Available online at www.sciencedirect.com

ScienceDirect

Online Social Networks and Media 01 (2017) 1–17

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, name.surname@kcl.ac.uk
[b]University of Edinburgh, franceshaw@gmail.com and julie.brownlie@ed.ac.uk

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.

Keywords: social media; crowd-sourcing; crowdflower; natural language processing; social science; emotional distress; high-profile suicides; public empathy

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 inter-related 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. 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 labeled 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. Crowd-sourcing has been recognized as a valuable research tool in numerous previous works [11][12][13] [14][15]. A comprehensive review of this literature has been provided in [13] which - among others - recognizes the impact of the job design on the efficiency of crowd-computations. For instance, Willett et al˙ in [15] describes a crowd-sourcing design for collecting surprising information in charts, [14] proposes 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 multi-variate 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] [19] [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 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 death:

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 bullying, with Swartz's activism argued to have played a role in his treatment [1].

[1] https://en.wikipedia.org/wiki/Aaron_Swartz

Case Study	From	Size	Sampled
Amanda Todd	2012-10-11	553,664	full
Leelah Alcorn	2014-12-30	390,561	full
Charlotte Dawson	2014-02-22	40,149	full
Robin Williams	2014-08-11	749,422	sampled
Aaron Swartz	2013-01-12	84,126	sampled

Table 1: Description of the Case Studies and Datasets. All datasets consist of tweets in English language for the first 20 days from the date indicated in the table.

2. Amanda Todd died by suicide at the age of 15 in 2012 in British Columbia, Canada. Her death was widely publicised as a result of a video detailing her experiences of cyberbullying which she had published on YouTube, and which went viral following her death, accumulating more than 1.6 million views in three days. Part of her cyberbullying experience was the abusive and ongoing sharing of images of her without her consent. An adult male was implicated in this abuse [2].
3. Charlotte Dawson was a New Zealand-Australian television personality, and former model, most famous for her roles on Australia's Next Top Model, New Zealand Getaway, and The Contender Australia. She was heavily involved in social media, and was a target of cyberbullying for several years prior to her death, with one incident in 2012 occurring around the same time as a previous suicide attempt. She died by suicide in 2014, aged 47. Prior to her death she was an ambassador against cyberbullying [3].
4. Leela Alcorn was an American transgender girl whose parents had reportedly refused to accept her female gender identity and sent her to Christian-based conversion therapy. Her suicide note, posted on Tumblr, attracted wide attention. Since her death, Alcorn's parents have been strongly criticised. Vigils and other activist events have taken place internationally to commemorate her life [4].
5. Robin Williams was a very well-known Hollywood actor and comedian. His suicide attracted an enormous amount of commentary from fans online. At the time of his death he had reportedly been suffering from severe depression and had recently been diagnosed with early-stage Parkinson's disease [5].

We collected five datasets of related Twitter posts for 20 days following each death. We were able to obtain the full dataset of tweets for three deaths (Amanda Todd, Leelah Alcorn and Charlotte Dawson) and sampled datasets for the remaining two (Robin Williams and Aaron Swartz). The number of tweets across the different cases ranged from 40K (Charlotte Dawson) to 749K (Robin Williams) and constituted a total of 1.8M tweets. The datasets are summarised in Table 1.

3.2. Analysis approach: semi-automated coding

For each of these deaths by suicide, from a social science perspective, we were interested in understanding the types of responses that were elicited on public conversations, a question that would traditionally be answered through manual coding, or classification of the responses through a frame developed after intensive qualitative analysis. Although coding has been a mainstay of social science research, it becomes difficult to apply this at scale given the volume of Tweets in Table 1[6]. The social scientist's typical alternative would be to select and focus on a small sample of the dataset. Unfortunately, this is not a fully satisfactory solution for two reasons: First, it is not a priori clear which parts of the dataset would be most interesting and should be selected for intensive analysis. Second, focusing on a small sample misses aggregate characteristics, such as the relative volumes and temporal dynamics of different classes of responses, which can provide a new

[2] https://en.wikipedia.org/wiki/Suicide_of_Amanda_Todd
[3] https://en.wikipedia.org/wiki/Charlotte_Dawson
[4] https://en.wikipedia.org/wiki/Death_of_Leelah_Alcorn
[5] https://en.wikipedia.org/wiki/Robin_Williams
[6] For comparison, analysing a sample of ≈ 200 tweets to develop an initial coding frame (Step 1 below) was an ≈ 1 person-day job.

