Written Representation 96

Name: Kalina Bontcheva
Professor of Text Analytics

Received: 1 Mar 2018

Written Representation to the Singapore Select Committee on Deliberate Online Falsehoods - Causes, Consequences and Countermeasures

Kalina Bontcheva
Professor of Text Analytics

University of Sheffield
Department of Computer Science

The phenomenon of using digital technology to deliberately spread falsehoods online

The past few years have heralded the age of ubiquitous disinformation - aka fake news - which poses serious questions over the role of social media and the internet in modern democratic societies. These emerge particularly strongly around high profile political events (e.g. elections), natural disasters, and controversial topics (e.g. climate change, vaccines)[1,2]. Typically the alt-facts posted on alternative media sites give rise to alternative narratives (e.g. rumours) on social media sites, where they are promoted often by automated bots and sockpuppet accounts.

Topics and examples abound, ranging from the Brexit referendum and the US presidential election to medical misinformation such as miraculous cures for cancer.

Social media now routinely reinforces people's confirmation bias, so often, little to no attention is paid to opposing views or critical reflections. Blatant lies often make the rounds, are re-posted and shared thousands of times and sometimes even jump successfully into mainstream media. Debunks and corrections, on the other hand, receive comparatively little attention and can easily be dismissed.

---

[1] M. Mendoza, B. Poblete, and C. Castillo. 2010. Twitter under crisis: can we trust what we RT?. In *Proceedings of the First Workshop on Social Media Analytics* (SOMA '10). ACM, New York, NY, USA, 71-79.

[2] Starbird, K. (2017). Examining the alternative media ecosystem through the production of alternative narratives of mass shooting events on twitter. In ICWSM, pages 230–239. AAAI Press.

Motivations and reasons for the spreading of such falsehoods, and the types of individuals and entities, which engage in such activity

So why is this happening? My short answer is the **4Ps of the modern disinformation age**: post-truth politics, online propaganda, polarised crowds and partisan media.

- **Post-truth politics**: The first societal and political challenge comes from the emergence of post-truth politics, where politicians, parties and governments tend to frame key political issues in propaganda, instead of facts. Misleading claims are continuously repeated, even when proven untrue through fact-checking by media or independent experts. This has a highly corrosive effect on public trust.
- **Online propaganda and fake news**: State-backed (e.g. Russia Today), ideology-driven (e.g. misogynistic or Islamophobic), or for-profit clickbait websites and social media accounts are all engaged in spreading misinformation, often with the intent to deepen social division and/or influence key political outcomes, such as elections and referenda. However, taken on their own, they should not be regarded as the sole source of online disinformation[3].
- **Partisan media**: The pressures of the 24-hour news cycle and today's highly competitive online media landscape have resulted in poorer quality journalism and worsening opinion diversity, with misinformation, bias and factual inaccuracies routinely creeping in. Many outlets also resort to highly partisan reporting of key political events, which, when amplified through social media echo chambers, can have acrimonious and divisive effects[4].
- **Polarised crowds**: As more and more citizens turn to online sources as their primary source of news, the combination of hyper-partisan media on one side and social media platforms and their advertising and content recommendation algorithms on the other, have facilitated the creation of partisan camps and polarised crowds, characterised by flame wars and biased content sharing, which in turn, reinforces their prior beliefs (typically referred to as *confirmation bias*).

The following **typology of the misinformation ecosystem** has been proposed by the non-profit coalition First Draft News[5]:

• **Satire or parody**: No intention to cause harm but with potential to fool;

• **Misleading content:** misleading use of information to frame an issue or an individual;

• **Imposter content**: when genuine sources are impersonated;

• **Fabricated content**: news content is 100% false, designed to deceive and do harm;

• **False connection**: when headlines, visuals or captions do not support the content;

• **False context**: when genuine content is shared with false contextual information;

---

[3] https://www.nytimes.com/2018/01/25/opinion/russian-trolls-fake-news.html

[4] M. Moore and G. Ramsay. UK media coverage of the 2016 EU Referendum campaign. https://www.kcl.ac.uk/sspp/policy-institute/CMCP/UK-media-coverage-of-the-2016-EU-Referendum-campaign.pdf

[5] https://firstdraftnews.com

• **Manipulated content**: when genuine information or imagery is manipulated to deceive.

