Processing Social Media in Real-Time

Damiano Spina[a], Arkaitz Zubiaga[b], Amit Sheth[c], Markus Strohmaier[d,e]

[a]*RMIT University, Melbourne, Australia*
[b]*University of Warwick, Coventry, UK*
[c]*Kno.e.sis Center, Wright State University, Dayton, Ohio, USA*
[d]*RWTH Aachen University, Aachen, Germany*
[e]*GESIS – Leibniz Institute for the Social Sciences, Cologne, Germany*

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

1. Introduction

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

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

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)

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

2. Papers in the Special Issue

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

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

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

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

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

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

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

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

3. Conclusion and Future Research Directions

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

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

Acknowledgments

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

This special issue has been in part possible thanks to the financial support of the Australian Research Council (projects nr. LP130100563 and LP150100252).

References

[1] Hila Becker, Dan Iter, Mor Naaman, and Luis Gravano. Identifying content for planned events across social media sites. In *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining (WSDM'12)*, pages 533–542. ACM, 2012.

[2] Joshua Brunty and Katherine Helenek. *Social media investigation for law enforcement*. Routledge, 2014.

[3] Narimene Dakiche, Fatima Benbouzid-Si Tayeb, Yahya Slimani, and Karima Benatchba. Tracking community evolution in social networks: A survey. *Information Processing & Management*, 2018. ISSN 0306-4573. doi: https://doi.org/10.1016/j.ipm.2018.03.005.

[4] Nicholas Diakopoulos, Munmun De Choudhury, and Mor Naaman. Finding and assessing social media information sources in the context of journalism. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'12)*, pages 2451–2460. ACM, 2012.

[5] Mauro Dragoni, Marco Federici, and Andi Rexha. An unsupervised aspect extraction strategy for monitoring real-time reviews stream. *Information Processing & Management*, 2018. ISSN 0306-4573. doi: https://doi.org/10.1016/j.ipm.2018.04.010.

[6] Fabio Franch. (Wisdom of the crowds) 2: 2010 uk election prediction with social media. *Journal of Information Technology & Politics*, 10(1): 57–71, 2013.

[7] Jorge David Gonzalez Paule, Yashar Moshfeghi, and Yeran Sun. On fine-grained geo-localisation of tweets and real-time traffic incident detection. *Information Processing & Management*, 2018. ISSN 0306-4573. doi: https://doi.org/10.1016/j.ipm.2018.03.011.

[8] Mahmud Hasan, Mehmet A. Orgun, and Rolf Schwitter. Real-time event detection from the Twitter data stream using the TwitterNews+ framework. *Information Processing & Management*, 2018. ISSN 0306-4573. doi: https://doi.org/10.1016/j.ipm.2018.03.001.

[9] Muhammad Imran, Carlos Castillo, Fernando Diaz, and Sarah Vieweg. Processing social media messages in mass emergency: A survey. *ACM Computing Surveys (CSUR)*, 47(4):67, 2015.

[10] Amir Javed, Pete Burnap, and Omer Rana. Prediction of drive-by download attacks on Twitter. *Information Processing & Management*, 2018. ISSN 0306-4573. doi: https://doi.org/10.1016/j.ipm.2018.02.003.

[11] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. What is twitter, a social network or a news media? In *Proceedings of the 19th International Conference on World Wide Web (WWW'10)*, pages 591–600. ACM, 2010.

[12] Stuart E Middleton, Lee Middleton, and Stefano Modafferi. Real-time crisis mapping of natural disasters using social media. *IEEE Intelligent Systems*, 29(2):9–17, 2014.

[13] Charles W Schmidt. Trending now: Using social media to predict and track disease outbreaks. *Environmental Health Perspectives*, 120(1):a30, 2012.

[14] Ovidiu Şerban, Nicholas Thapen, Brendan Maginnis, Chris Hankin, and Virginia Foot. Real-time processing of social media with SENTINEL: A syndromic surveillance system incorporating deep learning for health classification. *Information Processing & Management*, 2018. ISSN 0306-4573. doi: https://doi.org/10.1016/j.ipm.2018.04.011.

[15] Amit Sheth, Ashutosh Jadhav, Pavan Kapanipathi, Chen Lu, Hemant Purohit, Gary Alan Smith, and Wenbo Wang. *Twitris: A system for collective social intelligence*, pages 2240–2253. Springer New York, 2014. ISBN 978-1-4614-6170-8. doi: 10.1007/978-1-4614-6170-8_345.

[16] Arkaitz Zubiaga, Heng Ji, and Kevin Knight. Curating and contextualizing twitter stories to assist with social newsgathering. In *Proceedings of the 2013 International Conference on Intelligent User Interfaces (IUI'13)*, pages 213–224. ACM, 2013.

[17] Arkaitz Zubiaga, Damiano Spina, Raquel Martinez, and Victor Fresno. Real-time classification of twitter trends. *Journal of the Association for Information Science and Technology*, 66(3):462–473, 2015.

[18] Arkaitz Zubiaga, Ahmet Aker, Kalina Bontcheva, Maria Liakata, and Rob Procter. Detection and resolution of rumours in social media: A survey. *ACM Computing Surveys (CSUR)*, 51(2):32, 2018.