RAND EUROPE

# Human–machine detection of online-based malign information

William Marcellino, Kate Cox, Katerina Galai, Linda Slapakova, Amber Jaycocks, Ruth Harris

For more information on this publication, visit www.rand.org/t/RRA519-1

Published by the RAND Corporation, Santa Monica, Calif., and Cambridge, UK



RAND Europe is a not-for-profit research organisation that helps to improve policy and decision making through research and analysis. RAND's publications do not necessarily reflect the opinions of its research clients and sponsors.



www.rand.org

www.randeurope.org

# Preface

This report is the final output of a study commissioned by the UK Ministry of Defence's (MOD) Defence Science and Technology Laboratory (Dstl) via its Defence and Security Accelerator (DASA). The study was contracted as part of DASA's call for proposals that could help the UK MOD to develop its behavioural analytics capability.

The final output of the study includes a proof-of-concept approach to human–machine detection and analysis of malign use of information on social media. RAND Europe's methodology is based on the application of a publics theory-driven approach to influence operations and the innovative use of lexical stance comparison technology. This methodology has been applied to understand how specific tactics have been used in disinformation efforts aimed at different publics. The study aims to enable the development of a method for detecting malign use of information online and cultivating resilience to it.

This report establishes the malign information operations threat context (Task A) in Chapter 1. Chapter 2 then provides a summary of findings learned through the application of proof-of-concept machine detection in a known troll database (Task B) and tradecraft analysis of Russian malign information operations against left- and right-wing publics (Task C). Finally, Chapter 3 explores the development and application of the tool and considers components of a plausible strategic framework for building resilience to malign information in targeted populations (Task D).

RAND Europe is a not-for-profit policy research organisation that helps to improve policy and decision making in the public interest through evidence-based research and analysis. RAND Europe's clients include European governments, institutions, NGOs and firms with a need for rigorous, independent, interdisciplinary analysis.

For more information, please contact:

Ruth Harris
Research Group Director
Defence, Security and Infrastructure
RAND Europe
Westbrook Centre, Milton Road
Cambridge CB4 1YG, United Kingdom
Tel. +44 (0)1223 353 329
ruthh@rand.org

# Table of contents

# Figures

# Boxes

# Abbreviations

| | |
|---|---|
| AI | Artificial Intelligence |
| API | Application Programming Interface |
| BERT | Bidirectional Encoder Representations from Transformers |
| CSO | Civil Society Organisation |
| COCA | Corpus of Contemporary American English |
| DASA | Defence and Security Accelerator |
| Dstl | Defence Science and Technology Laboratory |
| EEAS | European External Action Service |
| EU | European Union |
| IRA | Internet Research Authority (Russian 'Troll' Factory) |
| ISIS | Islamic State of Iraq and Syria |
| KOM | Kick-Off Meeting |
| L2D | Learn to Discern |
| MCC | Matthews Correlation Coefficient |
| MMI | Man–Machine Interaction |
| ML | Machine Learning |
| MOD | Ministry of Defence |
| MODREC | Ministry of Defence Research Ethics Committee |
| NATO | North Atlantic Treaty Organization |
| NGO | Non-Governmental Organisation |
| NLP | Natural Language Processing |
| PCA | Principle Components Analysis |
| QCRI | Qatar Computing Research Institute |
| RRU | Rapid Response Unit |
| SA | Sentiment Analysis |
| SNA | Social Network Analysis |
| URL | Uniform Resource Locator |

# Executive summary

## This study explores human–machine approaches to detecting online-based malign information

Today, some 44 per cent of UK adults use social media for news consumption, making it the most popular type of online news outlet.[1] Social media has revolutionised political discourse and enabled popular participation in new ways, allowing users to engage directly with societal influencers and political leaders.[2] However, the increasing use of social media, absence of human editors in news feeds, and growth of synthetic online activity may also make it easier for actors to manipulate social networks.[3]

In this context, RAND Europe was commissioned to undertake a study on human–machine approaches to detecting malign information online. This study was contracted by the UK Ministry of Defence's (MOD) Defence Science and Technology Laboratory (Dstl) via its Defence and Security Accelerator (DASA). A publics theory-driven approach was applied, using linguistic stance technology in support of two study objectives:[4]

1. Improve understanding of how specific rhetorical tactics have been used in disinformation[5] efforts to engage different publics.
2. Enable the development of a method for detecting malign use of information online and for cultivating resilience to disinformation efforts.

The RAND study team undertook the following tasks, which correspond to the study objectives:

- **Task A:** Establishing the malign information operations threat context in the UK, Europe and the transatlantic community.
- **Task B:** Applying proof-of-concept machine detection in a known Russian troll database, with a focus on online discourse relating to the 2016 US presidential election.
- **Task C:** Conducting tradecraft analysis of Russian malign information operations against left- and right-wing publics.

1 Jigsaw Research (2018).

2 Zeitzoff (2017).

3 Bradshaw & Howard (2018); Robinson (2018).

4 Publics theory refers to public sphere argument; linguistic stance focuses on attitudinal layers of language.

5 'Disinformation' refers to the dissemination of deliberately false information with the intention of influencing the views of those who receive it; 'misinformation' refers to the dissemination of false information while not necessarily knowing it is false. See IATE (2019).

- **Task D:** Presenting options for further development of the tool, and outlining considerations for the development of a framework for building resilience to malign information.

## Social media is increasingly being used by human and automated users to distort information, erode trust in democracy and incite extremist views

In recent years, technological developments have enabled increasing manipulation of information through social media use in the UK, other NATO member states and globally. Social media can be used to generate false support or opposition at significant political, security and defence-related decision points, such as elections or referenda. Social media can also be used to erode trust in government institutions and democratic processes, and to incite extremist views among targeted groups.

The impact of disinformation has been recognised in the UK and internationally. In 2014, the World Economic Forum labelled the spread of false information online as one of the top ten perils to society.[6] The UK has acknowledged the threat of disinformation in several government documents and actions, including the creation of a Rapid Response Unit – a social media capability designed to encourage fact-based public debate.[7] Despite these efforts, the issue continues to grow in importance and impact,[8] making it more difficult for individuals and organisations to discern between true and false information.

