BroDyn'18: Workshop on Analysis of Broad Dynamic Topics over Social Media*

Tamer Elsayed[1], Walid Magdy[2], Mucahid Kutlu[1], Maram Hasanain[1], and Reem Suwaileh[1]

[1] Computer Science and Engineering Department, Qatar University, Doha, Qatar
[2] School of Informatics, University of Edinburgh, UK
telsayed@qu.edu.qa, wmagdy@inf.ed.ac.uk, mucahidkutlu@qu.edu.qa, maram.hasanain@qu.edu.qa, reem.suwaileh@qu.edu.qa

Abstract. Social media streams are flooded with posts related to topics that are *broad* (i.e., cover several sub-topics) and *dynamic* (i.e., develop over time) which attract long-standing user interests. Posts about topics like "Brexit" or "UK elections" are hard to miss any day while these topics are "hot", yet research on identifying and analyzing posts on such type of topics is still in its infancy. The BroDyn workshop aims at building a community interested in developing and exchanging ideas and methods for analyzing social media for broad dynamic topics. It also aims at understanding the limitations of existing techniques in answering emerging information needs for such topics, and proposing new techniques, evaluation methods, and test collections to address these limitations. The workshop is designed to bring together audience at all levels, including researchers from academia and industry as well as potential users, to create a forum for discussing recent advances in this area. The workshop, in its first version, featured three papers that span different aspects of the target research area.

1 Introduction

For long time, information retrieval (IR) research has mostly focused on information needs that are short-term and narrow in scope. That was manifested in a large body of work on the traditional ad-hoc search task [3], which naturally belongs to that type. Although tasks such as information filtering and topic tracking (which are both long-term by definition) have earlier attracted IR researchers [1, 16], mostly on news streams, the limited volume and low frequency of documents have limited the scope of information needs.

Since the emergence of social media platforms, new scenarios of IR were required due to the shift in the information needs of users on those platforms [13, 19]. That indeed motivated the proposal of new tasks in the IR research community represented in the TREC Microblog track [15] and its subsequent versions [8, 9]. Several attempts for modeling practical IR tasks were introduced through the

* https://sites.google.com/view/brodyn2018/

tracks between 2011 and 2016, e.g., tweet filtering [18], tweet timeline generation [7], and real-time summarization [9].

Over time, users developed a widespread perception of social media as a news source that they follow (almost all day long) to get updated on their topics of interest. Many of those topics are of *long-term* interest (i.e., span or stay active for long period of time), very *broad* (i.e., cover many aspects or sub-topics), and *dynamic* (i.e., develop and change focus over time). For example, following tweets related to a topic such as "UK Elections" requires tracking posts about several sub-topics (such as candidates, campaigns, political views, election process, etc.) for a long period of time [6]. Moreover, the sub-topics change over time, e.g.,"debates" is an important one before the elections, while "election results" is the most important one during voting; and a sub-topic sometimes spans a very short period of time, e.g., press statements by candidates and leaked content about them. Such kind of broad and dynamic topics can be running for few weeks, such as crisis events (e.g., "Irma hurricane") [14], or up to years, such as "the Syrian conflict". Other examples of broad dynamic topics include"Refugees in Europe", "GCC crisis", "Brexit", "North Korea and US conflict" to name a few. The diversity of such topics and the importance of meeting the needs of following and analyzing them warrant a matching focus from the IR research community [10, 11].

Analyzing broad dynamic topics over social media has several research challenges. Systems that analyze social media for that type of topics require *adaptive* techniques to effectively capture the different and changing aspects of the topics. They also have to be *scalable* to cope with the large volume of posts and the diversity of the sub-topics, *real-time* to be responsive to the high velocity of the data, and *reliable* to perform effectively over long periods of time. The process might indeed cover several steps including retrieval and filtering, topic/sub-topic detection, topic modeling, and summarization among many. Advanced spam and bot detection techniques might also be required to tackle the changes in content and techniques of spammers. Furthermore, a new evaluation framework for such domain is also needed with novel evaluation measures that capture the nature of topics (and thus systems), and new large reusable datasets that enable the researchers to run meaningful and representative experiments.

To advance the research work in that area, we organized the **BroDyn** workshop. BroDyn aims to engage with the IR community interested in different technologies applied to social media such as filtering, summarization, spam detection, *but focusing on broad and dynamic topics*. Moreover, the theme of the workshop concerns, besides IR researchers, a wide spectrum of potential users of the needed technology, such as journalists, historians, politicians, social scientists and analysts [17]. This shows the potential impact of the research needed in this area. While there have been many workshops on social media, to the best of our knowledge, this workshop is the first that focuses on broad and dynamic topics in that domain.

