers to how well it can perform a given task; however, the notion of "wellness" has yet to be defined, and the measure that can capture it has yet to be developed. For instance, the ability to detect coherent groups might demand a different measure than the ability to detect well-separated groups. We have found no comprehensive comparisons of the quality of algorithms; for separate algorithms, quality is often assessed from the experiments, but different measurements and different data sets are applied in different research papers, and these are only applied for a narrow set of algorithms. Furthermore, there is no consensus on what "quality of topic modeling" actually is, although a function termed "perplexity" is most often used. In general, the evaluation of topic models is a new and still underdeveloped area of inquiry (Wallach, Murray, Salakhutdinov, & Mimno, 2009). Software correctness refers to whether a software application implements a given mathematical algorithm in the correct way. Such tests are *never* mentioned in the articles we have searched.

Determining the number of clusters/topics is one of the most severe and unresolved problems of cluster analysis and of topic modeling. What the "right" number of groups should be is a highly debated question. A typical way to deal with this problem is to obtain multiple cluster/topic solutions (i.e., to run an algorithm several times with different numbers of topics/clusters), to measure the quality of each solution with a chosen quality measure, and then to choose the best. The problem is that the quality function usually grows monotonically with the growth of the number of clusters/topics; although the speed of this growth decreases and at some point becomes negligible, this point is virtually indiscernible. Therefore, additional criteria to stop this growth are needed. These "stopping criteria" are highly debated.

The labeling functionalities of software are the means given to a researcher to understand what the resulting clusters/topics are about (i.e., to assign a label to them). Automatically generated labels can take the form of lists of the most "typical" words and/or texts. If they are insufficient or absent, no matter how good and computationally efficient an implemented algorithm is, this software cannot be used for obtaining meaningful results. Text preprocessing is the preparation of texts for analysis, which includes word count. In the course of preprocessing, software should be able to "understand" the respective script (in our case, Cyrillic), clean it of html tags, punctuation and other impediment symbols, and recognize different forms of the same word (e.g., "go/went;" this process is called lemmatization) to provide the correct word count. These functions, except for word count, are almost never present in academic clustering or topic modeling software.

Finally, the problem of the applicability of algorithms to certain types of text stems from the variability of human language(s): what works well for English research articles may not work for Chinese poetry or for Russian blogs. We have not found any cross-language comparisons of algorithm performance in the research literature. Furthermore, most existing algorithms have been tested on texts that presumably can be easily discriminated on the basis of their lexical composition—that is, either on research articles or media texts. Blog posts, however, have highly overlapping vocabularies and are often too short to have specific lexical compositions, which makes their division into groups a more difficult task (Perez-Tellez, Pinto, Cardiff, & Rosso, 2010). At the same time, writing styles in blogs differ much more than the styles in scientific texts, so some topics are formed based on style, which is evident from our own research. We have found no comparisons of the performance of the same algorithms on blogs and other types of texts in the research literature. Finally, we have not found any comparison of clustering and topic modeling algorithms.

Algorithms, Software, and Manual Methods Used in This Research

The lack of comparative tests deprives social scientists and other potential users of these algorithms of criteria that could help them choose between existing software. With this limitation in mind, we have gathered information about

several dozen software implementations of clustering and topic modeling algorithms and chosen one for each approach: gCLUTO from Karypis Lab, University of Minnesota, for clustering, and Stanford Topic Modeling Toolbox (TMT) for topic modeling. In doing so, we have tried to find those that would do the best job in all seven areas of our concern, despite the near impossibility of fulfilling all the requirements.

gCLUTO[2] performed well in terms of algorithmic transparency and quality (see Rasmussen & Karypis, 2004; Zhao & Karypis, 2004, 2005), but the scarce labeling functionalities made this sophisticated tool virtually useless for our needs. We stopped using it early on, so here we only present the results of its use for topic modeling. TMT and its algorithms are equally well described in the academic literature (Asuncion et al., 2009; Griffiths & Steyvers, 2004; Ramage et al., 2009a, 2009b), and the software has also been applied to blog texts (Ramage, Dumais, & Liebling, 2010). It is less user-friendly, but its rich output is sufficient for interpretation of the results. Neither of these two packages does Russian text preprocessing. For this reason, we have developed a number of scripts and used the Russian-language lemmatizer called MyStem.

