

# Processing Social Media in Real-Time

Damiano Spina<sup>a</sup>, Arkaitz Zubiaga<sup>b</sup>, Amit Sheth<sup>c</sup>, Markus Strohmaier<sup>d,e</sup>

<sup>a</sup>*RMIT University, Melbourne, Australia*

<sup>b</sup>*University of Warwick, Coventry, UK*

<sup>c</sup>*Kno.e.sis Center, Wright State University, Dayton, Ohio, USA*

<sup>d</sup>*RWTH Aachen University, Aachen, Germany*

<sup>e</sup>*GESIS – Leibniz Institute for the Social Sciences, Cologne, Germany*

---

*Keywords:* social media mining, information retrieval, natural language processing, data mining

---

## 1. Introduction

1    Social media provide a wealth of information that reveals insights into  
2    current affairs and ongoing events [11, 1]. A careful exploitation of this  
3    information can be of help to enrich numerous applications with fresh in-  
4    sights. Example applications include news reporting [4, 16], disaster coordi-  
5    nation [9], law enforcement [2], health emergencies such as infectious disease  
6    outbreaks [13], and prediction of election or referendum outcomes [6]. In  
7    situations like these, where access to timely information is key, an ability to  
8    process social media in a real-time fashion becomes an important require-  
9    ment [15, 17]. This presents new challenging issues for the research com-  
10    munity in order to quickly make sense of torrential social streams as they  
11    come out, and to make the most from the fresh knowledge available from  
12    these streams. This special issue brings together contributions that report  
13    on novel techniques and applications that make the most of the information  
14    gathered from social media in a (near) real-time fashion.

15    Streams collected from social media have attracted a wide community of  
16    researchers to study the knowledge that can be garnered from information  
17    shared on social media [12]. Research in this direction has focused on many  
18

---

*Email addresses:* damiano.spina@rmit.edu.au (Damiano Spina),  
a.zubiaga@warwick.ac.uk (Arkaitz Zubiaga), amit@knoesis.org (Amit Sheth),  
markus.strohmaier@humtec.rwth-aachen.de (Markus Strohmaier)

19 different aspects such as search, summarization, trend detection, prediction,  
20 and recommendation, among others. While research that exploits social me-  
21 dia data offline gained popularity in the first decade of social media growth,  
22 processing this data in real-time is now picking up momentum. Processing  
23 social media in real-time involves a number of novel challenges to the area  
24 of social media research, such as processing streams of data online due to  
25 the inability to store and index everything; having to make early decisions  
26 observing only the first bits of a growing trend, where the next part of the  
27 trend is yet to be observed; or collecting time-aware relevance judgments  
28 for information whose relevance drops or whose validity expires over time.  
29 This special issue was edited to further research in this direction, by putting  
30 together state-of-the-art methods for real-time social media mining.

## 31 **2. Papers in the Special Issue**

32 For this special issue, we received 20 articles – excluding desk rejects.  
33 As a result of the review process, 6 papers were selected for inclusion in the  
34 special issue. The collection of articles included in this special issue show  
35 the many directions in which real-time processing is crucial in social media  
36 mining, including detection of evolving communities, aspect-based opinion  
37 mining, automated geo-location of social media posts, detection of cyberat-  
38 tacks, detection of news and events, and prediction, tracking and detection  
39 of diseases.

40 Dakiche et al. [3] present a survey discussing and summarizing work on  
41 the problem of tracking community evolution over time in dynamic social  
42 networks. The paper discusses existing methods organized in four types of  
43 approaches: (1) approaches based on independent successive static detection  
44 and matching; (2) approaches based on dependent successive static detec-  
45 tion; (3) approaches based on simultaneous studies of all stages of commu-  
46 nity evolution; and (4) approaches working directly on temporal networks.  
47 The paper also provides an introduction discussion into basic concepts about  
48 social networks to facilitate reading of the existing methods.

49 Dragoni et al. [5] describe a set of unsupervised strategies for aspect-based  
50 opinion mining together with a monitoring tool supporting users in visual-  
51 izing analyzed data. The proposed system is scalable and able to process a  
52 high volume of opinion-based documents in real-time. The effectiveness of  
53 the platform has been tested on benchmarks provided by the SemEval cam-  
54 paign and have been compared with the results obtained by domain-adapted

55 techniques. The effectiveness of the system is comparable to the supervised  
56 systems that participated in the SemEval 2015 and 2016 challenges.

