Preface

#Microposts2015, the 5th Workshop on *Making Sense of Microposts*, was held in Florence, Italy, on the 18th of May 2015, during (WWW'15), the 24th 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), and moved to WWW in 2012, where it has stayed, for the fourth year now. *#Microposts2015* continues to highlight the importance of the medium, as we see end users appropriating Microposts, small chunks of information published online with minimal effort, as part of daily communication and to interact with increasingly wider networks and new publishing arenas.

The #Microposts workshops are unique in that they solicit participation not just from Computer Science, but encourage interdisciplinary work. We welcome research that looks at computational analysis of Microposts, as well as studies that employ mixed methods, and also those that examine the human generating and consuming Microposts and interacting with other users via this publishing venue. New to #Microposts2015 is a dedicated Social Sciences track, to encourage, particularly, contribution from the Social Sciences, to harness the advantages that approaches to analysing Microposts from this perspective bring to the field.

The term *Micropost* now rarely needs definition. Microposts are here to stay, and have evolved from text only, to include images, and now, audio and video. New platforms are developed each year to serve specific markets, and niche services compete with each other for a share of the audience. Twitter's *Periscope* is a new service similar to *Meerkat*, both of which use microblogging platforms to alert a network to a live video stream. Microposts now often serve also as a portal, and are harnessed by recommendation services, marketing and other enterprise to advertise or push information, products and services on other platforms. This is a not surprising means to access potential users, who now exchange Microposts round the clock, using a variety of publishing platforms. Media trends show that users are doing so increasingly from personal, mobile devices, as a preferred/convenient option that started to overtake usage on PCs in 2014. To extend reach, both in developed and emerging markets, services for publishing Microposts from feature phones are being developed – these include the usual suspects, Twitter and Facebook, who employ native apps or the mobile web, and also newer entrants with dedicated services and apps such as *Saya*. Country and language-specific platforms such as Sina Weibo, while not as widespread, serve a specific region and market, especially where any of a number of reasons prevent access to the more well-known microblogging platforms. Political movements such as the Arab Spring have been reported to have increased the use of social media services and microblogging particularly in regions concerned, as the quick, low-cost means for sharing, in the moment, breaking news, local and context-specific information and personal stories, resulted in an increased sense of community and solidarity. Interestingly, in response to emergencies, mass demonstrations and other social events such as festivals and conferences, when regular access to communication services is often interrupted and/or unreliable, developers are quick to offer alternatives that end users piggyback on to post information. *Line* was born to serve such a need, to provide an alternative communication service and support emergency response during a natural disaster in Japan in 2011. Its popularity continued beyond its initial purpose, and Line has grown into a popular (regional) microblogging service.

The #Microposts workshop was created to bring together researchers in different fields studying the publication, analysis and reuse of these very small chunks of information, shared in private, semi-public and fully open, social and formal networks. Microposts collectively make up a vast knowledge store, contained in what is today described as "big data" – heterogenous, increasing at phenomenal rates, and with multiple, unbridled authors, covering myriad topics with varying degrees of accuracy and veracity. With each year we have seen submissions tackling different aspects of Microposts, with new methods and techniques developed to analyse this valuable dataset and also its publishers, human or bot, and examining the different ways in which the medium is used. With the increase in the use of Microposts as a portal to other services, we saw, this year, studies on the detection and analysis of spam, and the use of open posting as a cover for disseminating extremist opinions or to swamp dissenting views. Reflecting the very social nature of the publishing platform, submissions also covered analysis of the human reaction to recent, provoking news events.

We thank all contributors and participants: each author's work adds to research that continues to advance the field. Submissions to the two research tracks came from institutions in ten countries around the world. 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, and spanning an even larger number of countries. Most of our PC have reviewed for more than one, and a good percentage, all five #Microposts workshops. Very special thanks to our committee, without whom we would not be able to run the workshop – their dedication is seen in the feedback provided to us and to authors. Thanks also to the chairs of the Social Sciences Track and the NEEL Challenge, whose work has been invaluable in pulling the three parts together into a unified, successful workshop.

