Building the 'Truthmeter': Training algorithms to help journalists assess the credibility of social media sources

Richard Fletcher, Steve Schifferes, and Neil Thurman

Abstract

Social media is now used as an information source in many different contexts. For professional journalists, the use of social media for news production creates new challenges for the verification process. This article describes the development and evaluation of the 'Truthmeter' – a tool that automatically scores the journalistic credibility of social media contributors in order to inform overall credibility assessments. The Truthmeter was evaluated using a three-stage process that used both qualitative and quantitative methods, consisting of (1) obtaining a ground truth, (2) building a description of existing practices, and (3) calibration, modification and testing. As a result of the evaluation process, which could be generalized and applied in other contexts, the Truthmeter produced credibility scores that were closely aligned with those of trainee journalists. Substantively, the evaluation also highlighted the importance of 'relational' credibility assessments, where credibility may be attributed based on networked connections to other credible contributors.

Keywords

Algorithms, computational journalism, credibility, fake news, journalistic practice, news, social media, sourcing practices, verification

Introduction

Social media is now regularly used for the production, consumption and dissemination of information in a wide variety of different contexts (Pew, 2015). It has proven particularly popular for news. In terms of news consumption, over one-third of the online population of the United Kingdom, the United States, France, Spain and the Netherlands say they now use social media platforms such as Facebook and Twitter to access the news (Newman et al., 2016). In terms of production, many journalists and news organizations now see social media as a primary news source (see Lecheler and Kruikemeier, 2016 for an overview). Yet, despite the fact that social media provides convenient access to a plethora of potential news sources, journalists primarily use it to access 'elite' sources or those with privileged access to events. Though they will make use of eyewitness accounts of important events (Vis, 2013), they will typically source news from social media when it is used by politicians and other public figures to make newsworthy announcements (Broersma and Graham, 2012). For the time being at least, it appears that non-elite or alternative sources are primarily used to add depth or colour to existing stories (Broersma and Graham, 2013), or in situations where there is a dearth of elite sources (Bruno, 2011).

One of the reasons for the apparent reluctance of journalists to fully embrace social media concerns the challenges it creates for the journalistic verification process. For Kovach and Rosenstiel (2007: 79), verification is the 'essence' of journalism, and its practice is 'what separates journalism from entertainment, propaganda, fiction, or art'. A key part of the verification process is assessing credibility (Powers and Fico, 1994). Traditionally, journalists have relied on a small number of elite sources, such as governments and institutions, because they are perceived to have a high degree of credibility that ultimately stems from their power and authority within society (Gans, 1979). However, this approach may no longer be necessary

in a world where journalists have easy and convenient access to a very large number of other potential sources.

When journalists have attempted to source newsworthy material from social media, they have sometimes encountered problems. As Schifferes et al. (2014) have highlighted, there now exists a growing library of case studies that document news reports featuring ‘fake’ content sourced from social media, including ‘Photoshopped’ pictures following the death of Osama bin Laden in 2011, fake photographs of Hurricane Sandy in 2012 and the identification of innocent people as suspects following attempts to ‘crowdsource’ the capture of the Boston Marathon bomber in 2013. Additionally, there are also examples of governments and official sources exploiting the difficulties journalists face when verifying digital content by deliberately releasing misleading information in the hope that it will be reported as genuine (e.g. Oliver, 2008).

Some of these mistakes were likely rooted, at least in part, in a lack of digital media literacy within the profession (Tylor, 2015). Others, however, are likely to have been caused by problems of scale. Following the integration of social media into some newsrooms, professional journalists found that they were required to collect, filter, assess and contextualize vast amounts of information in short spaces of time (Newman, 2009). In these situations, it is not necessarily the case that traditional or existing verification practices are entirely inappropriate for checking information from social media, but rather that the contextual information required to carry them out is not readily available, or it is not possible to adequately assess all of the content in the time that is available. Put differently, although journalists may be able to apply modified versions of traditional verification practices to information from social media, the standard difficulties associated with verification are compounded by the sheer volume of potentially newsworthy information that social media makes available.

This challenge has prompted scholars to explore new concepts and approaches. Hermida (2012) has argued that journalists should base their verification on open collaboration with the public to harness their collective knowledge. Whereas this prescription retains a fundamentally human approach, others have attempted to use data from social media platforms to automate the verification process (Diakopoulos et al., 2012). Yet, empirical research into emerging journalistic practices has revealed that, although there is no universal strategy for verifying information from social media, journalists have adopted hybrid approaches that combine the human with the automated (see e.g. Larsen, 2016 and Heravi and Harrower, 2016). Traditional journalistic methods and practices continue to be valued, but there is also a desire for specialized verification tools that speed up the process and apply those methods and practices at scale (Brandtzaeg et al., 2016). This suggests that although traditional approaches to verification will persist in the newsroom, the automation of certain aspects of the process will be used to complement them when information from social media is being dealt with.

For journalists, news organizations and others outside of news production who aim to make the most of the huge amount of information that social media makes available, the introduction of automated processes raises important questions. In the first instance, which new or existing practices, if any, can be encoded into software? Following on from this, how can improvements to automated processes be identified and evaluated? In this article, we use the development of the 'Truthmeter' – a tool designed to automatically assess journalistic credibility – to outline a three-stage evaluation procedure that uses qualitative and quantitative methods to help answer these questions. The procedure consists of (1) obtaining a ground truth, (2) building a description of existing practices, and (3) calibration, modification and testing. As a result of the evaluation process the Truthmeter produced credibility scores that were closely aligned with those of trainee journalists. Due to the reliance on the validity of existing practices, this procedure is not necessarily appropriate for situations that require an automated

approach that is radically different from what is currently practised. However, this procedure, which could be generalized and applied in other contexts, is appropriate for situations where it is desirable to implement a broadly similar automated version of existing practices in order to speed them up or make them more manageable.

