Third International Workshop on Recent Trends in News Information Retrieval (NewsIR'19)

Dyaa Albakour
Miguel Martinez
Sylvia Tippmann
Signal AI
research@signal-ai.com

Ahmet Aker
University of Duisburg-Essen
aker@is.inf.uni-due.de

Jonathan Stray
Columbia Journalism School
jonathan.stray@columbia.edu

Shiri Dori-Hacohen
AuCoDe
shiri@controversies.info

Alberto Barrón-Cedeño
DIT Università di Bologna
a.barron@unibo.it

ABSTRACT

The journalism industry has undergone a revolution in the past decade, leading to new opportunities as well as challenges. News consumption, production and delivery have all been affected and transformed by technology. Readers require new mechanisms to cope with the vast volume of information in order to be informed about news events. Reporters have begun to use natural language processing (NLP) and (IR) techniques for investigative work. Publishers and aggregators are seeking new business models, and new ways to reach and retain their audience. A shift in business models has led to a gradual shift in styles of journalism in attempts to increase page views; and, far more concerning, to real mis- and dis-information, alongside allegations of "fake news" threatening the journalistic freedom and integrity of legitimate news outlets. Social media platforms drive viewership, creating filter bubbles and an increasingly polarized readership.

News documents have always been a part of research on information access and retrieval methods. Over the last few years, the IR community has increasingly recognized these challenges in journalism and opened a conversation about how we might begin to address them. Evidence of this recognition is the participation in the two previous editions of our NewsIR workshop, held in ECIR 2016 and 2018. One of the most important outcomes of those workshops is an increasing awareness in the community about the changing nature of journalism and the IR challenges it entails. To move yet another step forward, the goal of the third edition of our workshop is to create a multidisciplinary venue that brings together news experts from both technology and journalism. This would take NewsIR from a European forum targeting mainly IR researchers, into a more inclusive and influential international forum. We hope that this new format will foster further understanding for both news professionals and IR researchers, as well as producing better outcomes for news consumers. We will address the possibilities and challenges that technology offers to the journalists, the challenges that new developments in journalism create for IR researchers, and the complexity of information access tasks for news readers.

SIGIR '19, July 21–25, 2019, Paris, France

ACM ISBN 978-1-4503-6172-9/19/07.
https://doi.org/10.1145/3331184.3331646

ACM Reference Format:
Dyaa Albakour, Miguel Martinez, Sylvia Tippmann, Ahmet Aker, Jonathan Stray, Shiri Dori-Hacohen, and Alberto Barrón-Cedeño. 2019. Third International Workshop on Recent Trends in News Information Retrieval (NewsIR'19). In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '19), July 21–25, 2019, Paris, France.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3331184.3331646

1 BACKGROUND AND MOTIVATION

Text as a medium is a common ground for the domains of news and information retrieval (IR). While journalists make use of text for information gathering and create text as one medium of their output, IR aims to extract relevant resources to satisfy an information need. While traditional news gathering is largely manual, the field of IR seeks automation to process large amounts of text – and this automation is increasingly useful in journalism.

With the increasing popularity of personal computers, print publications went online, thereby removing the space limitation of the traditional newspaper. The proliferation of mobile devices accelerated this trend and, in 2008, more Americans reported consuming their news online rather than in print format [7]. Traditionally, news have been exclusively the remit of professional journalists, ensuring editorial standards, including fact checking and a coherence in writing style. The rise of online self-publication in the last decade(s), however, is challenging this definition of journalism. Blogging now allows anyone to write articles and publish them on the internet, to contribute and even steer public discourse. As a result there is more content, and more diverse content, which has been a boon to journalism in some respects. Unfortunately, this democratization of journalism has also led to more noise, along with more misinformation and disinformation[1].

Therefore, when consumers go online to satisfy a news information need, they will find a vast amount of news-like information, sometimes featuring alternative points of view and so-called 'alternative facts'. Consequently, trust and the credibility of news sources

[1] Misinformation is accidentally inaccurate content, whereas disinformation is false content written with a deliberate intent to mislead. The line between the two can be blurry and often difficult to judge.

are increasingly important metrics, and one of the intersections of IR and journalism.

The way we prioritize news has changed as well. Traditionally, making a news article a 'top story' was an editorial decision, based on perceived impact on the outlet target audience or wider society. With online journalism as a new style of publication, the 'front page' is not necessarily the most impactful news of the day anymore. Many news sources now rely on online advertising as a business model. The most read and shared article becomes the biggest story. Social media as a distribution channel has magnified this effect, thereby challenging editorial judgment with popularity. Indeed, two years ago, social media overtook print media as a source of news among US adults [6]. This hunt for clicks can in turn affect news text, writing style and click-bait headlines. Ranking and recommendation systems, which respond to these clicks, play a crucial role in news creation, consumption and propagation.