Matthew Rowe Lancaster University, UK
Milan Stankovic Sépage / Université Paris-Sorbonne, France
Aba-Sah Dadzie KMi, The Open University, UK
#Microposts2015 Organising Committee, May 2015

Introduction to the Proceedings

Main Track

The main workshop track attracted nine submissions, out of which two long papers and one short were accepted, in addition to an extended abstract and a poster. It should be noted that two of these crossed the boundary between Computer and Social Sciences, and were therefore assigned reviewers from both tracks. Topics covered ranged from machine learning and named entity recognition to Micropost classification and extraction. Applications were seen in topic, event and spam detection. We provide a brief introduction to each below.

De Boom, Van Canneyt & Dhoedt, in *Semantics-driven Event Clustering in Twitter Feeds*, present a novel perspective on event-detection in tweets, by associating semantics to tweets and hashtags. They demonstrate how an approach that combines machine learning with explicit semantics detection can yield considerable improvement over state of the art event clustering approaches.

In the paper *Making the Most of Tweet-Inherent Features for Social Spam Detection on Twitter*, Wang, Zubiaga, Liakata & Procter investigate the use of a variety of feature sets and classifiers for the detection of social spam on Twitter. These include user features (social network properties of the tweeter, such as their in- and out-degrees); content features (number of hashtags and mentions); n-gram features (mined from textual aspects); and sentiment features, based on both manually and automatically created semantic lexicons. Classifiers tested including naïve Bayes; k-Nearest Neighbours, Support Vector Machines, Decision Trees, and Random Forests. The paper presents an interesting investigation, classifying users as spammers (or not), as opposed to existing work which attempts to classify content as spam (or not).

In *User Interest Modeling in Twitter with Named Entity Recognition*, Karatay & Karagoz explore techniques for user profiling using Named Entity detection in tweets – a topic of increasing importance in the era of information overload, where filtering and personalising information is crucial for user engagement and experience. The in-depth view of appropriate techniques and issues related to Named Entity-based user profiling on Twitter will interest both academic and industrial audiences.

Within the broader area of spam, misconduct and automated accounts on Twitter, Edwards & Guy study the *Connections between Twitter Spammer Categories*. Unlike most other work in this area, they do not only distinguish spam from non-spam, but assume there are different types of spam accounts, which they categorise as "advertising", "explicit", "follower gain", "celebrity" and "bot". They show, in their extended abstract, that each type of spammer behaves differently with respect to establishing follower relations with other spam accounts. They also observe that genuine Twitter users can be found as followers of all types of spam accounts, but are more likely to connect with specific types of spammers.

Agarwal & Sureka, in *A Topical Crawler for Uncovering Hidden Communities of Extremist Micro-Bloggers on Tumblr*, discuss the use of microblogging systems such as Tumblr to promote extremism, taking advantage of the ability to post information anonymously. The poster paper describes a process that uses pre-identified keywords to flag relevant posts, and hence, identify suspect tags in textual posts. A random walk from a seed blogger is then used to identify further individuals and communities promoting extremism. The authors report misclassification of 13% and accuracy of 77% for predicting "hate promoting bloggers", with misclassification of unknown bloggers at 34%.

Social Sciences Track

The Social Sciences track attracted three submissions, of which two were accepted. In addition to data mining and/or statistical analysis over the very large amounts of data involved, each submission carried out in-depth, qualitative analysis to tease out nuanced information that is more difficult to identify with automated methods. The track was chaired by Katrin Weller and Danica Radovanović.

One of the major contemporary events that spiked user engagement on social media during the first months of 2015 was the Charlie Hebdo shooting in France on January 7th. Giglietto & Lee provide one of the first studies of Twitter users' reactions to this event, in *To Be or Not to Be Charlie: Twitter Hashtags as a Discourse and Counter-discourse in the Aftermath of the 2015 Charlie Hebdo Shooting in France*. In particular, they study the use of the hashtag #JeNeSuisPasCharlie, which was used in contrast to the initial #JeSuisCharlie hashtag. Using different approaches to data analysis (including activity patterns and word frequencies) the authors demonstrate how tweets including #JeNeSuisPasCharlie rather resemble crisis communication patterns, and at the same time support different expressions of self-identity such as grief and resistance.