Literature review

Social media is used as an information source in a wide variety of different contexts (Pew, 2015). Given the vast amount of information made available by social networks, automated or computerized processes are often required to collect and analyse this data in a comprehensive manner. These processes may replicate or complement actions or judgements ordinarily carried out by humans. An example can be seen in assessments of the credibility of information – where 'credibility' refers to 'believability' or 'offering reasonable grounds to be believed' (Castillo et al., 2011: 675). The use of the Internet as an information source has consistently generated an interest in, and anxieties over, credibility (e.g. Flanagin and Metzger, 2007; Morris and Ogan, 1996), and these have been reignited by the more recent rise of social media (Castillo et al., 2011). In response, scholars have investigated ways of automating the credibility assessment process in order to quickly filter out unreliable information.

Automated credibility assessment has often been investigated in a general sense, without a specific focus on journalism. In most cases, Twitter – a large social network with over 300 million active users – has been used as the research site, with findings broadly generalizable to other social networks. Typically, there is an assumption that the information made available on a contributor's Twitter profile page (or through the Twitter API) can be used, to some extent, to inform credibility assessments. The results of a survey by Morris et al. (2012) revealed that people typically associated the use of a cartoon profile image (including the default Twitter image), and the following of a large number of other contributors (especially

if the person in question has few followers themselves) with low credibility. Conversely, they found that a relevant location, official Twitter 'verification', the contributor being a recognized figure and a Twitter biography connoting 'relevant expertise' were associated with high credibility. Westerman et al. (2012) found that there was a curved relationship between the number of followers a contributor has and their perceived credibility, with both a very small and a very large number of followers indicating lower perceived credibility than a reasonably large number. Furthermore, they found that there existed a relationship between the contributor's ratio of followers to followings (the number of accounts that the contributor follows) and their perceived credibility, with a wide gap between the two indicative of low credibility. In a related study, Westerman et al. (2014) found that whilst there was no relationship between the 'recency' of a contributor's updates and their credibility, a high degree of 'cognitive elaboration' – defined as active engagement with other contributors through re-tweets and @-replies – was linked to high credibility. Finally, Edwards et al. (2013) found that providing Klout scores – an independently produced measure of 'influence' for social media contributors based on an assessment of over 400 signals across eight social networks – influenced credibility assessments of Twitter contributors, with contributors who had high Klout scores being deemed more credible than contributors who had medium to low scores.

A number of studies have investigated the possibility of using data generated in the course of social media activity to arrive at an automatic identification of credible news events. However, this has almost always been done without a specific focus on professional journalism. Castillo et al. (2011) found that more active users tend to spread more credible news, that new contributors with large numbers of followers spread more credible news and that credible news is propagated through contributors with a large number of tweets or re-tweets. In a later study, the same authors (2013) found that contributors with more followers are more likely to spread tweets containing credible news, and that credible tweets tended to

be longer, contain negative sentiments and contain URLs featured in the 10,000 most visited domains. Insights like these are now starting to be integrated into software in order to arrive at automated credibility assessments. Gupta et al. (2014) have described TweetCred, a real-time web-based system that automatically rates the credibility of individual tweets on a 7-point scale using data about both the contributor and the tweet text. Their evaluation procedure was based on a public release of TweetCred as a Chrome browser extension, with users encouraged to offer feedback on the automatically generated scores during a three-week period. They found that 40% of the 1,273 responses received agreed with the TweetCred score. Those that disagreed were asked to provide their own score for each tweet. Where such a score was provided, in 38% of the cases it was within 2 points of the TweetCred score on the 7-point scale.

A smaller number of studies have investigated how credibility on social media is assessed within the context of professional news production. Brandtzaeg et al. (2016) conducted interviews with 24 professional journalists from across Europe and found that they deemed elite individuals and institutions, such as celebrities, politicians and news organizations, to be credible social media sources largely due to the fact that they had typically built trust over many years. They found that non-elite sources were not typically seen as credible unless they could be verified using traditional journalistic practices, such as speaking with them over the telephone and checking with peers. In a survey of 421 professional journalists based in Ireland, Heravi and Harrower (2016) similarly found that most prefer to tap into ‘real-world’ networks to verify information from social media, with a smaller number choosing to verify information either online or though social media itself. In terms of information available on social media profiles, journalists reported that a link to an institutional or company website and the quality and number of posts were the most important factors, with the profile image and account age less important. Diakopoulos et al. (2012) conducted

interviews with four professional journalists to explore the methods they used to assess credibility on social media. They found that journalists typically saw eyewitnesses as credible sources, and thus used their Twitter location as a key indicator of credibility. They also used the contributor's level of 'conversational engagement' (in the form of @-replies), their use of hyperlinks and number of re-tweets as cues to arrive at credibility judgements. Diakopoulos et al. (2012) went on to describe Seriously Rapid Source Review (SRSR) – a tool that aims to identify credible eyewitnesses by categorizing Twitter contributors based on their professional status (using keyword identification), location and interactions, with respect to identified news stories. Their evaluation procedure was fundamentally qualitative and consisted of a hands-on session with seven professional journalists. The journalists responded positively to the contextual information that SRSR provided and related that 'source context including historical tweets, account age, website, interaction with others (@-reply behaviour), network properties, and Klout scores were all valuable cues that they routinely use to assess their trust of sources' (Diakopoulos et al., 2012: 2457).

