Seven Years of Social Sensors

Mario Cataldi
Universite Paris 8
m.cataldi@iut.univ-paris8.fr

Luigi Di Caro
University of Turin
dicaro@di.unito.it

Claudio Schifanella
University of Turin
schi@di.unito.it

ABSTRACT

The aim of this paper is to review seven years of research on a specific vision of social media which is that of *social sensors*, i.e., alternative information systems able to detect and characterize interesting and yet unreported information and events in real-time, crossing topics, locations and language barriers. In particular, we here present a computational exercise based on a Topic Modeling technique over a set of papers citing probably the first contribution about the conceptualization and formalization of the social sensor keyword. By extracting topics from 367 (English) titles and correlating them with metadata such as the year of publication and the number of received citations, we tried to light up interesting aspects and research directions in the social media mining community.

Keywords

Social Network Analysis, Data Mining, Social Media, Social Networks, Topic Detection, Event Detection, Social Sensors

1. INTRODUCTION

Nowadays, social platforms have become the most popular communication system all over the world. In fact, due to the short format of messages and the accessibility of these systems, users tend to shift from traditional communication tools (such as blogs, web sites and mailing lists) to social network for various purposes. Billions of messages are appearing daily in these services such as Twitter, Tumblr, Facebook, etc. The authors of these messages share content about their private life, exchanging opinions on a variety of topics and discussing a wide range of information news. Microblogging services also exploit the immediateness of handy smart devices.

In [8], and later in [6], we conceptualize the vision of this powerful communication channel as *social sensor*, which can be used to detect and follow interesting and yet unreported information and specifically unknown / interesting / anomalous events, facts, and topics in real time, crossing languages, domains, locations and language barriers. Future technologies on this connectivity may also provide applications with automatic techniques for the generation of news (filtered over user profiles), offering a sideways to the existing authoritative information media.

The quite high impact of such view in the literature motivated the organization of a workshop on its related aspects. The international workshop named *SIDEWAYS*, which currently counts three editions, received interesting materials ranging from socio-cultural contributions to computational approaches. In detail, the past two editions [7, 4] focused on the following subtopics:

- detect emerging events, facts, topics [21, 25, 20]
- track the evolution over time of events, facts and topics [27]
- enrich them with contextual information like categories and named entities [21]
- identify communities and analyse large scale online/offline social networks[22]
- unravel behaviours in social networks[19]
- retrieve partecipatory decision making on civic social networks [26]
- understand key social and psychological factors and problems [23, 10, 9, 11]
- find relationships with other events and sources of information[26]
- analyze privacy issues [16]

However, Social Sensor analysis may involve other fields and study such as visualization [13], collaboration networks [14], semantic annotation [3], influence analysis [5], Sentiment Analysis [15, 24], irony detection [17], TV content analysis [1], and others.

The aim of this paper is to review those research works that based their ideas, motivations and concepts on such social sensor view. In the light of this, we carried out a classic Topic Modeling exercise over the collection of papers that have cited our original conceptualization [8, 6, 14]. We thus

SIDEWAYS 2017, Prague, Czech Republic

Table 1: Topics extracted from the 367 English-based papers citing [8] on the *social-sensor* view.

TOPIC 1	TOPIC 2	TOPIC 3	TOPIC 4
social	topic	events	data
information	topics	event	time
media	text	twitter	model
users	news	detection	social
twitter	twitter	stream	networks
user	emerging	streams	detection
data	information	time	mining
research	paper	real-time	trends
analysis	clustering	information	network
network	data	tweets	patterns
content	results	real	problem
people	detection	detect	online
paper	microblog	temporal	topics
networks	tweets	sentiment	microblogs
social media: content and users	*emerging topic detection*	*real-time event detection*	*network mining*

collected around 368 publication titles with their relative metadata information such as the type of publication (journal or proceedings), the publication year and the number of received citations. We then extracted topics from titles and abstracts, correlating them along these dimensions, highlighting some useful insights and historical perspectives for future research.