2 Objectives

The BroDyn workshop aims to achieve the following objectives:

1. Directing the attention of IR researchers to the new emerging type of information needs that require following broad and dynamic topics and events over social media streams.
2. Better understanding the limitations of our current methods and inspiring research on new algorithms and techniques.
3. Encouraging the design of new evaluation measures, datasets, and test collections to support the research in that area.
4. Forming a new community that brings together IR researchers as well as potential users who are interested in the domain.

3 Topics of Interest

To reflect the large scope of work, we encouraged submissions that span the spectrum from retrieval and filtering to recommendation and summarization. Our workshop solicited contributions on all topics related to the theme, focused (but not limited to) on the following tasks:

- Adaptive high-recall high-precision filtering / topic tracking
- Adaptive summarization
- Topic/sub-topic or event/sub-event detection over time
- Retrospective generation of timelines
- Following controversial political events/crises: identification of decision makers, credibility/information source finding, stance/opinion mining, troll detection, fact checking
- Cross-media filtering (i.e., over heterogeneous sources)
- Multilingual topic/event detection
- Online/dynamic topic modeling
- Real-time/scalable techniques of processing high-volume streams
- Evaluation techniques and novel test collections (specific for broad dynamic topics)
- Spam and hashtag-spam detection
- Bot and automatically-generated content detection
- Data visualization
- Recommendation (e.g., of hashtags/topics/sub-topics)
- Learning techniques/deep learning over social streams

Example Datasets

We encouraged (but not required) submissions describing experiments using two datasets, GE2017 [5] and USPresElect2016 [6], as examples of datasets on broad dynamic topics. Researchers were free to define their own relevant tasks using the datasets if they elect to use any of them.

GE2017 is a dataset of around 18M tweets collected between April 28th and June 8th 2017 on the British General Elections 2017. A set of 56 keywords related to GE2017 was used to collect tweets on the topic. The Twitter streaming API was used to retrieve tweets containing any of these keywords over the period of study. The keywords consist of hashtags, accounts, and terms representing phrases on the elections (e.g. #GE2017, general elections), politicians involved in the elections (Theresa May, Corbyn, #jc4pm), and related topics (e.g., Brexit, NHS). Due to the restrictions of tweets redistribution, we only shared the tweet IDs of the dataset.

USPresElect2016 is a dataset of 3,450 labelled tweets representing the top 50 most retweeted tweets on the US presidential elections 2016 for every day during the period from 1 Sep 2016 to 8 Nov 2016 (the election day). The total number of retweets for these 3,450 tweets are over 26M times. Each tweet is labeled as: support/attack Trump/Clinton, or both, or neither (neutral).

4 Overview of Accepted Papers

We have three research papers accepted in our first workshop on broad and dynamic topics, two full and one short, covering different aspects of the main theme of the workshop.

Badache et al. [2] developed a system for detecting the intensity of contradicting views for a particular topic. They also introduced a measure for contradiction intensity and a dataset built by using reviews and courses in Coursera. Detecting intensity of contradicting reviews is particularly important for the analysis of broad and dynamic topics because it is very likely to have contradicting opinions in broad topics (e.g., supporting and opposing views about Brexit) and detecting the intensity of contradicting views help better understand the public opinion about a particular topic.

Mazoyer et al. [12] proposed two approaches to collect tweets discussing news in the French media. The first approach iteratively modifies the query sent to Twitter API to form a vocabulary-constrained collection. The other approach collects random tweets and dynamically clusters them in events. The approaches were designed such that the collected tweet datasets are representative of the true tweets activity on Twitter. Their approaches can be useful to analyze the activity in social media about news events in real-time.

Bulbul et al. [4] presented an approach to collect Twitter accounts of refugees residing in Turkey. The dataset covers tweets since Syrian Crisis started, allowing us to analyze how the opinions and feelings of Syrian refugees changed over time. Furthermore, the paper described an initial analysis of the topics covered by the tweets posted through these accounts. Acquiring such dataset can help in conducting social studies on the crisis of refugees or even acquire actionable knowledge to better understand and fulfill their needs.

Overall, the accepted papers provide novel methods to have better insights about broad and dynamic topics and construct datasets for further analysis.