The spread of online-based malign information is manifested in several ways, including the spread of junk news, the impersonation of individuals to manipulate audiences, cyber bullying activity, terrorist propaganda and recruitment, and political 'crowd-turfing' in the form of reputation boosting or smearing campaigns. These activities are perpetrated by synthetic accounts and human users, including online trolls, political leaders, far-left or far–right individuals, national adversaries and extremist groups.

Both human and synthetic online actors can target a wide range of audiences, depending on their mission. Malign online activity may be aimed at:

- Large groups of individuals, such as voters, political audiences or the younger generation
- Specific groups, such as military personnel targeted by adversaries through social media
- Markets, such as financial markets targeted to steal financial data or manipulate stock markets.

## A central output of this study is a theory-based machine learning model that can successfully detect Russian trolls

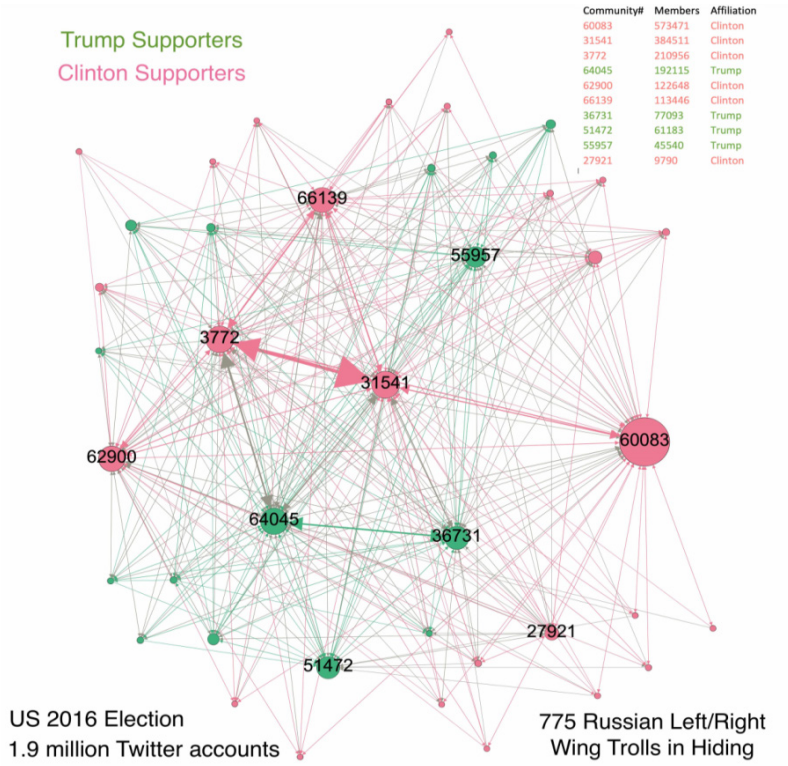
As the use of online disinformation to manipulate public opinion continues to grow, it is important to identify, characterise and ultimately curtail the influence of the malign actors behind these efforts. To support the detection and characterisation of malign actors, we developed a machine learning model designed to identify differences between authentic political supporters and

6 World Economic Forum (2018).

7 UK Government Communication Service (2019).

8 University of Oxford (2018).

**Figure ES-1. Network diagram of Trump and Clinton supporters on Twitter, 2015–2017**




Russian 'trolls' involved in online debates relating to the 2016 US presidential election. Our working assumption was that examining how the 'trolls' targeted this public debate will have applicability to other left/right political arguments worldwide likely targeted by Russia, including Brexit.

To develop the model, we used a publicly available dataset of Russian left- and right-wing troll accounts active on Twitter during the 2016 US presidential election, drawing on 2015–2017 Twitter data. The dataset contains 2,973,371 tweets from 2,848 handles operated by the Internet Research Agency.[9] We then used text mining to extract search terms for the publics targeted by the trolls, before harvesting tweets from 1.9 million user accounts. Further to this, we used a network analysis algorithm on the Twitter data which uses frequency of interaction to infer communities.

As anticipated, we identified two large meta-communities, with text analysis showing that one represented the supporters of Donald Trump and the other supporters of Hillary Clinton. Figure ES-1 visualises the Twitter argument space in relation to the 2016 US presidential election, highlighting the meta-communities that Russian trolls targeted: a liberal public (pink) supporting Clinton, and a conservative public (green) supporting Trump. Furthermore, the network shows that these

9 US House of Representatives Permanent Select Committee on Intelligence (2018).

large publics are composed of smaller sub-publics, which are allied but distinct.

What our algorithm labelled 'community 60083', we might instead call '#Black Lives Matter Advocates,' because text analysis shows that the language in that community includes '#BlackLivesMatter,' 'hip hop' and 'police brutality'. In contrast, within the conservative meta-community we see 'community 64045', a community marked by negative language relating to immigration, Islam and 'illegals'. In Figure ES-1, each dot (node) represents a community of Twitter users, sized according to the number of accounts in the community, from hundreds to hundreds of thousands of users. Each line (edge) indicates interactions between communities, with the thicker edges representing a higher number of interactions.

Our analysis identified 775 inauthentic Russian troll accounts masquerading as liberal and conservative supporters of Clinton and Trump, as well as 1.9 million authentic liberal and conservative supporters. With this data, we trained a machine learning model that could successfully distinguish between the four account categories: Russian left-wing trolls, Russian right-wing trolls, authentic liberal supporters and authentic conservative supporters.