SocialSensor and the Truthmeter

There are, then, only a small number of tools specifically designed to address the issues faced by professional journalists when attempting to assess the credibility of information sourced from social media. In response, the European Union SocialSensor project – a research consortium of 10 institutions across Europe (including Yahoo, IBM and Deutsche Welle) – has been working to develop new software tools to help journalists utilize social media more effectively. The SocialSensor project is fundamentally user-centred, with a particular use-case structured around meeting the specific needs of journalists.[1] Previous research carried out as part of the SocialSensor project highlighted a demand from journalists for verification tools (Schifferes et al., 2014). Early in the project, 22 practicing journalists were interviewed and

asked to score the relevance of certain proposed features. ‘Verifying social media content’ emerged as the second most relevant feature, behind ‘Predicting or alerting breaking news’.

Though tools that are able to automatically arrive at credibility assessments of information from social media would be of benefit to journalists, their production is far from straightforward. The range of different approaches to determining the credibility of information sourced from social media that is evident in the literature highlights the fact that credibility is complex and multifaceted. It follows that an overall assessment of the credibility of information from social media should reflect this. On this basis, the framework that underpins SocialSensor’s view of credibility is expressed in terms of three Cs: ‘contributor’, ‘content’ and ‘context’ (Schifferes et al., 2014):

1. Contributor: who the information came from;
2. Content: what is contained within the information;
3. Context: why the information was provided.

Though, as is clear from previous studies, an overall assessment of credibility should be based on a combined assessment of these facets, the assessment of each individual facet is likely to require quite different processes. Take the example of a tweet. An assessment of the contributor will be focused on the author of the tweet and may be based on an examination of the number of followers the author has, whereas an assessment of the content may be based on the identification of certain keywords contained within the tweet, and an assessment of the context may be based on the location where the tweet was sent from. Therefore, the automated assessment of each is likely to be based on fundamentally different computational tasks. Thus, though the outputs from each task may ultimately be combined to form an overall assessment, they can also be understood as separate from one another. Though SocialSensor aims to

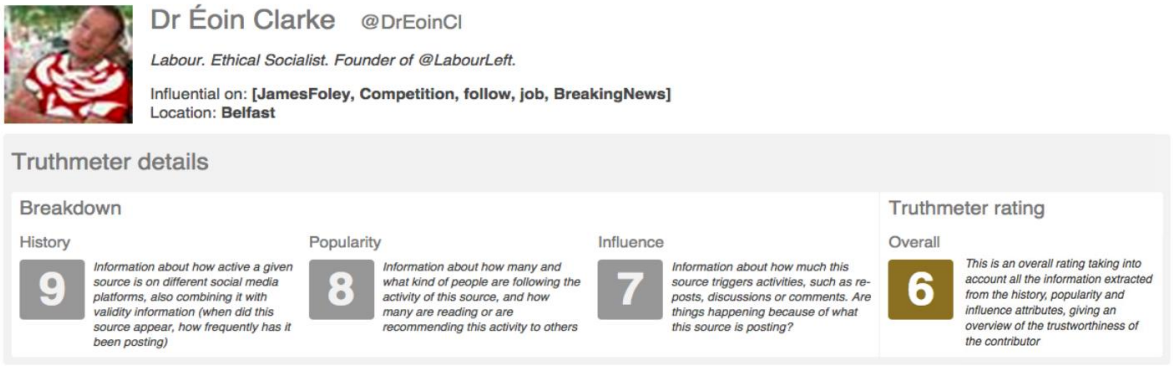
produce such an overall assessment, the Truthmeter component, which will be introduced in the next subsection, aims to arrive at an assessment of the contributor's credibility only.

The first version of the Truthmeter prototype was developed in collaboration with the Athens Technology Center (ATC). In short, the Truthmeter aims to score the journalistic credibility of Twitter contributors on a 0–9 scale, with nine indicating a high degree of credibility. Importantly, Truthmeter scores are not designed to be a definitive measure of contributor credibility, but rather a useful indicator, available in real-time, for journalists to consider alongside their other verification practices. The name of the tool is derived from an early prototype called 'Alethiometer', itself based on the Greek word for truth. However, this name is not meant to suggest that the tool is able to separate truth from falsehood.

Within the SocialSensor system – which as a whole aims to quickly surface trusted and relevant material from social media – contributor credibility scores are displayed next to all attributable content (see Figure 1). Clicking on the contributor's name provides a more detailed breakdown of their credibility, as well as information about the topics on which they are considered influential, and some basic information straight from their Twitter profile page (see Figure 2). Separate scores are provided for the contributor's 'history' – a measure of how active the contributor has been; 'popularity' – a measure of how many people follow the contributor; and 'influence' – a measure of how effectively the contributor triggers the activity of other contributors. These measures are then combined to form the overall score.

Figure 1: Contributor credibility displayed next to content

Figure 2: Contributor credibility page



The Truthmeter computes credibility scores based on data made available through the Twitter API. In common with other tools, therefore, credibility scores are based on a combination of Twitter contributor metrics, such as the 'number of followers' and the 'number of tweets'. The first version of the Truthmeter computed contributor credibility scores based on the metrics listed in Table 1. These metrics were then combined to arrive at an overall contributor credibility score.

