Online Social Networks and Media 1 (2017) 33–43

Contents lists available at ScienceDirect

Online Social Networks and Media

journal homepage: www.elsevier.com/locate/osnem

Bridging big data and qualitative methods in the social sciences: A case study of Twitter responses to high profile deaths by suicide

Dmytro Karamshuk[a,*], Frances Shaw[b], Julie Brownlie[b,*], Nishanth Sastry[a,*]

[a] King's College London, London WC2R 2LS, UK
[b] University of Edinburgh, Edinburgh EH8 9JU, UK

ARTICLE INFO

Article history:
Received 14 December 2016
Revised 23 January 2017
Accepted 25 January 2017

Keywords:
Social media
Crowd-sourcing
Crowdflower
Natural language processing
Social science
Emotional distress
High-profile suicides
Public empathy

ABSTRACT

With the rise of social media, a vast amount of new primary research material has become available to social scientists, but the sheer volume and variety of this make it difficult to access through the traditional approaches: close reading and nuanced interpretations of manual qualitative coding and analysis. This paper sets out to bridge the gap by developing semi-automated replacements for manual coding through a mixture of crowdsourcing and machine learning, seeded by the development of a careful manual coding scheme from a small sample of data. To show the promise of this approach, we attempt to create a nuanced categorisation of responses on Twitter to several recent high profile deaths by suicide. Through these, we show that it is possible to code automatically across a large dataset to a high degree of accuracy (71%), and discuss the broader possibilities and pitfalls of using Big Data methods for Social Science.

1. Introduction

Social science has always had to find ways of moving between the small-scale, interpretative concerns of qualitative research and the large-scale, often predictive concerns of the quantitative. The quantitative end of that spectrum has traditionally had two interrelated features: active collection of data and creating a suitable sub-sample of the wider population. To the extent that such methods have also captured open-ended or qualitative data, the solution has been to apply manual coding, using a frame developed on the back of intensive qualitative analysis or an exhaustive coding of a smaller sample of responses. Although labour-intensive, manual coding has been critical for obtaining a nuanced understanding of complex social issues.

Social media has created vast amounts of potential qualitative research material – in the form of the observations and utterances of its population of users – that social scientists cannot ignore. Unlike the responses to survey questions, such material is not elicited as part of the research process, nor is its volume limited by the constraints and practicalities of the sample survey. With social media, we now have so much information that it is impossible to process everything using either the detailed analysis methods of qualitative research or the application of manual coding approaches of the kind used in survey research. In short, there are exciting new possibilities but also significant challenges.

For instance, when celebrities die, or deaths become politicised or public in some fashion, hundreds of thousands or even millions of tweets may result. How can some of the traditional concerns of social science – with interpretation (nuance), meaning and social relationships – be pursued within this deluge of largely decontextualised communication? Whereas Big Data methods can easily count the number of tweets, or even attach a 'sentiment score' to individual tweets, it is less clear whether existing methods can identify issues such as the presence of or lack of empathy. And yet the application of traditional methods from qualitative social science, such as the close analysis of a small-scale sample of tweets relating to a public death, or the manual application of a coding frame to a larger volume of responses, are likely to miss crucial insights relating to the volume, patterning or dynamics. We therefore need a mechanism to train the social scientists' close lens on unmanageably large datasets – to bridge the gap between close readings and large scale patterning.

This paper develops a possible approach, that we term semi-automated coding: Our three-step method first manually bootstraps a coding scheme from a micro-scale sample of data, then uses a crowdsourcing platform to achieve a meso-scale model, and finally applies machine learning to build a macro-scale model.

[*] Corresponding authors.
E-mail addresses: Dmytro.Karamshuk@kcl.ac.uk (D. Karamshuk), FrancesShaw@gmail.com (F. Shaw), Julie.Brownlie@ed.ac.uk (J. Brownlie), Nishanth.Sastry@kcl.ac.uk (N. Sastry).

http://dx.doi.org/10.1016/j.osnem.2017.01.002

The bootstrapping is carefully done by trained researchers, creating the nuanced coding scheme necessary for answering social science questions, and providing an initial 'golden set' of labelled data. Crowdsourcing expands the labels to a larger dataset using untrained workers. The quality of crowd-generated labels is ensured by checking agreement among crowdworkers and between the crowd workers' labels and the golden set. This larger labelled dataset is then used to train a supervised machine learning model that automatically labels the entire dataset.

We argue that this approach has particular potential for the study of emotions at scale. Emotions have a mutable quality [1] and this is especially true in the context of social media. Thus, intensive manual coding over a small-scale sample may miss some of the temporal and volume dynamics that would be critical for a full sociological understanding of public expressions of emotion, in contrast to the semi-automated coding we propose here, which captures the entire dataset and its dynamics.

