Last author version before proofs of

Berendt, B. (in press). Text mining for news and blogs analysis. To appear in C. Sammut & G.I. Webb (Eds.), *Encyclopedia of Machine Learning and Data Mining*. Berlin etc.: Springer.

Text Mining for News and Blogs Analysis

Bettina Berendt

KU Leuven, Belgium

15 February 2015

Note: All terms in red are index terms from the first edition of the Encyclopedia of Machine Learning; assuming that the index terms have remained the same, these can be used as links within the new Encyclopedia. If the list of entries has grown, of course additional links can be made. (I assume this will be modified during the processing of this chapter).

Definition

News is "the communication of selected information on current events", where the selection is guided by "newsworthiness" or "what interests the public". News are also stories, from which the reader usually expects answers to the five Ws: who, what, when, where and why, to which a "how" is often added. News-style writing – as opposed to, for example, commentary writing – generally strives for objectivity and/or neutrality (the representation of different views on the event).

In this content-centric sense, news can be written/authored and published by professional journalists and news outlets (such as newspapers or radio or TV stations), but also by anyone else and in any other form, often called *citizen journalism*: "an alternative and activist form of newsgathering and reporting that functions outside mainstream media institutions, often as a response to shortcomings in the professional journalistic field, that uses similar journalistic practices but is driven by different objectives and ideals and relies on alternative sources of legitimacy than traditional or mainstream journalism." (Radsch, 2013, p. 159). However, news, or mainstream(-media) news, is also often thought of in a source-centric way: reports authored by professional journalists in mainstream media institutions, as opposed to reporting from citizen journalists (or anyone else) who generally publish on the Web, in the form of blogs with a certain form of periodicity.

A *blog* is a (more or less) frequently updated publication on the Web, sorted in reverse chronological order of the constituent blog posts. Blog content may reflect any interest including journalistic, personal, corporate, and many others. Early blogposts (late 1990s) tended to be published on content management platforms without length restrictions; with the success of Twitter and similar *microblogging* platforms, much blogging (and of blog mining) has shifted to short posts (e.g. 140 characters on Twitter.com and Weibo.cn, although the latter's Chinese characters allow for much more complex messages). Twitter in particular has attained a major worldwide role in the fast diffusion of news (or short summaries and statements, enriched by hyperlinks to more text and other

media), with citizen journalists, mainstream media themselves, politicians, and others being the publishers (Kwak et al., 2010). Current research in blog mining and the remainder of the present article reflect this dominance of (a) news or news-related content and (b) microblog format. In addition, blog mining overlaps with *social media mining* (Zafarani et al., 2014). In particular, the *social graph* of a microblogger allows the mining analyst to track the blogger's sources and readers/"followers" along with the contents.

News and blogs consist of textual and (in some cases) pictorial content, and, when Web-based, may contain additional content in any other format (e.g., video, audio) and hyperlinks. They are indexed by time and structured into smaller units: news media into articles, blogs into blog posts. In most news and blogs, textual content dominates. Therefore, text analysis is the most often applied form of knowledge discovery. This comprises tasks and methods from data/text mining, information retrieval, and related fields. In accordance with the increasing convergence of these fields, this article refers to all of them as *text mining*. The present entry will illustrate the overlap with / use of these fields and highlight the specifics that derive from the domain, including data, tasks, users and use cases.

Motivation and Background

News and blogs are today's most common sources for learning about current events and also, in the case of blogs, for uttering opinions about current events. In addition, they may deal with topics of more long-term interest. Both reflect and form societies', groups' and individuals' views of the world, fast or even instantaneous with the events triggering the reporting. However, there are differences between these two types of media regarding authoring, content, and form. News is generally authored by people with journalistic training who abide by journalistic standards regarding the style and language of reporting. Topics and ways of reporting are circumscribed by general societal consensus and the policies of the news provider. In contrast, everybody with Internet access can start a blog, and there are no restrictions on content and style (beyond the applicable types of censorship). Thus, blogs offer end users a wider range of topics and views on them.

