Gresham College

23rd Nov 2019

Taming the Trolls of Social Media

Professor Richard Harvey FBCS

Marshall McLuhan seems to be generally credited with first using the term "Moral panic" to encapsulate mass concern about something which was perceived as harmful. The Sociologist Stanley Cohen put the concept of a firmer foundation with a description of the five stages of moral panic:

1. Someone, something or a group are defined as a threat to social norms or community interests;
2. The threat is then depicted in a simple and recognizable symbol/form by the media;
3. The portrayal of this symbol rouses public concern;
4. There is a response from authorities and policy makers;
5. The moral panic over the issue results in social changes within the community.

I'm not sure where we are on the scale with regard to Social Media. Probably with Facebook at Level 4? Note that moral panics are often associated with new technology and usually based on precious little evidence. The evidence base is something which particularly concerns me with Social Media. It is thin. By which I mean there are a few noble researchers trying very hard to come to firm conclusions but, for reasons we shall explore in this lecture, many of the conclusions are tentative. And into this space of tentative conclusions and "messy" science ride charlatans, deniers and contrarians who are given too much air-time due to the prevalence of "false balance" or "bothsideism"[1].

The messiness of the science starts with definitions. What exactly *is* social media? There are surprisingly few definitions. Many modern authors are citing [1] which is a list of social media characteristics and was in no way intended as a definition[2]. They posit that:

1. Social Media are characterised by Web 2.0 Internet-based applications;
2. User-generated content is the lifeblood of social media;
3. Social Media allows creation of user-specific profiles for a site or app designed and maintained by a social media service;
4. Social Media facilitates the development of social networks online by connecting a profile with those of other individuals and/or groups.

Leaving aside that Item 1 is a tautology since Web 2.0 is defined to be web interfaces that incorporate social media, this is a pretty unsatisfactory situation. As an alternative, I would claim that the essence of social media is that, firstly, anyone can join; secondly anyone can publish anything to anyone who will listen; thirdly that publication is instant and fourthly that all content is unmediated (there are no censors or editors). These four features contain the essence of social media and, as we shall see, each aspect creates its own problems. My definition also excludes, for example, work-based social media platforms such as Yammer since such platforms are usually subject to strict editorial policies. The definition also deliberately exposes the philosophical problem that faces Facebook – if they censor the publications too much then they cease to be what they want to be.

That said, it is very unusual for a Gresham Lecture to cover material that is not properly defined. Consequently, unlike my other lectures, we are going to be dealing with, to use some jargon, *factoids* or "facts" that are not yet fully established. The jargon of social media is delightfully colourful and fun to know. A *flame* is counterargument but usually made with a very considerable degree of ad-personam commentary. Indeed, classic flaming may consist

[1] This is when journalists, in the interests of balance, present the issue as more balanced between opposing viewpoints than the evidence suggests. An example would be the inclusion of anti-vaccination proponents in debates about public health.

[2] The list appears in an introduction to special issue in an online journal. The introduction is quite short and doubt very much if the authors intended it to serve as a cited definition of social media.

of considerable personal insult and character assassination. A series of flames and counter arguments are known as *flame wars*. Flame wars are often more vicious than ordinary disagreements since the internet has a well-documented disinhibiting effect. Sometimes people post hoping to incite flame wars – this *flame bait*. Flame wars are often over arcane subjects, such as whether Blue-ray discs were superior to HD DVDs[3], but more recently flame wars have spread into public for a such the election battle between Donald Trump and Hilary Clinton. People who post flame-bait are usually designated *trolls* but a troll has a wider definition and includes posters who wish to upset the apple cart (or indeed upset anyone). Some trolls spread misinformation (when they do this under a false flag then they will usually be referred to as a *sock puppet* or *meat puppet*). The subject of trolls has received considerable sociological attention and there are taxonomies of trolls including the recent *concern troll* – a concern troll pretends to agree with you but will attempt to sway the argument "I agree with you but … I have concerns …" is the usual syntax for a concern troll. Another recent innovation has been corporate trolls – large numbers of people in the pay of corporations paid to influence opinion. If they appear to be a grass-roots movement then this is called *astroturfing*. Typically, these are in the pay or influence of authoritarian governments and appear to post suspiciously praiseworthy material along with fierce criticism of governmental critics. That said, I am writing this in the run-up to British general election and both the Labour and Conservative parties appear to have mobilised twitter activists who will be regarded by the other side as trolls[4].

I hope it is evident from the above discussion that Social Media is sometimes associated with polarised and therefore simple opinions. And polarised opinion means that it is amenable to readers giving it the "thumbs up" or the "thumbs down". Thus, it is common, particularly in social media to measure *sentiment*. In my other job at UEA, I am Director of Admissions, and it is commonplace for me to see a "sentiment analysis" of the University's twitter stream – it is a routine part of monitoring our communication effectiveness.