Although both packages have in-built quality functions, neither of them offers stopping criteria to decide when to stop increasing the number of clusters/topics. To solve this problem, we have implemented a recent approach called distortion theory (Sugar & James, 2003) and developed software that helps to find jumps in the quality function—that is, to indicate when the quality gain decreases abruptly. This point shows the "right" number of clusters for gCLUTO (based on its ISim quality function) and the "right" number of topics for TMT (based on the perplexity function). Because the optimal numbers were slightly different for our three time periods, we chose the second best solutions, which coincided at the number of 100 and made our datasets more easily comparable to each other. Unlike in our other work (Maslinsky, Koltsov, & Koltsova, 2013), we performed a manual comparison of the topic al composition of the three different periods. Labeling was based on the reading of the 20 most probable words and the 20 most probable blogposts for each topic by two researchers, who eliminated discrepancies through discussion of their results. Earlier, different quantities of words and texts had been tried. Smaller numbers were insufficient for human interpretation; larger numbers seldom added to the interpretability. Some preliminary analysis concerning authorship has also been made.

The main principles of the topic modeling approach can be outlined as follows. Metaphorically, any such algorithm "assumes" that a human creates texts by picking words from "bags" called topics. Having received the texts that resulted from this type of word-picking process, the algorithm tries to backtrack this process and to restore the word composition of each bag. Since it cannot identify it for sure, it assigns each word to each bag/topic with some probability, based on how often pairs of words meet in texts. Therefore, the first output of this type of algorithm is a table (matrix) of the probabilities of words belonging to topics (which can later be sorted by their probabilities to form lists of top-words, as shown in Table 2). Next, the algorithm assigns each text to each topic with

Table 2. Example of Sorted Lists of the 20 Most Probable Words in Two Topics, December 2011

Topic 000	Topic Weight: 25,397	Topic 001	Topic Weight: 29,420
Ukraine	1,003	Nemtsov[a]	905
President	611	Putin	477
Russian	532	Navalny[b]	456
Unkrainian	440	Sobchak[b]	286
Lukashenko[a]	234	Opposition	219
Head	229	Conversation	209
Minister	211	Bolotny[c]	204
Viktor[a]	180	Bojen[b]	200
Union	178	The people	197
Yanukovich[a]	175	To come out	196
State	170	Sakharov[c]	195
Timoshenko[a]	166	Boris[b]	184
Vladimir[a]	164	To give a speech	171
Claim	163	Kurginyan[b]	166
Belarus	163	Street	163
Council	161	To come	161
Nuclear	161	Xenia[b]	158
Security	158	Leader	152
Moscow	147	To hiss of	152
Belorussian	147	To name	139

[a]Names of Russian, Ukrainian, and Belorussian political leaders.
[b]Names of street actions' leaders.
[c]Locations of main street actions.

some probability, based on how similar the lexical composition of the text is to that of the topic. Thus, the second output of the algorithm is a matrix of texts belonging to topics. These text groups (also called clusters) are overlapping; that is, each text is considered to be multi-topical. Furthermore, by summing the probabilities of all of the words assigned to a topic, a relative weight of the topic's importance in a given collection of texts can be calculated (see weights placed next to topics' IDs in Table 2). It may be used as an indicator of the topic's salience in a given collection.

In this research, we have used a topic modeling algorithm called Latent Dirichlet Allocation (LDA) with Gibbs sampling, introduced by Griffiths and Steyvers (2004) and present in the TMT software. It was chosen on the basis of (albeit somewhat scattered) evidence of its better performance in terms of speed and quality for large text collections as compared to other topics modeling algorithms (Daud et al., 2010, p. 15; Griffiths & Steyvers, 2004, p. 5230; Steyvers & Griffiths, 2007, p. 436).

Results

The composition of the top 2,000 bloggers demonstrates some degree of stability across the three time periods as defined in Table 1: over 1,000 appear to have messages in all periods; nearly 400 have posts in two periods. Of the top 2,000 bloggers, not all have posts in each period, which indicates that popularity

in terms of the number of followers is not directly connected to activity. The largest proportion of the top 2,000 (approximately 1,800) wrote messages in December 2011, the month of the initial politicization; the lowest (approximately 1,500) wrote messages in August 2011, the control period.