These application characteristics lead to various linguistic and computational challenges for text-mining analyses of news and blogs:

- *Indexing, taxonomic categorization, partial redundancy, and data streams:* News is indexed by time and by source (news agency or provider). In a multisource corpus, many articles published at about the same time (in the same or in other languages) describe the same events. Over time, a story may develop in the articles. Such multiple reporting and temporal structures are also observed for popular topics in blogs.
- *Language and meaning:* News is written in clear, correct, "objective," and somewhat schematized language. Usually, the start of a news article summarizes the whole article (feeds are a partial analogue of this in blogs). Information from external sources such as press agencies is generally reused rather than referenced. In sum, news makes fewer assumptions about the reader's background and context knowledge than many other texts.
- *Nonstandard language and subjectivity:* The language in blogs ranges from high-quality, "news-like" language through poor-quality, restricted-code language with many spelling and grammatical errors, to creative, sometimes literary, language. A blog may employ high-

quality language, but operate outside the news genre or across journalistic genres (e.g. combining current-events reporting with commentary and background information). Jargon is very common in blogs, and new linguistic developments are adopted far more quickly than could be reflected in external resources such as lexica. Many blog authors strive not for objectivity, but for subjectivity and emotionality.

- *Thematic diversity and new forms of categorization:* News are generally categorized by topic area ("politics," "business," etc.). In contrast, a blog author may choose to write about differing, arbitrary topics. When blogs are labelled, it is usually not with reference to a stable, taxonomic system, but with an arbitrary number of tags: free-form, often informal labels chosen by the user.
- *Context and its impact on content and meaning:* The content of a blog (post) is often not contained in the text alone. Rather, blog software supports "Web" and "Social Web" behaviour, and bloggers practice it: multiway communication rather than broadcasting, and semantics-inducing referencing of both content and people. Specifically, hyperlinks to other resources provide not only context but also content; as do links to and from cited resp. citing people/sources. The latter evolved from "blogrolls" resp. "trackback links" in early blog software to "followees" and "retweet" links resp. "followers" in platforms such as Twitter.

Structure of the Learning System

Tasks

From a text-mining point of view, tasks can be grouped by different criteria:

- *Basic task and type of result:* description, classification and prediction (supervised or unsupervised, includes for example topic identification, tracking, and/or novelty detection; spam detection); search (ad hoc or filtering); recommendation (of blogs, blog posts, or (hash-)tags); summarization
- *Higher-order characterization to be extracted:* especially topic or event; opinion or sentiment
- *Time dimension:* nontemporal; temporal (stream mining); multiple streams (e.g., in different languages, see cross-lingual text mining)
- *User adaptation:* none (no explicit mention of user issues and/or general audience); customizable; personalized

Real-world applications increasingly employ selections or, more often, combinations of these tasks by their intended users and use cases, in particular:

- *News aggregators* allow laypeople and professional users (e.g. journalists) to see "what's in the news" and to compare different sources' texts on one story. Reflecting the presumption that news (especially mainstream news – sources for news aggregators are usually whitelisted) are mostly objective/neutral, these aggregators focus on topics and events. News aggregators are now provided by all major search engines.
- *Social-media monitoring tools* allow laypeople and professional users to track not only topical mentions of a keyword or named entity (e.g. person, brand), but also aggregate sentiment towards it. The focus on sentiment reflects the perceptions that even when news-related, social media content tends to be subjective and that studying the blogosphere is therefore an inexpensive way of doing market research or public-opinion research. The whitelist here is

usually the platforms (e.g. Twitter, Tumblr, LiveJournal, Facebook) rather than the sources themselves, reflecting the huge size and dynamic structure of the blogosphere / the Social Web. The landscape of commercial and free social-media monitoring tools is wide and changes frequently; up-to-date overviews and comparisons can easily be found on the Web.

- *Emerging application types* include text mining not of, but for journalistic texts, in particular natural language generation in domains with highly schematized event structures and reporting, such as sports and finance reporting (e.g. Allen et al., 2010; narrativescience.com) and social-media monitoring tools for helping journalists find sources (Diakopoulos et al., 2012).

Some tools have dashboard-style interfaces and complex data graphics, which may be most interesting for some professional users. However, the increasing move especially of casual users towards mobile devices with small screens has led to most applications showing original content and mining output that consists of (especially short) texts and a small number of (especially numeric) analytics.

