Preface

#Microposts2016, the 6th Workshop on *Making Sense of Microposts*, was held in Montréal, Canada, on 11th Apr 2016, during WWW'16, the 25th International Conference on the World Wide Web. The #Microposts journey started at the 8th Extended Semantic Web Conference (ESWC 2011), as #MSM, with the change in acronym from 2014. The workshop moved to WWW in 2012, where it has stayed for five years. *#Microposts2016* continues to highlight the importance of the medium, as we see end users appropriating these small chunks of information published online with minimal effort, as part of daily communication, to interact with increasingly wider networks and from new publishing arenas.

Microposts persist as a popular means of communication, from tweets to check-ins to pins and vines. Microblogging apps for the ubiquitous personal device simplify dissemination of photos and short videos through Microposts, annotated with hashtags and other "signals". Mobile-based services such as WhatsApp and Saya, piggybacking on SMS and augmented with social media features, continue to grow in popularity, especially in emerging markets where Internet access is often via mobile networks. Microposts also serve as portals, pointing to, for instance, micro-snaps on Snapchat.

Individual Microposts typically focus on a single thought, message or theme. Collectively, Microposts contribute to today's *big data* – very large scale, heterogeneous data and a source of valuable collective intelligence that feeds into opinion mining and crowd tracking, emergency response and community services. Microposts are used to generate feelings of community and inspire reaching toward a common goal, as seen in online activism during the Arab Spring and the 2013 protests in Turkey and Brazil. Microposts are used to drive social crowdfunding across several hops in a social network, contributing to, e.g., victims of the Boston Marathon terror attack. Beyond purely social impact, Microposts are used to foster innovation and generate new online and offline markets, drive targeted advertising and promote ideas and services. By providing a public, shared information space, insight may be obtained into how products and services are used by individuals and communities, where services fail, and on what innovation may be required to serve a particular need or fill a gap in the market.

The #Microposts workshops are unique in that they encourage inter-disciplinary work across Computer Science and Information, Web and the Social Sciences, providing a forum to enable discussion from multiple perspectives, and hence, improve understanding of the social and cultural phenomena that influence the publication and reuse of Microposts. Following on from 2015, therefore, the workshop also includes a special (Computational) Social Science track, to harness the benefits of approaches to research in this field. #Microposts2016 focused on topics across three main areas:

Making Sense/Understand – focusing on the human in Micropost data generation and analysis as the publisher or the focus, to better understand how situation and context drive Micropost generation and reuse.

Discover – to extract and feed the information content of Microposts into analysis, to aid pattern and trend discovery, to guide further knowledge acquisition and application development.

Case Studies & Applications – describing real-world cases demonstrating the use of Micropost data and Microposts as a tool, e.g., in targeted ads and for crisis management.

Submissions in 2016 examined: data mining for effective Micropost data content reuse; media & politics; cultural, generational and regional differences in access and use of Microposts, with humans as sensors and a trigger on the public pulse; network analysis and community detection; user behaviour & interaction across networks; influence detection; user profiling, personalisation & recommendation; data cleaning and content verification; security, crisis management and emergency response; identification and use of geo-location information embedded in or attached to Microposts.

A key aim of the workshop is to promote formal evaluation of text extraction tools for Micropost data. We have therefore included since 2013 an information extraction challenge, with a new, pertinent topic addressed each year that builds on previous challenges. The #Microposts2016 **N**amed **E**ntity r**E**cognition and **L**inking (NEEL) Challenge required participants to recognise and type entities, and link them to corresponding DBpedia 2015-04 resources. NEEL encourages innovation in the development of approaches and tools for information extraction and contributes to benchmark datasets for research that leads to making sense of Microposts.

We thank all contributors and participants; each adds to research that advances the field. Submissions to the research tracks came from institutions in seven countries. The challenge also continues to see wide interest, with final submissions from academia and industry across six countries. Our programme committee is even more varied, working in academia, independent research institutions and industry, spanning even more countries. Most of our PC have reviewed for more than one, and a good percentage, all six #Microposts workshops. Very special thanks to our committee, whose feedback is vital to running the workshop. Thanks also to the chairs of the (Computational) Social Sciences Track and the NEEL Challenge, whose work has been invaluable in pulling the three tracks together into a unified, successful workshop.

