http://dx.doi.org/10.14236/ewic/HCI2016.65

Constructing Social Media Knowledge Graphs with Social Scientists

John Paul Vargheese, Peter Travers, Jeff Pan
Computing Science
University of Aberdeen
Kings College
Aberdeen, AB24 5UA, UK
jpvargheese@acm.org
peter.travers@abdn.ac.uk
jeff.z.pan@abdn.ac.uk

Kathryn Vincent, Claire Wallace, Anna Kabedeva
School of Social Science
University of Aberdeen
Kings College
Aberdeen, AB24 3QY, UK
k.l.vincent@abdn.ac.uk
claire.wallace@abdn.ac.uk
anna.kabedeva.11@aberdeen.ac.uk

The increasing adoption and widespread use of social media provides significant opportunities for social scientists to discover novel insights of varying aspects of human behaviour. In response to increasing interest and research in this area, a wide variety of theoretical, methodological frameworks, guidelines and software tools have emerged. However, tools for collecting and analysing social media data are often inaccessible or unsuitable for social scientists. This is often due to interdisciplinary challenges that conflict with social scientists' research aims, objectives and methodological approaches towards collecting and analysing social media. To address this, we are developing an extensible open source platform to support social scientists' research in this area. This platform provides the means to collect and annotate social media data which can then be used to construct a knowledge graph. The knowledge graph provides social scientists with the means to consider their analysis within a broader context that may yield further insights.

social scientists, social media data, knowledge graph

1. INTRODUCTION

Social media may be defined as an internet-based means of facilitating dynamic social interactions through the creation and distribution of user generated content, widely accessible and unconstrained by location (Kaplan and Haenlein 2010). The diverse range of these social interactions include but are not limited to social, personal, political expression, commentaries and alignments in addition to the varying factors that may affect them, such as personality, political opinions, reactions and personal relationships (Shah et al. 2015; Golbeck et al. 2011). What distinguishes social media from other forms of media is the dynamic and egalitarian nature of the social interactions that take place which are typically unconstrained by any single group or individual (Peters et al. 2013). While this increases the richness and volume of data available for study, it also raises a number of challenges including how to truly consider the representativeness of a sample and how this may impact upon analysis, interpretation of results and the overall validity of any conclusions derived from such studies (Kietzmann et al. 2011; Shah et al. 2015). To address this, a number of software tools (Hopkins and King 2010; Settles 2011; Guille et al. 2013; Cao et al. 2015), theoretical and methodological frameworks have emerged (Kietzmann et al. 2011; Peters et al. 2013; Stieglitz and Dang-Xuan 2013; Jaafar et al. 2015), intended to support social media research. Many tools share common features including the facility to capture social media data from different sources and incorporate a range of analytical methods. These vary from providing the means of manual annotation or thematic coding of content to more automated means of analysis such as natural language processing and machine learning.

2. MOTIVATION

Despite the wide range of tools, theoretical and methodological frameworks available, not all are accessible, known to, suitable or are used by social scientists working with social media. This is often because existing methods for exploiting the richness of social media available, lack the reflexivity required by social scientists to interpret and analyse social media from a qualitative research perspective. Analytical techniques such as machine learning and natural language processing that may enhance social scientists' research with social media are often perceived negatively and in conflict with their research aims, objectives and methodology. Furthermore, increasing monetisation of social media data provides further challenges and obstacles for social scientists acquiring data for analysis (Batrinca and Treleaven 2015). To address these challenges,

we are developing an open source extensible platform that provides social scientists with the means of collecting and annotating social media data, to produce a knowledge graph. The knowledge graph complements the traditional social science approach towards qualitative analysis by highlighting relationships within a sample thematically coded by the user.

3. APPROACH

We conducted a semi-structured interview study with 12 social scientists to discover what tools are currently used to capture and analyse social media data. The results of this study indicate that the primary challenge for the participants was how to capture social media data. Current methods involve manually copying and pasting into either Word or Excel documents with analysis conducting using a thematic coding approach. Other methods involved the provision of a summary of all content distributed by a particular social media channel which would also be analysed via a thematic coding approach. Three participants we interviewed used a custom designed data capture tool which used API access to download social media data to an Excel file. However, participants expressed frustration at a lack of control as to how the data collected was specified. Furthermore, these participants argued that this approach conflicted with their research aims, objectives and methodological approach. Although none of the participants applied any theoretical, methodological framework or guidelines in how social media was captured and analysed, participants drew upon a range of theories within the social sciences for how they interpreted the content acquired.