Solution approaches

Standardisation: Tasks, datasets, APIs

The development of methods for mining news, blogs, and social media in general has profited from *standard datasets* and *standard tasks* and *competitions*. Prominent examples are the Reuters-21578 dataset, which is not only a collection of newswire articles but also the most classical dataset for text mining in general (https://archive.ics.uci.edu/ml/datasets/Reuters-21578+Text+Categorization+Collection), the larger and also multilingual RCV1, RCV2 and TRC2 datasets (http://trec.nist.gov/data/reuters/reuters.html), the blog datasets provided by the International Conference on Weblogs and Social Media (ICWSM, http://www.icwsm.org), and the SNAP datasets (https://snap.stanford.edu/data). The Topic Detection and Tracking (TDT) research program and workshops (http://www.itl.nist.gov/iad/mig/tests/tdt; Allan, 2002) were essential in the formation of news mining as a research topic. Important tasks and competitions that are ongoing, and that also offer important datasets, include the Text Retrieval Conference (TREC, http://trec.nist.gov) and the Text Analysis Conference (TAC, http://www.nist.gov/tac), formerly Document Understanding Conference DUC (http://duc.nist.gov). The history of tracks/tasks over time in these conferences also illustrates how fields have matured or become less relevant; for example, “blog tracks” have been replaced since 2010 by “microblog tracks”, and “topic detection” has given way to “event detection”.

Standard datasets are one answer to a central problem in news, blogs and social media mining in general. Since most platforms are commercial, they restrict access to their current or archived editions. Other platforms offer a free API but make it return a sample whose representativeness and/or even sampling criteria are not known; this can affect mining results severely (Morstatter et al., 2013). In addition, the terms of use present a challenge for creating re-usable datasets (for a solution approach, see McCreadie et al., 2012).

A further caveat concerns all social-media mining results: In general, APIs only give access to “public” posts and not to posts that users have set to “private” or otherwise limited to a restricted audience.

In addition, having gained access to an individual's online communication does not mean one may use or process it. Thus, privacy and data protection considerations limit the uses of social media for research; and they require careful interpretations of the results: these may be representative of the public utterances of users, but not of all of their online communication.

The modelling phase of text mining

Solution approaches are based on general data-mining methods and adapted to the conceptual specifics of news and blogs and their mining tasks (see list of tasks above). Methods include (document) classification and clustering, latent-variable techniques such as (P)LSA or LDA (cf. feature construction; specifically for an overview of topic models see Blei, 2012), mixture models, time series, and stream mining methods.

Named-entity recognition (e.g. Atkinson & Van der Goot, 2009; Ritter et al., 2011; Li et al., 2012) is an important part or companion of tasks such as topic detection or text enrichment (e.g. Štajner et al., 2010). Topic tracking and event threading are used to follow a news story unfolding over time (e.g. Shahaf and Guestrin, 2010), and especially for the purposes of summarization over time, special attention is paid to *bursty* topics or events (term introduced by Kleinberg, 2002; see Subašić & Berendt, 2013 for further references and empirical comparison), i.e. those that are marked by "spikes" in the frequency or other weight of reporting at certain points in time.

Information extraction can help to extract the *event(s)* of a news story. Events involve named entities (e.g. people and locations), a time, and a characterisation of what the event is about. Information extraction can leverage background ontologies (e.g. Kuzey et al., 2014). This covers the first four of the "five Ws" of a news story; the "why" and "how" at present remain to be extracted by human readers from the original text (which is therefore generally accessible from platforms, see remarks on semi-automatic sensemaking below). Clustering can be useful for the extraction of events from multilingual sources (Leban et al., 2014). Regularities in how reporting (or the world?) evolves have also been used for predicting events from news (Radinsky & Horvitz, 2013). The brevity of microblogs combined with the speed and volume of their streams pose special challenges for event detection (McCreadie et al., 2013).

Sentiment analysis and *opinion mining* are key especially for analysing blogs and other social media (see overviews in Feldman, 2013; Pang & Lee, 2007; Potts, 2013), and they are evolving towards more sophisticated methods that take syntactic structure and background knowledge / semantics into account (e.g. Gangemi et al., 2014). Sentiment analysis and opinion mining is designed to detect and classify "subjective" content and as such describes (some) social-media content well. It can also be appropriate for "subjective" journalistic genres such as commentary. However, this does not mean that news is really – or can ever be truly – objective. The often subtle and often subconscious structures, backgrounds and convictions that express themselves in how a news story is told are referred to as media bias or framing, and text mining has begun to address them (e.g. Recasens et al., 2013; Pollak et al., 2011; Odijk et al., 2013).