Table 1: Metrics used in the first version of the Truthmeter

Metric	**Category**
Number of tweets	History
Frequency[2]	History
Number of followers	Popularity
Number of followings	Popularity
Number of re-tweets[3]	Influence
Number of @-mentions	Influence

Evaluating the Truthmeter

Once the Truthmeter was able to produce credibility scores for Twitter contributors, it could be evaluated. The evaluation procedure aimed to combine elements of the processes used by Diakopoulos et al. (2012), in that it is grounded in qualitative research into the practices used by professional journalists, whilst also adding a quantitative dimension through the use of the performance measures described by Gupta et al. (2014). As such, the evaluation was designed to be multi-stage and based on both quantitative and qualitative data, with feedback from journalists informing future technical modifications. The evaluation was divided into three stages:

1. Ground truth and benchmark;
2. Qualitative description of practice;
3. Calibration, modification and testing.

The rationale for the first stage was to obtain a 'ground truth' for journalistic credibility and to benchmark how well the Truthmeter scored by comparison. Once a baseline indication of how well the Truthmeter was performing had been established, the rationale for the second stage of the evaluation was to develop a more detailed practical understanding of how journalists assess the credibility of social media contributors. Finally, the rationale for the third stage of the evaluation was to use this data to modify Truthmeter such that it was more closely aligned with the credibility assessment practices journalists use.

First stage: Ground truth and benchmark

As has been stated, the rationale for the first stage of the evaluation was to identify whether there were any differences in the credibility scores assigned to social media

contributors by the original version of the Truthmeter and by journalists. Additionally, it was decided that it would be useful to compare the Truthmeter's and journalists' scores with those assigned by Klout, given that the reliability of Klout has been informally established through common use.

For the first stage of the evaluation the independent variable was the source of the credibility score (Truthmeter, Klout and journalists). The dependent variable was the credibility scores produced by each source. A stratified random sample of 150 social media contributors was identified. The sample was drawn from contributors already known to the SocialSensor system in order to ensure that it was possible to quickly obtain a Truthmeter score (for information on the contributors utilized by the SocialSensor system, see Thurman et al., 2016). The sample was stratified by Truthmeter score. Klout scores for the same list of 150 contributors were obtained through the Klout API. Scores for the same 150 contributors were assigned simultaneously by a panel of eight trainee journalists during a specially designed credibility-scoring task.

The journalists that made up this panel were all master's students from the Department of Journalism at City, University of London. Students were recruited from five different MA programmes within the Department: 'Investigative', 'Newspaper', 'Magazine', 'Broadcast', and 'International' Journalism. Although students at the time of the evaluation, we considered their knowledge and experience of the practice of journalism was sufficient for their evaluations to be professionally credible for two reasons. Firstly, because the master's programmes on which they were enrolled are highly practical. On the courses students "learn how to gather and report [news] in various styles … [becoming] adept at print, broadcast and online journalism" and they are "encouraged to complete [a journalism] internship" (see e.g. City, University of London 2017a). Secondly, because in order to be admitted onto the programmes, students are required to have work experience in journalism (see e.g. City,

University of London 2017b). All described themselves as experienced social media users. The panel was made up of a mixture of males and females, half aged between 25 and 34 and half aged under 25. It should be acknowledged that there might have been some biases created by recruiting students from the same department. In particular, it is possible that they may have learnt similar verification practices and received similar training in other practical aspects of journalism, which may have resulted in particular patterns of credibility scoring. The panel would have benefitted from being more varied in terms of professional experience.

For the credibility-scoring task, each trainee journalist was provided with a list of links to 150 Twitter profiles. They were then asked to provide a score for each contributor on a 0–9 scale, with nine being the most credible. Journalists were given brief instructions on how to calibrate their scoring (e.g. the prime minister of the United Kingdom should be scored nine), but were ultimately free to use whatever processes they would normally use. Importantly, the journalists were told not to base their credibility assessments upon the content of the contributor's tweets, as the study was exclusively concerned with the credibility of the contributor in this instance. The trainee journalists provided their data remotely over the course of a week via an electronic scoring sheet. For this stage of the evaluation, and those that followed, the scores provided by journalists were thought of as a 'ground truth'.

Once this data was collected, Klout scores – which range from one to 99 – were collected and then straightforwardly adjusted to be comparable with the Truthmeter and journalist scores. The mean of the scores from the eight journalists for each contributor was calculated and rounded to one significant digit. Scores for two Twitter contributors were removed during the analysis, as they had deleted their profiles during the task.

The mean score produced by the Truthmeter ($\overline{x}$ = 5.71, SD = 2.45) was close to the mean score assigned by the journalists ($\overline{x}$ = 5.67, SD = 2.10) and by Klout ($\overline{x}$ = 5.40, SD = 2.17). Although the measure used by Gupta et al. (2014) to evaluate TweetCred was used to

evaluate the scores assigned to individual tweets, we can use a very similar measure here. When we do so, we see that 82% of the scores produced by the Truthmeter were within 2 points (on the 0–9 scale) of the mean journalists' score, compared to 93% of the Klout scores. To complement this measure, a Kruskal–Wallis one-way analysis of variance test was applied to the influence scores assigned by Klout (Mean Rank = 206.92), the credibility scores assigned by Truthmeter (Mean Rank = 233.20) and the credibility scores assigned by journalists (Mean Rank = 227.39). Though these results should not be interpreted as proof of statistical equivalence, the test nonetheless failed to find any statistically significant difference between the three sets of scores: $X^2(d.f.\ 2) = 3.49$, $p > .05$. Additionally, a Spearman's correlation test showed that there was a strong positive correlation between the Truthmeter scores and those assigned by journalists: $r_s(148) = .69$, $p < .01$. There was also a strong positive correlation between Klout scores and those assigned by the journalists: $r_s(148) = .80$, $p < .01$. Finally, there was a very strong positive correlation between Klout scores and those assigned by Truthmeter: $r_s(148) = .86$, $p < .01$.

In sum, these results suggest that the first version of Truthmeter was able to produce credibility scores that were a good match with those assigned by journalists. Despite this, it was also clear that the Truthmeter scores could be improved. In particular, the Gupta et al. (2014) measure highlighted the fact that for a minority of contributors the score assigned by the journalists was 4 or 5 points higher on the 0–9 scale than the score assigned by Truthmeter. A closer look at these contributors suggested that, although they had a small number of tweets or a small number of followers (and thus may be considered either 'inactive' or 'new'), they had ultimately been deemed credible by journalists. This highlighted a problem caused by an over-reliance on certain metrics, but at this stage it remained unclear how the journalists had come to this conclusion, and therefore also unclear how it could be addressed.