As a case study in applying semi-automated coding, this paper looks at public empathy – the expression of empathy that, even if it is imagined to be directed at one other person [2], can potentially be read by many – in the context of high-profile deaths by suicide. Five cases were chosen which had a high rate of public response on Twitter, with the aim of exploring what types of response were more or less common in the space of public Twitter, and what factors might affect these responses.

This paper primarily focuses on the methodological challenges of this research through an engagement with emergent findings and concludes by considering its potential use for interdisciplinary computational social science. A key issue, both within the case study, and more generally, for the success of semi-automated coding as an approach, is the accuracy of the automatically generated labels. One source of error is the quality of crowd-generated labels. As mentioned above, we control for this using different forms of agreement, among crowd workers, and with a curated golden set. However, our initial attempts on Crowdflower did not generate a good level of agreement. On closer analysis, we discovered that the crowdworkers were confused by the nuanced classification expected of them. To help them, we developed a second innovation, giving them a decision tree (Fig. 1) to guide their coding. This resulted in around 60% of tweets with agreement. Our tests show that the final machine generated labels agree with the crowd labels with an accuracy of 71%, which permits nuanced interpretations. Although this is over 5.6x times the accuracy of random baseline, we still need to reconcile the social side of research interpretations with the potentially faulty automatic classification. We allow for this by explicitly quantifying the errors in each of the labels, and drawing interpretations that still stand despite a margin of safety corresponding to these errors.

2. Related literature

The transformative potential of Big Data for social science is now widely recognised, [3,4] with social and emotional phenomena ranging from suicidal expression [5] and cyber hate [6] investigated through computational social scientific approaches. However, epistemological and methodological challenges [7,8] remain, and there is an active debate about several aspects of the use of Big Data methods in social science. One critical question is whether and how Big Data methods can scale up from small samples to big data in relation to complex social practices that may require close analysis and nuanced interpretation.

Our proposed solution for scaling up is to automate some of the manual research process involved in social science coding practices. Although previous efforts have looked at assisting social science through automated coding of dates and events in data [9] and even open-ended survey responses [10], coding of social media-data creates new challenges because of its temporality and breadth (unlike, for example, survey data which tends to be in response to specific questions). The main contribution of this paper is the proposed methodology, mixing machine-learning and crowd-sourcing, and using multiple levels of validation and refinement, to achieve a high degree of accuracy in coding nuanced concepts such as mourning and lack of empathy.

The practice of employing crowd-workers to manually label tweets has a short but rich history. *Crowdsourcing* has been recognised as a valuable research tool in numerous previous works [11–15]. A comprehensive review of this literature has been provided in [13] which – among others – recognises the impact of the *job design* on the efficiency of crowd-computations. For instance, Willett et al. in [15] describe a crowd-sourcing design for collecting surprising information in charts, [14] propose a design for online performance evaluations of user interfaces, etc. Our paper contributes to this body of work by proposing a *decision tree-based design* for crowd-sourcing typologies of social-media posts with built-in prioritisation of the coding process to meet the aims of the social inquiry being carried out.

Last, but not least, the methods developed here build on recent advances in applying artificial neural networks for *natural language processing* of short texts [16]. Specifically, we investigate how to adapt this approach for automating nuanced multivariate classification of public mourning related social media posts.

The underlying social science research is informed by work in social science and media studies on *public mourning and grieving*, particularly on social media. Previous studies have, for example, looked at the discussion of death and grief on Twitter following a violent tragedy [17]. Social media responses to the deaths of celebrities, and to deaths that have received public attention for other reasons, have also been examined [18–20]. Whereas previous studies have looked at communal grief and individual mourning in untimely deaths such as that of Michael Jackson [18,21], this paper aims to interrogate discourses and practices around suicide in mediated mourning, an area in which there has been much less of a focus to date.

3. Background and approach

As mentioned, we use the study of public expression of empathy in the face of high-profile suicides as a case study for testing the feasibility of semi-automated coding. Below we first describe the suicides we study, and the datasets that we examine relating to these deaths. Then we outline our philosophy and approach to developing semi-automated coding.

3.1. Datasets

To analyse public discourses on social media relating to high-profile death by suicides, we chose five such deaths which were highly publicised, either because the person was famous before their death or because of the circumstances of their death. We were interested in the range of reactions, from mourning and tributes, to activism and actions, that were elicited in public Twitter conversations relating to these deaths. Below, we provide some context about each of the cases:

1. Aaron Swartz, at the time of his death by suicide in 2013, was under federal indictment for data theft, relating to an action he undertook to automatically download academic journal articles from the online database JSTOR at MIT. Prosecutors and MIT were criticised by his family and others after his death. Some critics engaged in hacktivist activities, others suggested the federal prosecutors had engaged in