2. BACKGROUND ON TOPIC MODELING

Topic models are fundamental tools for the extraction of regularities and patterns providing automatic ways to organize, search and give sense to large data collections. The shared basic assumption is that documents have a latent semantic structure that can be inferred from word-document distributions.

Latent Semantic Analysis (LSA) [12] is a linear algebra-based method that reduces the a word-document co-occurrences matrix into a reduced space such that words which are close in the new space are similar. Its probabilistic and generative version (pLSA) [18] adds a latent context variable to each word occurrence which explicitly accounts for polysemy.

Latent Dirichlet Allocation (LDA) [2] is a fully Bayesian probabilistic version of LSA. Given a corpus of documents, the idea underlying LDA is that all documents share the same set of topics, but each document exhibits those topics in different proportions depending on words which are present in that document. Topics, in turn, are defined as different probability distributions over the words of a fixed vocabulary, but they are interpreted by restricting attention to words with the highest estimated frequency. Only documents are observed, while the topics, per-document topic distributions and the per-document per-word topic assignments are latent structures inferred from the data.

3. TOPICS FROM SOCIAL-SENSORS LITERATURE

In this section, we show the results of a LDA topic modeling exercise applied on the abstracts of the papers citing [8]. As already mentioned, this paper represents one of the first work which recognized (and formalized) the role of social sensor of social media.

Table 1 shows the 4 most significant topics[1], which we tried to label on the last row. We decided not to pre-process the texts with advanced natural language techniques (such as for example lemmatization, Named Entity Recognition and Word Sense Disambiguation) in favor of a simple experiment bringing to light the naturally-observed linguistic variability. Only English stopwords have been filtered out to highlight topics comprehension.

The results seem to show a quite clear map, where the main scientific effort is divided on 1) the analysis of social media (role, impact, contents, and user profiles), 2) the detection of emerging topics or 3) events, and 4) network mining approaches involving community detection techniques.

4. SOCIAL-SENSOR TOPICS TRENDS

In this section, we present some correlation study between the extracted topics (see previous section) and metadata such as the year of publication, the number of received citations and the type of publication (journal or not). Figure 1 shows the whole result of the study.

4.1 Social-Sensor Topics and Time

As it can be noticed, the total amount of research in the field has been growing from 2011 to 2015, when it reached a kind of convergence (year-2017 had few data records only). However, the topic ”Event Detection” is the only one that kept growing also in 2016. It is possible to think that part of the community working on topic detection then focused on events at a certain point, since Social Media is known to contain much more event-based information rather than other sources of information. This is actually one of the key motivation of the *social sensor* view.

