PHEME: Computing Veracity — the Fourth Challenge of Big Social Data

Leon Derczynski, Kalina Bontcheva, Michal Lukasik, Thierry Declerck, Arno Scharl, Georgi Georgiev, Petya Osenova, Toms Pariente Lobo, Anna Kolliakou, Robert Stewart, Sara-Jayne Terp, Geraldine Wong, Christian Burger, Arkaitz Zubiaga, Rob Procter, and Maria Liakata

Coordinator email: K.Bontcheva@sheffield.ac.uk

The veracity of information spreading through social media can sometimes be hard to establish and the deliberate or accidental spread of false information, especially during natural disasters or emergencies, is quite common [2]. We coined the term *phemes* to describe fast spreading memes which are enhanced with truthfulness information. The PHEME project (http://www.pheme.eu) attempts to identify in real-time four kinds of phemes: controversy, speculation, misinformation and disinformation. This brings challenges in modelling the social network spread of and the online conversations around phemes; developing rumour detection methods; and using historical data to model trustworthiness of the information source.

To detect such phemes automatically, large volumes of user-generated content need to be analysed quickly, yet complex real-time analytics are a major outstanding challenge. The PHEME FP7 project aims to model, identify, and verify phemes as they spread across media, languages, and social networks. Modelling is based on the information content of the posts, the reputation of the originating information sources (if known), the information diffusion pattern, similarities to historical data, and temporal dynamics.

The rest of this abstract describes what is to be presented at ESWC.

Ontological Models of Rumours: PHEME is building a new ontology to model veracity, misinformation, social and information diffusion networks, rumours, disputed claims and temporal validity. It draws a distinction between content authors, receivers, and diffusers. This includes modelling the temporal validity of statements (e.g. Lenin was born in the Soviet Union vs. in Russia) and lexicalisations (e.g. Kaliningrad vs. Konigsberg), based on work on adding temporal arguments to RDF triples.

New Rumour Annotated Datasets: In the first project year we also annotated sets of tweets [3] pertaining to major events, such as the Ferguson unrest, a shooting in Ottawa, Charlie Hebdo, etc. Journalists from SwissInfo chose those events, since amongst the factually correct tweets there were many rumours too. A new annotation scheme was designed and first piloted with journalists and other users within the project. In order to scale up the annotation, we also used paid-for crowdsourcing.

Detecting Misinformation: Being able to detect which tweets agree, disagree, or question a given pheme can reveal important information regarding a rumour. For instance, false rumours tend to have significantly larger percentage

of tweets rejecting the rumour than true stories. Consequently automatic estimation of collective trust in a circulating pheme can be a good proxy for its truthfulness. Since rumours have some shared and some individual characteristics, both supervised and unsupervised domain adaptation is considered and rumour predictions are made on the basis of other annotated rumours. Features are derived from the information content of the individual tweets, user profiles, and diffusion patterns.

Trustworthiness and Rumour Diffusion: Information diffusion plays a crucial role in a range of phenomena, including the spread of rumours within and between social networks. Tracking information flow in implicit networks is a more challenging task because it involves: identifying identical information units; determining when this information was published; tracking the flow within this network; and inferring the implicit diffusion network. Implicit networks created by dialogue can also be found over explicit social networks such as Twitter. For example, those contributing to a hashtag conversation are not bound by explicit links such as follower or friend relations, but instead cause information to diffuse among those interested in that topic.

In the second project year PHEME will also be working on models of trustworthiness of the information source (e.g. tweet authors), based on historical data, as well as account-specific information (e.g. user profile description, age of the account, social connections). User privacy will be respected, firstly by using only public social media posts and profile information, and secondly, by not making these individual models public, but using them in aggregate as features for the veracity classification algorithms.

Applications: PHEME is prototyping an open-source digital journalism tool, to support the cross-linking, verification, analysis, and visualisation of veracity, operating across media and languages. This will be based on the open-source Ushahidi platform. In such context, spatio-temporal knowledge plays an important role [1]. A key challenge is to identify the regionality of events (e.g., neighbourhood, city, or country level).

The second use case is health-related, e.g. identifying controversies around certain drugs or treatments or spread of medical misinformation.

Acknowledgements This project has received funding from the European Unions Seventh Framework Programme for research, technological development and demonstration under grant agreement No. 611233, PHEME.

References

1. Derczynski, L., Bontcheva, K.: Spatio-temporal grounding of claims made on the web, in Pheme. In: Proceedings of the 10th joint ACL-ISO workshop on Interoperable Semantic Annotation. ACL (2014)
2. Procter, R., Crump, J., Karstedt, S., Voss, A., Cantijoch, M.: Reading the riots: What were the police doing on twitter? Policing and society 23(4), 413–436 (2013)
3. Zubiaga, A., Liakata, M., Procter, R., Bontcheva, K., Tolmie, P.: Crowdsourcing the annotation of rumourous conversations in social media. In: WWW 2015 Companion (2015)