Information trustworthiness as a solution to the misinformation problems in social media

Jason R.C. Nurse[1], Ioannis Agrafiotis[1], Michael Goldsmith[1], Sadie Creese[1], Koen Lamberts[2], Darren Price[3], and Glyn Jones[3]

[1]Cyber Security Centre, Department of Computer Science,
University of Oxford, Oxford, UK
{*firstname.lastname*}@cs.ox.ac.uk

[2]University of York, UK
koen.lamberts@york.ac.uk

[3]Thales UK Limited, Research and Technology
{*firstname.lastname*}@uk.thalesgroup.com

Abstract

The advent of the Internet has reshaped the way we communicate and interact in our daily lives. It provides an ideal medium through which we can share information and ideas, form groups, and contribute to a variety of discussions. In this position paper, we focus specifically on the information now available online – especially content from social media – to consider in detail the challenges that such information poses to modern-day society. Typical examples of challenges include the prevalence of mistaken information and deliberate misinformation and rumours. With an understanding of these challenges, we then introduce the notion of information-trustworthiness measures as a potential solution to the problem of misinformation in social media. The idea here is to use quality and trust metrics to assess information, and then, based on values attained, advise users whether or not they should trust the content. This paper extends our previous research in the field by assessing the misinformation problem in much greater detail, and also presenting our current agenda for future work.

Introduction

The Internet has revolutionised the way that we, as humans, communicate and interact with each other. It provides a ubiquitous and, in many ways, ideal medium, through which we can share information and ideas, contribute to a range of discussions, and discover more about the world around us. Given its suitability for communication there should be little surprise at the extent to which it is currently used and the vast amount of data being shared every day, in particular via social media. To take Facebook as an example, each day 2.5 billion content items are shared (including information, photos, posts) [1].

Similar to their more traditional predecessors, social media have become a critical tool in influencing people's perceptions and decisions. Whilst this influence is often positive and well-intended (e.g., a tweet from a local council informing motorists of a blocked road), real-world cases continue to show instances of individuals poisoning information for their own malicious ends, and unfortunately, with serious consequences. The case of the London Riots in 2011, where deliberate rumours led to confusion over where emergency services should be deployed [2], is one example.

In this position paper, we reflect on the problem of misinformation on social media, and the use of information quality and trust metrics to help address it. This paper considers and also builds on previous research to outline an agenda for our future research aimed at addressing these outstanding issues. The main aim is developing a full system that is able to consume social content, assess the trustworthiness that should be associated with it, and generally help understand what might be happening in on-going scenarios such as emergencies or crises.

The misinformation problems with social media

Social media have provided us with many opportunities to discover, learn and interact. Unfortunately, however, there are several problems accompanying this capability, one of the largest being the misinformation or information poisoning problem (i.e., the posting of inaccurate or misleading information) and its use to negatively influence individuals. Take the examples below.

In the summer of 2011 several British cities experienced a significant period of unrest with spates of rioting and looting. The violence originated in London, but rapidly spread across the UK mostly affecting large cities including Manchester and Birmingham. In the aftermath, social media were put in the spotlight. Governmental authorities claimed that information poisoning facilitated the spread of rioting, either via circulating rumours presenting an overly chaotic situation or via sharing photos of police officers who remained indifferent while looting was taking place in their presence [3]. The role of social media in encouraging the riots and disrupting essential response was deemed so critical, that even the prospect of temporarily blocking access to Twitter and Blackberry

Messenger was raised by authorities. This gives some insight into the significance of the problem faced and challenges to official responders.

One of the main avenues in which rumours were spread during the riots was the micro-blogging platform, Twitter. According to retrospective reports, thousands of individuals re-tweeted dubious content leading to a sea of misinformation as the incident unfolded [4]. What is of great interest, though, is the extent to which people appeared to question their knowledge and common sense to embrace the rumours. For instance, an image portraying the London Eye in flames was heavily re-tweeted initially, and only after being online for a while did someone expressed doubts about the trustworthiness of the tweet; they rightly noted that the London Eye is made of iron and thus, it was difficult to imagine it ablaze. Even after the tweet debunking the rumour, more than 700 people within the next three hours re-tweeted the image expressing their anger at the destruction of the London attraction [5].

An additional problem was the enormous amount of data generated as a response to such tweets. This had a direct negative impact on police efforts to analyse the situation in the places where riots were taking place and to respond accordingly. Chris Sims, chief constable of West Midlands police, said his “force was actively engaged in trying to dispel information it believed to be untrue”, thus wasting valuable police resources [2]. In addition, the gold commander of Greater Manchester police described the amount of data from social media as overwhelming, recognising that “police struggled to analyse it even in the most basic way”, and also calling for innovative systems to elicit actionable intelligence from social media in an effective and quick manner [2].

