WSDM 2017 Workshop on Mining Online Health Reports

WSDM Workshop Summary

Nigel Collier[♠], Nut Limsopatham[♠], Aron Culotta[♥],
Mike Conway[◊], Ingemar J. Cox[♦,♣] and Vasileios Lampos[♦]
{nhc30,nl347}@cam.ac.uk, culotta@cs.iit.edu, mike.conway@utah.edu, {i.cox,v.lampos}@ucl.ac.uk

[♠] Department of Theoretical and Applied Linguistics, University of Cambridge, UK
[♥] Department of Computer Science, Illinois Institute of Technology, US
[◊] Department of Biomedical Informatics, University of Utah, US
[♦] Department of Computer Science, University College London, UK
[♣] Department of Computer Science, University of Copenhagen, Denmark

ABSTRACT

The workshop on Mining Online Health Reports (MOHRS) draws upon the rapidly developing field of Computational Health, focusing on textual content that has been generated through various activities on the Web. Online user-generated information mining, especially from social media platforms and search engines, has been in the forefront of many research efforts, especially in the fields of Information Retrieval and Natural Language Processing. The incorporation of such data and techniques in a number of health-oriented applications has provided strong evidence of the potential benefits, which include better population coverage, timeliness and applicability to places with less established health infrastructure. The workshop provides an opportunity to present relevant state-of-the-art research, and a venue for discussion between researchers with cross-disciplinary backgrounds. It will focus on the characterisation of data sources, the essential methods for mining this textual information, as well as potential real-world applications and the arising ethical issues. MOHRS '17 will feature 3 keynote talks and 4 accepted paper presentations, as well as a panel discussion.

Keywords

Natural Language Processing; Machine Learning; Computational Health; User-Generated Content

1. INTRODUCTION

The workshop on Mining Online Health Reports (MOHRS)[1] is part of the programme of the Web Search and Data Mining (WSDM) conference that will be held in February 6-10, 2017 at Cambridge, UK. The workshop's main focus lies in the area of Computational Health, with a particular interest in applications that are based on online text processing. We aim to bring together a cross-disciplinary audience of researchers from academia, industry and the health sector to share experience of techniques, resources and best practices, and to exchange perspectives in order to prioritise future research directions.

Online health information is widely published by individuals in social media, chat rooms and discussion boards. At the same time search query logs and various forms of text messaging contain a vast amount of textual information that can be directly or indirectly linked to health conditions. This informal evidence about our individual health, attitudes and behaviours has the potential to be a valuable source for health applications ranging from real-time disease monitoring [2–4], to offline assessment of health interventions [5], to understanding opinions [8], to situation awareness during disasters [11] and detecting adverse drug events [9].

Informal patient data on the Web is increasing, accessible, low cost, real-time and seems likely to cover a significant proportion of the population. Coupled with wearable body sensor data and the wealth of structured hospital data, it has the potential to offer insights leading to new lines of clinical investigation. However, in order to understand and integrate this data, researchers in academia and industry must grapple with theoretical, practical and ethical challenges that require immediate attention.

For example, how can we achieve fine-grained analysis and understanding of health-oriented language? How can we better engage with health experts to assure relevance? How can we assess the impact of online health data in real-world health applications? How can we integrate online health data with other data sources such as databases and ontologies? and, What are the legal issues and privacy trade-offs around the use of online health data?

We expect the workshop to develop a community of interested researchers, build future collaborations and develop understanding of best practice, e.g. with respect to ethical standards.

2. OPEN QUESTIONS

The focus of the workshop has been framed around four key open questions that we describe below.

[1] MOHRS '17, http://sites.google.com/site/mohrs2017

WSDM 2017 February 06-10, 2017, Cambridge, United Kingdom

ACM ISBN 978-1-4503-4675-7/17/02.
DOI: http://dx.doi.org/10.1145/3018661.3022761

Identification and characterisation of sources. Within the context of online health report mining, it is expected that major parts of the specified research aims and the subsequent analysis are driven by user-generated data (UGC). Unfortunately, only a few works are based on a common data set, something that is often caused by the proprietary nature of the input sources. Furthermore, different data sources and sub-samples of them encapsulate variable biases (e.g. demographic or behavioural) and hence, come with different limitations. Therefore, it will be of great benefit to (1) identify a solid base of potentially shareable sources of online UGC, (2) characterise and quantify the main attributes of each source, and (3) where necessary, incorporate the quantified biases into the computational analysis.

Data mining and integration. The transformation of raw textual data into useful information is a core element of most research paradigms that fit into the specification of this workshop. We are interested in applications that use state-of-the-art developments from the fields of Natural Language Processing, Machine Learning or Information Retrieval to achieve better interpretations of textual contents, aiming for better abstraction and generalised solutions. The integration of these online extracts into more traditional forms of health data is also something that has not received a lot of attention. We would like to identify the main challenges, discuss potential limitations of data fusion, and propose preliminary experiments, where hypothesis testing will be possible.