57 Gonzalez Paule et al. [7] propose a geo-location inference method for  
58 microblog streams based on a ranking approach and majority voting. Given  
59 a tweet, the method retrieves relevant geo-tagged tweets based on content  
60 similarity, and then combines the credibility of the source with the locations  
61 of relevant tweets to infer the location of the given tweet. Experiments in  
62 three datasets show the effectiveness of the proposed approach.

63 Javed et al. [10] look into the early detection of cyberattacks through  
64 Twitter. They particularly look into drive-by download attacks, which con-  
65 sists in obfuscating URLs leading to malicious web pages; this way, a user  
66 clicks on a URL whose final destination is unknown, and puts the user in  
67 a vulnerable situation. As a consequence, an attacker can gain control of  
68 the user’s system by exploiting unpatched system vulnerabilities. The paper  
69 focuses on the analysis of machine learning methods that leverage machine  
70 activity data and tweet metadata, achieving accurate predictions of mali-  
71 cious URLs with an F-measure of 0.833 when applied to an unseen test set.  
72 The ultimate goal of this classifier is to provide a basis from which to kill  
73 the connection to the server before an attack has completed and proactively  
74 blocking and preventing an attack, rather than reacting and repairing at a  
75 later date.

76 Hasan et al. [8] address the efficiency problem of detecting newsworthy  
77 events in real-time from social media streams such as Twitter. The proposed  
78 end-to-end event detection framework TwitterNews+ outperforms the com-  
79 pared state-of-the-art approaches in terms of effectiveness, while maintains  
80 real-time processing capabilities. The reported extensive parameter sensitiv-  
81 ity analysis provides insights into the impact of the different parameters, as  
82 well as into the identification of optimal configurations leading to the best  
83 performance in detecting newsworthy events.

84 Şerban et al. [14] propose SENTINEL, an end-to-end software system  
85 built with open source tools for real-time processing of publicly available  
86 social media data. The system ingests data from multiple sources that is  
87 automatically processed and presented to users who seek updates on situ-  
88 ational awareness or aim to perform nowcasting analyses in real-time (e.g.,  
89 syndromic surveillance and detection of disease outbreaks). Preliminary re-  
90 sults validate the effectiveness of the classifier for isolating health-related  
91 social media messages, as well as feasibility of the prediction of disease out-  
92 breaks.

### 93 3. Conclusion and Future Research Directions

94 This special issue covers a wide range of applications that rely on real-time  
95 processing of social media, including event detection [8, 14, 7], cybersecurity  
96 [10], opinion mining [5] and automatic geo-localization [7]. The diversity  
97 of submissions received shows the need for furthering research in processing  
98 social media streams in a (near) real-time. We anticipate that this line of  
99 research will keep gaining importance, as the use of social media in our daily  
100 activities keeps growing.

101 We envision that future research directions will include more sophisticated  
102 techniques to increase effectiveness while maintaining efficiency [9], as well as  
103 novel applications such as real-time detection of fake news or rumours [18],  
104 among others.

### 105 Acknowledgments

106 The guest editors wish to thank Elsevier and the Editorial Board of IP&M  
107 and especially the Editor in Chief, Jim Jansen, for their support to edit this  
108 special issue. We also wish to thank the authors and reviewers for their  
109 exceptional cooperation during the reviewing and revision process.

110 This special issue has been in part possible thanks to the financial sup-  
111 port of the Australian Research Council (projects nr. LP130100563 and  
112 LP150100252).

### 113 References

- 114 [1] Hila Becker, Dan Iter, Mor Naaman, and Luis Gravano. Identifying  
115 content for planned events across social media sites. In *Proceedings*  
116 *of the Fifth ACM International Conference on Web Search and Data*  
117 *Mining (WSDM'12)*, pages 533–542. ACM, 2012.
- 118 [2] Joshua Brunty and Katherine Helenek. *Social media investigation for*  
119 *law enforcement*. Routledge, 2014.
- 120 [3] Narimene Dakiche, Fatima Benbouzid-Si Tayeb, Yahya Slimani, and  
121 Karima Benatchba. Tracking community evolution in social networks:  
122 A survey. *Information Processing & Management*, 2018. ISSN 0306-  
123 4573. doi: <https://doi.org/10.1016/j.ipm.2018.03.005>.