Further classification tasks that are specifically relevant for news and blogs are generally solved with features that are characteristic of the domain and/or can be easily extracted from its data. They include (a) geolocation (e.g. Hale et al., 2012); (b) recommendation (e.g. tracking multiple topics over time in news, personalized to a user whose interests may change over time was developed by Pon et

al., 2007; an approach for microblogs was proposed by Ren et al., 2013); and (c) spam detection and blocking (Kolari et al., 2006; for a general overview see Castillo & Davison, 2011).

Text summarization (for an overview, see Fiori, 2014; specifically for microblogs, see Mackie et al., 2014) is a key technique for helping users to get an overview of (a) a single document's key messages or (b) a multitude by different documents, often from different sources that in turn may have copied from one another. Today, most summarization is extractive, either extracting key sentences or non-sentence structures such as graphs. In real-world applications, even simpler forms are still predominant, including the extraction of single terms based for example on frequency and their display in tag clouds, and the use of the first sentences of news articles that, by journalistic writing conventions, are designed to summarise the text. Abstractive summarization involves the generation of natural language, which remains a hard problem. Today, it is used mostly for text genres that are highly schematized, such that templates can be used and filled with the entities/constants relevant to the story at hand (see "Emerging application types" above).

Texts, or text summaries, can be represented not only as bags of words, sets of topics or events, but also as graphs in which words and/or named entities stand in multiple relations to one another (see Berendt et al., 2014, for examples and further references). (Shallow) semantic parsing is often used to extract triples (e.g. subject-predicate-object statements) (e.g. Štajner et al., 2010; Sudhahar et al., 2015).

Text-based modeling can be enhanced by (e.g., social) *network* structure (e.g. Mei, Cai, Zhang, & Zhai, 2008) (cf. link mining and link discovery). The analysis of how the actors in a network influence one another is important for the domain of news and social media (Guille et al., 2013). Such analyses are applied not only to individual text producers, but more often to whole domains. One general question is how blogs and news, viewed in the aggregate, refer to and contextualize each other (e.g., Gamon et al. 2008; Berendt & Trümper 2009; Leskovec, Backstrom, & Kleinberg 2009).

Specifics of data understanding, data cleaning and data preparation

Data cleaning is similar to that of other online documents; in particular, it requires the provision or learning of wrappers for removing mark-up elements. Analysis methods that focus on text mining usually ignore hypermedia elements such as photographs and videos, or use only their metadata.

While news texts employ standard language and can be handled with general-purpose text-analysis software, the *language* of (micro-)blogs requires specific lexica (e.g., containing the frequently-used emoticons), abbreviation expansion and grammatical rules, and similar techniques (see "Noah's ARK" at http://www.ark.cs.cmu.edu/TweetNLP/ for a suite of tools and references); and linguists have found that rather than being "wrong" and ungrammatical, microblogs are evolving towards new systems that resemble spoken language and indicate nuances such as geographical region (Eisenstein, 2015). Like other social media, they often contain irony and other indirect uses of language for expressing appreciation or discontent (e.g. Veale & Hao, 2010), and this remains a major stumbling block for the machine understanding of these texts.

The *semi-structured* nature of blogs and news can give valuable cues for understanding. For example, the format elements "timestamp" and "number of comments" can be treated as indicators of increased topical relevance and likelihood of being opinionated, respectively (Mishne, 2007). A

combination of text clustering and *tag* analysis can serve to identify topics as well as the blogs that are on-topic and likely to retain this focus over time (Hayes et al., 2007). Twitter *hashtags* have been used for example as indicators of sentiment (Wang et al., 2011).

Like other online texts, news and blogs make frequent use of *hyperlinks*, and the content of linked materials may be necessary even for a human reader to understand a post. This is particularly true for microblogs that are often mere pointers to a URL, or a URL plus a short comment. Many mining methods therefore enrich the text by, for example, the contents of referenced URLs (e.g. Abel et al., 2011). *Semantic enrichment* can also utilize external (semi-)structured data; for example, Wikification can add context information to microblogs by drawing on Wikipedia or DBPedia (e.g. Cheng & Roth, 2013). All these methods can help to enrich and to disambiguate meaning.