Ideas that are spread by Social Media are often called *meme*s. The term meme was coined by Richard Dawkins who I am sure had in mind interest in how grand ideas spread across humanity – how they were preserved and modified across generations. He also postulated that either Darwinian or Lamarkian traits can be applied to the evolution of memes[5]. I'm fairly sure that he did not have in mind "planking" or the "ice bucket challenge" when he was thinking about memes but of course they are cultural memes too. And although such memes might seem trivial, they are lot easier to study than complex ideas.

Since it is simple to identify social media posts about similar subjects by looking for hashtags, it is great way to spot the spread of memes. The best model for the spread of memes are those provided by epidemiologists – an epidemic spreads the same way [2,3]. A technical question is the number of contacts required. Most diseases spread with a single contact or a "simple" contagion. Social media requires multiple contacts before a meme is picked-up which is the so-called "complex contagion" model[6]. In summary, these models show that these internet fads can spread widely and quickly.

The speed, and cheapness of contagion, is obviously of great interest to commercial entities. In [4] researchers downloaded and examined 2000 Facebook posts from Audi, BMW, Chevrolet, Ford, Honda, Hyundai, Mercedes-Benz, Nissan, Toyota, and VW. They then analysed all the likes from the millions of followers to try to identify the things that worked. The first pattern is co-branding – mentioning other brands in a post. The second, the wow-effect, is harder to quantify but seems to amount to showing something unusual enough to cause some admiration. The third pattern is to include a cognitive task or question such as how many balloons are in this car? Many posts were timed to coincide with some major event – maybe a sports event. Thus, timing is the fourth pattern. And the fifth pattern is to be part of a campaign – multiple exposures to a message have more impact particularly from multiple channels.

One of the great attractions of social media is not only the cheapness of the system and its tremendously large distribution but also that several of the proprietors, particularly Facebook and Google will sell you lots of personal

[3] Yes, Dear Reader, this, according to Wikipedia was a genuine flame war. The website AVS Forum was temporality closed while the police investigated threats made by one poster to another.

[4] Indeed, I was pretty surprised when the Conservative Party Press Office relabelled it's Twitter feed as a fact checking service. This is classic sock puppetry. I was even more surprised when leading Conservative politicians failed to apologies.

[5] Darwinian evolution is analogous to copying the instructions whereas Lamarkian evolution is copying the thing.

[6] For those of you who wish to increase your social media cred then reference [2] includes a description of 25 well known internet fads.

data about your readers. This allows advertisers to not only know, not only if an advert is liked, but ultimately if it led to sales. Knowing if a comment is positive or negative is a tricky problem in natural language processing (see previous lectures). Of course, it is simple to detect words that a positive such as "great!" and negative such as "dire" but human reviewers are complex and frequently use constructions such as "I would have really enjoyed this film if only the director was not a moron." Such phrases and other more complicated rhetorical forms are a remain a real challenge.

One of the great surprises in the early decade of 2010 was a paper from Google. They announced in the top journal, *Nature*, that they were able to predict the number of visits people would make to Physicians (or General Practitioners as they would be called in UK) by an analysis of Google search terms [4]. Although not strictly social media, the paper was interesting because it implied that humankind's interactions with the web could be used to predict demand – if more people are searching for flu, then more people would have flu and there would be more visits to the doctor. In other words, a flu epidemic could be predicted before it became a crisis. Unfortunately, they got their methods completely wrong. They had searched through a huge number of possible Google search terms looking for correlates with doctor visits. They got a good match on historic data but as the data rolled forward into the future the predictions became too poor and much worse than really simple predictions. That had "over fitted" the data and in 2013 the project was shut down [5].

But when the idea was transferred to twitter there is some potential. One particularly challenging problem is predicting the box office takings of a movie in its first weekend, or week. No-one is quite sure what constitutes a good movie so the best-known predictor is the Hollywood Stock Exchange (HSX) in which players buy virtual shares in movies. It is an efficient market so the "wisdom of the crowd" can lead to impressive predictive power. However, monitoring twitter activity before the movie releases is even more effective [8]. If you prefer natural phenomena then instead of tracking earthquakes or typhons, one can track mentions of them on twitter [9].

Instagram, the photo-based social media service can also be used for prediction. In [8] researchers analysed around 44,000 photographs from 166 people. They found that they could detect depression in the individuals by analysing the photographs. Impressively they could detect depression even before the individuals had received a diagnosis of depression – your Instagram feed tells others how you feel before you know it yourself. This is a rich vein of future analysis – detecting mental health conditions, particularly suicidal tendencies before it is too late. Likewise, your Instagram feed provides clues as to whether you might later become addicted to alcohol [8].