- 124 [4] Nicholas Diakopoulos, Munmun De Choudhury, and Mor Naaman. Find-  
125 ing and assessing social media information sources in the context of jour-  
126 nalism. In *Proceedings of the SIGCHI Conference on Human Factors in*  
127 *Computing Systems (CHI'12)*, pages 2451–2460. ACM, 2012.
- 128 [5] Mauro Dragoni, Marco Federici, and Andi Rexha. An unsupervised as-  
129 pect extraction strategy for monitoring real-time reviews stream. *In-*  
130 *formation Processing & Management*, 2018. ISSN 0306-4573. doi:  
131 <https://doi.org/10.1016/j.ipm.2018.04.010>.
- 132 [6] Fabio Franch. (Wisdom of the crowds) 2: 2010 uk election prediction  
133 with social media. *Journal of Information Technology & Politics*, 10(1):  
134 57–71, 2013.
- 135 [7] Jorge David Gonzalez Paule, Yashar Moshfeghi, and Yeran Sun. On fine-  
136 grained geo-localisation of tweets and real-time traffic incident detection.  
137 *Information Processing & Management*, 2018. ISSN 0306-4573. doi:  
138 <https://doi.org/10.1016/j.ipm.2018.03.011>.
- 139 [8] Mahmud Hasan, Mehmet A. Orgun, and Rolf Schwitter. Real-time event  
140 detection from the Twitter data stream using the TwitterNews+ frame-  
141 work. *Information Processing & Management*, 2018. ISSN 0306-4573.  
142 doi: <https://doi.org/10.1016/j.ipm.2018.03.001>.
- 143 [9] Muhammad Imran, Carlos Castillo, Fernando Diaz, and Sarah Vieweg.  
144 Processing social media messages in mass emergency: A survey. *ACM*  
145 *Computing Surveys (CSUR)*, 47(4):67, 2015.
- 146 [10] Amir Javed, Pete Burnap, and Omer Rana. Prediction of drive-by down-  
147 load attacks on Twitter. *Information Processing & Management*, 2018.  
148 ISSN 0306-4573. doi: <https://doi.org/10.1016/j.ipm.2018.02.003>.
- 149 [11] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. What  
150 is twitter, a social network or a news media? In *Proceedings of the*  
151 *19th International Conference on World Wide Web (WWW'10)*, pages  
152 591–600. ACM, 2010.
- 153 [12] Stuart E Middleton, Lee Middleton, and Stefano Modafferi. Real-time  
154 crisis mapping of natural disasters using social media. *IEEE Intelligent*  
155 *Systems*, 29(2):9–17, 2014.

- 156 [13] Charles W Schmidt. Trending now: Using social media to predict and  
157 track disease outbreaks. *Environmental Health Perspectives*, 120(1):a30,  
158 2012.
- 159 [14] Ovidiu Șerban, Nicholas Thapen, Brendan Maginnis, Chris Hankin, and  
160 Virginia Foot. Real-time processing of social media with SENTINEL:  
161 A syndromic surveillance system incorporating deep learning for health  
162 classification. *Information Processing & Management*, 2018. ISSN 0306-  
163 4573. doi: <https://doi.org/10.1016/j.ipm.2018.04.011>.
- 164 [15] Amit Sheth, Ashutosh Jadhav, Pavan Kapanipathi, Chen Lu, Hemant  
165 Purohit, Gary Alan Smith, and Wenbo Wang. *Twitris: A system for  
166 collective social intelligence*, pages 2240–2253. Springer New York, 2014.  
167 ISBN 978-1-4614-6170-8. doi: 10.1007/978-1-4614-6170-8\_345.
- 168 [16] Arkaitz Zubiaga, Heng Ji, and Kevin Knight. Curating and contextu-  
169 alizing twitter stories to assist with social newsgathering. In *Proceed-  
170 ings of the 2013 International Conference on Intelligent User Interfaces  
171 (IUI'13)*, pages 213–224. ACM, 2013.
- 172 [17] Arkaitz Zubiaga, Damiano Spina, Raquel Martinez, and Victor Fresno.  
173 Real-time classification of twitter trends. *Journal of the Association for  
174 Information Science and Technology*, 66(3):462–473, 2015.
- 175 [18] Arkaitz Zubiaga, Ahmet Aker, Kalina Bontcheva, Maria Liakata, and  
176 Rob Procter. Detection and resolution of rumours in social media: A  
177 survey. *ACM Computing Surveys (CSUR)*, 51(2):32, 2018.