The importance of interactive tools for semi-automatic sensemaking

Like most of text mining, machine analyses of news, blogs and other social media are a first step in a process of human sense-making, whether for news consumers or for news producers. It is therefore imperative to provide them with interfaces that support further steps. Thus, tools for news consumers (such as news aggregators) typically provide links to the original articles. Tools for news producers show statistics (such as aggregate opinions of "the crowd" or properties of one potential source) as an information for journalists, and topics or events detected in corpora are generally a starting point for a story, but not a story in and of themselves. Reading, understanding and writing news and blogs can probably never be totally automated. One reason for this is that different people read a given text differently, which is well-known in social-science media research but still often neglected in computational research – maybe because it requires us to question key methodological concepts of text mining such as "the ground truth". Interactive tools for story detection and tracking have been proposed as an answer to this dilemma (Berendt et al., 2014), and drag-and-drop story editors are used to create one's own new story (storify.com).

In addition, text mining as a method for dealing with large data volumes is often in competition with or combined with human intelligence for doing the same. Thus, for example, the contributions from many (often unpaid) volunteers and interface elements such as voting constitute the "social news aggregator" reddit.com, and Twitter's "retweeting" is a major, and human-led, way in which tweets are fed into, and develop influence across, multiple sub-networks formed by the platform's users. In these human-machine collaborations, the algorithms employed by a platform however are not neutral companions, but shape how users perceive others' opinions, which in turn affects their further posting behaviour. For example, Twitter's "trending topics" algorithm rewards bursty topics (cf. Wilson, 2013). This implies that even a topic contained in many tweets can, if the interest over time remains stable, disappear from the trending topics and thereby from public visibility. The implications of such algorithmic decisions on user choices and perceptions as well as public decisions and policy are a new research topic that will be relevant not only for text mining.

References

Abel, F., Gao, Q., Houben, G.-J., & Tao, K. (2011). Semantic Enrichment of Twitter Posts for User Profile Construction on the Social Web. In Proc. ESWC (2) (pp. 375-389).

Allan, J. (Ed.). (2002). Topic detection and tracking: Event-based information organization. Norwell, MA: Kluwer Academic Publishers.

Allen, N.D., Templon, J.R., McNally, P.S., Birnbaum, L., & Hammond, K. (2010). StatsMonkey: A Data-Driven Sports Narrative Writer. In Proc. 2010 AAAI Fall Symposium Series. AAAI Press. http://www.aaai.org/ocs/index.php/FSS/FSS10/paper/view/2305

Atkinson, M. & Van der Goot, E. (2009). Near real time information mining in multilingual news. In Proceedings of the 18th international conference on World wide web (WWW '09). ACM, New York, NY, USA, 1153-1154.

Berendt, B., Last, M., Subašić, I., & Verbeke, M. (2014). New Formats and Interfaces for Multi-Document News Summarization and its Evaluation. In Fiori (2014) (pp. 231-255).

Berendt, B., & Trümper, D. (2009). Semantics-based analysis and navigation of heterogeneous text corpora: The Porpoise news and blogs engine. In I.-H. Ting & H.-J. Wu (Eds.), Web mining applications in e-commerce and e-services Berlin: Springer.

Blei, D.M. (2012). Probabilistic topic models. Communications of the ACM, 55(4), 77-84.

Castillo, C. & Davison, B.D. (2011). Adversarial Web Search. Foundations and Trends in Information Retrieval. 4, 5 (May 2011), 377-486. DOI=10.1561/1500000021 http://dx.doi.org/10.1561/1500000021

Cheng, X. & Roth, D. (2013). Relational Inference for Wikification. In Proc. EMNLP 2013 (pp. 1787-1796).

Diakopoulos, N., De Choudhury, M., & Naaman, M. (2012). Finding and assessing social media information sources in the context of journalism. In Proc. CHI 2012 (pp. 2451-2460). ACM.

Eisenstein, J. (2015). Identifying regional dialects in online social media. To be published in the Handbook of Dialectology.

Feldman, R. (2013). Techniques and applications for sentiment analysis. Communications of the ACM, 56(4), 82-89.

Fiori, A. (Ed.) (2014). Innovative Document Summarization Techniques: Revolutionizing Knowledge Understanding. IGI Global.

Gamon, M., Basu, S., Belenko, D., Fisher, D., Hurst, M., & König, A. C. (2008). BLEWS: Using blogs to provide context for news articles. In E. Adar, M. Hurst, T. Finin, N. Glance, N. Nicolov, B. Tseng, & F. Salvetti (Eds.), Proceedings of the second international conference on weblogs and social media (ICWSM'08). Seattle, WA. Menlo Park, CA. http://www.aaai.org/Papers/ICWSM/2008/ICWSM08-015.pdf

Gangemi, A., Presutti, V., & Reforgiato Recupero, D. (2014). Frame-Based Detection of Opinion Holders and Topics: A Model and a Tool. IEEE Computational Intelligence Magazine, 9(1), 20-30.