## Our pilot modelling effort had high accuracy, enabling us to understand the tradecraft used by trolls to create online dissent

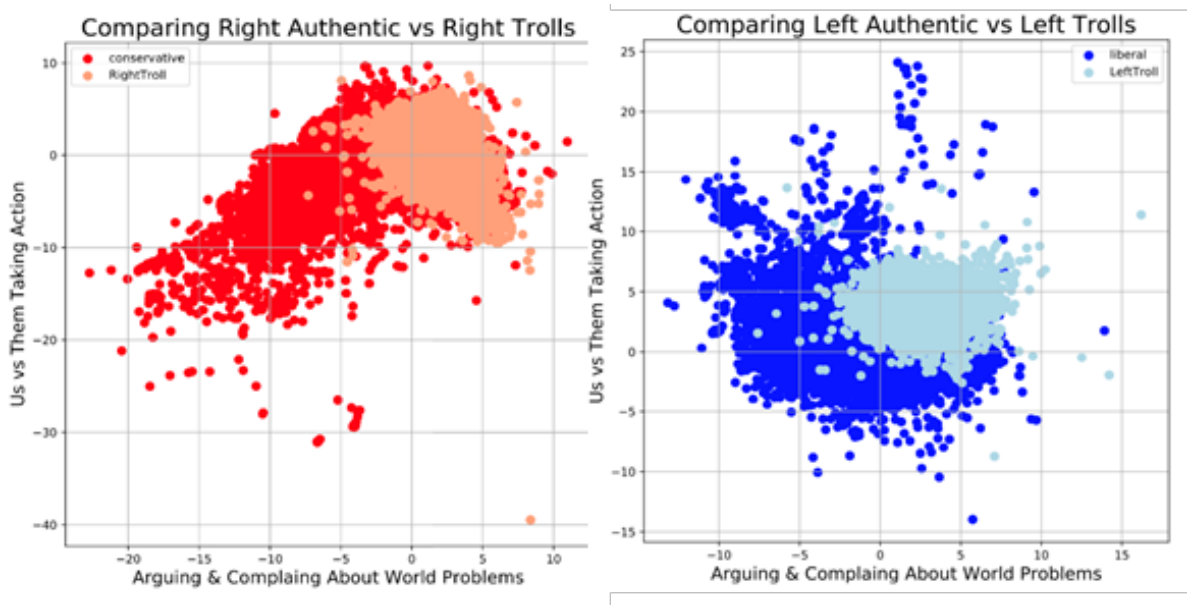
Our modelling effort had high accuracy (71 per cent on average), using highly interpretable algorithms that allowed us to understand the approaches used by trolls to spread misinformation and deepen divisions between online communities. The model was particularly effective at the top-level task of discerning trolls from authentic supporters (87 per cent). Our four-way classification was more effective in classifying right-wing trolls and authentic conservatives (74 per cent correctly classified) than left-wing trolls and authentic liberals (69 and 70 per cent correctly classified).

When analysing whether Russian trolls are better at imitating liberals or conservatives, we found that liberal supporters of Hillary Clinton tended to write in a relatively homogenous way, while conservatives were much more varied in their linguistic stance and thus more difficult to imitate. Figure ES-2 below illustrates how far narratives shared on authentic accounts overlap with those of trolls, indicating that left-wing trolls imitate authentic liberals more effectively than right-wing trolls emulate authentic conservatives. We found that left- and right-wing trolls use a common strategy of complaining about problems and using socially divisive language intended to inspire action (the respective horizontal and vertical axes below).

Our analysis indicates that Russian trolls imitated the most divisive, confrontational elements of liberal and conservative US discourse: a 'far-left' liberal discourse that constructs the US as fundamentally oppressive, and a 'far-right' conservative discourse that presents the US as under attack from external forces.[10] This appears to be consistent with the wider Russian practice of creating discord and division in other populations. Our analysis of the specific rhetorical tactics used by Russian trolls opens the door to identifying trolls in real time, while also highlighting the targets of these manipulation tactics.

10 This is evidenced in the PCA analysis.

**Figure ES-2. Visualisation of trolls and authentic accounts**




## Our study offers four key findings that could be operationalised by UK government entities concerned with foreign malign information campaigns

Building on our analysis, we present four key takeaways for UK government actors in this space:

Publics theory and network analytics can help map out the rhetorical battlefields that Russia (and others) attempt to influence in malign ways. To support counter-efforts, government entities should consider adopting publics-based analytics that identify meta-communities.

Russian trolls aggravate both sides of a controversy, stoking discord by highlighting emotive issues for each side through the use of repeated rhetorical patterns. To raise awareness and build resilience to these tactics, government entities should make this visible to members of targeted publics.

To develop detection capacity, using less powerful but interpretable shallow models is an important preliminary step. Interpretability allows for insights on tradecraft, and can inform subsequent efforts to engineer a more powerful deep algorithm.

Stance features can detect rhetorical patterns across various topics, and can powerfully enhance machine learning classification. This could be operationalised by UK government actors concerned with countering foreign malign information campaigns.

## There are several possibilities for future work on this issue, and the component technologies piloted in this study offer broad applications

Options for further developing this work include building on the current findings and extending the modelling capabilities to leverage other state-of-the-art deep learning and artificial intelligence approaches. There would also be merit in leveraging the sociolinguistic and derived features as inputs into partner models, as well as extending and generalising findings to other datasets and platforms. Beyond the specific problem of detecting malign information efforts, the component technologies piloted in this study have wider applications. The community detection, text analysis, machine learning and visualisation components of our model could be reconfigured to create a robust general purpose social media monitoring tool. Unlike monitoring tools from the commercial world, we could provide a system that accounts for aggregated structures like publics and for the socio-cultural purpose of social media talk. These are gaps in current off-the-shelf solutions, and would be of use to Dstl customers.

## To trial the model's portability, a promising next step in our effort could be to test our model in a new context such as the online Brexit debate

A feasible next stage in our effort could be to test our model for detecting malign troll accounts on social media in a new context in order to explore the model's portability. A relevant choice for this could be the online debate over Brexit, in anticipation of the UK's departure from the European Union. As a preliminary step – for proposed further analysis – we collected Twitter data from 1 January 2019 to 31 March 2019, conducting an initial network analysis on this dataset and identifying three online communities: 'Brexiteers', 'Remainers' and an international (i.e. non-English speaker) community engaged in pro-Remainer discourse. Building on this exploratory research, as a next step our aspiration would be to prove the robustness and utility of our model by applying it to the online Brexit argument.

# 1 Background and context

This report is the final deliverable of the Dstl/DASA-commissioned RAND Europe study on *Human–machine detection of online-based malign information.* This chapter explains the background to the study and the overall project approach, before presenting its initial findings on the state of play, identified through a literature review. It concludes with a summary of existing methodologies and approaches used to identify social media manipulation.