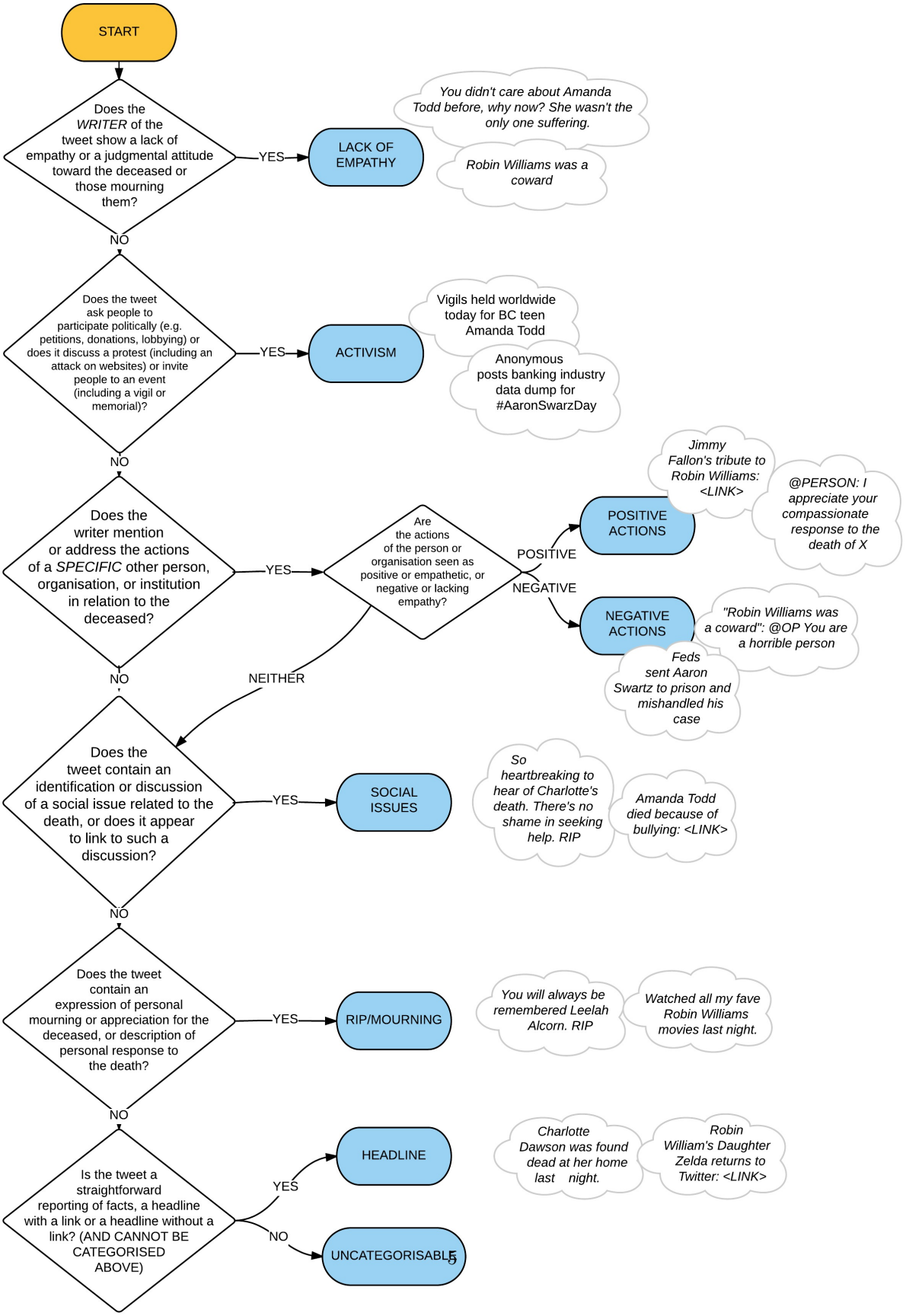

Figure 1: CrowdFlower job designed as a Decision Tree. CrowdFlower workers were asked to follow a sequence of binary decisions from the Decision Tree to label each tweet.

dimension to many social science questions, including ours, as it focuses on public, or aggregate, expressions of empathy.

We argue therefore that manual coding needs to be adapted using computational methods, to scale up the volumes of data created by social media platforms. Our solution, which we term semi-automated coding, works as follows: We start by noting that manual inspection cannot be avoided, because a) social scientists need to come up with a coding frame that makes sense for the research questions that are of interest and b) given that the classes of interest encompass nuanced, higher-order human- and social-interaction concepts, it is easiest to define these by example rather than develop complicated rules or heuristics that can identify tweets belonging to the class. Therefore, as a first step, researchers can identify the concepts/classes of interest, and provide examples. Subsequently, our goal would be to build a machine learning model that can learn these concepts based on the examples given.

In order for the above approach to work, we need two refinements: First, the machine learning model needs a sufficient number of labelled examples. This can still be difficult due to the labour-intensive nature of coding. Therefore, we adopt a two step approach to generate examples: First, trained researchers create a coding frame and a carefully curated set of example tweets. Next, an untrained set of workers on a crowd-sourcing platform is used to label a larger set of tweets, and controlling for the quality of labelling using agreement between crowd-workers, and agreement between crowd-workers' labels and the labels associated by researchers for the curated set of tweets.

Second, as with any application of machine learning, the automatically generated set of labels is bound to have a few errors. These should be taken into account in any large-scale analysis based on semi-automated coding. We observe that if we are able to quantify the extent of the errors, we can reason about the validity of results within a margin of safety, and ensure that the sociological insights drawn stand despite any shortcomings of the model.

4. Bootstrapping coding using manual effort

The main social science objective of the study was to analyze the typology and dynamics of messages on public Twitter following a high profile death by suicide. To tackle scalability issues, we designed a hybrid methodology in which our coding typology was applied manually and gradually on different data scales. We began by manually coding a few hundred tweets which were subsequently used to guide the execution of a large-scale labeling experiment on Crowdflower, a crowd-sourcing platform. We then used twelve thousand labeled tweets obtained from the results of the CrowdFlower experiment to train a state-of-the-art machine learning algorithm for short text analysis and to automatically label the full dataset (discussed in §5). Below, we describe the design of the manual part, and how it feeds in to bootstrap the machine learning model.

4.1. Coding typology using trained researchers

Initially, a random sample of 200 tweets from each of the five cases was coded qualitatively to identify patterns in communication, a method building on previous Twitter research that divides tweets according to content [22] [23] [24] [25]. The initial coding frame was developed from this subset of the dataset.

To begin with, we made lists of all content types emerging from the dataset, made observations on which content types were most common, and found ways to differentiate appropriately between tweet types based on emotional content (blame vs. grief, for example) and whom the tweet was directed at (other Twitter users, the deceased, certain people in particular, or society in general). Our coding frame was then inductively and iteratively developed using cross-validation between two coders, in a manner consistent with previous studies examining emotional content in online settings [25].

These aspects of tweets (emotional content, and to whom the emotions were directed) most strongly shaped the codes chosen. The reason for this focus was that the coding of these tweets was shaped by our research interest in empathy as a concept and as a social practice within the dataset, so we paid particular attention to tweets that either displayed empathic feeling or a lack of empathy toward the deceased or those mourning them. We also identified other strongly apparent communicative practices in the dataset. Such an approach to developing a coding frame requires analytical insight (as opposed to empirical knowledge) about the potential and likely feelings of tweeters, and the diversity of responses within the dataset.

Characteristic	Exp #1	Exp #2
Tweets Labeled	$\approx 2K$	10K
Test Questions	64	64
Judgments per Tweet	2	2
Speed vs Quality	quality	speed
Workers Quality Threshold	66%	65%
Number of Selected Workers	13	61
Selected Workers Quality	82%	78%
Workers Agreement	67%	59%
Workers Feedback	3.1/5	3.6/5

Table 2: The Summary of the CrowdFlower Experiments. The table indicates the parameters and the main performance indicators from each experiment.