Second stage: Qualitative description of practice

The rationale for the second stage of the evaluation was to develop a more detailed practical understanding of how journalists actually assess the credibility of social media contributors, in order to suggest how the Truthmeter's metrics might be improved. To achieve this, a short online questionnaire was used to elicit feedback from the journalist panel on how they went about the credibility-scoring task.

The questionnaire asked journalists about their background, social media use, understanding of credibility and the processes they used when completing the credibility-scoring task. For this final topic, the questionnaire used open-ended questions to elicit responses, and requested that the journalists provide long, detailed answers that one might expect from a qualitative interview. This was deemed necessary, as it was important for the participants to be able to communicate methods of assessing credibility that had not yet been identified by previous research. The final section placed a particular emphasis on understanding the large disparity between Truthmeter and journalist credibility scores for some contributors. The five contributors who produced the greatest difference in Truthmeter and journalist scores were identified, and the journalists were asked to describe the reasoning behind their scoring of each. For the open-ended questions, thematic analysis was used to interpret the responses, and a process of iterative coding was used to identify key themes.

The results of the questionnaire were analysed one week after the completion of the credibility-scoring task. The journalists were asked to specify contributors from the task for whom they found the process of assigning a credibility score to be easy. Typically, they found it easy when they believed that the contributor had a high degree of journalistic credibility, as in the case of famous individuals (e.g. @BarackObama), traditional mainstream news outlets (e.g. @BBCNews), and institutions (e.g. @FA). When asked about the specific indicators used when it was easy to assess credibility, journalists referred to the number of followers a

contributor had, and whether they had been ‘verified’ by Twitter. The journalists also, although less frequently, referred to the information contained in the contributor’s biography, the credibility of the contributors following the contributor being assessed, the contributor’s avatar and the general appearance of their profile.

Journalists were also asked to specify instances where they found the process of assigning a credibility score to be difficult. Typically, they found it difficult when they believed that the contributor had a relatively low degree of credibility. They reported that they found it difficult to assess the credibility of other journalists (e.g. @peteclifton), public relations professionals (e.g. @NicoleLoveLloyd) and businesses (e.g. @moneyclaims4u). In terms of specific indicators, they referred to the content of a contributor’s biography and their number of followers. The journalists also mentioned the use of Internet searches, the total number of tweets a contributor had made, the credibility of the contributors following the contributor being assessed, whether the account had been verified by Twitter and the contributor’s avatar.

The journalists were also asked to give detailed descriptions of how they arrived at their credibility scores for the five contributors whose scores differed the most from those assigned by Truthmeter. In these cases, thematic analysis of the responses revealed that journalists appeared to rely heavily on an assessment of who was following a contributor. In general, if credible or trusted contributors followed a contributor, then that contributor was to some extent imbued with their credibility. For example, of one contributor the journalists observed that ‘although he works for the BBC, he has only posted 11 tweets suggesting that he isn’t an active Twitter user’, but also that ‘some of the accounts I follow, follow this one, which makes me inclined to trust it’. Likewise, in commenting on a different contributor, one journalist observed that ‘the fact that the account is not verified always doubts me, and it does even more when the “bio” section doesn’t have a proper description of who the person is’, but conversely, ‘I also saw that people that I trust follow him and used that to confirm he was who he says he was.’

Similarly, it also emerged that if a contributor was followed by contributors who in some way corroborated the information in their biography, they were deemed more credible. For example, when asked about a contributor who had mentioned Sky News in their biography, a journalist commented that ‘her bio is short and to the point suggesting she uses Twitter seriously’ but ‘she doesn’t have many followers and her tweet count is extremely low’. However, this was offset because ‘I then checked with the people I know who work at the same organization to confirm they followed her.’

To sum up, the responses suggested that, in many cases, it was possible to use the data collected by Twitter during the course of a contributor’s activity to arrive at an assessment of their credibility. The journalists placed a particular emphasis on the number of followers a contributor had, and whether or not they were verified by Twitter. It was only when the information provided by Twitter did not prove to be sufficiently revealing, and credibility assessments became difficult, that they looked to other sources of information (such as Internet searches). That the information made available by Twitter was, in many cases, sufficient to make credibility judgements appeared to confirm the validity of the Truthmeter approach.

However, the responses also suggested that, in some cases, the metrics that the first version of Truthmeter used were not able to arrive at accurate credibility judgements if the contributor was inactive, new to Twitter or credible due to their connections to other credible contributors. It emerged that they were unable to capture how journalists interpreted the relationships between contributors, and in this sense the metrics overlooked a ‘relational’ dimension of how credibility functions on social media.

Third stage: Calibration, modification and testing

The rationale for the third and final stage of the evaluation was to use the findings from the online questionnaire (and other published research discussed in the ‘Literature review’

section) to inform modifications to the Truthmeter, and to then evaluate whether these changes had resulted in an improvement. At this point, given that the results of the first two stages of the evaluation had partially confirmed the Truthmeter approach, it was decided that future modifications should take the form of calibrations, with particular attention focused upon the metrics that Truthmeter used, as well as how they were weighted.