Real-world case studies. The translation of ongoing research into real-world, practical applications that improve health and well-being is one of our ultimate goals. This usually requires a close collaboration with experts from the target health domain. Without this inter-disciplinary research setting, we cannot ascertain that the problem definition is correct, the obtained solution is practically useful, or be in the position to deploy it. The workshop is interested in examples where online data driven applications have been used in a real-world setting as well as potential failures in this process. Characteristic examples include HealthMap,[2] a tool that collects and visualises various forms of health reports at a global scale, Flu Detector,[3] a tool that quantifies influenza-like illness rates using Twitter or Google search data, or the now terminated platform of Google Flu Trends.[4]

Ethical and legal issues. In addition the workshop intends to consider the evolving issues around the ethical standards of processing social media data to discover biomedical and behavioural insights. Moreover we will consider the professional practice of research and decision making that takes place based on social media data.

3. FORMAT AND PROGRAMME

The workshop will feature three keynote speakers. Elad Yom-Tov (Microsoft Research) will present a selection of computational health applications that take advantage of social media and web search data; Munmun De Choudhury (Georgia Tech) will provide insights on the identification and characterisation of data sources relevant to themes of the workshop; Daniel O'Connor (Wellcome Trust) will elaborate on the various ethical issues around this line of research.

From the 8 submissions that were received, we have accepted 4 papers (2 long and 2 short papers). Among the accepted papers, Liu et al. focus on the task of suicide prediction from social media activity aiming to create a more accurate labelled data set of suicide-related messages [6], and Norval and Henderson provide insights on the ethical issues arising when inferences based on social media content are focusing at the user-level [10]. Chandrashekhar et al. look at identifying a cohort of pregnant women and correlating medications by trimester [1]. Mowery et al. classify Twitter messages for evidence of depression using lexical and emotional features [7]. All papers will be presented. A full report of the workshop, including a summary of a panel discussion, will appear in June 2017 issue of the SIGIR Forum.

[2] HealthMap, `http://www.healthmap.org`

[3] Flu Detector, `http://fludetector.cs.ucl.ac.uk`

[4] Google Flu Trends, `https://www.google.org/flutrends`

References

[1] P. Chandrashekhar, A. Magge, A. Sarker, and G. Gonzalez. Social Media Mining for Cohort Identification and Exploration of Health-Related Data. In *Proc. of the 1st Workshop on Mining Online Health Reports (MOHRS) at WSDM '17*, 2017.

[2] N. Collier, S. Doan, A. Kawazoe, et al. BioCaster: detecting public health rumors with a Web-based text mining system. *Bioinformatics*, 24(24):2940–2941, 2008.

[3] A. Culotta. Towards Detecting Influenza Epidemics by Analyzing Twitter Messages. In *Proc. of the First Workshop on Social Media Analytics*, pages 115–122, 2010.

[4] J. Ginsberg, M. H. Mohebbi, R. S. Patel, et al. Detecting influenza epidemics using search engine query data. *Nature*, 457(7232):1012–1014, 2009.

[5] V. Lampos, E. Yom-Tov, R. Pebody, and I. J. Cox. Assessing the impact of a health intervention via user-generated Internet content. *Data Mining and Knowledge Discovery*, 29(5):1434–1457, 2015.

[6] T. Liu, Q. Cheng, C. Homan, and V. Silenzio. Learning from Various Labeling Strategies for Suicide-Related Messages on Social Media: An Experimental Study. In *Proc. of the 1st Workshop on Mining Online Health Reports (MOHRS) at WSDM '17*, 2017.

[7] D. Mowery, C. Bryan, and M. Conway. Feature Studies to Inform the Classification of Depressive Symptoms from Twitter Data for Population Health. In *Proc. of the 1st Workshop on Mining Online Health Reports (MOHRS) at WSDM '17*, 2017.

[8] M. Myslín, S.-H. Zhu, W. Chapman, and M. Conway. Using twitter to examine smoking behavior and perceptions of emerging tobacco products. *Journal of medical Internet research*, 15(8):e174, 2013.

[9] A. Nikfarjam, A. Sarker, K. O'Connor, R. Ginn, and G. Gonzalez. Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features. *Journal of the American Medical Informatics Association*, page ocu041, 2015.

[10] C. Norval and T. Henderson. Contextual Consent: Ethical Mining of Social Media for Health Research. In *Proc. of the 1st Workshop on Mining Online Health Reports (MOHRS) at WSDM '17*, 2017.

[11] J. Rogstadius, M. Vukovic, C. Teixeira, V. Kostakos, E. Karapanos, and J. A. Laredo. Crisistracker: Crowdsourced social media curation for disaster awareness. *IBM Journal of Research and Development*, 57(5):4–1, 2013.