Many tweets contain web addresses and links. This presents a challenge in Twitter analysis, particularlyin relation to historical data, because of the likelihood of broken links and the difficulty of verifying content. We considered coding according to whether or not a tweet contained a link, or automatically coding these as headlines or informative tweets. However we ultimately decided that this would strongly skew the dataset, and found that many tweets containing links were not only about information sharing, but also contained emotional content that was relevant to our research.

This initial coding suggested that empathy manifested in a number of different ways. Through a process of detailed coding followed by the building of a coding frame with a smaller number of representative categories, a typology of responses was generated: mourning, where people expressed their personal reactions to the death including sadness or shock; social issues, where people drew attention to or discussed social issues related to the death such as bullying or depression; activism, where people discussed taking action in relation to the aforementioned social issues or attending a candlelit vigil; positive actions and negative actions, where people discussed what others were doing or had done in relation to the death; lack of empathy, where people judged either the person who died or those mourning them; and headline, which denotes a straightforward news headline or statement of facts relating to the death. Tweets that did not fit comfortably with any of these classes were coded as uncategorisable.

4.2. Scaling the coding using crowd-sourcing

Next, we created jobs on the Crowdflower crowdsourcing platform to expand the list of human labelled tweets. Providing instructions for crowdworkers in a brief and descriptive way has been identified as one of the main challenges in conducting crowd-sourcing experiments [13] and this was the case in this research. Tweets are often ambiguous, containing multiple communicative acts, and might be coded 'correctly' in multiple ways. However, we wanted to sensitise coders to particular forms of communication over others. For example, if someone shared a news story about a death but also expressed shock or sadness alongside the sharing of the link, we wanted that tweet to be coded primarily in the RIP/mourning category. In order to help with this, we require each Tweet to be coded with exactly one label, and created a decision tree to help coders make decisions about how to code a particular tweet (Figure 1).

The prioritisation built into the coding process through the decision tree attempts to lessen problems caused by such ambiguities between multiple categories, such as a case in which a tweet identifies a social issue then calls for activism on that basis. It also aims for consistency within and between the datasets in terms of how different types of tweet are understood. However, we still expected some ambiguity within the overall dataset, and allowed for coder disagreement in our initial analysis of the data. The need for a decision tree to focus the work of coders reminds us that working with big data requires interpretation in the same way as qualitative analysis [8].

4.3. Fine-tuning Execution Parameters

We chose CrowdFlower as a platform for executing our experiments because it provided enough flexibility to fine-tune our experiment and coders from specific countries – a requirement imposed by our ethics board.

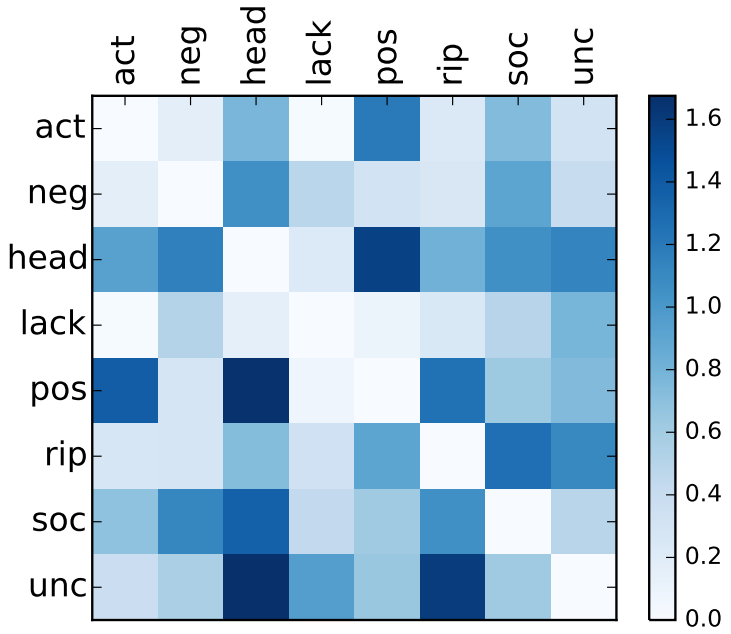

Figure 2: Confusion Matrix from The CrowdFlower Experiment. The percentage of tweets coded differently by two workers (columns and rows represent the higher- and lower-quality coders, respectively). The tweets coded similarly (i.e., diagonal elements) are excluded.

More specifically, we employed workers from the 15 (a limit imposed by Crowdflower) European Union countries with the largest populations.

CrowdFlower provided several mechanisms to control the quality of coders for the experiment, of which the coders' agreement with a short scale golden set of pre-coded answers proved the most effective. Two researchers labelled a sample of 200 tweets (40 from each use case) for the golden set experiment and refined this after three iterations of test runs on CrowdFlower. Further, we removed ambiguous tweets to ensure every possible chance for crowd workers to agree with the golden set. Finally, we followed the CrowdFlower's recommendations [7] and balanced the number of tweets in each class ending up with a golden set of 64 tweets, with 8 tweets from each class. These tweets, rather than being representative of the entire dataset, functioned as a benchmark to test the accuracy and agreement among coders in the experiment, and allowed us to ensure that tweets were coded by Crowdflower workers who had the best understanding of the appropriateness of a particular code for a particular tweet. Coding by those who showed an accuracy of less than 65 and 66% in relation to the golden set was excluded from the results of the first and the second experiments, respectively. Note, that – although consistent with some of the previous works [26] – these thresholds are slightly lower than a more frequently used value of 70% [27, 28]. Our choice has been motivated by an observation that most of the workers with accuracy between 66% and 70% in the first experiment (and between 65% and 70% in the second experiment) have provided reasonable feedback for their fails on the test questions and so we do not expect their contributions to introduce a systematic error in the results.