Coelho, Lapa, Ramos & Malini, in *A Research Design for the Analysis of Contemporary Social Movements*, present a research method to identify elements that promote social empowerment in the political vitality present in digital culture. They developed a model of investigation that allows discursive analysis of posts generated within net activist groups. Methods, instruments and resources were created and articulated for the collection and treatment of big data and for further qualitative analysis of content. In addition to contributing to ICT, by proposing a qualitative investigation of social networks, this research design contributes to the field of Education, as the results of its application can be used be used to develop guidelines for teachers, to support critical appropriation and education of social networks.

Named Entity rEcognition & Linking (NEEL) Challenge

The #Microposts2015 NEEL challenge again increased in complexity, to address further challenges encountered in the analysis of Micropost data. This year's challenge required participants to recognise entities and their types, and also link them, where found, to corresponding DBpedia resources.

The challenge attracted good interest from the community, with 29 intents to submit, out of which 21 applied for the final evaluation. Seven took part in the quantitative evaluation and six completed submission (including a written abstract). Of these three were accepted for presentation and a further three as posters. All accepted submissions also took part in the workshop's poster session, whose aim is to exhibit practical application in the field and foster further discussion about the ways in which knowledge content is extracted from Microposts and reused.

The NEEL challenge was chaired by A. Elizabeth Cano and Giuseppe Rizzo, with Andrea Varga and Bianca Pereira as dataset chairs. As in previous years, the challenge committee prepared a gold standard from the challenge corpus, which covered events in 2011, '13 & 14 on, for example, the London Riots, the Oslo bombing and the UCI Cyclo-cross World Cup. Changes to the submission and evaluation protocols included wrapping submissions as a publicly accessible, REST-based service. Up to ten runs were allowed per submission, of which the best three were used in computing the final rankings, using four weighted metrics: tagging (0.3), linking (0.3), clustering (0.4) and latency (computation time) to sort in case of a tie.

We provide here a brief introduction to participants' abstracts describing their submissions, and more detail about the preparation and evaluation processes in the challenge summary paper included in the proceedings.

Yamada, Takeda & Takefuji, in *An End-to-End Entity Linking Approach for Tweets*, present a five stage approach: (1) preprocessing, (2) candidate mention generation, (3) mention detection and disambiguation, (4) NIL mention detection and (5) type prediction. In preprocessing, they utilise tokenisation and POS tagging based on state of the art algorithms, along with extraction of tweet timestamps. Yamada *et al.* tackle candidate mention generation and disambiguation using fuzzy search of Wikipedia for candidate entity mentions, and popularity of Wikipedia pages for ranking the set of candidate entities. Finally, they tackle selection of NIL mentions and entity typing as supervised learning problems.

In *Entity Recognition and Linking on Tweets with Random Walks*, Guo & Barbosa present a sequential approach to the NEEL task by, first, recognising entities using *TwitIE*, and then linking them to corresponding DBpedia entities. Starting from the (DBpedia) candidate entities, Guo & Barbosa build a subgraph by adding all adjacent entities to the candidates. They execute a personalised PageRank, giving more importance to unambiguous entities. They then measure semantic relatedness between entity candidates and the "unambiguous" entities for the "document", and employ threshold and name similarity for NIL prediction and clustering.

In the submission *Combining Multiple Signals for Semanticizing Tweets: University of Amsterdam at #Microposts2015*, Gârbacea, Odijk, Graus, Sijaranamual & de Rijke employ a sequential approach composed of four stages: (1) candidate mention detection, (2) candidate typing and linking, (3) NIL clustering and (4) overlap resolution. The first stage is tackled with an annotation-based process that takes as input the lexical content of Wikipedia and an NER classifier trained using the challenge dataset. To resolve candidate mention overlaps, the authors propose an algorithm based on the results of the linking stage and the *Viterbi* path resolution output. A "learning to rank" supervised model is used to select the most representative DBpedia reference entity, and, therefore, type of each candidate mention, normalising the type via manual alignment from the DBpedia ontology and the NEEL taxonomy. Finally, Gârbacea *et al.* solve the NIL using a clustering algorithm operating on the lexical similarity of the candidate mentions for which no counterparts are found in DBpedia.