Amongst these, state-of-the-art automatic detection and/or verification tools have focused primarily on identifying manipulated content (e.g. whether an image has been tampered with).

Detection of satire, imposter, and fabricated content have also been studied, in particular hoaxes, fake news, and conspiracy theories. For instance, recent computational research on disinformation detection in Wikipedia[6] showed that some hoaxes remained undetected for long periods and were widely cited. They also showed that humans were significantly worse at detecting hoax articles than the machine learning algorithm.

Researchers have also studied network-based visualisations of claim and misinformation spread, e.g. the Hoaxy system[7].

Research on identification of key sources of disinformation and propaganda has primarily focused on spam bot detection. State-of-the-art bot detection methods[8] are predominantly based on social behaviour features (e.g. tweet frequency, hashtag use). The short lifespan of political bot accounts and fake news sites and the fast emergence of new ones remain a key challenge, especially with respect to assessing the trustworthiness of the information source and its textual content.

Once identified, another key challenge is containment of disinformation spread in social networks. Computer science models have focused on identifying the key nodes that need to be "decontaminated"[9], e.g. using epidemiological models[10] or nodes that can be recruited to spread debunking information through the network[11]. However, most of these models fail to account for the effect of partisan nodes and alternative media, as well as lack empirical validation on real social network data. In addition, recent journalism research has found that exposing victims of disinformation to factual, non-partisan debunking may change their knowledge, but not their beliefs[12].

Consequences that the spread of online falsehoods can have on Singapore society

These have been argued for clearly already in the Green paper.

---

[6] Kumar, S., West, R., and Leskovec, J. (2016). Disinformation on the web: Impact, characteristics, and detection of wikipedia hoaxes. In WWW, pages 591–602. ACM.

[7] Shao, C., Ciampaglia, G. L., Flammini, A., and Menczer, F. (2016). Hoaxy: A platform for tracking online misinformation. In WWW'2016, pages 745–750.

[8] Varol, O., Ferrara, E., Davis, C. A., Menczer, F., and Flammini, A. (2017). Online human-bot interactions: Detection, estimation, and characterization. In ICWSM, pages 280–289. AAAI Press.

[9] NP Nguyen, G Yan, MT Thai, and S Eidenbenz (2012): Containment of misinformation spread in online social networks. In Proceedings of the 4th Annual ACM Web Science Conference.

[10] M Tambuscio, G Ruffo, A Flammini and F Menczer (2015): Fact-checking effect on viral hoaxes: A model of misinformation spread in social networks. In Proceedings of the 24th International Conference on World Wide Web, ACM

[11] C Budak, D Agrawal and A El Abbadi (2011): Limiting the spread of misinformation in social networks. In Proceedings of the 20th international conference on World wide web, ACM.

[12] B Swire, AJ Berinsky, S Lewandowsky, UKH Ecker (2017): Processing political misinformation: comprehending the Trump phenomenon. Royal Society Open Science 4(3).

How Singapore can prevent and combat online falsehoods

Promote National Fact Checking Efforts

In order to counter subjectivity, post-truth politics, disinformation, and propaganda, many media and non-partisan institutions worldwide have started fact checking initiatives – 114 in total, according to Poynter[13]. These mostly focus on exposing disinformation in political discourse, but generally aim at encouraging people to pursue accuracy and veracity of information (e.g. Politifact, FullFact.org, Snopes). A study by the American Press Institute has shown that even politically literate consumers benefit from fact-checking as they increase their knowledge of the subject.[14]

Professional fact checking is a time-consuming process that cannot cover a significant proportion of the claims being propagated via social media channels. To date, most projects have been limited to one or two steps of the fact checking process, or are specialized on certain subject domains: Claimbuster[15], ContentCheck[16] and the ongoing Fake News Challenge[17] are a few examples.