In our test runs, we noted very diverse results in the level of coders' conformity with our golden set: Whereas over 40% of test questions were missed or contested by low-quality coders, a significant set of high-quality coders exhibited more than 96% of agreement with the golden set. The average level of accuracy among selected coders (i.e., among those who scored more than 65% on the golden set) reached 78-82%. To encourage participation of high-quality coders we doubled the default pay for the job and noted in the

[7] https://success.crowdflower.com/hc/en-us/articles/202702985-How-to-Create-Test-Questions

description that the job required extra attention and that a good performance would be rewarded with bonuses. We then ran two experiments trading off between speed and quality (i.e. level of conservatism in selecting new coders) and labeled an overall sample of around 12K tweets, with each tweet coded by two CrowdFlower workers. We opted to collect more data points at the cost of having fewer judgments for each label; at the same time, we were conservative in selecting only consensus votes for the next - machine learning - step of our analysis (in Section 5. Afew factors contributed to this decision. On the one hand, we had already imposed several measures to control the quality of the labeling process – by choosing only high-quality coders and opting for consensus votes from two coders. On the other hand, we expected our machine learning algorithm to benefit more from a diversity of data points rather than from a diversity of judgments. Since we were interested in analysing the temporal evolution of the discourse in our datasets, we sampled an equal number of tweets from each of the first twenty days in each considered use case. The parameters of our CrowdFlower experiments are summarised in Table 2.

4.4. Validation of crowdsourced labels

The results of the experiments suggested a reasonably high level (over 60%) of agreement between coders. In Figure 2 we characterize the cases when workers disagreed in their classification. Each cell in the matrix represents the percentage of tweets which were coded differently by two workers, and the columns and rows represent a higher- and lower-quality workers (as indicated by their level of agreement with the golden set). The firt thing to note is that the matrix is predominantly symmetrical, indicating that disagreements have little correlation with difference in the quality of workers: disagreement between a specific pair of classes A and B can similarly happen when a higher-quality voted A as well as when she voted B (recall, however, from the previous section that we only included those workers who matched over 65% of the golden set tweets). Secondly, some pairs of classes are confused much more frequently than others: The disagreements are most likely between tweets labeled as "positive action" and "headline" ($\approx 1.6\%$), and between tweets labeled "uncategorised" and "mourning" or "headline" ($\approx 1.1-1.6\%$). This result can be probably explained by the fact that many tweets about people's actions in response to the death came in the form of headlines, and by the fact that there was some misunderstanding among coders about when the headline code should be used.

We next validated the crowdsourced labels by analysing the sentiments of the tweets for which the labels were generated. Most sentiment analysis tools typically attach a positive or negative 'sentiment score', and therefore are less specific and nuanced than the coding frames typically used in social science. However, understanding the general sentiment scores of different classes that the crowd has identified provides us with a coarse-grained assurance in the validity of the results. To this end, we used the SentiStrength library [29], considered to be one of the best tools for short texts [30], and associated each tweet with a score between 1 and 5 for positive and negative sentiments.

Figure 3 presents the mean positive and negative scores for each class in our CrowdFlower dataset. Firstly, we note that the highest negative and the highest positive sentiment scores are observed among the tweets from the most polarized classes – that of Negative and Positive Action. Similarly, the Mourning/RIP and Lack of Empathy classes in our dataset are associated with expectedly high negative sentiments. Because both results are intuitively expected given the classes, we obtain some assurance about the quality of crowd labels.

We also observe a striking difference between the sentiment scores of tweets in Activism and Social classes, which are semantically close: Whereas the Activism related tweets have relatively neutral sentiment – as indicated by a low negative and positive sentiments – the tweets from the Social Issues class show average negative scores of over 2.5 – the second most negative result among all classes in the dataset. This suggests not only that sentiment analysis and coding are complementary analysis (and thus both can add different dimensions when used on the same dataset), but also that the crowd-workers are able to distinguish closely related semantic classes in a way that reflects expected differences, such as sentiment scores.

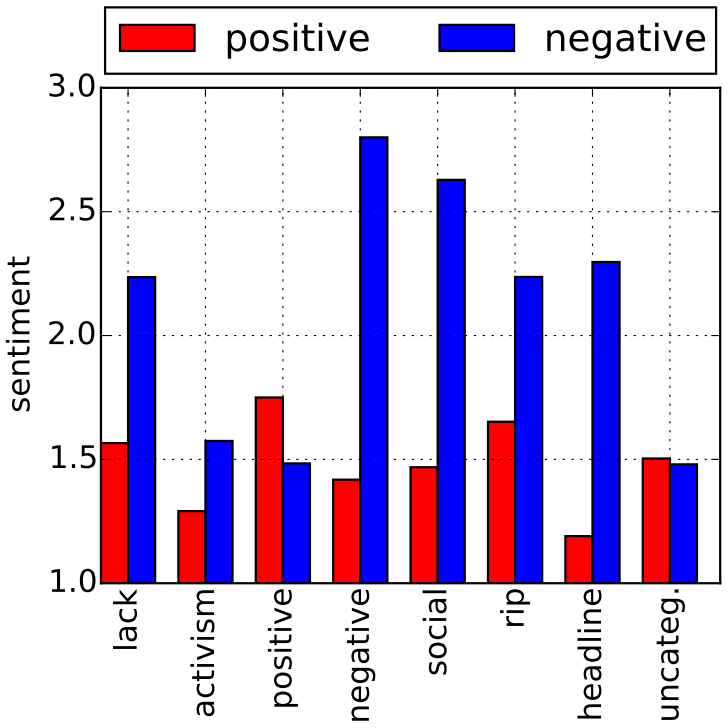

Figure 3: Relation between Sentiment Scores and the Classes of the Proposed Topology. The mean positive and negative sentiment scores – as measured by SentiStrength library – for each class in the dataset labeled by CrowdFlower workers.

5. Machine Learning Approach to Understanding Online Mourning

In order to scale up our analysis from twelve thousand to a million tweets, we used a supervised machine learning algorithms for processing short texts. We describe the model and its performance evaluation on our dataset in the rest of this section.

5.1. Algorithm

The goal of the machine learning model is to mimic the human researcher who codes (i.e., classifies) tweets based on their content. To recreate this effect, we exploited and adapted a state-of-the-art deep convolutional neural network architecture CharSCNN for short text classification proposed in [16] that was designed to operate at a word-level to capture syntactic and semantic information, and at a character-level to capture morphological and shape information. As argued in [16] the latter is particularly important for short texts such as Twitter posts that contain abbreviations, misspellings, emoticons and other word forms not common within traditional texts. As a result, CharSCNN showed significant improvement over alternative – recursive deep neural networks [31] and traditional bag-of-words models – when applied for fine-grained classification of Tweets.