Guille, A., Hacid, H., Favre, C., & Zighed, D.A. (2013). Information Diffusion in Online Social Networks: A Survey. SIGMOD Record, 42(2).

Hale, S., Gaffney, D., & Graham, M. (2012). Where in the world are you? Geolocation and language identification in Twitter. In Proceedings of ICWSM'12 (pp. 518-521).

Hayes, C., Avesani, P., & Bojars, U. (2007). An analysis of bloggers, topics and tags for a blog recommender system. In B. Berendt, A. Hotho, D. Mladeni, & G. Semeraro (Eds.), From web to social web: Discovering and deploying user and content profiles. LNAI 4737. Berlin: Springer.

Kleinberg, J.M. (2002). Bursty and hierarchical structure in streams. In Proc. SIGKDD 2002 (pp. 91-101).

Kolari, P., Java, A., Finin, T., Oates, T., & Joshi, A. (2006). Detecting spam blogs: A machine learning approach. In Proceedings of the 21st national conference on artificial intelligence. Boston: AAAI.

Kuzey, E., Vreeken, J., & Weikum, G. (2014). A Fresh Look on Knowledge Bases: Distilling Named Events from News. In Proc. CIKM 2014 (pp. 1689-1698).

Kwak, H., Lee, C., Park, H., & Moon, S. (2010). What is Twitter, a social network or a news media? In Proc. WWW (pp. 591-600). ACM.

Leban, G., Fortuna, B., Brank, J., & Grobelnik, M. (2014). Event registry: learning about world events from news. In Proc. WWW 2014 (Companion Volume) (pp. 107-110).

Leskovec, J., Backstrom, L., & Kleinberg, J. (2009). Meme-tracking and the dynamics of the news cycle. In J.F. Elder IV, F. Fogelman-Soulié, P.A. Flach, & M.J. Zaki (Eds.), Proceedings of the 15th ACM SIGKDD international conference on knowledge discovery and data mining, Paris, France. New York, NY.

Li, C., Weng, J., He, Q., Yao, Y., Datta, A., Sun, A. & Lee, B.-S. (2012). TwiNER: named entity ecognition in targeted twitter stream. In Proceedings of the 35th international ACM SIGIR conference on research and development in information retrieval (SIGIR '12). ACM, New York, NY, USA, 721-730. DOI=10.1145/2348283.2348380 http://doi.acm.org/10.1145/2348283.2348380

Mackie, S., McCreadie, R., Macdonald, C., & Ounis, I. (2014). Comparing Algorithms for Microblog Summarisation. In Proc. CLEF 2014 (pp. 153-159).

McCreadie, R., Macdonald, C., Ounis, I., Osborne, M., & Petrovic, S. (2013). Scalable distributed event detection for Twitter. In Proc. BigData Conference 2013 (pp. 543-549).

McCreadie, R., Soboroff, I., Lin, J., Macdonald, C., Ounis, I., & McCullough, D. (2012). On building a reusable Twitter corpus. In Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval (SIGIR '12). ACM, New York, NY, USA, 1113-1114. DOI=10.1145/2348283.2348495 http://doi.acm.org/10.1145/2348283.2348495

Mei, Q., Cai, D., Zhang, D., & Zhai, C. (2008). Topic modeling with network regularization. In J. Huai & R. Chen (Eds.), Proceeding of the 17th international conference on world wide web (WWW'08) Beijing, China. New York, NY. 10.1007/978-0-387-30164-8_827

Mishne, G. (2007). Using blog properties to improve retrieval. In N. Glance, N. Nicolov, E. Adar, M. Hurst, M. Liberman, & F. Salvetto (Eds.), Proceedings of the international conference on weblogs and social media (ICWSM). Boulder, CO. http://www.icwsm.org/papers/paper25.html

Morstatter, F., Pfeffer, J., Liu, H., & Carley, K.M. (2013). Is the Sample Good Enough? Comparing Data from Twitter's Streaming API with Twitter's Firehose. In Proc. ICWSM 2013. http://www.aaai.org/ocs/index.php/ICWSM/ICWSM13/paper/view/6071

Odijk, D., Burscher, B., Vliegenthart, R., & de Rijke, M. (2013). Automatic Thematic Content Analysis: Finding Frames in News. In Social Informatics 2013 (pp. 333-345). Berlin etc.: Springer. LNCS 8238.