Based on the results of the online questionnaire, the ‘number of followings’ metric was replaced with a metric based on the ratio of followers to followings, and the ‘number of mentions’ metric was replaced with a metric based on whether or not the contributor had been verified by Twitter (see Table 2). Importantly, the first version of the Truthmeter assumed that each of the metrics it used were of equal importance when calculating credibility scores. However, it was decided that this was unlikely to align well with the relative importance attached to each metric by journalists. Therefore, based on the findings from our own research and those described in other studies, a set of weighted metrics was proposed. Individual metrics were weighted according to a 1–5 scale, with five indicating very high importance. The importance assigned to each metric was then used to determine how significant each metric would be when computing credibility scores. Furthermore, for the weighted set, an additional ‘popularity’ metric (the number of days since the account was created divided by the number of followers since then) was included. This metric was included as it was seen as able to address one of the issues identified in earlier stages of the evaluation, namely that new contributors with a low number of tweets and a relatively low number of followers would be deemed wrongly to have low credibility. In other words, the ‘popularity’ metric can also be thought of as a measure of a contributor’s popularity outside of their activity on social media, as it reflects their ability to accumulate a large number of followers without necessarily being a long-time social media user.

Table 2: Weighted metrics used by the modified version of the Truthmeter

Metric	Category	Unweighted	Weighted
Number of tweets	History	3	1
Number of re-tweets	Influence	3	2
Number of followers[4]	Popularity	3	4
Ratio of followers to followings	Popularity	3	3
Verified	None[5]	3	5
Frequency	History	3	2
Popularity[6]	History	-	5

The third stage of the evaluation used a similar structure to the first stage. A subsample of the original 150 contributors was identified. Then, credibility scores for these contributors were produced using each metric set, and were compared to the same scores (for the same contributors) assigned by journalists during the credibility-scoring task.

The mean score produced using the weighted metrics ($\overline{x}$ = 6.37, SD = 0.91) was more closely aligned with the mean score assigned by journalists ($\overline{x}$ = 6.71, SD = 1.31) than that produced using the unweighted set ($\overline{x}$ = 7.74, SD = 1.24). In terms of the Gupta et al. (2014) measure, 89% of the scores produced by the unweighted metrics were within 2 points (on the 0–9 scale) of the mean journalist scores, compared to 100% of the weighted scores. Furthermore, 85% of the weighted scores were within 1 point of the mean journalist scores. As with the first stage of the evaluation, Spearman's correlation tests were applied to the journalists' scores and each of the metric sets. There was a positive correlation between the scores produced using the unweighted metrics and those assigned by journalists: $r_s(62) = .54$, $p < .01$. There was also a strong positive correlation between the scores produced using the weighted metrics and those assigned by journalists: $r_s(62) = .64$, $p < .01$. The tests therefore

suggested that the scores produced using the weighted metrics were more strongly correlated with the scores assigned by journalists than those produced using the unweighted metrics.

Based on these results, the weighted metrics were selected to produce the credibility scores for the next version of the Truthmeter. Though the weighted set was not designed to test a particular hypothesis, its relative superiority over both the version of Truthmeter from the first stage of the evaluation and the unweighted version from the third – particularly in terms of the Gupta et al. (2014) measure – suggested that the addition of the 'popularity' metric had partially addressed the problem of new but credible contributors, thus producing scores that were more closely aligned with the journalists' assessments.

Discussion

The emerging literature on journalism and social media highlights the fact that the traditional verification practices used by journalists to assess credibility may not be able to deal with the amount of information made available on social media. This points to a need for new tools to assist journalists with this task. In response, the SocialSensor project has developed the Truthmeter – a tool that uses data from social media contributors' Twitter profiles to automatically assign to them a credibility score on a 0–9 scale.

Following a three-stage evaluation procedure, the Truthmeter was able to produce contributor credibility scores that aligned well with the credibility scores assigned by journalists. The first stage of the evaluation essentially provided a ground truth against which to judge the Truthmeter prototype, as well as a benchmark of how well the Truthmeter performed. The results suggested that it was possible to produce reasonable contributor credibility scores based on data generated during the course of their Twitter activity. However, they also highlighted the fact that the Truthmeter produced poorly aligned credibility scores if the contributor was either inactive or a new user.

The results of the questionnaire from the second stage showed that journalists do make credibility assessments based on Twitter profile information. They also revealed that in most cases they attach greater importance to certain metrics, in particular the number of followers a contributor has and whether they are verified by Twitter. The results also showed that if a contributor was a new user, the relatively low number of followers and tweets might provide a misleading indication, with the contributor's offline presence actually indicating a relatively high degree of credibility. Finally, the results highlighted the fact that contributors can be imbued with credibility if they are followed by other credible contributors, thus demonstrating a relational dimension to credibility assessments.

The third stage of the evaluation attempted to integrate these insights into the Truthmeter by adjusting the weighting (or relative importance) of existing metrics, and introducing new ones. More specifically, the number of followers and the verified metrics were weighted strongly, and the popularity metric was introduced to allow the Truthmeter to arrive at more closely aligned scores for new but credible contributors. When compared to the unweighted metrics, the weighted set appeared to perform better. Furthermore, given that the questionnaire revealed that journalists do not attach equal importance to all metrics, the weighted metrics improved the Truthmeter in a way that was justifiable.

Limitations and further work

There are nonetheless some weaknesses with the approach outlined. Firstly, the conclusions from each stage were based on feedback from eight trainee journalists. As such, their particular credibility assessment practices informed the development of the Truthmeter, and their credibility scores were used to measure these modifications. There is undoubtedly a certain circularity to this process, and a lot rested on the quality of the data they provided. Although we considered their evaluations to be professionally credible, including older

journalists with more professional experience in the panel would have been desirable. It should be noted, however, that the process described in this article was just part of what is a larger scheme of testing and iteration. The Truthmeter module is being further developed as part of a follow-on EU FP7 project, REVEAL.[7] Secondly, due to practical constraints, it was not possible in this case to rigorously test hypotheses about the improvements associated with using weighted metrics, or the use of the 'popularity' metric. Constraints based on the imperatives of the SocialSensor project meant that it was not possible to test these independently. Therefore, it can only be suggested that these modifications were, to varying degrees, responsible for the Truthmeter's improved performance. However, there is no reason why other evaluations using the three-stage approach could not make use of statistical hypothesis testing during the final stage.