News recommendation techniques can help consumers identify relevant, more personalized, content —to find the needle in the haystack. However, it can also create a feedback loop that is echoed and enhanced by the readers own social network and isolate us into disjoint 'news filter bubbles'. Bots amplifying certain articles on social media can also shape the perception of which articles are worthy of our time, and indeed recent studies have estimated that 9-15% of Twitter profiles are bots [9] and that 66% of links shared on Twitter are shared by bots [3]. A more informed dialog between IR specialists and professional journalists may help to tune the techniques available for a more purposeful information gain and exchange that empowers real humans in society, ideally one that fosters nuanced understanding rather than exacerbating falsehoods and division.

With the same goal in mind, investigative journalists have turned towards NLP and IR techniques to support their reporting. The sheer volume of contemporary data – whether from leaks or open government sources – requires computational techniques, and journalists increasingly employ search, text summarization, automation in fact-checking, entity recognition and topic classification to support information gathering.

The aforementioned opportunities have attracted much research recently in the IR community and many other related communities, such as NLP and Machine Learning (ML). Indeed, IR experts have long relied on news corpora to study and learn from, since it reflects generalist information, e.g. [4]. News is multilingual, mostly free of domain specific jargon, and covers a wide range of topics and entities. News text has been used in IR and NLP to improve and evaluate models for a variety of tasks, such as text classification [4], entity recognition [8], and machine translation [2]. Now more than ever, a large volume of news text resides in the public domain or is otherwise abundantly available for research.

Moreover, previous NewsIR workshops run at ECIR 2016 and 2018 have demonstrated the continuing interest in news retrieval challenges and the applications to journalism. Both workshops were considered a great success according to various indicators: the large number of submissions, the quality of the papers, the diversity of topics, the large number of attendees and the high level of interactions [1, 5]. Taking this forward, we propose to run yet another NewsIR workshop. We detail the goals of our proposed workshop in the next section.

2 WORKSHOP PURPOSE

Now in its third version, we aim to organize NewsIR in an even more multidisciplinary format and taking the conversation from a European forum to the global IR community.

Building on the success of its previous versions [1, 5], the main goal of this workshop is to bring together news experts from both the technology side (mainly experts in IR, NLP and ML), and the journalism side. The objective is to stimulate discussion around the current challenges in the journalism and news processing environment and to combine the expertise of these communities. Technology experts and, in particular, the IR, NLP and ML communities, will present the latest breakthroughs in news processing and related subjects of trust and misinformation, with an emphasis on the application of their findings. The journalistic community will complement this by offering a better understanding of the needs and challenges they are facing, and suggestions for future research directions. We aim to have a substantial representation from industry, from small start-ups to large enterprises, to strengthen the relationships. This also represents a unique opportunity to understand the different problems and priorities of each community and to recognize areas that are not currently receiving much academic attention but are nonetheless of considerable commercial interest.

3 WORKSHOP FORMAT AND STRUCTURE

To help us achieve the goals set out in Section 2, we aim for an interactive and dynamic *full-day* workshop that will be a mixture of keynote, paper and demo presentations as well as discussion breakout sessions. We propose to start with an introductory session on the overall workshop topics led by the workshop organizers. For invited talks, we have approached a number of keynote speakers who can provide insights into the topic from both a journalistic industry and an information retrieval point of view. Finally, the workshop also includes a panel discussion with members drawn from academia, publishers, large companies and SMEs. For papers, we will solicit submissions of technical papers (4 pages), demos and position papers (2 pages) in ACM double-column style to cover the challenges discussed in Section 1. In particular, the topics of interest to the workshop may include (but is not limited to):

- Credibility, controversy and fact-checking
- Bias and plurality in news
- Summarization of opinions in news
- Information silos and the effect of algorithms on news consumption
- Event and anomaly detection
- IR/NLP in Data Journalism
- News-related user-generated content (e.g. using comments to enhance news retrieval or to detect controversy)
- Media monitoring and Reputation Management
- News ranking and online learning-to-rank
- News recommendation and personalization
- The "filter bubble" and de-personalization
- Entity recognition and entity linking in news
- De-duplication and clustering of news articles
- Author identification and disambiguation
- Evaluation of news retrieval systems
- Conversational journalism and chat bots

- Mobile-first journalism
- IR in investigative reporting
- Data visualization and story-telling
- Traditional and social media integration
- Multiple document and temporal summarization

Manuscripts will be reviewed by at least three members of the PC, with final acceptance decisions made by the organizers. Participants of accepted papers will be given 15 minutes for short presentations. All papers will also be presented in an interactive poster session to encourage more discussion and engagement. We plan to publish the proceedings with CEUR workshop proceedings.[2] This guarantees timely publication (specifically, in time for the workshop) and a wide dissemination due to CEUR's open-access policy.

It is crucial for the success of the workshop to attract submissions on relevant topics outlined in Section 1 from multiple communities. The key for that is the promotional activities that we have conducted to reach out to the relevant researchers and stakeholders in those communities. We believe that the diverse setup of the organization and the paper reviewers described in the next section will help us immensely in the reach out and to attract a large number of paper submission as in the previous versions of the workshop [1, 5]. Indeed, as a result of these promotional activities, we have attracted a total of 21 paper submissions.