There are two ways to lower the overheads and I believe both are worth pursuing: 1) create a national fact-checking initiative that promotes collaboration between different media organisations, journalists, and NGOs; 2) fund the creation of automation tools for analysing disinformation, to help the human effort. We discuss the latter in more detail next.

Fund open-source research on automatic methods for disinformation detection

In the PHEME research project we focused specifically on studying rumours associated with different types of events—some were events like shootings and others were rumours and hoax stories like "Prince is going to have a concert in Toronto"—and how those stories were disseminated via Twitter or Reddit. We looked at how reliably we can identify such rumours: one of the hardest tasks is how to group all the different social media posts like tweets or Reddit posts around the same rumour together. In Reddit it is a bit easier thanks to threads. Twitter is harder because often there are multiple originating tweets that refer to the same rumour.

That is the real challenge: to piece together all these stories, because the ability to identify whether something is correct or not depends a lot on evidence and also on the discussions around that rumour, that the public are carrying out on social media platforms. By seeing one or two tweets, sometimes even journalists cannot be certain whether a rumour is true or false, but as we see the discussion around the rumours and the accumulating evidence over time, the judgment becomes more reliable.

---

[13] https://www.poynter.org/2017/there-are-now-114-fact-checking-initiatives-in-47-countries/450477/

[14] American Press Institute. "New studies on political fact-checking: Growing, influential; but less popular among GOP", 2015. https://www.americanpressinstitute.org/fact-checking-project/new-research-on-political-fact-checking-growing-and-influential-but-partisanship-is-a-factor/

[15] http://idir-server2.uta.edu/claimbuster/

[16] https://team.inria.fr/cedar/contentcheck/

[17] http://www.fakenewschallenge.org/

Consequently, it becomes easier to predict the veracity of a rumour, but the main challenge is identifying reliably all these different tweets that are talking about the same rumour. If sufficient evidence can be provided across different tweet posts, it becomes possible to determine the veracity of that rumour with around 85% accuracy.

In the wider context, there is emerging technology for **veracity checking and verification of social media content** (going beyond images/video forensics). These include tools developed in several European projects (e.g. PHEME, REVEAL, and InVID), tools assisting crowdsourced verification (e.g. CheckDesk, Veri.ly), citizen journalism (e.g. Citizen Desk), and repositories of checked facts/rumours (e.g. Emergent, FactCheck). However, many of those tools are language specific and would thus need adaptation and enhancement to new languages. Besides, further improvements are needed to the algorithms themselves, in order to achieve accuracy comparable to that of email spam filter technology.

It is also important to invest in establishing ethical protocols and research methodologies, since social media content raises a number of privacy, ethical, and legal challenges. The latter, in particular, are country-specific.

Dangers and pitfalls of relying purely on automated tools for disinformation detection

Many researchers, including myself, are researching automated methods based on machine learning algorithms, in order to identify automatically disinformation on social media platforms. Given the extremely large volume of social media posts, key questions are can disinformation be identified in real time and should such methods be adopted by the social media platforms themselves?

The very short answer is: Yes, in principle, but we are still far from solving many key socio-technical issues, so, when it comes to containing the spread of disinformation, we should be mindful of the problems which such technology could introduce:

- **Non-trivial scalability**: While some of our algorithms work in near real time on specific datasets such as tweets about the Brexit referendum - applying them across all posts on all topics as Twitter would need to do, for example, is very far from trivial. Just to give a sense of the scale here - prior to 23 June 2016 (referendum day) we had to process fewer than 50 Brexit-related tweets per second, which was doable. Twitter, however, would need to process more than 6,000 tweets per second, which is a serious software engineering, computational, and algorithmic challenge.
- **Algorithms make mistakes**, so while 90 per cent accuracy intuitively sounds very promising, we must not forget the errors - 10 per cent in this case, or double that at 80 per cent algorithm accuracy. On 6,000 tweets per second this 10 per cent amounts to 600 wrongly labeled tweets per second rising to 1,200 for the lower accuracy algorithm. To make matters worse, automatic disinformation analysis often combines more than one algorithm - first to determine which story a post refers to and second - whether this is likely true, false, or uncertain. Unfortunately, when algorithms are executed in a sequence, ***errors have a cumulative effect***.
- **These mistakes can be very costly**: broadly speaking algorithms make two kinds of errors - false negatives in which disinformation is wrongly labelled as true or bot accounts wrongly identified as human and false positives, correct information is