Labeling has shown that approximately two-thirds of the topics were easily interpretable by a human and could be called domain clusters, that is, clusters that contain texts on a specific subject area or event. The remaining one-third was shared between "language," "style," and "noise" topics. Language topics contained texts in languages other than Russian (in our case, Ukrainian and English). Style topics were centered on writing styles, for example, offensive vocabulary, or excessive use of specific terms, proper names, digits, measures, or computer terms. Noise contained un-interpretable texts or un-interpretable combinations of meaningful texts. The presence of noise might indicate an excessive number of topics in the solutions. However, some domain topics obtained from these solutions contained important political events that were not visible in solutions with fewer topics. Solutions with fewer groupings draw a better overall picture and detect larger topics that change less with time (e.g., "politics and state," "recreational activity," or "private life"). Solutions with more subdivisions detect smaller topics, while larger topics are split into subtopics, so their structure may be better studied. Thus, "Islam," as one of the test topics, did not appear unless the number of subdivisions reached 100, and even then, it appeared as a component of neighboring topics (e.g., "terrorism," "Libyan conflict," or "Israeli–Palestinian relations"). The elections and protests topic, on the other hand, at the level of 100 clusters, is well divided into several meaningful subtopics.

The general topic structure is very stable across all three periods, although its proportions change moderately (Table 3). Two large groups can be discerned within the cluster of domain topics. The "culture and private issues" group demonstrates manifest stability. The composition of its four subgroups—"culture," "recreational activity," "consumption," and "private life"—virtually does not change, except for seasonal holidays. The "private life" group includes personal relations, family, and everyday issues. Our first attempts to link topics to authors shows that there are virtually no mono-topical authors. On average, each blogger addresses half of the topics, but the more posts an author makes, the more topics they cover. Active bloggers cover 100 percent or a little less, including cultural and private issues. This means that among the top 2,000 bloggers, virtually no one concentrates solely on political or social issues, although their belonging to the top might be attributed to the presence of these topics. Whether those who may be called opinion leaders are actually more concerned about politics is a question to be answered in future research, by comparing them with "ordinary" bloggers. It is already clear, however, that private and cultural issues are no less important than politics for opinion leaders. This may support Bishop's (2005, p. 187) assumption about people having firmer opinions about everyday matters than about the public affairs about which they are most frequently asked.

Table 3. Topic Structure for the Three Sample Periods

	Topics' Weight Shares and Quantity		
	August–September 2011	December 2011	March 2012
Social and political issues (public affairs)	32% (37 topics)	33.7% (33 topics)	41% (40 topics)
	International relations: U.S.–Middle East relations; Russian–Ukrainian–Belorussian relations; Russia–Caucasus–Central Asia relations; Arab–Israeli relations & Islam; Asia; Caucasus; Moldova; Syria; 9/11	*International relations*: U.S.–Middle East relations; Russian–Ukrainian–Belorussian relations; Russia–Israel & Russian–Jewish relations; history of Russia–Caucasus relations; Kim Chen Ir's death	*International relations*: U.S.–Middle East relations & UN; Belorussian politics, Baltics & Poland; Ukrainian politics; U.S. politics
	Russian politics & history: general politics, governmental projects & decision making; parties & elections; protest actions; Putin & Medvedev; nazi crime; military; aircrafts & tanks; World War 2; Soviet history; general history	*Russia politics & history*: governmental projects & decision–making; regional administration; political slogans; Word War 2 & aviation; military	*Russian politics & history*: power, politics & economics; political ideologies; governmental policies; Moscow & regional administration; military transport; military & Ulianovsk; World War 2; Soviet history; mixed opinionated texts
	Economics: oil & Russian economy; industry; world economic crisis	***Elections & protests* (14.9%): opinions for & against protests; protests at Sakharova & Bolotnaya—personal accounts; protests—media reports; protest participants—media reports; protest participants—opinions; authorities' reaction on protests—media reports; authorities' reaction on meetings—opinions; protesters' arrests; firing *Kommersant* journalists; parliamentary elections results; observers' reports on parliamentary elections; anticipating presidential elections; registration of candidates for presidency**	***Elections & protests* (8.8%): Putin's victory; voting & falsifications; elections, including mayoral; March protests & Navalny; protesters' arrests; protests, TV & Sobchak; Shein's hunger strike; law, society & freedom**
	Social issues: law & statehood; Motherland—friends & foes; courts & trials; crime & police; church & orthodoxy; faith, God, sin; transport; space; population problems; higher education; new academic year; air crush in Yaroslavl; Perm events; Dagestan university incident	*Economics*: industry; world finance;	*Economics*: financial markets & crisis; large business & policy on it; state budget & wages; employment, companies & their clients
			Social issues: space & nuclear energy; public transport; health care & medicine; terrorism & related crime; nationalism & migration; Jews, Arabs, Judaism & Islam; orthodoxy; Pussy riot arrest;

