| | | |
|---|---|---|
| **Institution:** University of Sheffield | | |
| **Unit of Assessment:** B-11 Computer Science and Informatics | | |
| **Title of case study:** Truth and trust: combating disinformation and abuse in social media | | |
| **Period when the underpinning research was undertaken:** 2001–2020 | | |
| **Details of staff conducting the underpinning research from the submitting unit:** | | |
| **Name(s):** | **Role(s) (e.g. job title):** | **Period(s) employed by submitting HEI:** |
| Bontcheva, K. | Professor of Text Analysis | 1999–present |
| Cunningham, H. | Professor of Internet Computing | 1995–present |
| Maynard, D. | Research Fellow | 2000–present |
| Gorrell, G. | Post Doctoral Researcher | 2008–present |
| Tablan, V. | Research Fellow | 1999–2014 |
| **Period when the claimed impact occurred:** 2016–2020 | | |
| **Is this case study continued from a case study submitted in 2014?** N | | |

**1. Summary of the impact** (indicative maximum 100 words)

Sheffield's big data analytics has probed the veracity, sentiment, and sharing patterns of social media posts and exposed the ways social media can be used and abused to shape opinions about significant political events, such as elections or the Brexit referendum. The methods and findings have been used to promote truth in public discourse, underpinning UK and international policy responses to misinformation and the misuse of social media in relation to various issues, including COVID-19 and online abuse directed at politicians around elections and major national events. The work has also fuelled extensive media coverage on the misuse of social media that has raised public awareness of its risks.

**2. Underpinning research** (indicative maximum 500 words)

By the beginning of this REF assessment period, work on information extraction (IE) involving Sheffield had established a new orthodoxy for automated data capture from unstructured and semi-structured text. Shallow analysis methods that combined finite state and statistical pattern matching made it possible to identify entities, relations, and events in sources such as news articles to a much higher level of accuracy than previously. At this time, the Sheffield research programme centred around GATE (a General Architecture for Text Engineering, our open-source software architecture for IE **[R1, R2]**) and published results that have attracted more than 10,000 citations.

In the 2000s and 2010s, Professor Kalina Bontcheva and her team adapted and extended these shallow analysis methods to the specific case of social media, showing how the structured and contextual elements of this new textual form can boost the accuracy and richness of data capture with IE. In the same way that hypertext allowed new applications to exploit the web's link structure (e.g. Google's PageRank algorithm), we showed how social media hashtags, @mentions, likes, association networks and location data opened up a wide range of possible analytics. Bontcheva's research on micropost indexing, semantic annotation, search and visualisation demonstrated real-time analytics of web-scale Twitter streams **[R3]**. As before, the published results were complemented by open-source data and processing infrastructure that enabled experimental repeatability and boosted external take-up.

The team applied the same approach in work on sentiment analysis and rumour and veracity measurement that coincided with the rise of fake news, disinformation, and online abuse **[R4, R5, R6]**. The new work built on the team's previously developed taxonomies and application programming interfaces for flexibly and efficiently modelling text processing components, achieving the best compromise between expressive power and efficiency for pattern matching over textual annotations, thereby effectively allowing the combination of statistical counts with derivations based on linguistic intuition. Exploiting the ability to mine vast quantities of consumer-generated media, learn from datasets and recognise groups of words with similar mood/meaning rather than individual terms, Bontcheva's team created datasets of hundreds of millions of tweets and commissioned journalists to annotate the tweets regarding their veracity in relation to a set of rumours prevalent at the time. The team then used crowdsourcing to classify the stance of each tweet towards these rumours (agree, disagree, comment, question, etc.). The datasets the team created were among the very first of their kind available, offering fine detail. Their usefulness is illustrated by their subsequent uptake in other research and applications, with 58 citations for the original dataset paper, 173 citations for the 2017 dataset, and already 48 citations for the dataset published in 2019.

**3. References to the research** (indicative maximum of six references)

Sheffield staff and students in **bold**.