Danica Radovanović	Basic Internet Foundation, Norway
Amparo E. Cano Basave	Aston University, UK
Daniel Preoţiuc-Pietro	University of Pennsylvania, US
Katrin Weller	GESIS, Germany
Aba-Sah Dadzie	The Open University, UK

#Microposts2016 Organising Committee, April 2016

Introduction to the Proceedings

Invited Talk

Mihajlo Grbovic[1], a senior research scientist at Yahoo! Labs, presented his work on '*Leveraging Blogging Activity on Tumblr to Infer Demographics and Interests of Users for Advertising Purposes*'. Mihajlo joined the Targeting Science group in Yahoo! in September 2012 and has since worked on projects on behavioural targeting. aim is to monetise the results of research into behaviour on Tumblr, influenced by user gender and interest, through targeted advertising. He received his PhD from the Department of Computer Science at Temple University, Philadelphia, USA. His thesis examined machine learning applications in "Decentralized Fault Detection and Diagnosis". During his PhD Mihajlo spent time as a research intern at Xerox Research Labs Europe and ExxonMobil and Akamai Technologies in the US. His current research interests are machine learning, ad targeting, monetisation, web search and data mining.

The paper, co-authored with Vladan Radosavljevic, Nemanja Djuric, Narayan Bhamidipati and Ananth Nagarajan, presents a framework for a set of key targeted advertising components for Tumblr, focusing specifically on gender and interest profiling. Grbovic *et al.,* describe the main challenges involved in the development of this framework, discussing first the creation of a ground truth for training gender prediction models and mapping Tumblr content to an interest taxonomy. The authors propose a novel semi-supervised neural language model for categorisation of Tumblr content and train this model on a large-scale data set consisting of 6.8 billion user posts. The model has been successfully deployed with gender and interest targeting capability for Yahoo! production systems, delivering inference for users that covers more than 90% of their daily activities on Tumblr.

Main Research Track

The main workshop track attracted three submissions, of which two long papers were accepted. Topics covered included information diffusion within diverse online communities, topic detection, automatic determination of veracity and validation of the myriad sources that contribute to big data and location identification from contextual information in Microposts.

The study of the veracity of information in online social media has very recently garnered important research momentum in the Social Media and Natural Language Processing communities. *ClaimFinder: A Framework for Identifying Claims in Microblogs*, by Wee Yong Lim, Mong Li & Lee Wynne Hsu, represents a step toward automatically dealing with *veracity*, the so-called 4th 'V' of big data. The study looks in particular at identifying claims in microblogs using open relation extraction and clustering of similar claims. Lim *et al.,* present a novel approach which, going beyond shallow linguistic information, may serve as a first step in a larger pipeline that automatically validates claims in Microposts. Two practical utilities are provided by the authors: a streaming version of their method and a new annotated, real-world dataset crawled from Twitter, which, with a third party dataset, should be of interest to the wider research community.

The paper *Birds of a Feather Tweet Together: Computational Techniques to Understand User Communities in Social Networks* by David Burth Kurka, Alan Godoy & Fernando J. Von Zuben, analyses the effects of homophily in Twitter using automatic community detection and topic analysis. Their use case looks at the Twitter followers of the most popular Brazilian news account, Folha de Sao Paulo. Kurka *et al.,* study the different communities formed by their social networks and information diffusion within these communities. Text analysis reveals common topics specific to each community, which range from local, contextual interest (e.g., a distinct community that emerged due to the discussion of events specific to the Brazilian state of Pernambuco) to topical interest (e.g., partisan political supporters).

Kelly Geyer, Kara Greenfield, Alyssa Mensch & Olga Simek, in their late-breaking paper, *Named Entity Recognition in 140 Characters or Less*, explore how adaptable the open-source MIT Information Extraction Toolkit (MITIE) is to addressing the particular challenges inherent in Micropost data, among others, inconsistent use of grammar and the lack of context around named entities. The authors ran their tests by retraining on this year's NEEL Challenge data, with mixed results.

Computational Social Science Track

This special track, chaired by Katrin Weller of the GESIS Leibniz Institute for the Social Sciences in Cologne, Germany, was introduced to encourage closer work with the Social Sciences. The SocSci track saw three submissions, with a fourth that crossed the boundary between Computer Science and Social Science approaches to research. A topic that saw large coverage in this track and across the workshop was user profiling, in this track looking at ordinary and *elite* users. One submission found in a study using linguistic analysis, marked differences in personality profiles on and off social media. A second topic that has attracted attention in the last few years is the analysis of often highly charged or emotive information generated due to crises triggered by terrorist activity.