Being the first workshop of its kind, we believe that the accepted papers will pave the way for further development in this emerging research area.

5 Program

BroDyn was held on March 26, 2018 in Grenoble, France in conjunction with the 40th European Conference on Information Retrieval (ECIR'18). The program started with the keynote speech given by Michalis Vazirgiannis featuring new techniques for event detection over social media. The speech was followed by three presentations of the accepted papers. We allowed 30 minutes for presenting full papers and 20 minutes for the short paper. After each presentation, we moderated a discussion for 10 minutes. Once all papers are presented, we also moderated an open discussion session in which we discussed the current challenges and future direction for research on broad and dynamic topics. The detailed program of the workshop is given below.

2:30 - 2:35	**Opening**
2:35 - 3:35	**(*Keynote Speech*) Graph-Based Event Detection in Streams: The Twitter Case** *Michalis Vazirgiannis*
3:35 - 4:00	**Social Media based Analysis of Refugees in Turkey** *Abdullah Bulbul, Salah Haj Ismail, Çağr Kaplan*
4:00 - 4:30	**Coffee Break**
4:30 - 5:10	**Contradiction in Reviews: is it Strong or Low?** *Ismail Badache, Sebastien Fournier, Adrian-Gabriel Chifu*
5:10 - 5:50	**Real-time Collection of Reliable and Representative Tweets Datasets Related to News Events** *Béatrice Mazoyer, Céline Hudelot, Marie-Luce Viaud, Julia Cagé*
5:50 - 6:20	**Open Discussion**

6 Reviewing Process and Program Committee

All submitted papers were peer-reviewed through a double-blind reviewing process by at least three program committee members[3]. We would like to deeply thank all members of the committee for their great work. The committee consists of the following members:

- Dyaa Albakour, Signal Media
- Mossaab Bagdouri, Walmart Labs
- Mohand Boughanem, IRIT University Paul Sabatier Toulouse
- Fabio Crestani, University of Lugano (USI)

[3] a meta review by the workshop chair was also added in some cases.

– Kareem Darwish, Qatar Computing Research Institute
– Hui Fang, University of Delaware
– Saptarshi Ghosh, Indian Institute of Technology Kharagpur
– Maram Hasanain, Qatar University
– Gareth Jones, Dublin City University
– Andreas Kaltenbrunner, NTENT
– Preslav Nakov, Qatar Computing Research Institute
– Lynda Tamine, University of Toulouse
– Ingmar Weber, Qatar Computing Research Institute
– Peilin Yang, University of Delaware

7 Organizing Committee

The workshop organizing committee has two chairs who are responsible for the managing the reviewing process, planning the program, and developing proceedings: **Tamer Elsayed** (assistant professor of Computer Science at Qatar University) and **Walid Magdy** (lecturer at the School of Informatics at the University of Edinburgh). It also has three members who are responsible for publicity (website, social media, mailing lists, etc.) and helping with developing the proceedings: **Mucahid Kutlu** (post-doctorate researcher at Qatar University), **Maram Hasanain** (Computer Science PhD candidate at Qatar University), and **Reem Suwaileh** (MSc student and research assistant at Qatar University).

Tamer Elsayed received his PhD in Computer Science from the University of Maryland, College Park. His main research interests are information retrieval, text mining, and big data analytics. He has over 50 publications in top-tier journals (e.g., JASIST and IP&M) and conferences (e.g., SIGIR and ICWSM). He received two best paper awards at AIRS 2015 and HCOMP 2016 conferences. He is a member of the editorial board of IP&M Elsevier journal and served as a PC member in top IR conferences (e.g., ACM SIGIR and ACM CIKM). His research team has regular participation in microblog track at TREC since 2011, including ad-hoc search, filtering, and real-time summarization tasks.

Walid Magdy received his PhD from the School of Computing at Dublin City University. His main research interests include computational social science, information retrieval, and data mining. Before joining UoE in 2016, He worked for about five years as a scientist at Qatar Computing Research Institute (QCRI). He also worked at his early career for IBM and Microsoft as a research engineer between 2005 and 2008. He has over 60 publications in top-tier conferences and journals, in addition to 9 patents filed under his name. Some of his work was featured in popular press, such as CNN, BBC, Washington Post, the Independent, Daily Mail, and Mirror.

Mucahid Kutlu received his PhD in Computer Science and Engineering from Ohio State University. His main research interests are big data and information retrieval with emphasis on evaluation and more specifically building scalable test collections over the Web and social media.