**R1.** **Cunningham, H., Maynard, D., Bontcheva, K. & Tablan, V.** (2002). GATE: an architecture for development of robust HLT applications. *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics - ACL'02*, Philadelphia, July 2002, 168-175. http://doi.org10.3115/1073083.1073112. Cited by 2293.

**R2.** **Cunningham, H.**, **Tablan, V.**, Roberts, A., & **Bontcheva, K.** (2013). Getting More Out of Biomedical Documents with GATE's Full Lifecycle Open Source Text Analytics. *PLoS Computational Biology*, *9*(2), e1002854. https://doi.org/10.1371/journal.pcbi.1002854. Cited by 326.

**R3.** **Maynard, D., Roberts, I., Greenwood, M. A., Rout, D.,** & **Bontcheva, K.** (2017). A framework for real-time semantic social media analysis. *Journal of Web Semantics*, *44*, 75–88. https://doi.org/10.1016/j.websem.2017.05.002. Cited by 46.

**R4.** **Gorrell, G., Greenwood, M.A., Roberts, I., Maynard, D. & Bontcheva, K.** (2018). Twits, Twats and Twaddle: Trends in Online Abuse towards UK Politicians. *Proceedings of The Twelfth International Conference on Web And Social Media (ICWSM)*, 600-603. eprints.whiterose.ac.uk/133570/. Cited by 11.

**R5.** **Gorrell, G., Roberts, I., Greenwood, M.A., Bakir, M.E., Iavarone, B. & Bontcheva, K.** (2018). Quantifying Media Influence and Partisan Attention on Twitter during the UK EU Referendum. *International Conference on Social Informatics*, 274-290. https://doi.org/10.1007/978-3-030-01129-1_17. Cited by 7.

**R6.** **Gorrell, G., Bakir, M.E., Roberts, I., Greenwood, M.A.,** Iavarone, B. & **Bontcheva. K.** (2019). Partisanship, Propaganda and Post-Truth Politics: Quantifying Impact in Online Debate. *Journal of Web Semantics*, 7. DOI: https://doi.org/10.34962/jws-84. Cited by 15.


**4. Details of the impact** (indicative maximum 750 words)

The ability to analyse large volumes of social media streams for sentiment and subject in real time has underpinned national and international policy and informed public debate.

**National policy**

Bontcheva has contributed to parliamentary inquiries and policy fora on abuse of MPs and disinformation in the submission period.

In December 2017, the **House of Commons Digital, Culture, Media, and Sport Committee** (DCMS) launched an inquiry into disinformation and fake news. Bontcheva was one of two academics invited to introduce the committee to social media analytics to provide context for them to better understand the subsequent evidence. In addition, she submitted supplementary written evidence. Both were cited in the final report **[S1]**.

In 2019, Bontcheva provided an analysis of abusive tweets directed at MPs from the first six months of that year to the **Joint Human Rights Committee inquiry on democracy, free speech and freedom of association**, who used the evidence in their report to show the scale of the problem **[S2]**.

In November 2020, Bontcheva was invited to a virtual expert roundtable with social media companies and experts hosted jointly by the Digital and Health Secretaries, focussing on the threat that mis- and disinformation pose to the acceptance and uptake of a COVID-19 vaccine **[S3a]**. The group explored ways to build long-term processes for working together to tackle disinformation related to COVID-19 and beyond, and understand the impact of existing interventions. This resulted in an agreement with Facebook, Twitter, and Google on new measures to limit the spread of vaccine misinformation and disinformation, and to help people find information about any COVID-19 vaccine **[S3b]**. Bontcheva was then invited to attend the first **Counter-Disinformation Policy Forum** in December 2020, which was set up by the Minister for Digital and Culture and brought together industry, civil society, and academia to develop a collective response to the evolving threats to the information environment **[S4]**.