The paper *Comparing Social Media and Traditional Surveys Around the Boston Marathon Bombing* by Cody Buntain, Erin McGrath, Jennifer Golbeck & Gary LaFree fits well in the intersection of computer science and social science, as it compares research methods from both areas, namely social media data analysis and surveys. Buntain *et al.,* show how Micropost data such as tweets complement survey data, and compare advantages and disadvantages over the use of either or both data formats. The comparative analysis is based on a case study around the 2013 Boston Marathon bombings, for which both types of data are available. Buntain *et al.,* consider several factors, including representativeness of each data type, data quality and costs in obtaining and analysing social media data as against survey data. The paper concludes with lessons learnt from triangulating the outcomes of analysis of the two data types.

Alex Jeongwoo Oh & Pramuan Bunkanwanicha in the poster paper *CEOs on Twitter* explore, with a specific focus on Twitter, the impact of CEOs' social media activity on corporate performance. The study seeks to determine if the behaviour of executives on the public channels provided by social media impacts performance in the external, offline world. Oh & Bunkanwanicha found differences in CEOs that use Twitter and those who don't, providing new insight into economic implications for firms and shareholders. Notable differences were found in age and compensation level, with positive impact of CEOs' use of Twitter on corporate performance,

[1]M. Grbovic research pages: http://astro.temple.edu/~tua95067

contrary to previous studies showing evidence of negative effect on performance. Results to date in the on-going study indicate that engagement with social media may have implications for top level management and corporate finance.

The poster *Studying the Role of Elites in U.S. Political Twitter Debates* by Sebastian Stier presents exploratory work on the analysis of "elites" on social media in the US political sphere – sitting members of congress, incumbent governors, national party accounts and presidential candidates – by examining two use cases in US politics. Stier's analysis uses follow count as the most basic element of influence, and by categorising Twitter users into different actors in the political arena, provides an initial look at how influence may be exerted in political debate through Twitter. The poster concludes with pointers toward further avenues of research and additional metrics to examine in analysis in this sphere. An outcome of the on-going study is a publicly available dataset[2], collected from related open, official data on the users and accounts studied.

Named Entity rEcognition & Linking (NEEL) Challenge

The #Microposts2016 NEEL challenge provides a forum in which to tackle the challenges associated with information extraction from textual Microposts, due to their brevity and user of non-standard language and abbreviations. In its fourth year the NEEL challenge again increased complexity in both task and dataset, in order to encourage the use of novel and/or adaptation of existing approaches for information extraction to this type of data. The #Microposts2016 NEEL challenge was chaired by Giuseppe Rizzo of the Istituto Superiore Mario Boella, Torino, Italy, and Marieke van Erp of Vrije Universiteit Amsterdam, The Netherlands.

Thirty-seven teams expressed an intent to participate, out of which five proceeded to the final evaluation process. We provide here a brief introduction to participants' abstracts describing their approach to solving the challenge, in order of ranking from the highest scored submission. A detailed description of the preparation and challenge evaluation processes can be found in the challenge summary paper included in the proceedings.

Jörg Waitelonis & Harald Sack in their submission, *KEA*, use joint Mention Extraction and Candidate Selection to map text ngrams to DBpedia entities. A preprocessing stage cleans and normalises tweets, following which they are resolved to DBpedia entities. Candidate selection is verified using a confidence score, to determine whether to annotate with the mention assigned or, on failing, to NIL.

Pablo Torres-Tramón, Hugo Hromic, Brian Walsh, Bahareh R. Heravi & Conor Hayes implement a linguistic pipeline that uses relevant data from DBpedia as a lookup for lexical values of mentions for Candidate Selection. This submission also makes use of a preprocessing stage that normalises tweet text to formal language. Disambiguation of entities from candidate DBpedia resources relies on entity relatedness reasoning. Hierarchical clustering is then used to resolve NILs – entities linked to mentions but for which no corresponding DBpedia referent is found.

Kara Greenfield, Rajmonda Caceres, Michael Coury, Kelly Geyer, Youngjune Gwon, Jason Matterer, Alyssa Mensch, Cem Sahin & Olga Simek make use of a joint graph-based and linguistic approach, but without tweet normalisation. Using DBpedia as a dictionary of entities, t his submission maps the DBpedia Ontology to the NEEL challenge taxonomy. Parallel candidate name generation is followed by the linking task, which is treated as a binary classification task with an extensive feature set. Entities are assigned using Named Entity Recognition, and clustered using the normalised Damerau-Levenshtein.

Souvick Ghosh, Promita Maitra & Dipankar Das employ a sequential linguistic pipeline, starting by enriching mentions using DBpedia types in the preprocessing stage, by exploiting also the training data to obtain additional mentions. This is followed by Named Entity Recognition, to extract proper nouns using a random forest with a rich feature vector. Named Entity Linking uses Babelfy to annotate tweets one by one, following which NILs are identified through clustering of non-linked entities.