Another case where misinformation from social media affected people's decisions with dramatic consequences was the Boston Marathon bombing in 2013 [6]. Within seconds of the first explosion, speculation, rumours and reactions from the masses dominated social media discussions. While first responders were on route to the incident, there were posts reporting additional explosions, library buildings being targeted, increased casualties, and even accusations against the Muslim community as being responsible for the attack [7]. Although the motives behind these rumours may not all have been malign, certainly such misinformation hindered authorities in allocating their resources effectively.

An example of the potentially devastating impact of such misinformation emerged from the rush to identify the perpetrators of the bombing attack. Once the FBI released photos from the scene where it took place, several social-media users responded by reviewing

ISSN 2052-8604

the information and naming anyone that looked similar as a potential suspect [7]. This took a dramatic turn when a tweet claiming that the Boston Police department had declared Sunil Tripathi and Mike Mulugeta as suspects, went viral, with thousands of individuals re-tweeting the names. Possibly as a result, Sunil Tripathi, who had nothing to do with the case, disappeared the same day and was found dead one month later [7].

From the cases above it is evident that social media can be exploited to misinform and to circulate inaccurate information with sometimes devastating consequences, even the loss of innocent lives. The need to develop mechanisms to evaluate the quality and trustworthiness of social-media information is therefore more urgent now than ever before.

Measuring the trustworthiness of online content

Previous research

Information quality and trustworthiness have been of interest to researchers for some time. To assess the quality of information, a typical question is, how fit is the information for its intended use. Trustworthiness can be thought of as an extension of quality, as it looks at the perceived likelihood that a piece of information will preserve a user's trust and belief in it [8]; presuming the information is of high quality therefore, the likelihood might arguably be high as well.

There have been numerous proposals that aim to utilise the quality and trust factors identified above to measure the trustworthiness of social content automatically. Agichtein et al., for instance, focus on the problem of finding high-quality content in social media and propose a classification framework for combining evidence (especially related to the quality factors discussed prior) from different sources of information [9]. As it pertains to the trustworthiness and credibility of online content, Castillo et al. draw on similar general factors (regarding features of the message, the information's source, and the topic) and use a supervised classifier (machine learning) to produce automated measurements of a tweet's credibility [10]. These are just two of the many approaches that aim towards this problem; space limits how much we can cover here, but readers are free to read more in [11]. Through the use of these automated techniques there is hope for a more general approach to tackle the misinformation problems plaguing online content.

Our work in the TEASE project

The TEASE research project was born out of the need to address the misinformation problems commonly faced with online social-media content. Our objective was to research and prototype a computer system that was able to measure the trustworthiness of information, and feed this back to users to assist them in making decisions. There were several significant contributions made by TEASE. The first was a novel methodology and framework for assigning trustworthiness measures to openly-sourced information, including tweets, Facebook posts, and news reports [12-13]. This approach considered key trustworthiness aspects, including provenance, intrinsic quality, and infrastructure integrity, and their related sub-factors such as the identity of a source, their reputation and competence, how timely the information was, and the vulnerabilities and threats to the infrastructure through which information traversed before reaching the user. Through an analysis of information (and its related metadata) in terms of these factors, we were able to produce trust scores (one per item) that could then be displayed along with the related content. These would therefore help to identify misinformation early on and hopefully prevent its spread.

With regard to the user interface and ensuring that it was highly usable, we engaged in numerous user experiments, both with the general public, and for specific use cases, with experts (e.g., in crisis management). There were several notable findings from our experimentation. For instance, traffic lights are much more effective communicators of trustworthiness than other visual means such as stars or transparency [13]; that is, lights were better able to direct individuals away from bad information and towards good information. Another crucial finding was that individuals are astoundingly capable of combining trustworthiness ratings and evaluative information to make efficient judgements [13]. The experiment in this case was based on the common assumption that individuals can easily combine sets of information (e.g., tweets describing what's happening in a scenario) and their respective trust scores (e.g., assignments of various trustworthiness levels to the tweets) to first, understand what might be happening in the scenario, and then to make decisions. Both these findings assisted in our interface design but also contributed to broader research in the field of communicating quality and trust.

Looking towards the future

This section looks towards the future and ways to extend current research to tackle the outstanding challenges of misinformation in social-media. We propose a research and development agenda which draws on our previous work, and is concentrated on the use of social information for official response purposes.