In addition to these methodological limitations, there also exists plenty of scope for further work in the form of improvements to the Truthmeter itself. For example, as previous studies have acknowledged (e.g. Castillo et al., 2011), credibility is multifaceted, and the journalistic credibility of the contributor is only one dimension. Though there may be some overlap, assessments of context and content are needed to complete the picture.

Perhaps most interestingly, more can be done to better incorporate the relational way in which journalists assess the credibility of social media contributors. This insight emerged from the questionnaire issued to the journalists during the second stage of the evaluation, and showed that contributors appear to be imbued with credibility if they are followed by credible contributors. This has also been briefly alluded to, but not made completely explicit, in other studies. For example, during an exercise with journalists and their SRSR tool (which foregrounded information from contributors that the journalist followed), Diakopoulos et al. (2012) observed that 'one participant noted that a source had much more credibility because it had the Red Cross as a follower', leading them to conclude that, in the future, 'network

representations can be developed to evaluate a source or confer credibility'. Here, we use 'relational' in the same broad sense as Scott (2000) in his description of the field of network analysis. As such, it refers to the connections between agents within a networked system, and can be contrasted with attribute data, which refers to the qualities or characteristics of agents, and is typically understood in terms of variables. Though platforms like Twitter can be readily understood in terms of network analysis, contributor credibility has been largely understood in terms of attributes. Even metrics such as 'number of followers' – despite representing the number of connections a contributor has to others within a network – are to a large extent treated as individual attributes, without proper consideration of what those connections might imply for how credibility flows around a network. A relational view of contributor credibility would therefore aim to extend this conceptualization by making an additional attempt to understand how credibility behaves within a system of networked contributors, as well as providing an understanding of the nature of connections and the context within which they are established.

Despite the existence of well-established concepts within the field of social network analysis, capturing the relational aspect of journalists' credibility judgements is likely to be challenging. However, as Singer (2012) has argued, for journalists, coming to terms with the networked aspect of news creation is now of primary importance. Some have argued that the adoption of social media should prompt a move towards 'collaborative verification' (Hermida, 2012), where news stories are allowed to develop in the open in response to contributions from non-journalists (e.g. in the form of live blogs). However, journalists may be reluctant to embrace practices that appear to dissolve the boundaries between them and the public. For good reasons, journalists are also unlikely to abandon traditional verification practices. Yet, at the same time, news is being written, distributed and shared automatically, with Facebook, Google and Twitter key players. The sheer volume of social media output means that sorting

is inevitable. For journalists to continue to be relevant in the online world, they will need to embrace digital tools, including those that allow them to highlight credible information. However, straightforwardly computing a contributor's credibility score based on a correlation with, say, their raw number of followers, may understandably be viewed sceptically by journalists. In contrast, assessments based on the nature of those followers, and the networks of which they are a part, are likely to chime with established verification practices. The substantive results of this evaluation therefore suggest that automated approaches to verification should be based on extracting and utilizing the collective intelligence located within social media by viewing credibility as transferable through the networked connections between contributors that social media makes visible.

Funding: Our thanks to the European Union for supporting this research as part of a SocialSensor consortium FP7 research grant (number 287975). The article was also supported by a Volkswagen Foundation Freigeist Fellowship.

Notes

1. See www.socialsensor.eu for more information about the SocialSensor project.

2. 'Frequency' is defined as the number of days since the account was created divided by the total number of posts since then.

3. 'Number of re-tweets' refers to the total number of re-tweets of all of a contributor's tweets rather than the number of re-tweets for one particular tweet.

4. The 'number of followers' metric has a curvilinear relationship with credibility scores. Therefore, a very high number of followers results in a lower credibility score than a high number of followers.

5. This is indicated on the contributor page (see Figure 2) by a ‘verified’ icon. Though this metric does not feed into any particular category, it does feed into the overall contributor credibility score.

6. The popularity metric gives a credibility boost to social media contributors in proportion to the number of followers they amass in a given period. The more followers, the greater the boost. Because of the way the metric is calculated (number of days since the creation of the account divided by the number of followers since then), it has an inverse relationship with credibility scores. Therefore, a high popularity metric does not indicate high popularity – it results in a low credibility score.

7. https://revealproject.eu

References

Brandtzaeg PB, Lüders M, Spangenberg J, Rath-Wiggins L and Følstad A (2016) Emerging journalistic verification practices concerning social media. *Journalism Practice* 10(3): 323–342.

Broersma M and Graham T (2012) Social media as beat: Tweets as a news source during the 2010 British and Dutch elections. *Journalism Practice* 6(3): 403–419.

Broersma M and Graham T (2013) Twitter as a news source: How Dutch and British newspapers used tweets in their news coverage, 2007–2011. *Journalism Practice* 7(4): 446–464.

Bruno N (2011) *Tweet first, verify later? How real-time information is changing the coverage of worldwide crisis events*. Oxford: Reuters Institute for the Study of Journalism. Available at: http://reutersinstitute.politics.ox.ac.uk/publication/tweet-first-verify-later (accessed 2 March 2016).

Castillo C, Mendoza M and Poblete B (2011) Information credibility on Twitter. In: *Proceedings of the 20th International Conference on World Wide Web* (eds S Sadagopan, K Ramamritham, A Kumar and MP Ravindra), Hyderabad, 28 March – 1 April, pp.675–684. New York: Association for Computer Machinery.