Davide Caliano, Elisabetta Fersini, Pikakshi Manchanda, Matteo Palmonari & Enza Messina also make use of a sequential approach. After preprocessing to remove special characters the employ T-NER for entity identification. Candidate selection is carried out using a learning to rank strategy to score mentions, weighting lexical similarity of mention against the corresponding Wikipedia title and contextual similarity within the tweet and the corresponding DBpedia abstract. A post-processing stage is used to resolve mention boundary issues.

Additional Material

The call for participation and all paper, poster and challenge abstracts are available on the #Microposts2016 website[3]. The full proceedings are also available at , as Vol-1691[4]. The gold standard for the NEEL Challenge is available for download[5].

Previous workshop proceedings are available online:

#Microposts2015 as CEUR Vol-1395[6]. The NEEL2015 challenge gold standard is available for download[7].

#Microposts2014 as CEUR Vol-1141[8]. The NEEL2014 challenge gold standard is available for download[9].

#MSM2013 main track proceedings available as part of the WWW'13 Proceedings Companion[10].

#MSM2013 Concept Extraction Challenge proceedings published as a separate volume, as CEUR Vol-1019[11]. The challenge gold standard is available for download[12].

#MSM2012 as CEUR Vol-838[13].

#MSM2011 as CEUR Vol-718[14].

[2]Stier, S., *Elite actors in the U.S. political Twittersphere.* GESIS datorium: http://dx.doi.org/10.7802/1178

[3]http://microposts2016.seas.upenn.edu

[4]**#Microposts2016 Proc.** http://ceur-ws.org/Vol-1691

[5]**#Microposts2016 NEEL gold standard.** http://ceur-ws.org/Vol-1691/microposts2016_neel-challenge-report/microposts2016-neel_challenge_gs.zip

[6]**#Microposts2015 Proc.** http://ceur-ws.org/Vol-1395

[7]**#Microposts2015 NEEL gold standard.** http://ceur-ws.org/Vol-1395/microposts2015_neel-challenge-report/microposts2015-neel_challenge_gs.zip

[8]**#Microposts2014 Proc.** http://ceur-ws.org/Vol-1141

[9]**#Microposts2014 NEEL gold standard.** http://ceur-ws.org/Vol-1141/microposts2014-neel_challenge_gs.zip

[10]**WWW'13 Companion.** http://dl.acm.org/citation.cfm?id=2487788

[11]**#MSM2013 CE Challenge Proc.** http://ceur-ws.org/Vol-1019

[12]http://ceur-ws.org/Vol-1019/msm2013-ce_challenge_gs.zip

[13]**#MSM2012 Proc.** http://ceur-ws.org/Vol-838

[14]**#MSM2011 Proc.** http://ceur-ws.org/Vol-718

Track Sponsors

Main Track. The best paper award for this track was sponsored by the MK:Smart project, a large, collaborative initiative for developing innovative solutions by mining vast amounts of data. MK:Smart is part-funded by HEFCE (the Higher Education Funding Council for England) and led by The Open University, UK. The award is to encourage research into and the development of novel applications based on data extracted from Microposts. Nominations were sought from reviewers, and a final decision agreed by the workshop chairs, based on the nominations and review scores.

The #Microposts2016 best paper award went to:

Wee Yong Lim, Mong Li Lee & Wynne Hsu
for their submission entitled:
ClaimFinder: A Framework for Identifying Claims in Microblogs

NEEL Challenge. The European H2020 project FREME sponsored the award for the best submission. FREME aims to develop an open framework of services for multilingual and semantic enrichment of digital content. FREME's technologies may be used to harvest and analyse content, to provide added value to content and data value chains across industry, countries and languages. By sponsoring the challenge, FREME acknowledges growing interest in automatic approaches for gleaning information from increasingly large scale social media data and reinforces the value of Micropost knowledge content for industry. The challenge award was also determined by the results of the quantitative evaluation.

The #Microposts NEEL Challenge award went to:

Jörg Waitelonis and Harald Sack
for their submission entitled:
Named Entity Linking in #Tweets with KEA

Computational Social Science Track. GESIS, the Leibniz Institute for the Social Sciences, sponsored the (Computational) Social Science track. GESIS is the largest infrastructure institution for the Social Sciences in Germany. At the GESIS department of Computational Social Science interdisciplinary research teams study society through new types of data, which often include different types of Microposts. Sponsorship of this special track at the #Microposts workshop helps to increasingly connect researchers from social sciences and computer sciences and to explore interdisciplinary approaches for making sense of Microposts.