Table 3. Continued

	Topics' Weight Shares and Quantity		
	August–September 2011	December 2011	March 2012
		China & world economics	scandal with patriarch's flat; misconduct & tortures in police; criminal proceedings; accusations against Shuvalov
		Social issues: church & orthodoxy; education; corruption; crime & courts; automobiles in the streets, space	
Cultural & private issues	30.4% (39 topics)	32.6% (39 topics)	31% (40 topics)
	Culture: architecture; music & concerts; literature & writing; book publishing; visual art; movies; photography; television; computers; Internet & blogs; LJ; Twitter; art auction in England, philosophy	*Culture*: architecture & churches; museums & art; theaters & concerts; book publishing; movie about Vysotsky; television; photography; online photo & video; blogs (2 topics)	*Culture*: memorials & buildings; museums & exhibitions; literature & books; ideas, knowledge & culture; science & research; concerts, albums & movies; movies & *Men in Black*; photography; Internet, blogs, social networks & infographics (5 similar topics)
	Recreational activity: cars & races; football; pets; cooking recipes; competitions; fortune-telling; tourism; eco-travel; nature & fishing; weather	*Recreational activity*: New Year & Christmas; transporting; football; animals; general cooking recipes; sweet cooking recipes; competitions; tourism to European cities; eco-travel & India; nature; weather; communication (letters, phones, Internet)	*Recreational activity:* Easter, cars, football, pets, animals & zoo; cooking recipes; seasons & weather; nature; travel & rest; tourism—Asia; miss Belarus competition
	Consumption & goods: fashion & consumption; wine consumption; on-line games; phones; Internet-trade	*Consumption & goods*: restaurants; clothes & body; mobile devices & Internet; housing—buying; housing—repair; goods through Internet (2 topics); announcements	*Consumption & goods*: restaurants & food; fashion & clothes; drinks; cars & traffic jams; banks; photos & toys; computers & mobile devices; Internet & Yandex; goods through Internet
	Private: love & personal relations; sex; sex, smoking & drinking; wedding; family; children; work & money gaining; personal health; doctors & hospitals	*Private*: poems & romantic photos; love & interpersonal relations; sex & beauty; family & children; home, family & work; work, salary & credits	*Private*: sex; love, coupling & wedding; family & relatives; family, children & parenting; home & everyday life

Table 3. Continued

	Topics' Weight Shares and Quantity		
	August–September 2011	December 2011	March 2012
Other	37.6% (24 topics) *Language*: Ukrainian & English language texts (3 clusters) "*Style*": offensive vocabulary; calendar; personal names; English-language Internet terms (2 clusters) *Noise*: uninterpretable & mixed (14 topics)	33.7% (28 topics) *Language*: Ukrainian & English language texts, English-Russian translation (4 topics) "*Style*": offensive vocabulary; calendar; measures; English-language scripts *Noise*: uninterpretable & mixed (18 topics)	28% (20 topics) *Language*: Ukrainian texts, English-Russian translations "*Style*": offensive vocabulary, personal names, calendar, digits & measures (8 topics) *Noise*: uninterpretable (10 topics)

Note Topics pertaining to elections and protests are in bold.