Castillo C, Mendoza M and Poblete B (2013) Predicting information credibility in time-sensitive social media. *Internet Research* 23(5): 560–588.

City, University of London (2017a) International Journalism. In: City.ac.uk. Available at: http://www.city.ac.uk/courses/postgraduate/international-journalism (accessed 19 April 2017).

City, University of London (2017b) Magazine Journalism. In: City.ac.uk. Available at: http://www.city.ac.uk/courses/postgraduate/magazine-journalism#during-your-course (accessed 19 April 2017).

Diakopoulos N, De Choudhury M and Naaman M (2012). Finding and assessing social media information sources in the context of journalism. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (ed JA Konstan), Austin, 5–10 May, pp.2451–2460. New York: Association for Computer Machinery.

Edwards C, Spence PR, Gentile CJ, Edwards A and Edwards A (2013) How much Klout do you have... A test of system generated cues on source credibility. *Computers in Human Behavior* 29(5): A12–A16.

Flanagin AJ and Metzger MJ (2007) The role of site features, user attributes, and information verification behaviors on the perceived credibility of web-based information. *New Media & Society* 9(2): 319–342.

Gans HJ (1979) *Deciding What's News: A Study of CBS Evening News, NBC Nightly News, Newsweek, and TIME*. New York: Pantheon.

Gupta A, Kumaraguru P, Castillo C and Meier P (2014) TweetCred: Real-time credibility assessment of content on twitter. In: Social Informatics: 6th international conference, SocInfo 2014 (ed LM Aiello and D McFarland), Barcelona, Spain, 11–13 November 2014, pp. 228–243. Springer International Publishing.

Heravi BR and Harrower N (2016) Twitter journalism in Ireland: Sourcing and trust in the age of social media. *Information, Communication & Society* 19(9): 1194–1213.

Hermida A (2012) Tweets and truth: Journalism as a discipline of collaborative verification. *Journalism Practice* 6(5–6): 659–668.

Kovach B and Rosenstiel T (2007) *The Elements of Journalism: What Newspeople Should Know and the Public Should Expect (Completely Updated and Revised)*. New York: Three Rivers Press.

Larsen, Anna Grondahl (2016) Investigative reporting in the networked media environment: Journalists' use of social media in reporting violent extremism. *Journalism Practice*. doi: 10.1080/17512786.2016.1262214

Lecheler S and Kruikemeier S (2016) Re-evaluating journalistic routines in a digital age: A review of research on the use of online sources. *New Media & Society* 18(1): 156–171.

Morris M and Ogan C (1996) The Internet as mass medium. *Journal of Computer-Mediated Communication* 1(4). Available at: http://doi.org/10.1111/j.1083-6101.1996.tb00174.x (accessed 2 March 2016).

Morris MR, Counts S, Roseway A, Hoff A and Schwarz J (2012) Tweeting is believing? Understanding microblog credibility perceptions. In: *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work* (eds S Poltrock and C Simone), Seattle, 11–15 February, pp.441–450. New York: Association for Computer Machinery.

Newman N (2009) *The rise of social media and its impact on mainstream journalism*. Oxford: Reuters Institute for the Study of Journalism. Available at: http://reutersinstitute.politics.ox.ac.uk/publication/rise-social-media-and-its-impact-mainstream-journalism (accessed 2 March 2016).

Newman N, Fletcher R, Levy DAL and Nielsen RK (2016) *Reuters Institute digital news report 2016*. Oxford: Reuters Institute for the Study of Journalism, University of Oxford. Available at http://www.digitalnewsreport.org (accessed 29 November 2016).

Oliver L (2008) Spot the difference: AFP withdraws 'digitally altered' missile shot. In: Journalism.co.uk. Available at: http://blogs.journalism.co.uk/2008/07/10/spot-the-difference-afp-withdraws-digitally-altered-missile-shot (accessed 6 December 2016).

Pew (2015) *Social media usage: 2005–2015*. Washington DC: Pew Research Center. Available at: http://www.pewinternet.org/2015/10/08/social-networking-usage-2005-2015 (accessed 2 March 2016).

Powers A and Fico F (1994) Influences on use of sources at large U.S. newspapers. *Newspaper Research Journal* 15(4): 87–97.

Schifferes S, Newman N, Thurman N, Corney D, Göker A and Martin C (2014) Identifying and verifying news through social media: Developing a user-centred tool for professional journalists. *Digital Journalism* 2(3): 406–418.

Scott J (2000) *Social Network Analysis: A Handbook (Second Edition)*. London: SAGE.

Singer JB (2012) Journalism in the network. In: Allan S (ed) *The Routledge Companion to News and Journalism (Revised Edition)*. Abingdon: Routledge, pp.277–286.

Thurman N, Schifferes S, Fletcher R, Newman N, Hunt S and Schapals AK (2016) Giving computers a nose for news: Exploring the limits of story detection and verification. *Digital Journalism* 4(7): 838–848.

Tylor J (2015) An examination of how student journalists seek information and evaluate online sources during the newsgathering process. *New Media & Society* 17(8): 1277–1298.

Vis F (2013) Twitter as a reporting tool for breaking news: Journalists tweeting the 2011 UK riots. *Digital Journalism* 1(1): 27–47.

Westerman D, Spence PR and Van Der Heide B (2012) A social network as information: The effect of system generated reports of connectedness on credibility on Twitter. *Computers in Human Behavior* 28(1): 199–206.

Westerman D, Spence PR and Van Der Heide B (2014) Social media as information source: Recency of updates and credibility of information. *Journal of Computer-Mediated Communication* 19(2): 171–183.