4 KEYNOTES AND PANEL

We aim to have at least two keynote speakers to share their experiences from the journalistic space and the application of IR technologies for news processing. Aron Pilhofer, Chair in Journalism Innovation at Temple University, has kindly confirmed to give a keynote at the workshop. In his keynote, he will focus on the application of technology to journalism in general and to local journalism in particular: "What can we do for the 99 percent of journalists, who are overworked, under-resourced at the local level"

In addition, we will also have a panel with representation from the journalism world, as well as the academic and commercial perspectives. For the panel we aim to have the following experts:

- Jochen Leidner, Director of Research at Refinitiv Ltd., formerly the Financial & Risk division of Thomson Reuters). He is also a Royal Academy of Engineering Visiting Professor of Data Analytics at the University of Sheffield, Guest Lecturer at the University of Zurich and Scientific Expert for the European Commission.
- Julio Gonzalo, a professor at UNED, Mardid. Julio has recently been co-organizer of the RepLab Evaluation Campaign for Online Reputation Management System.
- Tom Philips, Editor of automated factchecking at UK's leading fact-checking charity Full Fact.

5 ORGANIZERS AND PROGRAMME COMMITTEE

After running NewsIR for the first time in 2016, some co-organisers left space for new ones to join and lead the second workshop in 2018. This is inline with the spirit of the workshop; reaching out to new communities and increasing the diversity. For our proposed workshop, we continue this trend where five new co-organisers have joined us this time. We are excited to have a multi-disciplinary and a truly diverse committee in all different aspects. In particular, between us, we have expertise in multiple areas of research (IR, NLP, machine learning and journalism). We also represent both academia and industry, working in orgainsations spread around different parts of the world.

For the programme committee, we have approached many researchers from both academia and industry. All of whom are either experts in the topics of workshop or have been involved in news information retrieval in the past. The PC includes the following members:

- Marco Bonzanini, Bonzanini Consulting Ltd, UK
- Alejandro Bellogin Kouki, UAM, Spain
- Ricardo Campos, Polytechnic Institute of Tomar, Portugal
- Jon Chamberlain, University of Essex, UK
- Dario Garigliotti, University of Stavanger, Norway
- Anastasia Giachanou, Valencia, Spain
- Frank Hopfgartner, The University of Sheffield, UK
- Julia Kiseleva, University of Amsterdam
- Udo Kruschwitz, University of Essex, UK
- Jochen Leidner, Refinitiv Labs & University of Sheffield, UK
- Haiming Liu, University of Bedfordshire, UK
- Edgar Meij, Bloomberg, UK
- Elaheh Momeni, University of Vienna, Austria
- Raymond Ng, Signal AI, UK
- Arian Pasquali, University of Porto, Portugal
- Filipa Peleja, Vodafone Research, Portugal
- Barbara Poblete, University of Chile, Chile
- Damiano Spina, RMIT University, Australia
- Andreas Spitz, Heidelberg University
- Arkaitz Zubiaga, University of Warwick, UK

[2] http://ceur-ws.org

REFERENCES

[1] Dyaa Albakour, David Corney, Julio Gonzalo, Miguel Martinez1, Barbara Poblete, and Andreas Vlachos. 2018. Report on the 2Nd International Workshop on Recent Trends in News Information Retrieval (NewsIR'18). *SIGIR Forum* 52, 1 (Aug. 2018), 140–146. https://doi.org/10.1145/3274784.3274799

[2] Bill Dolan, Chris Quirk, and Chris Brockett. 2004. Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources. In *Proceedings of the 20th international conference on Computational Linguistics*. Association for Computational Linguistics, 350.

[3] Knight FoundationLLC. 2018. Disinformation, 'Fake News' and Influence Campaigns on Twitter. https://www.knightfoundation.org/reports/disinformation-fake-news-and-influence-campaigns-on-twitter

[4] David D Lewis, Yiming Yang, Tony G Rose, and Fan Li. 2004. RCV1: A new benchmark collection for text categorization research. *Journal of machine learning research* 5, Apr (2004), 361–397.

[5] Miguel Martinez, Udo Kruschwitz, Gabriella Kazai, Frank Hopfgartner, David Corney, Ricardo Campos, and Dyaa Albakour. 2016. Report on the 1st international workshop on recent trends in news information retrieval (NewsIR16). In *ACM SIGIR Forum*, Vol. 50. ACM, 58–67.

[6] Elisa Shearer. 2018. Social media outpaces print newspapers. http://www.pewresearch.org/fact-tank/2018/12/10/social-media-outpaces-print-newspapers-in-the-u-s-as-a-news-source/

[7] Aron Smith. 2009. The Internet as a Source of Political News and Information. http://www.pewinternet.org/2009/04/15/the-internet-as-a-source-of-political-news-and-information/

[8] Erik F Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*. Association for Computational Linguistics, 142–147.

[9] Onur Varol, Emilio Ferrara, Clayton A Davis, Filippo Menczer, and Alessandro Flammini. 2017. Online human-bot interactions: Detection, estimation, and characterization. *arXiv preprint arXiv:1703.03107* (2017).