Helping Social Scientists to Collect and Understand Social Media Data

Jeff Z. Pan[1], Stephen H. Muggleton[2], and John Paul Vargheese[1]

[1]University of Aberdeen, Aberdeen, UK
[2]Imperial College London, London, UK

The widespread use of social media provides significant opportunities for social scientists to discover novel insights of human behaviour. As the volume of individual users of social media increases, the need for organisations to maintain an active presence across the diverse range of social media channels also increases, due to the potential to communicate and address as broader and diverse audience as possible [9, 7].

In response to increasing interest and research in this area, an increasing number of tools [5, 12, 4, 1] and theoretical frameworks [8, 11, 13, 6] have been developed, to assist with capturing, analysing and understanding social media data. For example, the honeycomb framework aims to support understanding of the operational and functional aspects of social media and defines seven properties of social media as: Presence, relationships, reputation, groups, conversations, sharing and identity [8].

However, tools for collecting and analysing social media data are often inaccessible or unsuitable for social scientists. Results from our previous investigation of social scientists' experience and usage of tools for collecting and analysing social media data, highlighted a negative perception of automated analytical techniques. This is often due to interdisciplinary challenges that conflict with social scientists' research aims, objectives and methodological approaches towards collecting and analysing social media, as well as lacking reflectivity required for interpretation.

For example, these issues were emphasised by two of the study participants as follows:

'...Yes, the machine will do a wonderful job of counting ... So we know how many followers you have, we know how many people had this hashtag...Did we recognise the sarcasm? No, we didn't.'

'...I feel it doesn't pick up on sarcasm and irony and so on...I think you really needed that manual, the human knowledge.'

Overall, our study indicated that the main challenge faced by social scientists who participated in our study, was centred upon the means of capturing social media data [2]. The negative perception of automated analysis is due to the lack of mutual understanding between machine and users. Our proposed solution [2] is to use knowledge graphs as common grounds between social scientists and computing tools. We have developed two tools accordingly. The first tool provides a means of capturing tweets from Twitter and assigning user defined manual annotations or thematic codes to the content. These codes are embedded within the social media data captured to produce a knowledge graph. The second tool provides a means of visualising the knowledge graph created from the capture tool and provides controls for users to modify the visualisations available. In this way, we aim to bridge the gap between the human driven interpretation,

knowledge and understanding of social media content and increase accessibility to automated analytical techniques capable of exploiting the rich interconnected attributes that exist within the content captured for analysis.

We propose a four stage process for identifying relevant initial schema for knowledge graphs generated from our capture tool, driven by social scientists interpretation of social media data. Firstly, it is necessary to identify relevant sources and channels that are most suited for acquiring content with regards to social scientists' research aims and objectives. This is comparable to the Initiation phase of the social strategy cone [3]. It is important to note that this is an iterative and ongoing stage of the process as certain channels and sources may yield more relevant content than others. Secondly, understanding the structure of the social media data captured is required in order to identify what entities can be extracted from a sample. At this stage, it may be possible to incorporate elements from the honeycomb or other theoretical frameworks for analysing social media, with a view to providing an initial structuring of the knowledge graph. For example, associating categories of data, such as favourites, likes, retweets and shares with operational functions such as sharing, presence, relationships, reputation and identity [8, 2]. Thirdly, users select which entities to extract from a sample for analysis. For example these may consist of individuals, events, locations and organisations which further structure the knowledge graph. These may be user defined or predetermined theoretical elements that describe the structure and content of a sample. Finally, establishing relations amongst the extracted entities in order to identify relevant topics, agendas and arguments contained within the social media data captured. We anticipate this human-like approach to constructing a social media knowledge graph will provide a means of creating a user defined and structured knowledge graph that may queried using advanced reasoning and analytical techniques. At each stage of the process, users may assign manual annotations or thematic codes to the content acquired so that the users interpretation and understanding of each stage of the process is taken into account.

Although we have received positive feedbacks from social scientists about our approach and tools, we will need to address two further issues, so as to enable machines to consume the produced knowledge graphs.