## 1.1. Introduction

Some 44 per cent of UK adults use social media for news consumption, making it the most popular type of online news outlet.[11] People in the UK increasingly access news via the various search (20 per cent) and social media (25 per cent) services provided by US-based platform companies such as Google and Facebook.[12] This mirrors the high use of online and social platforms for news and information in other countries. According to 2019 research by the Reuters Institute for the Study of Journalism, 47 per cent of news consumers in 38 countries[13] report using broadcaster websites or newspapers for their news.[14] Meanwhile, in 2018, 68 per cent of US adults used Facebook, and 39 per cent of US adults did so to consume news.[15]

Social media and Twitter, in particular, have revolutionised political discourse[16] and enabled popular participation in novel ways, allowing users to create their own content and directly respond to elites and political leaders.[17] Giving ordinary citizens a voice in international online discourse, platforms like Twitter and Facebook are easily accessible worldwide to virtually any owner of a mobile phone or computer.[18] In late 2010, social media platforms empowered democracy advocates across North Africa and the Middle East, contributing to the social and political phenomenon of the Arab Spring.[19] But

11 Jigsaw Research (2018).

12 Jigsaw Research (2018).

13 Based on a YouGov survey of over 75,000 online news consumers, available at: http://www.digitalnewsreport.org/

14 Nielsen (2017).

15 According to the 2018 Reuters Digital News Report, quoted in Marchal et al. (2018).

16 Ruge (2013).

17 Zeitzoff (2017).

18 Nyabola (2017).

19 Bradshaw & Howard (2018).

the absence of human editors in news feeds as well as increased synthetic online activity may make it easier for political actors to manipulate social networks.[20] Furthermore, malign online activity can sow discord and confuse audiences.[21]

## 1.2. Methodology

The objectives of the study were to:

1. Explore how specific rhetorical tactics are used in malign information operations to engage different *publics*. Publics are groups of people with shared ways of speaking and a shared advocacy position on a given issue.[22]
2. Enable the development of an approach for detecting malign use of information and cultivating resilience to it.

The overall study proposes and tests a proof-of-concept use of a theory-based analytical approach and technology for identifying malign information operations. Recognising that malign information operations can be performed by a variety of actors, we focus on an analysis of Russian trolls as we understand their activity to have particular relevance and potential impact for the UK, and because we had available a ground truth dataset of identified Russian trolls to use as a basis for our research. We use 'theory-based' as opposed to purely empirical approaches to machine learning that are characterised by black-box models[23] tied to specific datasets. A theory-based approach means our use of machine learning is informed by publics and rhetorical theory at the front-end (the selection of data and the kinds of features used for machine learning), and then produces actionable insight at the back-end by showing how and why the machine model worked. It will, therefore, aim to support the UK's capability to identify and characterise the types of malign information online and help ensure that the UK continues to build and maintain its expertise in this area.

This report establishes the malign information operations and responses context in the UK, Europe and the transatlantic community (Task A).[24] It also sets out the research findings resulting from proof-of-concept machine detection in a known troll database (Task B) and tradecraft analysis of Russian malign information operations against left- and right-wing publics (Task C). Finally, it explores the development and application of the tool (Task D/1) and considers components of a plausible strategic framework for building resilience to malign information in targeted populations (Task D/2). Figure 1.1 provides an overview of the project methodology.

This report makes repeated use of several terms and concepts. For ease of understanding, we provide their definitions in Figure 1.2 below.

20 Bradshaw & Howard (2018).

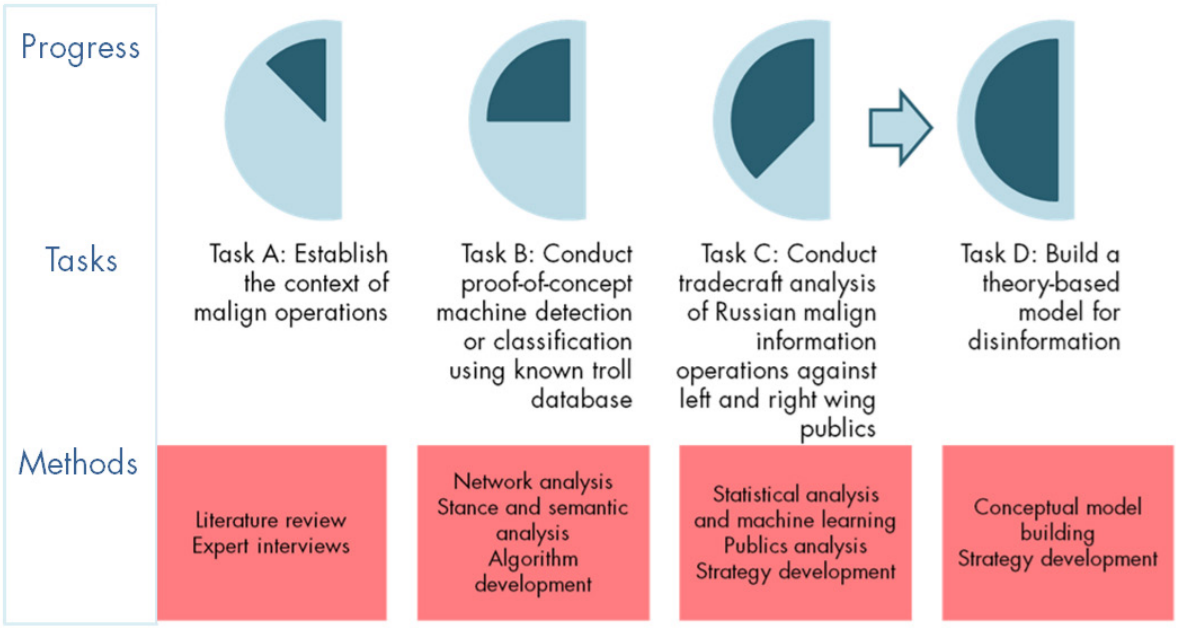
21 Robinson (2018).

22 Hauser (2002); Kaufer & Butler (2010).

23 Refers to a system that does not reveal its internal mechanisms; Molnar (2019).

24 The output of Task A is presented in Sections 1.3–1.7 and is based on a review of 35 academic and policy reports focusing on social media analysis and disinformation and misinformation in social media.