Bontcheva developed a close collaborative relationship with policymakers at the **DCMS**. She conducted a longitudinal study on the abuse of MPs and candidates leading up to the 2019 general election to inform the government response to this increasingly prevalent issue. The DCMS stated, "*Sheffield University's research in this area, and particularly their abuse work spanning across 2015, 2017, and 2019 elections has been extremely valuable and has contributed towards our initial policy development"* **[S5]**. The study also provided an intuitive web-based visualisation to enable policy makers to track online abuse directed at UK politicians in real time, *"DCMS will be the primary beneficiary, but other Government departments – including the Defending Democracy team in the* ***Cabinet Office****, who protect candidates during elections, and* ***Home Office*** *colleagues, who look after the physical security and protection of MPs and monitored MP abuse during the campaign – will also benefit*" **[S5]**.

Bontcheva continued her collaboration with the DCMS by using Twitter to compare the COVID-19 pandemic with an election in terms of the abuse of MPs, specifically examining the responses elicited by ministers' and MPs' communications. A November 2020 report outlined the topics attracting the most engagement/abuse during the early stages of national lockdown (up to 25 May), with research covering June 2020 scheduled to report in Q1 2021 **[S5]**.

**International policy**

In March 2020, due to the potentially fatal results of the spread of disinformation about the COVID-19 pandemic – what the WHO described as a "*massive infodemic*" – **UNESCO** commissioned Bontcheva to co-author two policy briefs. She drew on her previous research to: provide a detailed analysis of the types of viral disinformation helping to drive the pandemic; investigate how individuals, the news media, internet communications companies, and governments are responding to contamination of the information ecosystem; offer rich food for thought about actions undertaken to combat the disinfodemic; and assess the potential risks associated with restrictive measures and provide recommendations to align crisis responses to international human rights standards on access to information, freedom of expression, and privacy. On the 24 April 2020, these policy briefs were published in three languages on the UNESCO website, receiving 15,487 visits by the end of December 2020 **[S6a, S6b]**.

Bontcheva expanded on these briefs by co-editing and contributing to '*Balancing Act: Countering Digital Disinformation While Respecting Freedom of Expression*' for UNESCO in September 2020 **[S6c]**. The report features global-scale, comprehensive analyses as well as sector-specific actionable recommendations and a 23-point framework to test disinformation responses. It is available on the International Telecommunications Union and UNESCO websites and has received a combined 8,408 unique visits by the end of December 2020 **[S6b, S6d]**.

In June 2020, the team began working with **First Draft**, an influential international network of journalists, news organisations, policymakers and social media platforms (including the BBC, Facebook, Twitter, and Google) who promote integrity in the world's information ecosystem. Sheffield contributed technical knowledge and data visualisation expertise to create a novel conceptual framework centred around identifying vulnerabilities and areas requiring intervention, rather than looking for viral misinformation. Sheffield solved problems such as how to calculate a 'data deficit' number, and make that readable for humans in numbers and visualisations, resulting in the First Draft COVID-19 Debunk Dashboard **[S7a]** and accompanying in-depth report **[S7b**], which launched worldwide in July 2020. Sheffield also created a uniquely comprehensive database of fact checks that First Draft used for further editorial work **[S7c]**. First Draft's Impact and Policy Manager stated, "Throughout, they fed into our thinking around what data deficits are, how they can be measured, and how those measurements can be represented in an actionable way for users". He explains that the concept of 'data deficits' has been adopted in most major disinformation initiatives and says that the research has contributed to "an important, broader shift in the problem-definition of the misinformation field" **[S7d]**.

**Increased public awareness of the use and misuse of social media**

Bontcheva's research has made it possible to interrogate millions of tweets about different topics and extract more nuanced, granular insights and quantitative data than ever before**.** An ongoing collaboration with journalists at **Buzzfeed News** led to three key articles (below) being reported widely by national and international print, broadcast, and online media **[S8]**.

*Attitudes towards key issues in the Brexit referendum* (December 2016), an analysis of 3 million tweets about the Brexit referendum, gave valuable insights into the concerns and online behaviours of both "remain" and "leave" supporters, and showed that immigration was by far the principal concern of leave voters despite attempts by mainstream politicians to play down its importance.

*Abuse directed at politicians on Twitter* (July 2017) was picked up by BBC2 in a special report for its flagship programme **Newsnight**, watched by some 200,000 people. The article was reported widely in local, national, and international print and online media as well as national radio and television news, reaching an estimated 149.3 million people **[S8]**. In July 2017, the **Committee for Standards in Public Life** (tasked with investigating abuse of public office-holders through social media) included this Buzzfeed report in the evidence it examined **[S9]**.