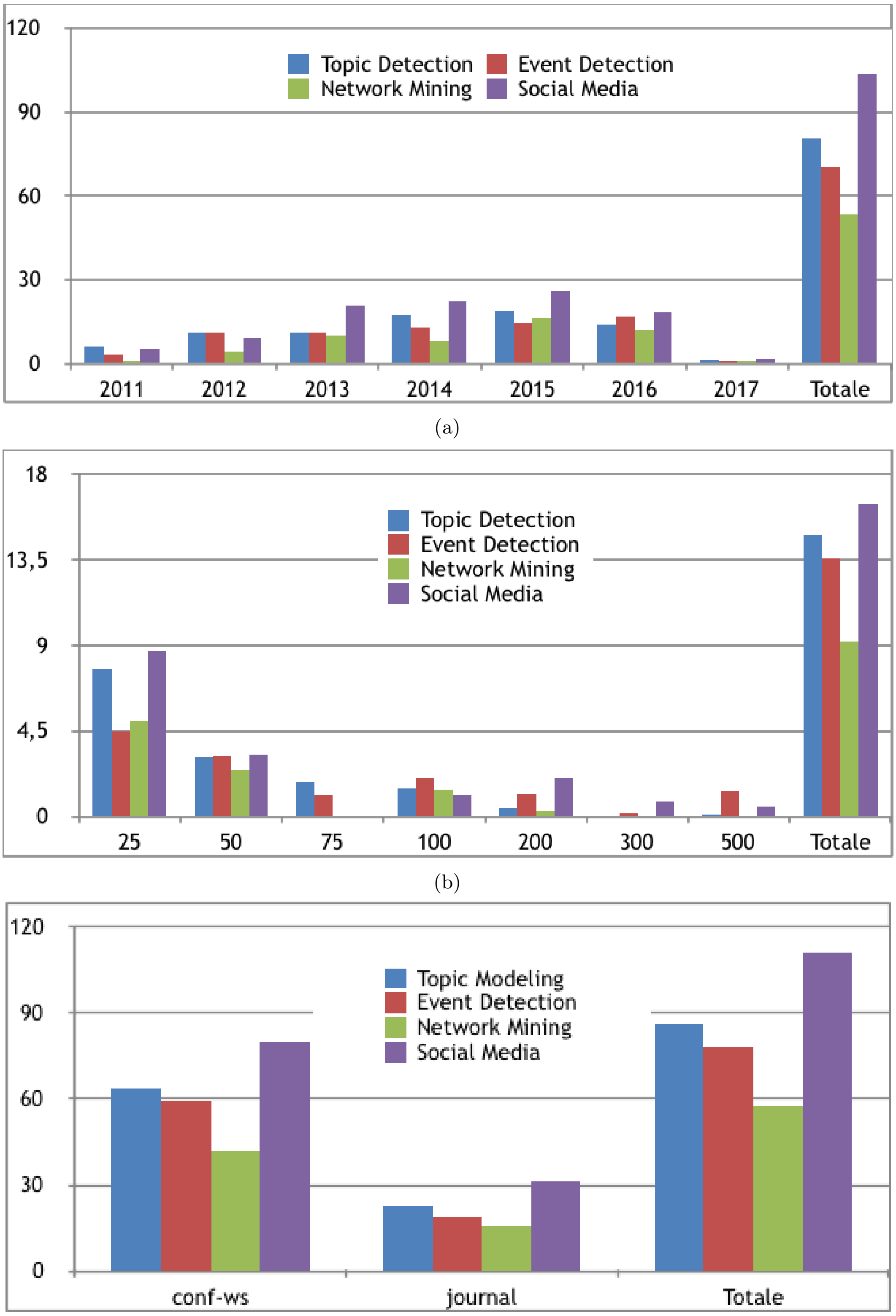
4.2 Social-Sensor Topics and Impact

Another interesting aspect was to analyze the impact of the extracted topics in terms of received citations from the research community. Figure 1 (b) shows that social-sensor papers with low citation numbers are more about topic detection and social media with respect to the other two topics. Instead, highly-cited papers are also about event detection, while topic detection papers disappear on the right side of the plot. This is quite interesting, since topic detection is the top-2 topic. In a sense, it seems that most of the work is on topic detection though it does not linearly impact on future and contextual research.

4.3 Social-Sensor Topics and Journals

With this analysis, we tried to understand if social-sensor topics have a similar distribution on conferences and workshops rather than on journals. What we found, as shown in 1 (c), is that the distribution on journals flatten the total number of papers on the different topics. This can be probably interpreted as a quality-based natural filtering.

[1] We experimented with other number of topics, showing less interpretable results.



(a)

(b)

(c)

Figure 1: Correlation between the 4 topics extracted from the 367 English-based papers citing [8] with (a) year of publication, (b) number of citations and (c) type of publication.

5. REFERENCES

[1] A. Antonini, L. Vignaroli, C. Schifanella, R. G. Pensa, and M. L. Sapino. Mesoontv: a media and social-driven ontology-based tv knowledge management system. In *Proceedings of the 24th ACM Conference on Hypertext and Social Media*, pages 208–213. ACM, 2013.

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

[3] G. Boella and L. Di Caro. Extracting definitions and hypernym relations relying on syntactic dependencies and support vector machines. In *ACL (2)*, pages 532–537, 2013.

