# Social Media Mining for Journalism

Arkaitz Zubiaga[1], Bahareh Heravi[2], Jisun An[3], Haewoon Kwak[3]
[1] University of Warwick, Coventry, UK
[2] University College Dublin, Dublin, Ireland
[3] Qatar Computing Research Institute, Doha, Qatar

The exponential growth of social media as a central communication practice, and its agility in capturing and announcing breaking news events more rapidly than traditional media, has changed the journalistic landscape: social media has been adopted as a significant source by professional journalists, and conversely, citizens are able to use social media as a form of direct reportage. This brings along new opportunities for newsrooms and journalists by providing new means for newsgathering through access to a wealth of citizen reportage and updates about current affairs, as well as an additional showcase for news dissemination.

In addition to being a big opportunity and having changed the day-to-day practices in the newsrooms, social media has introduced a number of challenges when it comes to news-gathering, verification, production, reporting and dissemination. These include real time monitoring of streams, event detection, noise filtering, contextualisation, source and content verification, fact checking, annotation and archiving. The development of more advanced algorithms and tools for journalists requires not only furthering research in computational techniques, but also engaging more closely with journalists to understand how they work, what problems they're facing when using social media, and how their day-to-day workflows can be improved.

Social media are increasingly becoming the go-to platforms to get the news. A 2018 survey by the Pew Research Center found that as many as 62% Americans use social media to get the news[1]. Likewise, news organisations are now employing full-time social media editors, and major news organisations such as Reuters[2] or the BBC[3] recommend their journalists to make frequent use of social media.

Research looking into social media use in journalism has also increased substantially in recent years. After Kwak et al.'s work highlighting the presence of news in social media (Kwak, Lee, Park, & Moon, 2010), now cited over 5,000 times, an increasing number of works have studied social media as a platform that can be leveraged, inter alia, for researching, gathering and verifying breaking news (Diakopoulos, De Choudhury, & Naaman, 2012; Heravi & Harrower, 2016; Tolmie et al., 2017; Zubiaga, Aker, Bontcheva, Liakata, & Procter, 2018; Konstantinovskiy, Price, Babakar, & Zubiaga, 2018), for broadening the audience by maximising the diffusion of news (Diakopoulos & Zubiaga, 2014; McCollough, Crowell,

[1] http://www.journalism.org/2018/09/10/news-use-across-social-media-platforms-2018/
[2] http://handbook.reuters.com/index.php?title=Reporting_From_the_Internet_And_Using_Social_Media
[3] http://news.bbc.co.uk/1/shared/bsp/hi/pdfs/26_03_15_bbc_news_group_social_media_guidance.pdf

& Napoli, 2017) or for news analytics (Castillo, El-Haddad, Pfeffer, & Stempeck, 2014; Zubiaga, Liakata, Procter, Hoi, & Tolmie, 2016).

This special issue provides a gateway to look into a variety of research questions from both theoretical and practical perspectives. We highlighted four major topics of interest to the special issue when we launched the call for papers. These include: (1) newsgathering from social media, aiming to study algorithmic approaches to facilitating collection and research around newsworthy content with social media as a source (Zubiaga, 2018; Khare, Torres, & Heravi, 2015; Heravi, Morrison, Khare, & Marchand-Maillet, 2014), (2) social media news analytics, where the objective is to analyse news readership from the perspective of social media as well as to perform additional analyses that give insights into how news circulates and is consumed online (Diakopoulos, Naaman, & Kivran-Swaine, 2010), (3) data and computational journalism, which aims to leverage social media data to enable computationally assisted production of journalistic content (Gray, Chambers, & Bounegru, 2012; Heravi & McGinnis, 2015; Heravi, 2018), and (4) ethics and digital citizenship, where ethical aspects of gathering eyewitness content from social media as well as other factors affecting diffusion of social media, such as gatekeeping and censorship, are explored (Frost, 2015).

This special issue puts together seven articles covering these subjects. Given recent trends in the research area, the topics that prevail among these articles include verification of newsworthy content and detection of fake news, as well as event detection. In what follows we provide brief summaries of these seven articles.

Opening up the discussion, the paper *"A Bibliometric Analysis of Event Detection in Social Media"* (Chen, Wang, Tang, & Hao, 2018) explores the research status and development trend of the field of event detection in social media through a bibliometric analysis of academic publications on Event Detection in Social Media research field between 2009 and 2017. The study suggests that the area of event detection in social media has received increasing attention and interest in academia with Computer Science and Engineering as two major research subjects. In terms of geographical contribution to the field, the paper identifies the United States and China to contribute the most to the publications on these topics. It further suggests that affiliations and authors researching this area tend to collaborate more with those within the same country. Finally the paper identifies 14 research themes in this area, as part of which a number of newly emerged themes, such as Pharmacovigilance event detection, are discovered.