#Microposts2016

Additional Sponsors

WWBP, the World Well-Being Project. is pioneering scientific techniques for measuring psychological well-being and physical health based on the analysis of language on social media. As a collaboration between computer scientists, psychologists, and statisticians, WWBP aims to shed light on the psychosocial processes that affect health and happiness, and explore the potential for unobtrusive well-being measures to supplement, and in part replace, expensive survey methods. The project is based out of the University of Pennsylvania's Positive Psychology Center. WWBP is supported by the Templeton Religion Trust.

EDSA, the European Data Science Academy. analyses sector-specific skill sets required for Data Scientists across the EU, in order to identify skill gaps, and therefore develop modular, context-specific courses to train "data workers" to fill these gaps. EDSA considers a number of big data sources, including social media, to extract assessment of expertise in different domains and for specific role types, as determined by policy makers, employers and job seekers.

#Microposts2016

Main Track Programme Committee

Nikolaos Aletras Amazon, UK
Pierpaolo Basile University of Bari, Italy
Julie Birkholz CHEGG, Ghent University, Belgium
Marco A. Casanova Pontifícia Universidade Católica do Rio de Janeiro, Brazil
Óscar Corcho Universidad Politécnica de Madrid, Spain
Guillaume Erétéo Vigiglobe, France
Miriam Fernandez KMi, The Open University, UK
Lucie Flekova Technische Universität Darmstadt, Germany
Anna Lisa Gentile Universität Mannheim, Germany
Dirk Hovy University of Copenhagen, Denmark
Jelena Jovanovic University of Belgrade, Serbia
Mathieu Lacage Alcméon, France
Maria Liakata Warwick University, UK
Vasileios Lampos University College London, UK
João Magalhães Universidade Nova de Lisboa, Portugal
Yelena Mejova Qatar Computing Research Institute, Qatar
José M. Morales del Castillo El Colegio de México, Mexico
Fabrizio Orlandi University of Bonn, Germany
Bernardo Pereira Nunes Pontifícia Universidade Católica do Rio de Janeiro / Federal University of the State of Rio de Janeiro, Brazil
Harald Sack HPI, University of Potsdam, Germany
Bernhard Schandl mySugr GmbH, Austria
Victoria Uren Aston Business School, UK
Andrea Varga Cube, UK
Svitlana Volkova Pacific Northwest National Laboratory, USA
Lyle Ungar University of Pennsylvania, USA
Alistair Willis The Open University, UK
Wei Xu University of Pennsylvania, USA
Ziqi Zhang The University of Sheffield, UK

Advisory Committee & Publicity

Milan Stankovic Université Paris-Sorbonne/ Sépage, France

Social Sciences Track Programme Committee

Gholam R. Amin University of New Brunswick, Canada
Julie Birkholz CHEGG, Ghent University, Belgium
Tim Davies University of Southampton, UK
Jordan Carpenter University of Pennsylvania, USA
A. Seza Doğruöz Tilburg University, The Netherlands
Fabio Giglietto Università di Urbino Carlo Bo, Italy
Athina Karatzogianni University of Leicester, UK
José M. Morales del Castillo El Colegio de México, Mexico
Raquel Recuero Universidade Católica de Pelotas, Brazil
Luca Rossi Università di Urbino Carlo Bo, Italy
Victoria Uren Aston Business School, UK
Alistair Willis The Open University, UK

NEEL Challenge Evaluation Committee

Ebrahim Bagheri Ryerson University, Canada
Pierpaolo Basile University of Bari, Italy
Grégoire Burel KMi, The Open University, UK
David Corney Signal Media, UK
Milan Dojchinovski Czech Technical University in Prague, Czech Republic / AKSW/INFAI, Leipzig University, Germany
Guillaume Erétéo Vigiglobe, France
Anna Lisa Gentile Universität Mannheim, Germany
Filip Ilievski Vrije Universiteit Amsterdam, The Netherlands
Mathieu Lacage Alcméon, France
Miguel Martinez-Alvarez Signal, UK
José M. Morales del Castillo El Colegio de México, Mexico
Enrico Palumbo Istituto Superiore Mario Boella, Italy
Bianca Pereira Insight Centre for Data Analytics, NUIG, Ireland
Bernardo Pereira Nunes Pontifícia Universidade Católica do Rio de Janeiro / Federal University of the State of Rio de Janeiro, Brazil
Julien Plu EURECOM, France
Giles Reger Otus Labs, UK
Irina Temnikova Qatar Computing Research Institute, Qatar