Maram Hasanain's research interests include information retrieval over tweets with special focus on Arabic text. Her work focuses on problems related to evaluation, ad-hoc search, filtering, summarization and question answering. Her publications appeared at top conferences and journals and she served as a reviewer for several of them.

Reem Suwaileh's research area is information retrieval with emphasis on topic tracking and summarization over tweets. She has been a regular member of Qatar University team participating at TREC since 2015, where her team was ranked second in 2015 and first in 2016 in real-time summarization track.

Acknowledgments

This work was made possible by NPRP grant# NPRP 7-1313-1-245 from the Qatar National Research Fund (a member of Qatar Foundation). The statements made herein are solely the responsibility of the authors.

References

1. Allan, J., Carbonell, J.G., Doddington, G., Yamron, J., Yang, Y.: Topic detection and tracking pilot study final report (1998)
2. Badache, I., Fournier, S., Chifu, A.G.: Contradiction in reviews: is it strong or low? In: Proceedings of the first International Workshop on Analysis of Broad Dynamic Topics over Social Media: BroDyn'18 (2018)
3. Baeza-Yates, R., Ribeiro-Neto, B., et al.: Modern information retrieval, vol. 463. ACM press New York (1999)
4. Bulbul, A., Kaplan, C., Haj Ismail, S.: Social media based analysis of refugees in turkey. In: Proceedings of the first International Workshop on Analysis of Broad Dynamic Topics over Social Media: BroDyn'18 (2018)
5. Cram, L., Llewellyn, C., Hill, R.L., Magdy, W.: UK general election 2017: a twitter analysis. CoRR abs/1706.02271 (2017), http://arxiv.org/abs/1706.02271
6. Darwish, K., Magdy, W., Zanouda, T.: Trump vs. hillary: What went viral during the 2016 us presidential election. In: Proceedings of the 9th International Conference on Social Informatics (SocInfo'17). pp. 143–161. Springer (2017)
7. Lin, J., Efron, M., Wang, Y., Sherman, G.: Overview of the trec-2014 microblog track. In: Proceedings of the 23rd Text REtrieval Conference. TREC '14 (2014)
8. Lin, J., Efron, M., Wang, Y., Sherman, G., Voorhees, E.: Overview of the trec-2015 microblog track. In: Proceedings of the 24th Text REtrieval Conference. TREC '15 (2015)
9. Lin, J., Roegiest, A., Tan, L., Richard, M., Voorhees, E., Diaz, F.: Overview of the trec 2016 real-time summarization track. In: Proceedings of the 25th Text REtrieval Conference. TREC '16 (2016)
10. Magdy, W., Elsayed, T.: Adaptive method for following dynamic topics on twitter. In: Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media (ICWSM) (2014)
11. Magdy, W., Elsayed, T.: Unsupervised adaptive microblog filtering for broad dynamic topics. Information Processing & Management 52(4), 513–528 (2016)

12. Mazoyer, B., Cagé, J., Hudelot, C., Viaud, M.L.: Real-time collection of reliable and representative tweets datasets related to news events. In: Proceedings of the first International Workshop on Analysis of Broad Dynamic Topics over Social Media: BroDyn'18 (2018)
13. Morris, M.R., Teevan, J., Katrina, P.: A comparison of information seeking using search engines and social networks. In: Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media. pp. 291–294 (2010)
14. Olteanu, A., Castillo, C., Diaz, F., Vieweg, S.: Crisislex: A lexicon for collecting and filtering microblogged communications in crises. In: Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media (ICWSM) (2014)
15. Ounis, I., Macdonald, C., Lin, J., Soboroff, I.: Overview of the trec-2011 microblog track. In: Proceedings of the 20th Text REtrieval Conference. TREC '11 (2011)
16. Robertson, S.E., Soboroff, I.: The trec 2002 filtering track report. In: TREC. vol. 2002, p. 5 (2002)
17. Ruths, D., Pfeffer, J.: Social media for large studies of behavior. Science 346(6213), 1063–1064 (2014)
18. Soboroff, I., Ounis, I., Lin, J., Macdonald, C.: Overview of the trec-2011 microblog track. In: Proceedings of the 21st Text REtrieval Conference. TREC '12 (2012)
19. Teevan, J., Ramage, D., Morris, M.R.: # twittersearch: a comparison of microblog search and web search. In: Proceedings of the fourth ACM international conference on Web search and data mining. pp. 35–44. ACM (2011)