The paper titled *"What the fake? Assessing the extent of networked political spamming and bots in the propagation of #fakenews on Twitter"* (Al-Rawi, Groshek, & Zhang, 2018) uses 14 million tweets with #fakenews posted from January 3 to May 7, 2018, investigating what and who are behind in promoting and propagating fake news as a national issue. In particular, the authors examine 1) the most associated users and hashtags mentioned in #fakenews tweets and 2) the most active Twitter accounts in spamming and disseminating the #fakenews tweets. A large portion of the tweets were attacks against CNN and other mainstream media outlets which is partly a result of the success of networked political spamming by conservative groups. Investigating the most active users promoting #fakenews tweets, it turned out the majority of the most active accounts likely came from spamming bots. This study has provided insight into Twitter usersâĂŹ networked spamming accounts that influenced the discussion on fake news on Twitter.

In another paper, titled *"A Corpus of Debunked and Verified User-Generated Videos"* (Papadopoulou, Zampoglou, Papadopoulos, & Kompatsiaris, 2018), the authors built an annotated dataset, called Fake Video Corpus 2018 (FVC-2018), of 380 user-generated videos that contain 200 debunked (fake) and 180 verified (real) videos uploaded in YouTube, Facebook, and Twitter. The dataset also contains 77,258 tweets that shared any of the 380 videos. The authors followed the definition of the fake videos proposed in (Teyssou et al., 2017) and extended the initial Fake Video Corpus dataset compiled in the same study. In addition to the efforts to build the annotated dataset, the authors also provide a detailed analysis of the descriptive statistics of the videos and helped to understand the characteristics of the dataset. The FVC-2018 dataset provides a valuable resource for a challenging benchmark and future studies on video verification.

The study in the paper titled *"Location impact on source and linguistic features for information credibility of social media"* (Aladhadh, Zhang, & Sanderson, 2018) investigates the impact of location on information source and credibility level in social media with tweets of a diverse set of events across multiple countries. In particular, the authors examine 1) the types of sources expected in different events from both in- and outside the country of events and 2) linguistic features among sources of different type, credibility level, and location. The authors found that the distribution of some sources differ between locations significantly and the tweets of the same credibility level have different linguistic features based on their distance from an event and the topic of an event. The results of this study provide insights for improving current credibility models when applied to different domains: most importantly, such models need to be trained on data from the same place of event.

In their paper on *"Event news detection and citizens community structure for disaster management in social networks"*, Toujani and Akaichi present a methodology that combines detection of natural hazards from social media with determination of endangered communities as a result of those natural hazards (Toujani & Akaichi, 2018). They present a methodology which consists of three steps: first, they perform a set of natural language processing methods to detect event triggers, extract named entities mentioned in those texts which helps identify the communities and areas involved in the natural hazards, and a dependency analysis which is intended to mitigate the ambiguity of social media posts; second, the apply fuzzy techniques on the extracted events to cluster related posts; and third, these clusters are leveraged in order to determine communities which are at risk owing to the effects of the natural hazard. The authors show the effectiveness of their methodology with experiments on 26 crisis events as well as a set of synthetic datasets, outperforming other baselines. The paper also shows a tool that enables visualisation of the events and communities identified by the system. The tool is intended to facilitate, among others, journalists' work of sifting through large collections of tweets posted during these natural hazards.

In the following paper, the authors explored the potential of NodeXL as a tool for analysis and visualisation in the context of news diffusion. This is the paper titled *"Social Media Analytics: Analysis and Visualisation of News Diffusion using NodeXL"* (Ahmed & Lugovic, 2018), where the authors first conducted a comprehensive literature review and showed how effective NodeXL is to understand reactions in social media. NodeXL, for example, can discover the most shared URLs, popular hashtags, or influential users from the stream of social media posts, and such features are helpful for newsrooms to cover

social media. As NodeXL is easy to use without any programming language, journalists also can easily include social media content to their stories and potentially attract more online readers by showing how online communities react to certain topics.

Wrapping up the articles in this special issue, Chio takes an exploratory approach to the quantification of journalistic values. In the paper titled *"An Exploratory Approach to the Computational Quantification of Journalistic Values"*, the author matches the textual indices extracted through automated content analysis, with human conceptions of journalistic values which were derived from surveying journalism grad students (Choi, 2018). The results of this paper suggest that the numbers of words and quotes news articles contain have a strong association with the survey respondent assessments of their balance, diversity, importance, and factuality. Additionally the paper suggests that the assessment of journalistic values influences the perception of news credibility. In terms of specific indicators, the paper suggests that linguistic polarisation is an inverse indicator of respondents' perception of balance, diversity, and importance. While linguistic intensity was shown to be useful for gauging respondents' perception of sensationalism, the paper suggests that it is an ineffective indicator of importance and factuality. Furthermore, the number of adverbs and adjectives in news articles appear to be useful for estimating respondents' perceptions of factuality and sensationalism. Finally the study suggests that the greater the numbers of quotes, pair quotes, and exclamation/question marks in a news headline, the lower the respondents' perception of journalistic values in that news article would be.

The papers presented in this special issue illustrate the extensiveness and potentials of social media mining for journalism and news industry, including news and event detection, analytics, verification, and journalistic values associated, and/or affected by the use of social media in this domain.