Each tweet in this approach is represented by a sequence of N words $[w_0, .., w_N]$ where each word vector $w_i = [r^{wrd}, r^{wch}]$ is composed of two sub-vectors r^{wrd} for word-level embeddings and r^{wch} for character-level embeddings. We use a one-hot vector representations for character-level embeddings. Whereas in principle the model should be able to extract reasonable word-level embeddings from a one-hot vector representation of words too (if the training dataset is sufficiently large), in practice, it proved to be much more efficient to use externally pre-trained word-level embeddings. Such unsupervised pre-training of word representations has significantly improved the classification accuracy in the original CharSCNN paper – a result which has been also confirmed in our experiments. In particular, we used the Glove word vectors pre-trained on the

Class	Frequent Terms	Precision	Recall	Type-1 Err.	Type-2 Err.
activism	sign, law, therapy, conversion, lgbtq+, ban, enact, petition	0.89	0.79	0.01	0.03
negative action	funeral, parents, death, best, friend, banned	0.60	0.41	0.02	0.05
headline	star, tv, found, dead, australian, suicide, dies	0.61	0.78	0.08	0.04
lack of empathy	bleach, like, shit, people, getting, commit, fuck	0.67	0.73	0.03	0.02
positive action	tribute, billy, crystal, dedicated, emmys, dedicates, transparent	0.67	0.62	0.04	0.05
rip/mourning	rip, sad, rest, miss, heart, piece, beautiful, missed	0.72	0.77	0.06	0.05
social issue	bullying, people, suicide, stop, sopbullying, cyberbullying, depression, society	0.69	0.61	0.03	0.04
uncategorized	liked, de, clt, welcome, youtube, amandashires	0.80	0.78	0.02	0.02

Table 3: Prediction Performance of the Classifier. The averaged values of precision (Prec.) and recall (Rec.) of the 10-fold cross validation are reported among with the most frequent terms from each class. We also report the rates of Type-1 and Type-2 errors which indicate the relative contribution of each class to the total error of the 8-way classifier. Note that the sum of Type-1 and the sum of Type-2 errors (rounded in the table for better presentation) each add up to 0.29 = 1 - reported accuracy (0.71).

dataset of 2B tweets from [32]. We used randomly generated values for a minority (25%) of words which did not appear in the Glove vocabulary.

The neural network we designed in the Theano machine learning package [8] was composed of two convolution layers with max pooling aggregation - one for character-level and one for word-level embeddings, respectively - followed by two fully connected layers with dropouts to control for over-fitting and a final softmax layer with eight outputs corresponding to each of the labels in our dataset. The network was trained using mini-batch gradient descent by minimizing the negative log likelihood of the training dataset.

5.2. Cross validation

We validated the performance of the algorithm over the dataset of tweets labeled by the CrowdFlower workers as described in the previous section. Specifically, we used all labels with agreement between the coders which resulted in a dataset of 7.1K tweets. We note that the modeled reached an average accuracy of 71% in a 10-fold cross validation with approximately 50 training epochs in each experiment and minor improvements thereafter.

Looking at the model capabilities of predicting individual classes of messages (Table 3) we note that the precision varied between 60% for predicting 'negative actions' to 89% for discriminating 'activism' with the average precision being over 70% across all classes. In terms of recall, the model was able to capture 69% of instances of each class on average with a maximum of 79% achieved for 'activism'.

5.3. Manual validation

To provide an intuitive understanding of the algorithm's strong performance in discriminating different classes of tweets, in Table 3 we present the words with the highest relative frequencies in each class with respect to the overall frequency of the words in the dataset. We note that the illustrations contain words

[8] http://www.deeplearning.net/software/theano/

that can be expected to signify each category (e.g. the 'activism' messages predominantly consist of highly relevant words such as 'sign', 'law', 'petition', 'ban', etc.).

But beyond most frequent word, an important question from the social science perspective is whether the machine learning model can interpret nuance in particular cases. In some cases, particularly tweets where people recommended, congratulated or praised what someone else had done or written in response to the death, it did very well in determining subtle changes in tweets and accurately identifying the rhetorical intent of the tweet. These were frequently correctly coded as Positive Action, even though the tweets were otherwise similar to tweet types such as Negative Action or Social Issues:

- Thank-you @xxxxxx[9] for this balanced article that illustrates the danger of a powerful state & those who resist it. <LINK>

However there were also several instances where it did less well, and the repetition of similar tweets or claims might then lead to inaccuracies in the overall volume of tweets in each category. In relation to the question of add-ons to quoted tweets, this proved problematic in some cases. For example, the following tweet was coded as an RIP tweet, though neither the quoted tweet nor the comment - #blocked - should have been coded in that way:

- #blocked RT @xxxxxx: I don't know much about the case, but what I do know is I don't feel sorry for Aaron Swartz' suicide.

The original tweet should have been coded as Lack Of Empathy, and the add-on comment as Negative Action. Clearly, however, this is a very complex tweet in terms of rhetorical intent and there are likely to be issues with the correct coding of single hashtagged words even in human coding.

In the next section we highlight the greater prevalence of a 'lack of empathy' in responses to the death of Amanda Todd. There were many clear examples of correctly identified 'lack of empathy' in this dataset. However, in some cases the machine learning approach appears to have misinterpreted complex constructions of empathy as a lack of empathy. Here are two examples:

- RT @xxxxxx: I hate that everyone is suddenly buzzing about Amanda Todd now. She doesn't need the sympathy now, she needed it before ...

In this example, although the phrase 'she doesn't need the sympathy' taken on its own would of course be read as a lack of empathy, the tweet taken as a whole might be understood as saying that suicide is preventable, and as the result (in this case) of a failure of empathy. Likewise:

- RT @xxxxxx: Amanda Todd's story breaks my fucking heart. She made a stupid mistake, and it followed her for all of the wrong reasons.

This is clearly an example of empathy, but it may have been interpreted as a lack of empathy because of the phrases 'stupid mistake' or 'wrong reasons'. The use of the word 'mistake' , however, actually refers to her being blackmailed and cyberbullied after sharing images of her body on video chat rather than to her death.

Despite these distortions, it is clear that the machine learning correctly identifies lack of empathy to be more prevalent in this case, and that this changes over time. However, in a multi-case study, we should be aware that individual circumstances surrounding events may have an impact on the accuracy of comparisons between cases, and any large-scale analysis would need to take into consideration that the machine learning model would have some erroneously labelled tweets.