The public affairs group includes "international relations," "Russian politics and public administration," "social issues," and, quite surprisingly, very little on "economics." If elections and protests were to be discounted, social issues would be the most volatile group, with changes either driven by unexpected events, such as social crises, or by new governmental policies and legislative initiatives. This suggests that this group of messages may function as a barometer of self-generated public opinion; even though these topics may be less salient than some from the recreational area, they reflect the bloggers' reactions on the changing social situation. When compared to each other and to different time periods, they indicate the relative importance of different social issues for the blogosphere community. Some more specific observations may be mentioned: international relations mostly concern Russia's neighbors, the Middle East, and the U.S. Europe is missing from this picture. It is for future research to reveal whether this is a permanent trend or a coincidence. In domestic affairs, World War II and Soviet history are present in all three periods—topics that we did not expect to be so persistent. From reading those texts, it can be preliminarily concluded that most of them are somehow connected with the construction and reconstruction of Russia's identity throughout its history.

The general topic structure obtained facilitated further fast selection of texts belonging to the topics of our special interest, and further in-depth manual analysis. Having cast a closer look at the "elections and protests" topics, we were able to describe the composition of this group of messages and their communicative roles. This area accounted for 13 topics in December 2011, when it reached almost 15 percent of the topic weight. Four groups of topics can be identified within this subject area. The first group of six subtopics concerns the retransmission of news from online media, while selecting those of oppositional content (e.g., the firing the *Kommersant* newspaper top-management, the arrest of political activists, or the registrations and refusals to register presidential candidates). A second group of four subtopics contained opinion messages: short emotional utterances about political characters, namely about participants in protest actions and presidential candidates, and long sophisticated speculations on the forthcoming presidential elections, as well as arguments for and against street actions. A third group of two subtopics were alternative news, that is, personal reports of street actions and reports from observers at the parliamentary elections. The last very specific topic contained the results of voting, turnout, exit polls and regular polls, and conclusions about the legitimacy of the parliamentary elections based on juxtaposition of these data. It is worth mentioning that none of the examined LiveJournal posts contained direct appeals for political action, such as "come to the street action." Neither were they used for practical coordination, such as arranging the time and place of meetings; rather they promoted specific agenda and discussion, while mobilization activity might presumably be found in Twitter and the social networks.

In March 2012, the proportional weight of the "elections and protests" subject area declined to about 9 percent—protests may have been considered a lost game at the moment Putin won the presidential elections on March 4. However, the

proportion of all social and political topics became larger than in both August–September 2011 (control) and December 2011 (parliamentary elections). While the "elections and protests" topic expanded at the expense of other public affairs in December, in March (presidential elections), the "public affairs" topic expanded at the expense of both "protests and elections," and "noise." There was a widening of the spectrum of issues addressed within the sociopolitical domain (32 nonelection social and political topics in March, compared to 20 in December). The enrichment of this sector of the LiveJournal agenda may be seen as one of the outcomes of the protests, which have been fruitless in the sense of any direct influence on the result of the elections, but which had other social consequences, such as mobilization of various issue-specific social movements and activities.

Conclusions

To a certain degree, this study is of an illustrative nature. With more complete data—for instance, including posts from ordinary bloggers and from different platforms—more generalizable conclusions could be derived. However, this article shows what results can be obtained through methods of topic modeling developed in computer science, and how these methods may serve the goals of social science for which they are applied for the first time. In particular, our research demonstrates how topic modeling can identify public agendas, their composition, structure, relative salience of different topics, and their evolution without prior knowledge of the issues to be sought. When methods for tracking attitudes to those agendas are added, a rich map of self-generated public opinion can be drawn. Such mapping cannot be directly compared to opinion polls and is not meant to replace them; rather, it is a new way of learning what people think and what they think about—a way that makes vast amounts of user-generated content about society available for social analysis. There is still a long way to go before the relevant instruments become mature, and this will demand the efforts of a whole community of researchers, not of a single research group. However, at this stage of the research, a number of important conclusions and implications for further research can be derived.

The large degree of stability of the public agenda was revealed, and this is a valuable finding, establishing a benchmark to trace on-going changes. The revealed stability of the overall volume of messages produced in a unit of time is also important: although online publishing possibilities are unlimited, the ability of bloggers to create agendas and opinions is limited and constant. Given the influence of the top bloggers in the Russian blogosphere (and elsewhere), it may be claimed that, like the traditional media, they serve as filters of issues to be thought about and as definers of their relative importance and salience.