**Figure 1.1. Overall project methodology**




*Source: RAND Europe.*

**Figure 1.2. Study definitions**

| | |
|---|---|
| **Malign use of online information** | Manipulation or falsification of online information performed via a wide variety of approaches and methods (e.g. manufacturing content, misuse of content, misappropriation, obfuscation, as well as use of deceptive identities).[25] |
| **Misinformation** | Dissemination of false information while not necessarily knowing it is false.[26] |
| **Disinformation** | Dissemination of deliberately false information, especially when supplied by a government or its agent to a foreign power or to the media, with the intention of influencing the policies or opinions of those who receive it; false information so supplied.[27] |
| **False information** | Untrue information of any kind. |
| **Troll** | Someone who posts content online to start an argument or be provocative.[28] |
| **Bot** | Automated agent in a social network.[29] |

25 Definition based on RAND's internal expertise.

26 IATE (2019).

27 As defined by the Oxford English Dictionary.

28 Williams (2012).

29 Zaman (2018).

## 1.3. Military consequences for the UK and its allies from the malign use of information on social media

Manipulation of information and opinion through social media has been rising in the UK, other NATO member states and globally in recent years, triggered by the proliferation of information technologies and various social media platforms. It has taken time for Western governments and experts to acknowledge that this may be considered a threat to democracy and national security.[30] Manipulation of information has become a 'force multiplier' used by many actors to influence the decisions and behaviours of democratic governments and their armed forces.[31] Social media may be used to generate and disseminate false support or opposition at significant political, security and defence-related turning points, such as elections or referenda. It may, additionally, be employed to degrade trust in government institutions and democratic processes, influence decisions on troop deployment and sow extremist views in selected population groups.

The extent and impact of disinformation have been recognised in the UK and internationally. In 2014, the World Economic Forum noted that the rapid spread of false information online is one of the top ten perils to society.[32] The UK has acknowledged the threat of disinformation in several government-level documents and actions, including the creation of a Rapid Response Unit (RRU) – a social media capability that will aim to encourage fact-based public debate.[33]

Although the UK and allied government, non-government and private actors have made efforts to mitigate and prevent malign use of information and manipulation of opinions on social media (see Section 1.5 below), the issue is continuing to grow in terms of its importance and potential consequences.[34] The Internet has turned into a 'vast chamber of online sharing and conversation and argumentation and indoctrination, echoing with billions of voices',[35] making it more difficult for individuals and organisations, civilian and military alike, to discern between true and false information.

This effect is exacerbated by evolving digital marketing techniques which seek to affect users' behaviour on social media. This includes clickbait, the use of sensationalist and otherwise exaggerated headlines and posts that aim to create a 'curiosity gap', 'baiting' users to click on links with potentially misleading content.[36] Though clickbait is often used as a marketing tool luring users to commercial sites, it can represent a form of fake news due to the use of links leading users to content of dubious value or interest.[37]

Social media is increasingly being weaponised across the globe. It has become a tool for terrorist recruitment and radicalisation, information and psychological warfare, military deception and strategy building, even becoming

30 University of Leiden (2019).
31 Van Oudenaren (2019); Theohary (2018).
32 World Economic Forum (2018).
33 UK Government Communication Service (2019).
34 University of Oxford (2018).
35 Brooking & Singer (2016).
36 CITS (2019); Chen et al. (2015).
37 Merriam Webster (2019).

a means for declaring war. Social media, when used by adversarial actors, often reinforces binary narratives of 'us versus them', generating panic, fear and violence, and exaggerating already existing societal divergences instead of emphasising the pursuit of peaceful and mutually agreeable solutions. For example, ISIS has made ample use of social media for a variety of purposes from recruiting foreign fighters to declaring war on the United States, and from sharing videos of executions to battlefield reporting (e.g. flooding Twitter with victorious announcements of their invasion of northern Iraq). In Syria, social media was used as a psychological warfare tool when Syrian-regime loyalists posted pictures of their dinner in response to starvation in a rebel-held town.[38]

Malign use of information on social media may have significant strategic, operational and tactical military impacts along with physical and soft or psychological effects. Social media campaigns have already been used to attempt to disrupt strategic military relations between allied countries.[39] For example, the US–Turkish relationship and the presence of American nuclear weapons at Incirlik Air Base, Turkey, have been targeted in several false reports and social media campaigns. One such campaign in 2016 used social media to disseminate false information that a small anti-US protest near the base was instead a large riot, and speculated that the American nuclear weapons would be repositioned to Romania.[40] It has also been claimed that social media (e.g. Facebook) has been used to coordinate violent attacks. One such example is the attacks on the Muslim Rohingya minority in Myanmar, where false information that Muslims were attacking Buddhist sites was amplified through social media.[41]

Social media has also been used for operational effect, for example by decreasing the morale of the UK and allied forces, and to create distrust in democracy and chain of command, making them more vulnerable to the adversary's information warfare. For example, during the ISIS invasion of Mosul in northern Iraq, the motivation of the Iraqi forces, already suffering from low morale and corruption, dropped even lower upon the receipt of false reports of ISIS viciousness and speed of progress. As a result, the 25,000-strong Iraqi forces, equipped with US-made tanks and helicopters, collapsed when facing only 1,500 lightly equipped ISIS fighters.[42] Social media may further be used to undermine the relationship between the UK and allied forces and the host nation and local communities.[43] For example, false information in text, photo and video form has been used unsuccessfully to undermine the perception of allied troops located in Poland and the Baltic states.