6. Analysing Dynamics of Public Empathy

In this section, we highlight the utility of a machine learning approach in assisting and supporting qualitative research by presenting some emergent findings. Specifically, we argue that a machine coding approach

[9] Because of the sensitive nature of the topic, all names and identifiable parts (e.g., URLs), have been anonymised or removed.

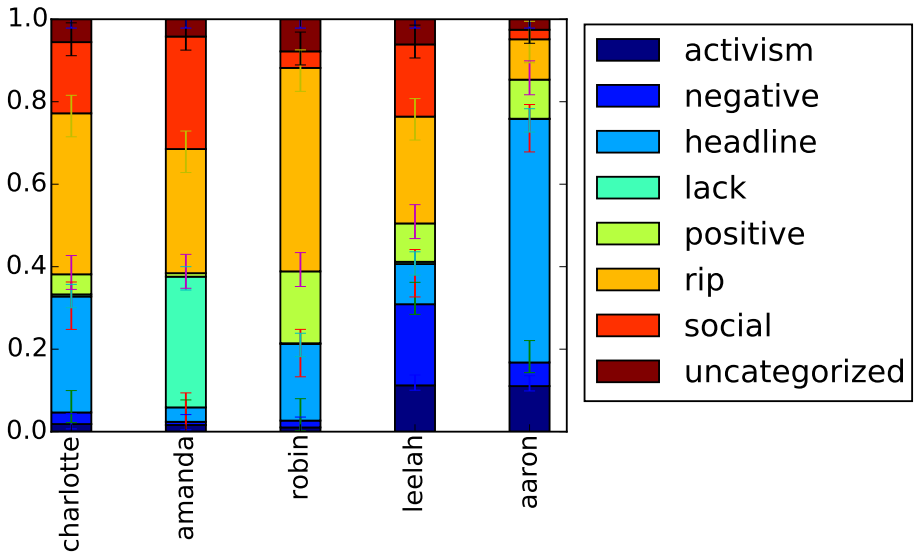

Figure 4: Relative Volumes of Classes across Use Cases. The relative volumes of tweets from each class among with the estimated error intervals are reported.

can contribute to a nuanced reading of subtle social and discursive changes as an event unfolds. There are still clearly instances using this approach where subtleties are missed and, as a result, qualitative analysis is important to fully understand what is being articulated. The dynamics of communication identified through the machine learning approach, however, provide a focus for this analysis. The combined use of iterative qualitative coding, crowd-coding, machine learning, and qualitative analysis can potentially help us to better understand complex and nuanced social discussions at scale.

6.1. On interpreting semi-automated coding

In the following, we focus on the analysis of the temporal dynamics of expressions of empathy (or lack of empathy) in our case studies. We do so through the analysis of the relative shares of tweets classified by our machine learning algorithm in each class at each day during the events. Both the crowd-coded and the machine-coded datasets allowed for the production of visualisations of the dynamics of each of the cases. However, the machine-coded datasets, because of the volume of tweets coded, allowed for a highly specific and complex reading of the interplay of each of the different tweet types, both across the set of tweets for each suicide and at particular times within each suicide.

Our main approach will be to compare relative volumes of different classes, both in the aggregate (c.f. Fig. 4), as well as over time (c.f. Fig. 5). To do so, it is important to estimate the error interval of the predictions made by our algorithm. Specifically, we need to understand how the error of mislabeling tweets in our experiments is distributed across individual classes and how that affects our estimates of relative volumes. To this end, for every class C we estimate the rates of Type-1 and Type-2 errors induced by mislabeling tweets in class C. More specifically, Type-1 error is measured as a share of all cases when a tweet from a class other than C has been labeled as C, whereas Type-2 error is measured as a share of all cases when a tweet from class C has been mislabeled as some other class (see the last two columns of Table 3). In other words, Type-1 (t^1_C) and Type-2 (t^2_C) errors asses the extent to which the share ρ_C of class C might have been over- or under-estimated in our calculations and, therefore, are represented as an interval $[\rho_C - t^1_C ... \rho_C + t^2_C]$ in Fig. 4 and Fig. 5.

Intuitively, the mislabeling error for each individual class contributes to the overall share of mislabeled cases and can be measured by the complement of accuracy, i.e., individual Type-1 errors (as well as individual Type-2 errors) in Table 3 sum up to 0.29 which is equivalent to 1 - accuracy (0.71).

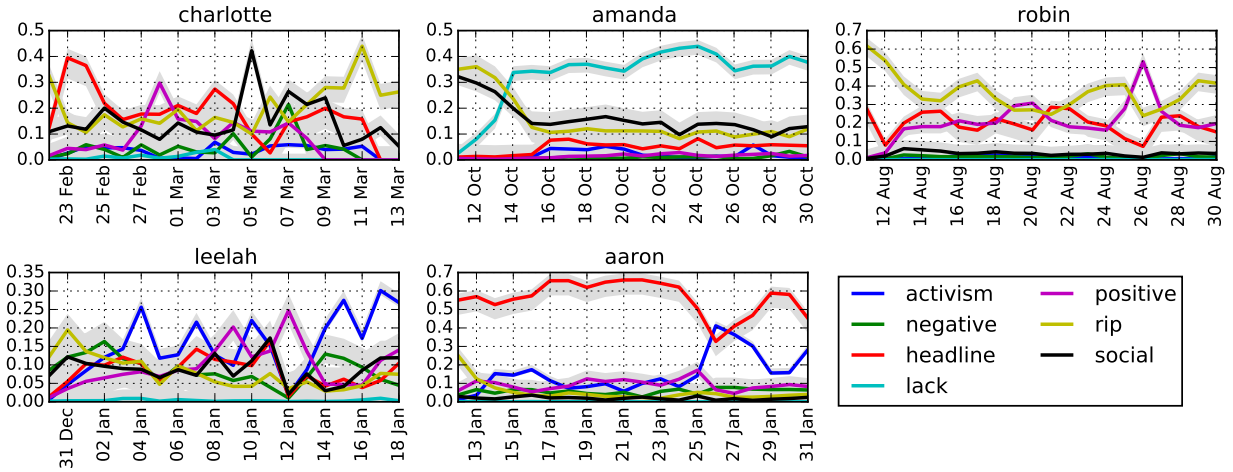

Figure 5: Dynamics of Public Response to High-profile suicides. The per-day relative volumes of tweets from each class are reported among with the estimated error intervals for each class drawn as gray areas, with values obtained from Table 3. Note, that the reported errors are static across individual days and are calculated from the last two columns of Table 3.