wrongly labelled as disinformation or genuine users being wrongly identified as bots. False negatives are a problem on social platforms, because the high volume and velocity of social posts (e.g. 6,000 tweets per second on average) still leaves with a lot of disinformation "in the wild". If we draw an analogy with email spam - even though most of it is filtered out automatically, we are still receiving a significant proportion of spam messages. False positives, on the other hand, pose an even more significant problem, as falsely removing genuine messages is effectively censorship through artificial intelligence. Facebook, for example, has a growing problem with some users having their accounts wrongly suspended.

Therefore, I strongly believe that the best way forward is to implement human-in-the-loop solutions, where people are assisted by machine learning and AI methods, but not replaced entirely, as accuracy is still not high enough, but primarily, for the censorship danger.

Establishing Cooperation and Data Exchange between Social Platforms and Scientists

Our latest work on analysing misinformation in tweets about the UK referendum[18,19] showed yet again a very important issue - when it comes to social media and furthering our ability to understand its misuse and impact on society and democracy, the only way forward is for data scientists, political and social scientists and journalists to work together alongside the big social media platforms and policy makers. I believe data scientists and journalists need to be given open access to the full set of public social media posts on key political events for research purposes (without compromising privacy and data protection laws), and be able to work in collaboration with the platforms through grants and shared funding (such as the Google Digital News Initiative).

There are still many outstanding questions that need to be researched - most notably the dynamics of the interaction between all these Twitter accounts over time - for which we need the complete archive of public tweets, images, and URL content shared, as well as profile data and friend/follower networks. This would help us quantify better (amongst other things) what kinds of tweets and messages resulted in misinformation spreading accounts gaining followers and re-tweets, how human-like was the behaviour of the successful ones, and also were they connected to the alternative media ecosystem and how.

The intersection of automated accounts, political propaganda, and misinformation is a key area in need of further investigation, but for which, scientists often lack the much needed data, while the data keepers lack the necessary transparency, motivation to investigate these issues, and willingness to create open and unbiased algorithms.

Policy Decisions around Preserving Important Social Media Content for Future Studies

Governments and policy makers are in a position to help establish this much needed cooperation between social platforms and scientists, promote the definition of policies for ethical, privacy-preserving research and data analytics over social media data, and also ensure the archiving and preservation of social media content of key historical value.

---

[18] https://www.buzzfeed.com/jamesball/3-million-brexit-tweets-reveal-leave-voters-talked-about-imm
[19] https://www.buzzfeed.com/tomphillips/we-found-45-suspected-bot-accounts-sharing-pro-trump-pro

For instance, given the ongoing debate on the scale and influence of Russian propaganda on election and referenda outcomes, it would have been invaluable to have Twitter archives made available to researchers under strict access and code of practice criteria, so it becomes possible to study these questions in more depth. Unfortunately, this is not currently possible, with Twitter having suspended all Russia-linked accounts and bots, as well as all their content and social network information. Similar issues arise when trying to study online abuse of and from politicians, as posts and accounts are again suspended or deleted at a very high rate.

Related to this is the challenge of open and repeatable science on social media data, as again many of the posts in current datasets available for training and evaluating machine learning algorithms, have been deleted or are not available. This causes a problem as algorithms do not have sufficient data to improve as a result and neither can scientists determine easily whether a new method is really outperforming the state-of-the-art.

Promoting Media Literacy and Critical Thinking for Citizens

According to the Media Literacy project[20]: "Media literacy is the ability to access, analyze, evaluate, and create media. Media literate youth and adults are better able to understand the complex messages we receive from television, radio, Internet, newspapers, magazines, books, billboards, video games, music, and all other forms of media."