Basile, Caputo & Semeraro in *UNIBA: Exploiting a Distributional Semantic Model for Disambiguating and Linking Entities in Tweets*, introduce an unsupervised approach which uses a modified version of their *Lesk* algorithm. Basile *et al.* use similarity of "distributional semantic spaces" for disambiguation, and two alternative and state of the art approaches for the candidate identification phase, based on either POS tagging or n-gram similarity. Entities are typed through inheritance of the type of the DBpedia reference entity pointed to, which is in turn manually aligned to the NEEL taxonomy.

In *AMRITA - CEN@NEEL: Identification and Linking of Twitter Entities*, Barathi Ganesh, Abinaya, Anand Kumar, Soman & Vinaykumar address the NEEL task sequentially by, first, tokenising and tagging the tweets using *TwitIE*. They then classify entity mentions by applying supervised learning using direct (POS tags) and indirect features (the two words before and after a candidate mention entity). Using a total of 34 lexical features, the authors experiment with three supervised learning algorithms to determine the recognition configuration that would achieve the best performance in the development test. Barathi Ganesh *et al.* tackle the linking task by looking up DBpedia reference entries; that maximising the similarity score between related entries and the named entities is designated the representative. Named entities without related links are assigned as NIL.

Finally, Sinha & Barik, in *Named Entity Extraction and Linking in #Microposts*, present a sequential approach to the NEEL task which recognises entities and then links them. The first stage is grounded on linguistic clues extracted from conventional approaches such as POS tagging, word capitalisation and hashtag in the tweet. They then train a CRF with the linguistic features and the contextual similarity of adjacent tokens, with the token window set to 5. Priyanka & Barik perform the linking task using an entity resolution mechanism that takes as input the output of the NER stage and that of *DBpedia Spotlight*. For each entity returned from DBpedia Spotlight found to be a substring of any of the entities extracted in the NER stage and for which a substring match is found, the corresponding URI is returned and assigned to it. Otherwise the entity is assigned as NIL.

Workshop Awards

Main Track. The #Microposts2015 best paper award went to:

Cedric De Boom, Steven Van Canneyt & Bart Dhoedt
for their submission entitled:

Semantics-driven Event Clustering in Twitter Feeds

Social Sciences Track. GESIS[1], the Leibniz Institute for the Social Sciences, sponsored the best paper award for the Social Sciences track. We teamed up with GESIS, the largest service and infrastructure institution for the Social Sciences in Germany, to highlight the role of interdisciplinary approaches in obtaining a better understanding of the users behind social media and Microposts. As in the main track, the decision was guided by nominations from the reviewers and review scores. The #Microposts2015 Social Sciences Track best paper award went to:

Fabio Giglietto & Yenn Lee
for their submission entitled:

To Be or Not to Be Charlie: Twitter Hashtags as a Discourse and Counter-discourse in the Aftermath of the 2015 Charlie Hebdo Shooting in France

NEEL Challenge. SpazioDati[2], an Italian startup who took part in the #Microposts2014 NEEL challenge, sponsored the award for the best submission. SpazioDati aim to provide access to a single source of common-sense knowledge, mined and synthesised from a large number of open and closed data sources. By sponsoring the challenge, SpazioDati reinforce the value in the content of the increasingly large knowledge source that is Micropost data. The challenge award was also determined by the results of the quantitative evaluation. The #Microposts NEEL Challenge award went to:

Ikuya Yamada, Hideaki Takeda & Yoshiyasu Takefuji
for their submission entitled:

An End-to-End Entity Linking Approach for Tweets

[1] http://www.gesis.org
[2] http://spaziodati.eu

Additional Material

The call for participation and all paper, poster and challenge abstracts are available on the #Microposts2015 website[3]. The full proceedings are also available on the CEUR-WS server, as Vol-1395[4]. The gold standard for the NEEL Challenge is available for download[5].

The proceedings for #Microposts2014 are available as Vol-1141[6] The proceedings for the #MSM2013 main track are available as part of the WWW'13 Proceedings Companion[7]. The #MSM2013 Concept Extraction Challenge proceedings are published as a separate volume as CEUR Vol-1019[8], and the gold standard is available for download[9]. The proceedings for #MSM2012 and #MSM2011 are available as CEUR Vol-838[10]. and CEUR Vol-718[11], respectively.