One of the most counterintuitive findings concerning the Russian-language blogosphere is that even among the top bloggers—therefore presumably the most politically active—public attention is relatively evenly shared between social/political issues and private/recreational topics. Therefore, even less attention to

public affairs may be expected among average users, whose study is the next step of our research. The correct interpretation of this relatively low political activity of the top-users can be only affirmed by comparison with the topical structures of the blogospheres of other language domains.

Finally, apart from introducing sentiment analysis and more sophisticated techniques of diachronic topic modeling, a key direction for future research is the enrichment of topic detection in posts with various concomitant data, such as bloggers' positions in various ratings and data on comment distributions and content. This will provide a wealth of valuable information about which topics attract the most commenting activity (positive and negative) and which topical "profiles" of bloggers are the most successful.

Olessia Koltsova, National Research University Higher School of Economics [olessia.koltsova@gmail.com]
Sergei Koltcov, National Research University Higher School of Economics

Notes

This research was carried out by the Laboratory for Internet Studies of National Research University Higher School of Economics (NRU HSE), St. Petersburg campus, supported by the NRU HSE Research Foundation, Moscow, Russia, grant no. 11040006, 2011–12. The authors are thankful to Kirill Maslinsky and Elizaveta Tereschenko for performing topic modeling and clustering, and for text pre-processing.

1. www.blogs.yandex.ru (accessed March 14, 2013).
2. http://glaros.dtc.umn.edu/gkhome/cluto/gcluto/overview (accessed April 19, 2012).

References

Alexanyan, K., and O. Koltsova. 2009. "Blogging in Russia Is Not Russian Blogging." In *International Blogging: Identity, Politics and Networked Publics*, eds. A. Russel, and N. Echchaibi. New York: Peter Lang, 65–84.

Andrews, N.O., and E.A. Fox. 2007. "Recent Developments in Document Clustering." October 16. http://eprints.cs.vt.edu/archive/00001000/01/docclust.pdf. Accessed April 17, 2012.

Asuncion, A., M. Welling, P. Smyth, and Y.W. Teh. 2009. "On Smoothing and Inference for Topic Models." In Proceedings of 25th Conference on Uncertainty in Artificial Intelligence (UAI '09), 27–34.

Bardes, A.B., and R.W. Oldendick. 2012. *Public Opinion: Measuring the American Mind*. Plymouth, UK: Rowman & Littlefield.

Bishop, G.F. 2005. *The Illusion of Public Opinion: Fact and Artefact in American Public Opinion Polls*. Lanham, MD: Rowman and Littlefield.

Blei, D.M., A.Y. Ng, M.I. Jordan, and J. Lafferty. 2003. "Latent Dirichlet Allocation." *Journal of Machine Learning Research* 3: 993–1022.

Blumer, H. 1948. "Public Opinion and Public Opinion Polling." *American Sociological Review* 13 (5): 542−549.

Carpineto, C., S. Osiński, G. Romano, and D. Weiss. 2009. "A Survey of Web Clustering Engines." *ACM Computing Surveys (CSUR)*, 41 (3): Article No. 17.

Cohen, B.C. 1963. *The Press and Foreign Policy*. Princeton: Princeton University Press.

Daud, A., J. Li, L. Zhou, and F. Muhammad. 2010. "Knowledge Discovery Through Directed Probabilistic Topic Models: A Survey." *Proceedings of Frontiers of Computer Science in China* 4 (2): 280–301.

Dey, L., and S.M. Haque. 2009. "Opinion Mining from Noisy Text Data." *IJDAR* 12: 205–26.

Diesner, J., and K.M. Carley. 2010. "A Methodology for Integrating Network Theory and Topic Modeling and Its Application to Innovation Diffusion." IEEE International Conference on Social Computing (SocComp), Workshop on Finding Synergies Between Texts and Networks, August 20–22, Minneapolis, MN.

Droba, D.D. 1931. Methods Used for Measuring Public Opinion. *American Journal of Sociology*, 37, 410–423.

Enikolopov, R., V. Korovkin, M. Petrova, K. Sonin, and A. Zakharov. 2013. "Field Experiment Estimate of Electoral Fraud in Russian Parliamentary Elections." *Proceedings of the National Academy of Sciences.* 110 (2): 448–52.

Erikson, R., and K. Tedin. 2011. *American Public Opinion: Its Origins, Content, and Impact*, 8th ed. Russia: Pearson.