Training citizens in the ability to recognise spin, bias, and mis- and disinformation are key elements. Due to the extensive online and social media exposure of children, there are also initiatives aimed specifically at school children, starting from as young as 11 years old[21]. There are also online educational resources on media literacy and fake news[22,23], that could act as a useful starting point of national media literacy initiatives.

Increasingly, media literacy and critical thinking are seen as key tools in fighting the effects of online disinformation and propaganda techniques[24,25]. Many of the existing programmes today are delivered by NGOs in a face-to-face group setting. The next challenge is how to roll these out at scale and also online, in order to reach wide audience across all social and age groups.

---

[1] http://framexframe.tumblr.com/

[2] https://www.bravenewtech.org/

[3] https://chrome.google.com/webstore/detail/reveye-reverse-image-sear/keaaclcjhehbbapnphnmpiklalfhelgf?hl=en

[4] http://exif.regex.info/exif.cgi

[5] https://chrome.google.com/webstore/detail/firstdraftnewscheck/japockpeaaanknlkhagilkgcledilbfk

---

[20] http://medialiteracyproject.org/learn/media-literacy/

[21] https://lie-detectors.org/

[22] https://www.edutopia.org/blogs/tag/media-literacy?gclid=CLu42_mu6NQCFU1MDQod7NEG0Q

[23] https://www.nytimes.com/2017/01/19/learning/lesson-plans/evaluating-sources-in-a-post-truth-world-ideas-for-teaching-and-learning-about-fake-news.html

[24] "Defending and Ultimately Defeating Russia's Disinformation Techniques" Centre for European Policy Analysis, August 2016 https://cepa.ecms.pl/files/?id_plik=2713

[25] https://www.nytimes.com/2017/09/25/opinion/the-only-way-to-defend-against-russias-information-war.html

[6] https://chrome.google.com/webstore/detail/storyful-multisearch/hkglibabhninbjmaccpajiakojeacnaf?hl=en

[7] https://distill.io/

[8] https://onpublico.com/

[9] https://coralproject.net/about.html

Establish/revise and enforce national code of practice for politicians and media outlets

Disinformation and biased content reporting are not just the preserve of fake news and state-driven propaganda sites and social accounts. A significant amount also comes from partisan media and factually incorrect statements by prominent politicians.

In the case of the UK EU membership referendum, for example, the Green paper has already mentioned one of the false claims regarding immigrants from Turkey, made on the front pages of a major UK newspaper[26]. Another widely known and influential example was VoteLeave's false claim that the EU costs £350 million a week[27]. Even though the UK Office of National Statistics disputed the accuracy of this claim[28] on 21 April 2016 (2 months prior to the referendum), it continued to be used throughout the campaign, including as a printed slogan on a red campaign bus.

Therefore, an effective way to combat deliberate online falsehoods must address such cases as well. Governments and policy makers could help again through establishing new or updating existing codes of practice of political parties and press standards, as well as ensuring that they are adhered to.

These need to be supplemented with transparency in political advertising on social platforms and a review process for political advertising, in order to eliminate or significantly reduce promotion of misinformation through advertising. These measures would also help with reducing the impact of all other kinds of disinformation already discussed above (i.e. fake news sites, Russian propaganda, etc).

---

[26] Turkey poll findings were flawed – clarification, THE DAILY EXPRESS (Jun 19, 2016, 12:00AM), https://www.express.co.uk/news/clarifications-corrections/681097/Turkeypoll-findings-were-flawed-clarification; https://www.theguardian.com/commentisfree/2016/may/19/inaccurate-pro-brexit-infacts-investigation-media-reports-eu-referendum

[27] M. Moore and G. Ramsay. UK Media Coverage of the 2016 EU Referendum Campaign https://www.kcl.ac.uk/sspp/policy-institute/CMCP/UK-media-coverage-of-the-2016-EU-Referendum-campaign.pdf

[28] https://www.ons.gov.uk/aboutus/transparencyandgovernance/freedomofinformationfoi/leavecampaignclaimsduringbrexitdebate