4. KNOWLEDGE GRAPH: WHAT AND WHY

A knowledge graph consists of a collated group of interconnected entities with rich attributes (Singhal 2012). For example, attributes associated with an entity such as a social media user's post may include demographic data such as age, nationality, current location in addition to more rich data such as followers, friends and likes. Knowledge is derived by establishing facts from meaningful relations amongst entities which are highlighted in the knowledge graph. The inferred knowledge graph produced from the users' thematic analysis of the sample may include concepts from existing social media analytics frameworks. For example, the Honeycomb framework defines seven components of social media. These components include presence, relationships, reputation, groups, conversations, sharing and identity (Kietzmann et al. 2011). These may be used to define relations within

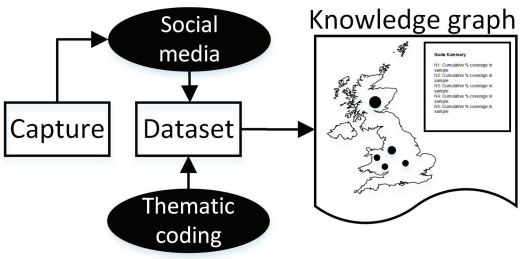

Figure 1: *Summary of knowledge graph creation*

a sample that has been annotated or assigned to various themes by a social scientist. Deriving knowledge from social media in this manner has the advantage of allowing social scientists to apply their own interpretation and analysis through thematic coding and further reflect upon these findings within a wider context such as those specified by the Honeycomb framework (Kietzmann et al. 2011). For example, consider the scenario whereby a social scientist (user) aims to investigate how narratives are developed by a political party through content (Tweets) posted on Twitter. Using our tool, the user collects a dataset which consists of all Tweets posted by a political party. The dataset also includes the number of followers for the political party's Twitter account and for each Tweet in the dataset, the number of retweets, favourites (likes), shares and what hashtags, mentions, links and geolocation tags were included. The user analyses the Tweets through a thematic coding approach. For example, the following tweet: *"Politician X has been involved in corrupt and illicit acts!"*, is assigned the following themes: *"accusation; corruption; rumours; opponent"* by the user. Upon completion of the analysis, a subsequent knowledge graph produced examines potential relationships between the thematic analysis of the Tweets by the user and associated attributes for each coded Tweet and retweet. The knowledge graph indicates that Tweets coded with the theme *Corruption* are significantly popular amongst users located within location *L*. This is inferred by examining the volume of retweets, likes and shares from other locations for Tweets assigned with the code *Corruption* and comparing these to those assigned with the same theme, retweeted, liked and shared at location *L*. The user may conclude that location *L* will become a focal point for the political party and potentially other opponents, given the influence of narratives posted through Twitter and their impact at location *L* (see Figure 1).

5. DISCUSSION AND FUTURE WORK

We have developed an open source extensible platform to support social scientists with capturing and

analysing social media. This platform incorporates a thematic coding feature from which a knowledge graph is produced to enable social scientists to exploit both the richness and volume of social media data. Due to the limited number of participants involved in our initial study, we are continuing to conduct interview studies with social scientists to verify and refine our initial requirements specification. We are also in the process of evaluating the platform with social scientists to further cross validate and verify our findings with a view to exploring the potential for incorporating existing social media analytical frameworks for structuring the knowledge graph. We anticipate that this work will allow us to extend existing social media analytical frameworks and help social scientists overcome existing interdisciplinary challenges towards research with social media.

This work was funded by a grant to the University of Aberdeen by the UK Economic and Social Research Council grant reference ES/MOO1628/1.

REFERENCES

Batrinca, B. and P. C. Treleaven (2015). Social media analytics: a survey of techniques, tools and platforms. *AI & SOCIETY 30*(1), 89–116.

Cao, N., L. Lu, Y.-R. Lin, F. Wang, and Z. Wen (2015). Socialhelix: visual analysis of sentiment divergence in social media. *Journal of Visualization 18*(2), 221–235.

Golbeck, J., C. Robles, and K. Turner (2011). Predicting personality with social media. In *CHI'11 Extended Abstracts on Human Factors in Computing Systems*, pp. 253–262. ACM.

Guille, A., C. Favre, H. Hacid, and D. A. Zighed (2013). Sondy: An open source platform for social dynamics mining and analysis. In *Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data*, pp. 1005–1008. ACM.

Hopkins, D. J. and G. King (2010). A method of automated nonparametric content analysis for social science. *American Journal of Political Science 54*(1), 229–247.

Jaafar, N., M. Al-Jadaan, R. Alnutaifi, et al. (2015). Framework for social media big data quality analysis. In *New Trends in Database and Information Systems II*, pp. 301–314. Springer.

Kaplan, A. M. and M. Haenlein (2010). Users of the world, unite! the challenges and opportunities of social media. *Business horizons 53*(1), 59–68.

Kietzmann, J. H., K. Hermkens, I. P. McCarthy, and B. S. Silvestre (2011). Social media? get serious! understanding the functional building blocks of social media. *Business horizons 54*(3), 241–251.

Peters, K., Y. Chen, A. M. Kaplan, B. Ognibeni, and K. Pauwels (2013). Social media metrics?a framework and guidelines for managing social media. *Journal of Interactive Marketing 27*(4), 281–298.

Settles, B. (2011). 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*, pp. 1467–1478. Association for Computational Linguistics.

Shah, D. V., J. N. Cappella, and W. R. Neuman (2015). Big data, digital media, and computational social science possibilities and perils. *The ANNALS of the American Academy of Political and Social Science 659*(1), 6–13.

Singhal, A. (2012). Introducing the knowledge graph: things, not strings. *Official Google Blog, May Official Google Blog, May*.

Stieglitz, S. and L. Dang-Xuan (2013). Social media and political communication: a social media analytics framework. *Social Network Analysis and Mining 3*(4), 1277–1291.