Firstly, we will need to learn more detailed schema for such knowledge graphs. Our plan is to try Metagol, an inductive logic programming (ILP) system based on the meta-interpretive learning framework (MIL) [15]. MIL is a form of ILP based on an adapted Prolog meta- interpreter. A standard Prolog meta-interpreter proves goals by repeatedly fetching first-order clauses whose heads unify with the goal. By contrast, a MIL learner proves goals by fetching higher-order metarules whose heads unify with the goal. The resulting meta-substitutions are saved, allowing them to be used as background knowledge by substituting them into corresponding metarules.

Secondly, in order to support some advanced reasoning, queries and analysis, we plan to exploit some novel and faithful approximate ontology reasoning techniques [10]. Approximate reasoning has been very popular for supporting the W3C standard Web Ontology Language OWL (version 2). The idea here is to approximate OWL 2 ontologies to those in its tractable sub-languages, so as to exploit more efficient and scalable reasoning algorighms. A successful example of approximate reasoner is the TrOWL

reasoner [14, 10], which outperformed some well known sound and complete OWL reasoners in time-constrained sound-and-complete OWL Reasoner Evaluations.

Acknowledgements This research has been funded by the UK Economic and Social Research Council grant reference ES/MOO1628/1 and the EU Marie Curie IAPP K-Drive project (286348).

References

1. N. Cao, L. Lu, Y.-R. Lin, F. Wang, and Z. Wen. Socialhelix: visual analysis of sentiment divergence in social media. *Journal of Visualization*, 18(2):221–235, 2015.
2. A. Cropper and S. Muggleton. Learning higher-order logic programs through abstraction and invention. In *IJCAI 2016*, pages 1418–1424, 2016.
3. R. Effing and T. A. Spil. The social strategy cone: Towards a framework for evaluating social media strategies. *International journal of information management*, 36(1):1–8, 2016.
4. A. Guille, C. Favre, H. Hacid, and D. A. Zighed. Sondy: An open source platform for social dynamics mining and analysis. In *Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data*, pages 1005–1008. ACM, 2013.
5. D. J. Hopkins and G. King. A method of automated nonparametric content analysis for social science. *American Journal of Political Science*, 54(1):229–247, 2010.
6. N. Jaafar, M. Al-Jadaan, R. Alnutaifi, et al. Framework for social media big data quality analysis. In *New Trends in Database and Information Systems II*, pages 301–314. Springer, 2015.
7. S. Kemp. Digital, social & mobile worldwide in 2015. *We are social*, 2015.
8. J. H. Kietzmann, K. Hermkens, I. P. McCarthy, and B. S. Silvestre. Social media? get serious! understanding the functional building blocks of social media. *Business horizons*, 54(3):241–251, 2011.
9. K. Larson and R. T. Watson. Tying social media strategy to firm performance: A social media analytics framework. In *ICIS*, 2011.
10. J. Z. Pan, Y. Ren, and Y. Zhao. Tractable approximate deduction for OWL. *Arteficial Intelligence*, 235:95–155, 2016.
11. K. Peters, Y. Chen, A. M. Kaplan, B. Ognibeni, and K. Pauwels. Social media metrics?a framework and guidelines for managing social media. *Journal of Interactive Marketing*, 27(4):281–298, 2013.
12. B. Settles. Closing the loop: Fast, interactive semi-supervised annotation with queries on features and instances. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1467–1478. Association for Computational Linguistics, 2011.
13. S. Stieglitz and L. Dang-Xuan. Social media and political communication: a social media analytics framework. *Social Network Analysis and Mining*, 3(4):1277–1291, 2013.
14. E. Thomas, J. Z. Pan, and Y. Ren. TrOWL: Tractable OWL 2 Reasoning Infrastructure. In *the Proc. of the Extended Semantic Web Conference (ESWC2010)*, 2010.
15. J. P. Vargheese, P. Travers, J. Z. Pan, K. Vincent, C. Wallace, and A. Kabedeva. Constructing Social Media Knowledge Graphs with Social Scientists. In *HCI 2016*, 2016.