*Suspect activity amongst Brexit Party Twitter followers* (May 2019) revealed interesting connections between suspicious Twitter accounts. In particular, it identified an extensive fake network whose component accounts worked together to "amplify" pro-Brexit messages, providing clear evidence of the existence and modus operandi of such social media networks. The article's author, the senior political correspondent for Buzzfeed News, stated that the underpinning research provided "*the case study I needed to illustrate the problem. It was the first time such a network had been mapped, and the discovery made a significant contribution to the final article. […] My article aimed to raise public awareness of such misinformation, and to expose the fact that, around a very important democratic event, certain anonymous actors were*

*creating a lot of noise on social media that was not genuine*" **[S10]**. The story was picked up by outlets with a combined estimated reach of 17.7 million **[S8]**.

**5. Sources to corroborate the impact** (indicative maximum of 10 references)

**S1.** Final report from House of Commons Digital, Culture, Media and Sport Committee inquiry on Disinformation and 'fake news' (2019). Corroborates the use of Sheffield's oral (page 101 Q1-51, 2017) and written (page 59 para 206) evidence. (Accessed 16th Jun 2020). https://publications.parliament.uk/pa/cm201719/cmselect/cmcumeds/1791/1791.pdf

**S2.** Report from Joint Committee on Human Rights inquiry on Democracy, freedom of expression and freedom of association: Threats to MPs (2019). Corroborates the use of Sheffield's written evidence (page 16 para 35, page 39 para 100 and page 62 item 24). (Accessed 1st Feb 2021). https://bit.ly/2OSPQ7I

**S3.** Combined: DCMS & DHSC vaccine disinformation roundtable:

a) Confidential personal invitation to the round table of experts (2020)
b) Government statement following the round table outlining the agreement reached (2020). (Accessed 29th Jan 2021). http://bit.ly/3bF6jF8

**S4.** Confidential personal invitation to the first Counter-Disinformation Policy Forum (2020).

**S5.** Confidential statement from Policy Advisor at DCMS (2020). Corroborates the value of Sheffield's research to DCMS.

**S6.** Combined: UNESCO information:

a) Two UNESCO policy briefs co-authored by Professor Bontcheva on COVID-19 disinformation (24th April 2020). (Accessed 7th Oct 2020). http://bit.ly/3vk88iE
b) Report for UNESCO website viewing figures for S6a & S6c up to 31st December 2020.
c) UNESCO report co-edited and contributed to by Professor Bontcheva "Balancing Act: Countering Digital Disinformation While Respecting Freedom of Expression" (18th Sept 2020). (Accessed 15th Dec 2020). https://en.unesco.org/publications/balanceact
d) Email confirming the International Telecommunications Union website viewing figures for S6c up to 31st December 2020.

**S7.** Combined First Draft information:

a) COVID-19 Debunk Dashboard (Accessed 29th Jan 2021). https://datadeficits.firstdraftnews.org/
b) In-depth report citing the use of Sheffield's research. (Accessed 29th Jan 2021).https://firstdraftnews.org/long-form-article/data-deficits/
c) First Draft article based on the fact checks database produced by Sheffield. (Accessed 18th Mar 2021). https://firstdraftnews.org/latest/the-first-six-months-of-the-pandemic-as-told-by-the-fact-checks/

**S8.** d. Confidential testimonial from the Head of Impact and Policy at First Draft (2021). Confirms Sheffield's contribution to the collaboration.

**S9.** Summary of media reach of all three Buzzfeed stories from

**S10.** Committee of Standards in Public Life 17th report, "Intimidation in Public Life" (2017). Cites the Buzzfeed article (p43). (Accessed 16th June 2020). https://bit.ly/3tdSCDm

**S11.** Confidential testimonial statement from the senior political correspondent for Buzzfeed News (2020). Corroborates that Sheffield's research provided the fundamental data this article was based on.