#Microposts2015

gesis Leibniz Institute for the Social Sciences

SPAZIODATI

[3] http://www.scc.lancs.ac.uk/microposts2015
[4] #Microposts2015 Proc. http://ceur-ws.org/Vol-1395
[5] http://ceur-ws.org/Vol-1395/microposts2015_neel-challenge-report/microposts2015-neel_challenge_gs.zip
[6] #Microposts2014 Proc. http://ceur-ws.org/Vol-1141
[7] WWW'13 Companion: http://dl.acm.org/citation.cfm?id=2487788
[8] #MSM2013 CE Challenge Proc. http://ceur-ws.org/Vol-1019
[9] http://ceur-ws.org/Vol-1019/msm2013-ce_challenge_gs.zip
[10] #MSM2012 Proc. http://ceur-ws.org/Vol-838
[11] #MSM2011 Proc. http://ceur-ws.org/Vol-718

Main Track Programme Committee

Pierpaolo Basile University of Bari, Italy
Julie Birkholz CHEGG, Ghent University, Belgium
John Breslin National University of Ireland Galway, Ireland
A. Elizabeth Cano KMi, The Open University, UK
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
Andrés Garcia-Silva Universidad Politécnica de Madrid, Spain
Anna Lisa Gentile The University of Sheffield, UK
Jelena Jovanovic University of Belgrade, Serbia
Mathieu Lacage Alcméon, France
Philipe Laublet Université Paris-Sorbonne, France
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, Brazil
Danica Radovanović University of Belgrade, Serbia
Guiseppe Rizzo Eurecom, France
Harald Sack HPI, University of Potsdam, Germany
Bernhard Schandl mySugr GmbH, Austria
Sean W. M. Siqueira Universidade Federal do Estado do Rio de Janeiro, Brazil
Victoria Uren Aston Business School, UK
Andrea Varga Swiss Re, UK
Katrin Weller GESIS Leibniz Institute for the Social Sciences, Germany
Alistair Willis The Open University, UK
Ziqi Zhang The University of Sheffield, UK

Sub Reviewers

Tamara Bobic HPI, University of Potsdam, Germany

Social Sciences Track Programme Committee

Gholam R. Amin Sultan Qaboos University, Oman
Julie Birkholz CHEGG, Ghent University, Belgium
Tim Davies University of Southampton, UK
Munmun De Choudhury Georgia Tech, USA
Ali Emrouznejad Aston Business School, UK
Fabio Giglietto Università di Urbino Carlo Bo, Italy
Simon Hegelich Universität Siegen, Germany
Kim Holmberg University of Turku, Finland
Athina Karatzogianni University of Leicester, UK
José M. Morales del Castillo El Colegio de México, Mexico
Raquel Recuero Universidade Católica de Pelotas, Brazil
Bianca C. Reisdorf University of Leicester, UK
Luca Rossi Università di Urbino Carlo Bo, Italy
Saskia Vanmanen The Open University, UK
Alistair Willis The Open University, UK
Taha Yasseri University of Oxford, UK
Victoria Uren Aston Business School, UK

NEEL Challenge Evaluation Committee

Gabriele Antonelli SpazioDati, Italy
Ebrahim Bagheri Ryerson University, Canada
Pierpaolo Basile University of Bari, Italy
Grégoire Burel KMi, The Open University, UK
Óscar Corcho Universidad Politécnica de Madrid, Spain
Leon Derczynski The University of Sheffield, UK
Milan Dojchinovski Czech Technical University in Prague, Czech Republic
Guillaume Erétéo Vigiglobe, France
Andrés Garcia-Silva Universidad Politécnica de Madrid, Spain
Anna Lisa Gentile The University of Sheffield, UK
Miguel Martinez-Alvarez Signal, UK
José M. Morales del Castillo El Colegio de México, Mexico
Bernardo Pereira Nunes Pontifícia Universidade Católica do Rio de Janeiro, Brazil
Daniel Preoţiuc-Pietro University of Pennsylvania, USA
Giles Reger Otus Labs, UK
Irina Temnikova Qatar Computing Research Institute, Qatar
Victoria Uren Aston Business School, UK