Social media present society with a plethora of opportunities, especially with regards to information to make decisions. The only way that these can be realised, however, is if the users of online content are able to identify inaccurate and misleading information, and have the tools to isolate high quality content. TEASE tackled this problem with notable success, in the creation of a flexible framework for measuring trustworthiness and an interface that emphasised usability. Nonetheless, there were important areas unable to be completely addressed in the lifetime of the project. One of these areas was the creation of a fully automated system, capable of working with live Internet feeds. The real challenge here is the research and design of a scalable system able to consume content about a specified topic (e.g., a bombing in Boston), use the TEASE methodology to measure the trustworthiness of all the items, and present information and annotated trustworthiness levels back to users in a timely manner. This is all with the understanding that in crises, there are typically hundreds of social-media posts per minute, a myriad of new users joining to contribute (thus, persons with unknown reputation levels), and metadata about content often missing (e.g., the location of an information source is key to assessing an eyewitness attribute).

Another feature that would be extremely valuable in such a system is the notion of World Views introduced in [12]. A World View is a cluster of social-media information (e.g., tweets and posts) that is related to each other (i.e., about the same topic) and is somewhat consistent, i.e., there is little discrepancy between the information items. Our research pursuit with respect to World Views therefore, would be defining how to create the clusters. We envisage an approach involving Natural Language Processing (to better understand the information and facilitate comparison) and formal modelling (to build consistent clusters). Even then, considering that the range of text is so expansive, it will be crucial to scope the problem – this is another reason that we have chosen crisis response. In this field, there are several existing encoding formats for content that will be invaluable. Additionally, we will be able to blend social-media content with closed-source intelligence (e.g., reports from emergency-service personnel) within World Views to create a more complete picture for responders.

With a fully functional system, the next aim will be evaluating it, and particularly its use in supporting decision-making during crisis situations. We propose a set of experiments where experts use the system first within a controlled context, where we can carefully monitor for any usage issues, and then, once any feedback has been incorporated, in the field. To clarify, we do not envisage a fireman with a tablet PC searching through rubble, but rather, a control centre directing first responders based on information now marked with trustworthiness scores. The utility of the system could be judged based on interviews and questionnaires after response to events.

Acknowledgements: TEASE was a collaborative research project that involved the University of Oxford, University of Warwick, HW Communications Ltd and Thales UK Research and Technology. The project was supported by Innovate UK's Trusted Services Competition (www.innovateuk.org) and the Research Councils UK Digital Economy Programme (www.rcuk.ac.uk/digitaleconomy).

References

[1] CNET News: Facebook processes more than 500 TB of data daily (2012) http://news.cnet.com/8301-1023 3-57498531-93/facebook-processes-more-than-500-tb-of-data-daily.
[2] Guardian News: Riot rumours on social media left police on back foot (2012) http://www.theguardian.com/uk/2012/jul/01/riot-rumours-social-media-police.
[3] Guardian News: UK riots 'made worse' by rolling news, BBM, Twitter and Facebook (2012) http://www.theguardian.com/media/2012/mar/28/uk-riots-twitter-facebook.
[4] Guardian News: How twitter was used to spread and knock down rumours during the riots (2011) http://www.theguardian.com/uk/2011/dec/07/how-twitter-spread-rumours-riots.
[5] Guardian News: How riot rumours spread on twitter (2011) http://www.theguardian.com/uk/interactive/2011/dec/07/london-riots-twitter.
[6] Los Angeles Times: Boston bombings: Social media spirals out of control (2013) http://articles.latimes.com/2013/apr/20/business/la-boston-bombings-media-20130420.
[7] The Atlantic: #bostonbombing: The anatomy of a misinformation disaster (2013) http://www.theatlantic.com/technology/archive/2013/04/-bostonbombing-the-anatomy-of-a-misinformation-disaster/275155/
[8] Kelton, K., Fleischmann, K.R., Wallace, W.A.: Trust in digital information. Journal of ASIST 59(3) (2008) 363-374
[9] Agichtein, E., Castillo, C., Donato, D., Gionis, A., Mishne, G.: Finding high-quality content in social media. In: International Conference on Web Search and Data Mining, (2008) 183-194
[10] Castillo, C., Mendoza, M., Poblete, B.: Information credibility on twitter. In: International Conference on World Wide Web, ACM (2011) 675-684
[11] Moturu, S.T., Liu, H.: Quantifying the trustworthiness of social media content. Distributed and Parallel Databases 29(3) (2011) 239-260
[12] Rahman, S.S., Creese, S., Goldsmith, M.: Accepting information with a pinch of salt:

handling untrusted information sources. In: Security and Trust Management Workshop. Springer (2012) 223-238

[13] Nurse, J.R.C., Agrafiotis, I., Goldsmith, M., Creese, S., Lamberts, K.: Two sides of the coin: measuring and communicating the trustworthiness of online information. Journal of Trust Management 1(1) (2014) 1-20