Etling, B., K. Alexanyan, J. Kelly, R. Faris, J. Palfrey, and U. Gasser. 2010. "Public Discourse in the Russian Blogosphere: Mapping RuNet Politics and Mobilization." Berkman Center Research Publication. October 19. http://cyber.law.harvard.edu/publications/2010/Public_Discourse_Russian_Blogosphere. Accessed July 31, 2012.

Godbole, N., M. Srinivasaiah, and S. Skiena. 2007. "Large Scale Sentiment Analysis for News and Blogs." ICWSM'2007, March 26–28, Boulder, Colorado.

González-Bailón, S., A. Kaltenbrunner, and R.E. Banches. 2012. "Emotions, Public Opinion and U.S. Presidential Approval Rates: A 5 Year Analysis of Online Political Discussions," *Human Communication Research* 38 (2): 121–43.

Gorny, E. 2004. "Russian LiveJournal: National Specifics in the Development of a Virtual Community." *Russian-cyberspace.org*. http://www.ruhr-uni-bochum.de/russ-cyb/library/texts/en/gorny_rlj.pdf. Accessed April 05, 2012.

Griffiths, T.L., and M. Steyvers. 2004. "Finding Scientific Topics." *Proceedings of the National Academy of Sciences* 101 (Suppl 1): 5228–35.

Hopkins, D., and G. King. 2010. "A Method of Automated Nonparametric Content Analysis for Social Science." *American Journal of Political Science* 54 (1): 229–47.

King, G., J. Pan, and M.E. Roberts. 2013. "How Censorship in China Allows Government Criticism but Silences Collective Expression." *American Political Science Review* 107 (2): 1–18.

Koltsova, O. 2011. "Coverage of Social Problems in St. Petersburg Press." In *Use and Views of Media in Sweden & Russia*, eds. C. Feilitzen, and P. von Petrov. Södertörn Academic Studies, no. 44, Mediestudier vid Södertörns högskola, no. 2011:1, Huddinge.

Ku, L.-W., Y.-T. Liang, and H.-H. Chen. 2006. "Opinion Extraction, Summarization and Tracking in News and Blog Corpora." *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs* 2006, 100–107.

Lazarsfeld, P., B. Berelson, and H. Gaudet. 1950. *The People's Choice*. New York: Duell, Sloan and Pearce.

Li, B., S. Xu, and J. Zhang. 2007. "Enhancing Clustering Blog Documents by Utilizing Author/Reader Comments." In *Proceedings of the 45th ACM Southeast Conference (ACMSE 2007)*, 94–9, Winston-Sale, North Carolina. http://portal.acm.org/citation.cfm?id=1233359. Accessed February 16, 2011.

Li, F., M. Huang, and X. Zhu. 2010. "Sentiment Analysis with Global Topics and Local Dependency." Twenty-Fourth AAAI Conference on Artificial Intelligence, July 11–15, Atlanta

Lin., C., and Y. He. 2009. "Joint Sentiment/Topic Model for Sentiment Analysis." In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*. November 2–6, Hong Kong, China, 375–84.

Liu, Y., A. Niculescu-Mizil, and W. Gryc. 2009. "Topic-Link LDA: Joint Models of Topic and Author Community." In *Proceedings of the 26th Annual International Conference on Machine Learning*, June 14–18, Montreal, Canada, 665–72.

Maslinsky, K., S. Koltsov, and O. Koltsova. 2013. "Changes in the Topical Structure of Russian-Language LiveJournal: The Impact of Elections 2011." National Research University Higher

School of Economics Working Papers of the Basic Research Program, Series: Sociology, WP BRP 14/SOC/2013.

McCombs, M.E., and D.L. Shaw. 1972. "The Agenda-Setting Function of Mass Media." *Public Opinion Quarterly* 36 (2): 176.

Mebane, W.R., and K. Kalinin. 2009. "Comparative Election Fraud Detection." APSA 2009 Toronto Meeting Paper. Available at SSRN: http://ssrn.com/abtract=1450078.

Mei, Q., D. Cai, D. Zhang, and C. Zhai. 2008. "Topic Modeling with Network Regularization." *Proc. of Int. World Wide Web Conf. (WWW'08)*, 101-10.

Monroe, A.D. 1975. *Public Opinion in America*. New York: Harper & Row Publishers.