Pang, B. & Lee, L. (2007). Opinion Mining and Sentiment Analysis. Foundations and Trends in Information Retrieval 2(1-2): 1-135 (2007).

Pollak, S., Coesemans, R., Daelemans, W., & Lavrač, N. (2011). Detecting contrast patterns in newspaper articles by combining discourse analysis and text mining. Pragmatics, 21 (4), 647-683.

Pon, R. K., Cardenas, A. F., Buttler, D., & Critchlow, T. (2007). Tracking multiple topics for finding interesting articles. In P. Berkhin, R. Caruana, & X. Wu (Eds.), Proceedings of the 13th ACM SIGKDD international conference on knowledge discovery and data mining. San Jose, CA. New York, NY.

Potts (2013). Introduction to Sentiment Analysis. (slide set). http://www.stanford.edu/class/cs224u/slides/2013/cs224u-slides-02-26.pdf [retrieved 2015-02-15]

Radinsky, K. & Horvitz, E. (2013). Mining the web to predict future events. In Proc. WSDM 2013 (pp. 255-264).

Radsch, C.C. (2013). Digital Dissidence & Political Change: Cyberactivism and Citizen Journalism in Egypt. Doctoral Dissertation, American University, School of International Service. Available at SSRN: http://ssrn.com/abstract=2379913

Recasens, M., Danescu-Niculescu-Mizil, C., & Jurafsky, D. (2013). Linguistic models for analyzing and detecting biased language. In Proceedings of ACL.

Ren, Z., Liang, S., Meij, E., & de Rijke, M. (2013). Personalized time-aware tweets summarization. In Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval (SIGIR '13). ACM, New York, NY, USA, 513-522. DOI=10.1145/2484028.2484052 http://doi.acm.org/10.1145/2484028.2484052

Ritter, A., Clark, S., Mausam, & Etzioni, O. (2011). Named entity recognition in tweets: an experimental study. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP '11). Association for Computational Linguistics, Stroudsburg, PA, USA, 1524-1534.

Shahaf, D. & Guestrin, C. (2010). Connecting the dots between news articles. In Proc. SIGKDD 2010 (pp. 623-632).

Sudhahar, S., de Fazio, G., Franzosi, R., & Cristianini, N. (2015). Network analysis of narrative content in large corpora. Natural Language Engineering, 21(1), 81-112.

Štajner, T., Rusu, D., Dali, L., Fortuna, B., Mladenic, D., & Grobelnik, M. (2010). A service oriented framework for natural language text enrichment. Informatica (Ljublj.), 34 (3), 307-313.

Subašić, I., Berendt, B. (2013). Story graphs: Tracking document set evolution using dynamic graphs, Intelligent Data Analysis, 17 (1), 125-147.

Veale, T. & Hao, Y. (2010). Detecting Ironic Intent in Creative Comparisons. In Proceedings of the 2010 conference on ECAI 2010: 19th European Conference on Artificial Intelligence, Helder Coelho, Rudi Studer, and Michael Wooldridge (Eds.). IOS Press, Amsterdam, The Netherlands, The Netherlands, 765-770.

Wang, X., Wei, F., Liu, X., Zhou, M., & Zhang, M. (2011). Topic sentiment analysis in twitter: a graph-based hashtag sentiment classification approach. In Proceedings of the 20th ACM international conference on Information and knowledge management (CIKM '11), Bettina Berendt, Arjen de Vries, Wenfei Fan, Craig Macdonald, Iadh Ounis, and Ian Ruthven (Eds.). ACM, New York, NY, USA, 1031-1040. DOI=10.1145/2063576.2063726 http://doi.acm.org/10.1145/2063576.2063726

Wilson, R. (2013). Trending on Twitter: A Look at Algorithms Behind Trending Topics. Ignite Social Media Blog. http://www.ignitesocialmedia.com/twitter-marketing/trending-on-twitter-a-look-at-algorithms-behind-trending-topics/ [retrieved 2015-02-15]

Zafarani, R., Abbasi, M.A., & Liu, H. (2014). Social Media Mining: An Introduction. Cambridge: Cambridge University Press.