Thus, we plot the relative volumes as well as the dynamics of classes identified by our machine learning algorithm on the full dataset, indicating the error of the estimate with the gray intervals around each class in Fig. 5 and with error bars in Fig. 4. We note that in general the errors of estimated shares are relatively smaller than the differences between the shares of the most prominent classes in the vast majority of cases across all suicides considered, allowing to make qualitative conclusions on the dynamics of classes dominating in the discourse. This in turn allows for the selection of particular moments to be investigated through a closer qualitative reading.

6.2. Qualitative Reading through semi-automated coding

To illustrate the benefits of the semi-automated coding approach, we now discuss some qualitative findings which would have been difficult to obtain if only a small subset of data had been used for close reading.

In terms of overall types of communication, Fig. 4 showed us that each case had a different profile in terms of the kinds of communication that dominated the discussion. We found that for Aaron Swartz, headlines and news articles dominated, but Leelah Alcorn and Aaron Swartz both had high levels of activism compared to the other cases: their deaths were more politicised – both deaths resulted in draft laws named after them. In both of these cases the negative actions of others are also strong in the dataset (particularly for Leelah Alcorn) suggesting that in both cases a sense of injustice or mistreatment drove the politicisation and the activism that followed. Further qualitative analysis will be carried out to understand the connection between this politicisation and the way each death was understood within the dataset and in the broader public sphere.

As previously mentioned, the Amanda Todd case had the highest levels of 'Lack of Empathy' coded tweets as a proportion of the data. Such tweets make up more than a quarter of the Todd data, whereas they form only a small proportion of the data for each of the other cases. This means that participants in the conversation following her death were more likely to judge her harshly or to judge others for caring about her death. Although we have previously discussed the possibility of some minor distortion in this case, lack of empathy remains a feature when we look closely at the data.

This raises the importance of understanding change over time in relation to each case, rather than relying only on comparison of volumes, and we turn to Fig. 5 for this. In the case of Amanda Todd, although 'Lack of Empathy' is present in the data from day one, it does not begin to dominate until the 4th or 5th day of discussion, showing something of a backlash effect. Prior to that, the dominant themes were mourning and social issues. Such issues were strong in these data because of the discussion of bullying and cyberbullying

in relation to the death of Amanda Todd. Over time, however, participants in the conversation increasingly make claims about Amanda being to blame for the bullying she endured. Further qualitative analysis is needed to understand the discourses at play here, though in the case of Amanda Todd this might be related to a continuation of bullying behaviour, as well as perhaps her age and gender [33] [34].

In the case of Aaron Swartz, there is a peak in the 'activism' code just over two weeks after his death which coincides with activities by the group Anonymous. Participants acting under this moniker launched attacks on government websites to protest Aaron's prosecution and death. As described in the section on overall patterns, activism in Aaron's case was linked to discussion of negative actions, in this case on the part of the United States Department of Justice, the FBI, and the institutions involved in Aaron's legal case. One example of this type of tweet was:

- Many many American Dissidents believe U.S. Officials DROVE #RedditFounder #AaronSwartz to suicide with capriciously aggressive prosecution.

These examples show the uses of machine learning for identifying, at scale, moments during the unfolding of events and public discussions on Twitter, where something significant occurs, minds are changed, or new arguments and claims are made. It also provides an opportunity for the examination of relationships between different communicative types, whether across a whole dataset or for individual Twitter users.

7. Discussion, conclusions and lessons

Social science has tended to use small-scale, intensive, qualitative methods to explore issues of nuance and emotion. However, if we are interested in the aggregated or social patterning or collective expression of such phenomena – as in the case of public empathy – we need methods that are capable of bridging from small-scale, intensive study to potentially very large volumes of data that lie beyond the capabilities of manual coding.

The analysis presented here suggests that the combination of qualitative analysis with machine learning can offer both a big picture view of public events and close analysis of particular turning points or key moments in discussions of such events. As such, it can potentially yield new insights not easily achievable through traditional qualitative social science methods.

Although our specific case study looked at emotions and empathy in relation to high-profile deaths by suicide, the overall approach of semi-automated coding could be adapted to other research questions. Our experience suggests, however, that such adaptation will not be as simple as using a tool or a library. Rather, it is an approach that needs to be tailored to the problem at hand – each research question may require specific tweaks. For instance, if crowdsourcing is used to increase the set of manual labels, slightly different approaches or different decision trees may need to be developed to enable adequate levels of agreement amongst crowd workers. We made a decision to assign each tweet to one unique class. Addressing other problems may lead to ambiguous tweets being treated differently, e.g., allowing simultaneous or fractional (weighted) membership in multiple classes.

With the kind of customization described above, big data-based methods can give us some purchase on aggregated and collective aspects of emotional expression online. This is increasingly necessary given the significance of social media in mediating and constituting emotional lives. At the same time, however, the analysis above also reminds us that, while decision trees and similar approaches aimed at guiding manual or automated coding can help to narrow differences in classification, the interpretive gap cannot be completely closed.

Our method aims to combine a conventional classification method used in qualitative social science (coding), with algorithmic classification using machine learning. Although the authors of this article included experts in both these approaches, significant challenges arose in merging the two: in particular, we underestimated the difficulty of creating a coding scheme that can be interpreted and applied by crowd workers to create reliable high quality labels. Our initial efforts were unsuccessful as different crowd workers assigned different priorities to the different labels, leading to inconsistency. In our second attempt, therefore, we provided a clear guide for crowd workers, using the decision tree in Fig. 1 to help to create greater consistency in labelling. This improvement, while simple, was instrumental to the success of our methodology.

We believe this example also illustrates the nature of potential pitfalls, and how they are more likely to be non-technical than technical. Paying attention to such human factors is likely to be an essential feature of future interdisciplinary research in computational social science.

Acknowledgements

This work was supported by the Space for Sharing (S4S) project (Grant No. ES/M00354X/1).