Last but not least, the use of social media by UK and allied military forces, while providing alternative ways of communication and connection among those who serve and their families, also exposes them to vulnerabilities. The challenges include maintaining operational security, the risk that information posted online by soldiers may be misrepresented or misused, and the risk that online military communities may be penetrated by adversarial users and

38 Brooking & Singer (2016).

39 Davis (2018).

40 Mueller-Hsia (2019).

41 Bradshaw & Howard (2018).

42 Brooking & Singer (2016).

43 IHS Markit (2018).

bots aiming to sow discord, distrust and extremism. Existing research has shown that there is interaction between current and former military personnel and various extremist and conspiracy theory groups, exposing the military community to being influenced by the false and purposefully destructive use of information and to the exploitation of the level of trust that exists within military communities.[44] For example, US military veterans and active soldiers were targeted with anti-government messages by Russian social media campaigns in 2016 before the US presidential elections, using junk news websites created for this specific population of Internet users.[45]

## 1.4. The nature of social media manipulation

Several factors have led to the widespread impact of social media manipulation on public life. The role of social media as a news consumption platform has grown significantly, leading to the inadvertent dependence of mainstream media on social media.[46] This has led to an increasing dominance of sensationalism and novelty over newsworthiness, while users become increasingly susceptible to possible media manipulation through clickbait. At the same time, the use of automation, algorithms and big-data analytics intensifies the volume of online content, which may also have a swaying effect on online user opinion.[47]

Despite the differences in content, social media manipulation is often characterised by the following factors:

### Public divide in the context of political or social issues (e.g. Brexit, presidential elections)

Much of the literature on social media manipulation and the spread of malign information online reviewed for this study[48] focuses research on socially or politically significant case studies. Although the main messages in the online space are primarily dependent on an event – such as the Salisbury poisoning, Russia's annexation of Crimea, the 2016 US presidential election, or Brexit – social media activity surrounding such events is characterised by a significant public divide.

### Radicalisation of vulnerable or ostracised groups

Social media may increase opportunities for radicalisation by equipping terrorists with a tool to recruit, train and communicate with their followers, sympathisers and other audiences.[49] Social media platforms can therefore facilitate radicalisation by promoting content with an emotional appeal that taps into the grievances of users and reinforces their frustrations.[50] Given that social media captures users' 'preferences' in its algorithms,[51] less extreme material may be phased out, which may reinforce the extremist ideas of vulnerable

44 Previous research has specifically focused on the US military; Gallacher et al. (2017).

45 Daily Beast (2017).

46 Marwick & Lewis (2017).

47 Bradshaw & Howard (2018).

48 NATO Stratcom (2019); Zaman (2018); Stewart & Dawson (2018); Ferrara (2015); Marwick & Lewis (2017); Thamm (2018); Mihaylov et al. (2015); Neudert (2018); Robinson (2018); Bradshaw & Howard (2018); Marchal et al. (2018); Zeitzoff (2017); Shao et al. (2017); Reuters (2017); Varol et al. (2017).

49 Zeitzoff (2017); Klausen et al. (2018); Varol et al. (2017).

50 RAND Europe interview with Innocent Chiluwa, academic expert, 18 September 2017.

51 Bloom et al. (2017).

users. For example, ISIS uses a wide range of social media platforms (including Facebook, Twitter, YouTube, WhatsApp, Telegram, JustPaste.it, Kik and Ask.fm) to spread propaganda in order to attract fighters, skilled labour and families.[52]

## Undermining of democratic systems and/or processes by adversaries

Another aspect of online-based malign activity directly relates to the undermining of fair and democratic processes. Sophisticated data analytics and political bots promote scepticism and distrust, and polarise voting constituencies in emerging and Western democracies.[53] Authoritarian regimes increasingly use social media manipulation to subvert elections, enforce censorship and control public opinion.[54] For example, both Hillary Clinton and Donald Trump used Twitter as a campaign tool, while Russia used social media and cyber attacks to build and control the narrative concerning Ukraine.[55]

Online-based malign information manifests itself in different formats and can be perpetuated in several ways (summarised in Figure 1.3):

- Spread of fake or junk news – propagating or retweeting false messages to attract attention, manipulate user opinion and spread disinformation.[56]
- Organised boosting of consensus on social and political issues – also known as 'crowd-turfing'[57] or astroturfing[58] in political campaigns. These activities can include reputation boosting or, on the contrary, smearing campaigns to alter or manipulate user opinion.
- Impersonation and honeypot pages[59] – these collect private or identifying information about a particular individual, which is then published to manipulate subjects.
- Spread of propaganda, incitement and recruitment by extremist and terrorist organisations – this type of online-based malign activity aims to radicalise users to support extremist groups and to influence their activity both online and physically.[60]
- Cyber-bullying – a form of bullying that is carried out in cyberspace usually by anonymous users (also known as 'trolls').[61]
- 'Robo-trolling' – refers to automated trolling and includes two main types of user manipulation: activity from automated accounts, and messaging from fake human accounts.[62] The latter may be operated by patriotically minded individuals or groups, or generated for profit by so-called troll factories. Unlike the traditional notion of trolling that is

52 Cox et al. (2018).

53 Bradshaw & Howard (2018).

54 Robinson (2018).

55 Zeitzoff (2017).

56 Marwick & Lewis (2017).

57 Refers to the activity of a malicious crowd sourcing system that exists on social media and on the internet; McAfee (2013).

58 Refers to the attempt to create an impression of a strong 'grassroots' support for a political event, individual or policy; Bienkov (2012).

59 A honeypot refers to a fake social media account used to solicit information and compromise accounts via approaches including malicious links and sexual exchanges (Helmus et al. 2018).

60 Varol et al. (2017); Klausen et al. (2018).

61 Ferrara (2015).

62 NATO Stratcom (2019).

**Figure 1.3. Types and use of online-based malign information**


*Source: RAND analysis.*

associated with largely benign activity, augmented or automated trolling is more likely to be driven by malicious intent, given the volume of activity and its ability to sway opinion.

- Spread of misleading content – includes propagating hoaxes, rumours, conspiracy theories, fabricated reports and clickbait headlines online.
- Undermining of trust in the media, public institutions and science – malign online activity aimed at creating further divides in society or between societies (for example, Brexit Leave vs Remain voters in the UK or Russia vs the West in the Salisbury attack). Such activity undermines public confidence and is aimed at destabilising democratic states and processes.[63]

63 Thamm (2018).

## 1.5. Actors behind social media manipulation

All of the abovementioned activities are perpetrated by two types of actors: human users and synthetic or automated accounts. Human-controlled or -influenced users can include online trolls,[64] politicians and influencers,[65] far-right or far-left individuals,[66] national and non-governmental adversaries,[67] radical and extremist groups,[68] and others. A range of government agencies and political parties can form their own groups of social media manipulators comprising users tasked with manipulating public opinion online.[69]

However, the shift in social media landscape from popular political participation (as observed during the Arab Spring) to the online activity seen in the 2016 US presidential election and the Brexit debate is largely characterised by the growing automation of user accounts. What sets automated accounts or 'bots' apart from human users is the ability to produce an immense amount of content. Although bots can generate a large number of benign tweets,[70] they can also create a high volume of content aimed at changing r manipulating user opinions through the propagation of spam and fake news.