[4] L. D. Caro, M. Cataldi, and C. Schifanella, editors. *Proceedings of the 2nd International Workshop on Social Media World Sensors, SIDEWAYS 2016, co-located with 10th International Conference on Language Resources and Evaluation (LREC 2016), Portoroz, Slovenia, May 24, 2016*, volume 1696 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2016.

[5] M. Cataldi and M.-A. Aufaure. The 10 million follower fallacy: audience size does not prove domain-influence on twitter. *Knowledge and Information Systems*, 44(3):559–580, 2015.

[6] M. Cataldi, L. D. Caro, and C. Schifanella. Personalized emerging topic detection based on a term aging model. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 5(1):7, 2013.

[7] M. Cataldi, L. D. Caro, and C. Schifanella, editors. *Proceedings of the 1st ACM Workshop on Social Media World Sensors, Guzelyurt, SIdEWayS@HT 2015, TRNC, Cyprus, September 1, 2015*. ACM, 2015.

[8] M. Cataldi, L. Di Caro, and C. Schifanella. Emerging topic detection on twitter based on temporal and social terms evaluation. In *Proceedings of the Tenth International Workshop on Multimedia Data Mining*, page 4. ACM, 2010.

[9] M. D. Choudhury. Social media for mental illness risk assessment, prevention and support. In Cataldi et al. [7], page 1.

[10] C. Colella. Distrusting science on communication platforms: Socio-anthropological aspects of the science-society dialectic within a phytosanitary emergency. In Caro et al. [4], pages 19–24.

[11] L. H. M. S. S. K. Dane Bell, Daniel Fried. Challenges for using social media for early detection of t2dm. In Caro et al. [4].

[12] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391, 1990.

[13] L. Di Caro, K. S. Candan, and M. L. Sapino. Navigating within news collections using tag-flakes. *Journal of Visual Languages & Computing*, 22(2):120–139, 2011.

[14] L. Di Caro, M. Cataldi, and C. Schifanella. The d-index: Discovering dependences among scientific collaborators from their bibliographic data records. *Scientometrics*, 93(3):583–607, 2012.

[15] L. Di Caro and M. Grella. Sentiment analysis via dependency parsing. *Computer Standards & Interfaces*, 35(5):442–453, 2013.

[16] C. Ellwein and B. Noller. Social media mining: Impact of the business model and privacy settings. In Cataldi et al. [7], pages 3–8.

[17] A. Gianti, C. Bosco, V. Patti, A. Bolioli, and L. Di Caro. Annotating irony in a novel italian corpus for sentiment analysis. In *Proceedings of the 4th Workshop on Corpora for Research on Emotion Sentiment and Social Signals*, pages 1–7, 2012.

[18] T. Hofmann. Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 50–57. ACM, 1999.

[19] R. Kikas, M. Dumas, and A. Saabas. Explaining international migration in the skype network: The role of social network features. In Cataldi et al. [7], pages 17–22.

[20] T. Kreutz and M. Nissim. Catching events in the twitter stream: A showcase of student projects. In Caro et al. [4], pages 14–18.

[21] K. S. C. R. P. M. L. S. Luca Vignaroli, Claudio Schifanella. Tracking and analyzing the ”second life” of tv content: a media and social-driven framework. In Caro et al. [4].

[22] P. S. Ludu. Inferring latent attributes of an indian twitter user using celebrities and class influencers. In Cataldi et al. [7], pages 9–15.

[23] C. F. U. K. Massimo Poesio, Ayman Alhelbawy. Exploiting social media to address fundamental human rights. In Caro et al. [4].

[24] L. Robaldo and L. Di Caro. Opinionmining-ml. *Computer Standards & Interfaces*, 35(5):454–469, 2013.

[25] E. D. Rosa and A. Durante. App2check: a machine learning-based system for sentiment analysis of app reviews in italian language. In Caro et al. [4], pages 8–13.

[26] A. Ruggeri and G. Boella. Gibsonian modeling of users in social networks. In Caro et al. [4], pages 25–31.

[27] G. Siragusa. Place as topics: Analysis of spatial and temporal evolution of topics from social networks data. In Caro et al. [4], pages 32–35.