We, the guest editors, would like to extend our appreciation to the authors who submitted to this special issue, as well as the reviewers who dedicate their time to furthering the research of our contributors. We are looking forward to the continued growth and evolution of this rapidly growing interdisciplinary field of research.

## References

Ahmed, W., & Lugovic, S. (2018). Social media analytics: analysis and visualisation of news diffusion using nodexl. *Online Information Review*.

Aladhadh, S., Zhang, X., & Sanderson, M. (2018). Location impact on source and linguistic features for information credibility of social media. *Online Information Review*.

Al-Rawi, A., Groshek, J., & Zhang, L. (2018). What the fake? assessing the extent of networked political spamming and bots in the propagation of# fakenews on twitter. *Online Information Review*.

Castillo, C., El-Haddad, M., Pfeffer, J., & Stempeck, M. (2014). Characterizing the life cycle of online news stories using social media reactions. In *Proceedings of the 17th acm conference on computer supported cooperative work & social computing* (pp. 211–223).

Chen, X., Wang, S., Tang, Y., & Hao, T. (2018). A bibliometric analysis of event detection in social media. *Online Information Review*.

Choi, S. (2018). An exploratory approach to the computational quantification of journalistic values. *Online Information Review*.

Diakopoulos, N., De Choudhury, M., & Naaman, M. (2012). Finding and assessing social media information sources in the context of journalism. In *Proceedings of the sigchi conference on human factors in computing systems* (pp. 2451–2460).

Diakopoulos, N., Naaman, M., & Kivran-Swaine, F. (2010). Diamonds in the rough: Social media visual analytics for journalistic inquiry. In *Visual analytics science and technology (vast), 2010 ieee symposium on* (pp. 115–122).

Diakopoulos, N., & Zubiaga, A. (2014). Newsworthiness and network gatekeeping on twitter: The role of social deviance. In *Icwsm.*

Frost, C. (2015). *Journalism ethics and regulation.* Routledge.

Gray, J., Chambers, L., & Bounegru, L. (2012). *The data journalism handbook: How journalists can use data to improve the news.* " O'Reilly Media, Inc.".

Heravi, B. (2018). Data journalism in 2017: A summary of results from the global data journalism survey. In G. Chowdhury, J. McLeod, V. Gillet, & P. Willett (Eds.), *Transforming digital worlds* (pp. 107–113). Springer International Publishing.

Heravi, B., & Harrower, N. (2016). Twitter journalism in ireland: Sourcing and trust in the age of social media. *Information, Communication & Society*, *19*(9), 1194–1213.

Heravi, B., & McGinnis, J. (2015). Introducing social semantic journalism. *The Journal of Media Innovations.*

Heravi, B., Morrison, D., Khare, P., & Marchand-Maillet, S. (2014). Where is the news breaking? towards a location-based event detection framework for journalists". In *Multimedia modeling* (pp. 192–204). Springer International Publishing.

Khare, P., Torres, P., & Heravi, B. (2015). What just happened? a framework for social event detection and contextualisation. In *2015 48th hawaii international conference on system sciences* (p. 1565-1574). doi: 10.1109/HICSS.2015.190

Konstantinovskiy, L., Price, O., Babakar, M., & Zubiaga, A. (2018). Towards automated factchecking: Developing an annotation schema and benchmark for consistent automated claim detection. *arXiv preprint arXiv:1809.08193*.

Kwak, H., Lee, C., Park, H., & Moon, S. (2010). What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on world wide web* (pp. 591–600).

McCollough, K., Crowell, J. K., & Napoli, P. M. (2017). Portrait of the online local news audience. *Digital Journalism*, *5*(1), 100–118.

Papadopoulou, O., Zampoglou, M., Papadopoulos, S., & Kompatsiaris, I. (2018). A corpus of debunked and verified user-generated videos. *Online Information Review.*

Teyssou, D., Leung, J.-M., Apostolidis, E., Apostolidis, K., Papadopoulos, S., Zampoglou, M., ... Mezaris, V. (2017). The invid plug-in: web video verification on the browser. In *Proceedings of the first international workshop on multimedia verification* (pp. 23–30).

Tolmie, P., Procter, R., Randall, D. W., Rouncefield, M., Burger, C., Wong Sak Hoi, G., ... Liakata, M. (2017). Supporting the use of user generated content in journalistic practice. In *Proceedings of the 2017 chi conference on human factors in computing systems* (pp. 3632–3644).

Toujani, R., & Akaichi, J. (2018). Event news detection and citizens community structure for disaster management in social networks. *Online Information Review.*

Zubiaga, A. (2018). Mining social media for newsgathering. *arXiv preprint arXiv:1804.03540*.

Zubiaga, A., Aker, A., Bontcheva, K., Liakata, M., & Procter, R. (2018). Detection and resolution of rumours in social media: A survey. *ACM Computing Surveys (CSUR)*, *51*(2), 32.

Zubiaga, A., Liakata, M., Procter, R., Hoi, G. W. S., & Tolmie, P. (2016). Analysing how people orient to and spread rumours in social media by looking at conversational threads. *PloS one*, *11*(3), e0150989.