Although considered wholly or partially automated, bots or proxy accounts are run on behalf of real or fake individuals.[71] An automated bot account usually manipulates user opinions online by pretending to be human.[72] Another type of automated activity can be executed by cyborg accounts[73] that spread fake news in a way that blends synthetic activity with human input.

Social and political bots may be operated by either by government or private actors, and from either the left or right wing.[74] Their activity is becoming increasingly relevant as technology becomes more sophisticated. New-generation bots with natural-language processing capacities (similarly to Alexa and Google Assistant)[75] are becoming harder to detect and are therefore more likely to confuse audiences and successfully influence public opinion and voters' decisions. Analysis of the 2016 US presidential election demonstrated that a small number of very active bots could significantly shift public opinion.[76]

In addition, machine learning and artificial intelligence (AI) has enabled the advancement of so-called 'deepfake' technology that maps real individuals' faces onto controlled avatars.[77] The 'deepfake' concept refers to audio-visual

64 Mihaylov et al. (2015).
65 Zeitzoff (2017).
66 Marwick & Lewis (2017).
67 Thamm (2018).
68 Klausen et al. (2018).
69 Bradshaw & Howard (2018).
70 Howard & Kollanyi (2016).
71 McAfee (2013).
72 Digital Forensic Research Lab (2016).
73 Shu et al. (2017); Reuters (2017).
74 Robinson (2018).
75 Neudert (2018).
76 Zaman (2018); el Hjouji et al. (2018).
77 Robinson (2018).

material which has been manipulated but appears authentic and realistic. Conventionally this involves processing methods utilising AI and other similar technology.[78]

Both human and synthetic online actors can target a wide range of audiences, depending on their mission. Malign online activity may be aimed at:

- Large groups of individuals, such as voters, political audiences or the younger generation.[79]
- Specifically defined groups, such as military personnel targeted by adversaries through a number of social media platforms to harvest data on the operations of the armed forces.[80]
- Platforms or markets, such as financial markets targeted to steal financial information, extract trade secrets and manipulate stock markets.[81]

## 1.6. Response to social media manipulation

As a result of growing malign online activity and social media manipulation, both governments and non-governmental stakeholders such as social media companies are employing technological, regulatory and educational countermeasures.

### Technological approaches

As synthetic social media activity increases, so do the technological responses, which become better at detecting automated bot-like behaviour.[82] As a result, it has become easier to suspend or block automated accounts. The problem with relying on malign account identification and suspension is that these measures, while effective in decreasing the overall volume of disinformation, are only temporary.[83] For example, the majority of the Iranian Twitter users whose accounts were blocked to stop the spread of malicious misinformation ahead of the US midterm elections subsequently switched to new accounts.[84]

Similarly, CAPTCHAs[85] or challenge-response tests have been deployed widely to determine whether a user is human. By limiting automatic re-posting of content, CAPTCHAs have been used successfully to combat email spam and other types of online abuse.[86] However, reliance on this approach could also lead to detrimental effects on benevolent applications using automation by legitimate entities, such as news media and emergency response coordinators.

There have also been more niche and targeted measures developed to counter malign online activity. For example, in their research focusing on finding online extremists in social networks,[87] Klausen, Marks & Zaman

78 Technopedia (2019d).
79 Ferrara (2015).
80 Bay & Biteniece (2019).
81 Stewart & Dawson (2018); Ferrara (2015); Varol et al. (2017).
82 Neudert (2018).
83 Robinson (2018).
84 Robinson (2018).
85 von Ahn et al. (2003).
86 Shao et al. (2017).
87 Klausen et al. (2018).

developed a set of capabilities that allow for more effective mitigation of extremist social media user threats. Their approach combines statistical modelling of extremist behaviour with optimal search policies, defined as search policies that '[minimise] the expected number of unsuccessful queries or the expected cost of the search.'[88]

## Regulatory approaches

Several countries have established new government agencies or mandated existing organisations to combat fake news and foreign influence operations. The response often involves generating and disseminating counter-narratives or creating reporting, flagging and fact-checking portals to support citizen awareness and engagement.[89]

According to a 2018 NATO Stratcom report on government responses to the malicious use of social media,[90] 43 governments have proposed or implemented legal or regulatory interventions in the area of social media manipulation since 2016. These include:

- Measures targeting social media platforms
- Measures targeting offenders
- Measures targeting government capacity
- Measures targeting citizens, civil society and media organisations.

At the European level, the European Commission has taken regulatory steps resulting in Facebook, Twitter, YouTube and Microsoft signing up to a code of conduct[91] that aims to tackle online hate speech and take down the majority of potentially illegal content within 24 hours.[92]

Another effort by the Poynter Institute[93] aims to demystify the policy effort in the area of fake news and online misinformation. It depicts different types of interventions, from law and regulation in countries such as France, Poland and the US, to a national policy in Saudi Arabia that seeks to influence the public through threats (see Figure 1.4).

Although there appears to be significant reported activity in terms of countering online-based malign information, it is difficult to ascertain the level of success of particular measures, given the variety of approaches and the difficulty of quantifying social media manipulation.