References

[1] J. Brownlie, Ordinary relationships. A Sociological Study of Emotions, Reflexivity and Culture, Palgrave MacMillan, 2014.
[2] D. Brake, Sharing our lives online: Risks and exposure in social media, Springer, 2014.
[3] A. Halavais, Bigger sociological imaginations: framing big social data theory and methods, Information, Communication & Society 18 (5) (2015) 583–594.
[4] D. V. Shah, J. N. Cappella, W. R. Neuman, Big data, digital media, and computational social science possibilities and perils, The ANNALS of the American Academy of Political and Social Science 659 (1) (2015) 6–13.
[5] G. B. Colombo, P. Burnap, A. Hodorog, J. Scourfield, Analysing the connectivity and communication of suicidal users on twitter, Computer communications 73 (2016) 291–300.
[6] M. L. Williams, P. Burnap, Cyberhate on social media in the aftermath of woolwich: A case study in computational criminology and big data, British Journal of Criminology 56 (2) (2016) 211–238.
[7] R. Tinati, S. Halford, L. Carr, C. Pope, Big data: methodological challenges and approaches for sociological analysis, Sociology (2014) 0038038513511561.
[8] d. boyd, K. Crawford, Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon, Information, communication & society 15 (5) (2012) 662–679.
[9] P. A. Schrodt, Automated coding of international event data using sparse parsing techniques, in: annual meeting of the International Studies Association, Chicago, 2001.
[10] A. Esuli, F. Sebastiani, Machines that learn how to code open-ended survey data, International Journal of Market Research 52 (6).
[11] A. D. Shaw, J. J. Horton, D. L. Chen, Designing incentives for inexpert human raters, in: Proc. ACM CSCW, 2011.
[12] V. S. Sheng, F. Provost, P. G. Ipeirotis, Get another label? improving data quality and data mining using multiple, noisy labelers, in: Proceedings of the KDDS, ACM, 2008, pp. 614–622.
[13] A. Kittur, J. V. Nickerson, M. Bernstein, E. Gerber, A. Shaw, J. Zimmerman, M. Lease, J. Horton, The future of crowd work, in: Proceedings of CSCW, ACM, 2013, pp. 1301–1318.
[14] S. Komarov, K. Reinecke, K. Z. Gajos, Crowdsourcing performance evaluations of user interfaces, in: Proc. CHI, ACM, 2013, pp. 207–216.
[15] W. Willett, J. Heer, M. Agrawala, Strategies for crowdsourcing social data analysis, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM, 2012, pp. 227–236.
[16] C. N. dos Santos, M. Gatti, Deep convolutional neural networks for sentiment analysis of short texts., in: International Conference on Computational Linguistics -COLING, 2014, pp. 69–78.
[17] K. Glasgow, C. Fink, J. L. Boyd-Graber, " our grief is unspeakable": Automatically measuring the community impact of a tragedy., in: ICWSM, 2014.
[18] J. Garde-Hansen, Measuring mourning with online media: Michael Jackson and real-time memories, Celebrity Studies 1 (2) (2010) 233–235. doi:10.1080/19392397.2010.482299.
[19] G. Terzis, et al., Death trends: Hashtag activism and the rise of online grief, Kill Your Darlings (22) (2015) 9.
[20] S. K. Radford, P. H. Bloch, Grief, commiseration, and consumption following the death of a celebrity, Journal of Consumer Culture 12 (2) (2012) 137–155. doi:10.1177/1469540512446879.
[21] C. Sian Lee, D. Hoe-Lian Goh, "gone too soon": did twitter grieve for michael jackson?, Online Information Review 37 (3).
[22] A. Bruns, J. Burgess, K. Crawford, F. Shaw, #qldfloods and @qpsmedia: Crisis communication on twitter in the 2011 south east queensland floods, Tech. rep., Brisbane, Australia (01 2012).
URL `http://eprints.qut.edu.au/48241/`
[23] F. Shaw, J. Burgess, K. Crawford, A. Bruns, Sharing news, making sense, saying thanks, Australian Journal of Communication 40 (1) (2013) 23.
[24] Z. Zhou, R. Bandari, J. Kong, H. Qian, V. Roychowdhury, Information resonance on twitter: watching iran, in: Proceedings of the first workshop on social media analytics, ACM, 2010, pp. 123–131.
[25] D. Hoe-Lian Goh, C. Sian Lee, An analysis of tweets in response to the death of michael jackson, in: Aslib Proceedings, Vol. 63, Emerald Group Publishing Limited, 2011, pp. 432–444.
[26] Z. R. De Kuthy, K., D. Meurers, Learning what the crowd can do: A case study on focus annotation., in: In Proceedings of the 6th Conference on Quantitative Investigations in Theoretical Linguistics., 2015.
[27] B. Ionescu, A.-L. Radu, M. Menéndez, H. Müller, A. Popescu, B. Loni, Div400: a social image retrieval result diversification dataset, in: Proceedings of the 5th ACM Multimedia Systems Conference, ACM, 2014, pp. 29–34.
[28] S. Mac Kim, S. Wan, C. Paris, Detecting social roles in twitter, in: Conference on Empirical Methods in Natural Language Processing, 2016, p. 34.

[29] M. Thelwall, Heart and soul: Sentiment strength detection in the social web with sentistrength, Proceedings of the CyberEmotions (2013) 1–14.
[30] P. Gonçalves, M. Araújo, F. Benevenuto, M. Cha, Comparing and combining sentiment analysis methods, in: Proceedings of ACM COSN, ACM, 2013, pp. 27–38.
[31] R. Socher, A. Perelygin, J. Y. Wu, J. Chuang, C. D. Manning, A. Y. Ng, C. Potts, et al., Recursive deep models for semantic compositionality over a sentiment treebank, in: Proceedings of the conference on empirical methods in natural language processing (EMNLP), Vol. 1631, Citeseer, 2013, p. 1642.
[32] J. Pennington, R. Socher, C. D. Manning, Glove: Global vectors for word representation., in: EMNLP, Vol. 14, 2014.
[33] R. Penney, The rhetoric of the mistake in adult narratives of youth sexuality: the case of amanda todd, Feminist Media Studies 16 (4) (2016) 710–725.
[34] J. Ringrose, L. Harvey, Boobs, back-off, six packs and bits: Mediated body parts, gendered reward, and sexual shame in teens' sexting images, Continuum 29 (2) (2015) 205–217.