We now turn to the rather delicate issue of reliable research in social media. It is a fact that almost all the social media platforms are in private hands. This has had a number of consequences of which the most damaging is that research that is likely to give negative publicity to social media is not likely to be supported. So, research that is critical of social media, is often supported by rather small studies. Also, social media is rather rapidly developing – a small change in Google's search algorithm (and it changes all the time) might have dramatic consequences for your experimental data[7]. We should also note the increasing levels of moral panic about social media hardly make for a dispassionate research environment.

At this stage is difficult to give precise guidance on which aspects of social media are harmful. Some authors claim vigorously that screen time is damaging to young people. Others claim the effect is negligible. Parents tend to be fond of the precautionary principle so ban screens near bedtime even though they themselves probably ready racy books under the bedcovers by torchlight which is hardly healthy on a school night. However, one aspect that almost everyone agrees with is that social media has potential to deliver damaging, demeaning or dangerously wrong information to a large number of people at startling speed. In the lecture I play a clip from a film made to persuade people that the European Union would soon be a Muslim state. The movie has been viewed 3.5M times but, as the BBC Radio 4 programme "More or Less" showed, the film contains a pack of lies. Indeed, one has to search hard to find a single truth in the film. Needless to say, the counter-argument film has been viewed only 350k times.

Dealing with misinformation, or "fake news" as President Trump would say, is a major scientific challenge of our times. In my view the best solutions are likely to be found from the analogous field of epidemiology. To stop an

[7] This is said to have affected Google Flu Trends – the search algorithm changed so people's searches gave different results. Furthermore, for reasons unknown, the original Nature paper did not list the search terms so the work was unrepeatable.

epidemic, the first action is to make sure you do not pass on your disease to anyone else – don't retweet material unless you are sure it is true. Retweeting is currently the most casual of actions, there is no forced time delay, there is no cost of retweeting. Imagine if propagators of fake news were punished by being forced to lose followers. It is entirely feasible to develop algorithms for trust and why should low-trust tweeters be given the same access to networks as high-trust ones? This leads us to the second point which is people who generate or propagate fake news should be punished. To analogise, people who are infected need to be quarantined to protect the rest of us. Of course, this might lead to an "echo chamber" of extremists all ranting at each other, but that is better than them having access to all of us. The third prospect is immunisation. There is some emerging work in immunisation [10] which I like very much. But I would go further, I suspect there are some ideas which are so powerful that they make you far less prone to misinformation. Could it be that a liberal arts or scientific education make you less prone to misinformation? Or maybe it is specialist skill, as implied in The Debunkers Handbook? Or maybe it is journalistic training? That, the inoculative power of ideas, is a topic for the future – doubtless it will be debated vigorously on social media.

1. Obar, Jonathan & Wildman, Steven. (2015). *Social Media Definition and the Governance Challenge: An Introduction to the Special Issue.* SSRN Electronic Journal. 10.2139/ssrn.2647377.
2. Sprague DA, House T (2017) *Evidence for complex contagion models of social contagion from observational data*. PLoS ONE 12(7): e0180802. https://doi.org/10.1371/journal.pone.0180802
3. Granovetter, M. S. (1973). *The Strength of Weak Ties*,The American Journal of Sociology. **78** (6): 1360–1380. doi:10.1086/225469. JSTOR 2776392.
4. Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, et al. (2008) *Detecting influenza epidemic using search engine query data.* Nature 457: 1012–10155
5. *The Parable of Google Flu: Traps in Big Data Analysis*, David Lazer, Ryan Kennedy, Garry King and Alessandro Vespignani, Science, 14 Mar 2014 : 1203-1205.
6. S. Asur and B. A. Huberman, *Predicting the Future with Social Media*, 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, Toronto, ON, 2010, pp. 492-499.
7. Reece, A.G., Danforth, C.M. Instagram photos reveal predictive markers of depression. *EPJ Data Sci.* **6,** 15 (2017) doi:10.1140/epjds/s13688-017-0110-z
8. Hassanpour, S., Tomita, N., DeLise, T. *et al.* Identifying substance use risk based on deep neural networks and Instagram social media data. *Neuropsychopharmacol* **44,** 487–494 (2019)
9. Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. 2010. *Earthquake shakes Twitter users: real-time event detection by social sensors.* In Proceedings of the 19th international conference on World wide web (WWW '10). ACM, New York, NY, USA, 851-860.
10. *Neutralizing misinformation through inoculation: Exposing misleading argumentation techniques reduces their influence*, Cook J, Lewandowsky S, Ecker,. 2017, PLOS ONE 12(5): e0175799.