Nallapati, R., and W.W. Cohen. 2008. "Link-PLSA-LDA: A New Unsupervised Model for Topics and Influence of Blogs." *Proceedings of ICWSM 2008*, March 30–April 2, Seattle.

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

Pang, B., L. Lee, and S. Vaithyanathan. 2002. "Thumbs Up? Sentiment Classification Using Machine Learning Techniques." *Proceedings of the Conference on Empirical Methods in Natural Language Processing of EMNLP*, 79–86.

Papacharissi, Z. 2007. "Audiences as Media Producers: Content Analysis of 260 Blogs." In *Blogging, Citizenship and the Future of Media*, ed. M. Tremayne. NY & London: Routledge, 21–38.

Perez-Tellez, F., D. Pinto, J. Cardiff, and P. Rosso. 2010. "Characterizing Weblog Corpora." In *NLDB 2009, LNCS 5723*, 299–300.

Ramage, D., S. Dumais, and D. Liebling. 2010. *Characterising Microblogs with Topic Models*. ICWSM 2010. http://research.microsoft.com/pubs/131777/twitter-icwsm10.pdf. Accessed August 4, 2012.

Ramage, D., D. Hall, R. Nallapati, and C.D. Manning. 2009a. "Labeled LDA: A Supervised Topic Model for Credit Attribution in Multi-Label Corpora." *Proceedings of EMNLP*, August 6–7, Singapore.

Ramage, D., E. Rosen, J. Chuang, C.D. Manning, and D.A. McFarland. 2009b. "Topic Modeling for the Social Sciences." Workshop on Applications for Topic Models, NIPS, December 11, Whistler, Canada.

Rasmussen, M., and G. Karypis. 2004. "gCLUTO: An Interactive Clustering, Visualization, and Analysis System." Tech Rep. 04–021 Univerisity of Minnesota.

Sayre, B., L. Bode, D. Shan, D. Wilcox, and C. Shah. 2010. "Agenda Setting in a Digital Age: Tracking Attention 8 in Social Media, Online News, and Conventional News." *Policy and Internet* 2 (2): 7–32.

Simon, R.J. 1974. *Public Opinion in America, 1936–1970*. Chicago: Rand McNally College Pub. Co.

Steyvers, M., and T. Griffiths. 2007. "Probabilistic Topic Models." In *Handbook of Latent Semantic Analysis*, eds. T. Landauer, D. McNamara, S. Dennis, and W. Kintsch. Hillsdale, NJ: Erlbaum, 427–48.

Sugar, C., and G. James. 2003. "Finding the Number of Clusters in a Data Set: An Information Theoretic Approach." *Journal of the American Statistical Association* 98: 750–63.

Thelwall, M., K. Buckley, and G. Paltoglou. 2012. "Sentiment Strength Detection for the Social Web." *Journal of the American Society for Information Science and Technology* 63 (1): 163–73.

Wallach, H., I. Murray, R. Salakhutdinov, and D. Mimno. 2009. "Evaluation Methods for Topic Models." Proceedings of the 26th International Conference on Machine Learning, June 14–18, Montreal, Canada.

Wallsten, K., 2007. "Agenda Setting and the Blogosphere: An Analysis of the Relationship Between Mainstream Media and Political Blogs." *Review of Policy Research* 24 (6): 567–87.

Wu, S., J.M. Hofman, W. Mason, and D.J. Watts. 2011. "Who Says What to Whom on Twitter." International WWW Conference 2011, Hyderabad, India. ACM 978-1-4503-0637-9/11/03.

Zhao, Y., and G. Karypis., 2004. "Empirical and Theoretical Comparisons of Selected Criterion Functions for Document Clustering." *Machine Learning* 55: 311–31.

Zhao, Y., and G. Karypis. 2005. "Hierarchical Clustering Algorithms for Document Clustering." *Data Mining and Knowledge Discovery* 10 (2): 141–68.

Zhong, S., 2005. "Efficient online spherical k-means clustering." *International Joint Conference on Neural Networks* 5: 3180–85.

Zhou, Y., and P. Moy. 2007. "Parsing Framing Processes: The Interplay Between Online Public Opinion and Media Coverage." *Journal of Communication* 57 (1): 79–98.