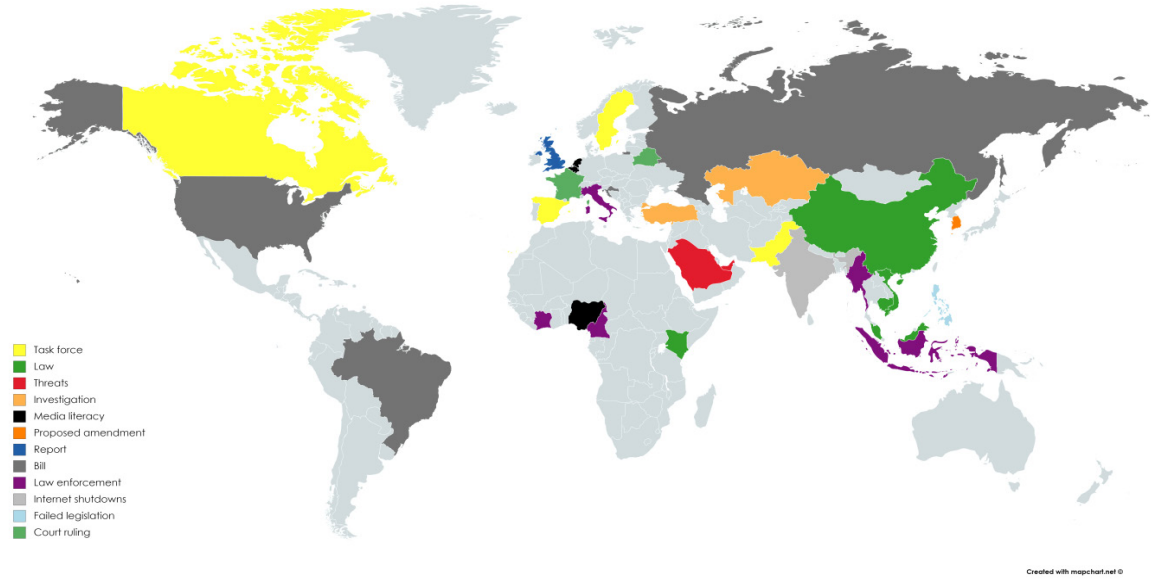
## Educational approaches

A 'soft' but equally important measure of countering social media manipulation falls into the education category. Addressing the actors and the platforms only offers a partial solution, leaving users vulnerable in the absence of relevant education and awareness-raising initiatives. There is a need to educate people on the power of both text and images to manipulate and persuade.[94]

In addition to improving news literacy, i.e. the ability to differentiate between opinion and hard news,[95] governments and civil society have been emphasising skills such as critical thinking,

88 Klausen et al. (2018, 969)
89 Bradshaw & Howard (2018).
90 Bradshaw et al. (2018).
91 European Commission (2016).
92 Wardle & Derakhshan (2017).
93 Poynter (2019).
94 Wardle & Derakhshan (2017).
95 Wardle & Derakhshan (2017).

**Figure 1.4. Government response to online-based malign information**




*Source: Adapted from Funke & Flamini (2019).*

fact-checking[96] and evaluating quantitative statements in the media. Finland offers an example of concerted action at a national and civil society level to address and actively combat disinformation and fake news (see Box 1).

## 1.7. Existing approaches to identifying social media manipulation

Although social media influence is a relatively recent phenomenon, a variety of technology-based approaches has been used to achieve a better understanding of the malign use of social media, including disinformation and manipulation of user opinions and behaviour. Recent research has shown that machine learning algorithms are better at detecting false information than humans, with the added benefit that they are able to process large amounts of data rapidly.[97] Technology-based analytical approaches (as opposed to policy responses) to countering the malign use of information have taken several forms. Our review shows that the principal aims of the existing approaches are:

- Fact checking/identification of false information and content verification

96 FactBar EDU (2018).

97 Elangovan et al. (2018).

**Box 1. Example of a government response to social media manipulation: Finland**

Although Finland has experienced Kremlin-backed propaganda since its independence from Russia in 1917, several factors help the country and its population reinforce their resilience against social media manipulation. There is a strong sense of national homogeneity and civic belonging shared amongst the Finns, with the country ranking in top five worldwide for freedom of press, happiness, transparency, social justice and gender equality.[98] A high level of media literacy,[99] coupled with a strong tradition for reading, speaks to the ability to recognise fake news.

In 2014, the Finnish government launched an anti-fake news initiative aimed at teaching residents, students, journalists and politicians how to counter false information designed to sow division. This included testing students' ability to recognise 'deepfake'[100] material and reinforce critical thinking.

The initiative is just one layer of a cross-sector approach that Finland is taking to prepare citizens of all ages for the complex digital landscape of today. For instance, in 2017 the European Centre of Excellence for Countering Hybrid Threats[101] was established in Helsinki to combat subversive threats and work in close partnership with other EU Member States.[102]

Most recently, a Finnish fact-checking agency, Faktabaari (FactBar), in cooperation with the French-Finnish School of Helsinki, has developed a digital literacy 'toolkit' for elementary to secondary school students learning about the EU elections.

*Source: RAND analysis.*

- Detection of fake accounts and bots
- Detection of computational amplification
- Detection of disinformation campaigns
- Understanding the use and impact of divisive techniques (e.g. hate speech).

An overview of practices used in the analysis of malign information on social media reveals several methodologies that interlace most of the social media analytical body. The contemporary landscape of 'social media analysis' is mostly founded on network-based and text-based analytical techniques:

**Social network analysis** (SNA) is widely used to identify bots, detect fake accounts, and understand the spread and amplification of disinformation. This method of measuring and mapping the 'flow of relationships and relationship changes between knowledge-possessing entities' or different groups and members of those groups (aka node entities) has evolved over the last two decades along with the development of modern

98 Mackintosh (2019).

99 Finland ranked first out of 35 countries in a study measuring resilience to the post-truth phenomenon.

100 Refers to audio-visual material which has been manipulated but appears authentic and realistic. Conventionally this involves processing methods utilising AI and other similar technology; Technopedia (2019d).

101 Hybrid CoE (2019).

102 Krelim Watch (2019).

computing into an interdisciplinary method, allowing the mapping of highly complex node relationships.[103] On Twitter, for example, node relationships may be formed through following, mentioning and replying functions and using hashtags, as well as sharing, liking, tagging and commenting with the context that has been created by a specific user or user group. SNA makes use of computer-based tools such as NodeXL, NetworkX and PARTNER.[104] This methodology has been used in detecting and analysing disinformation in different types of social media (including Twitter and Facebook) as well as the networks and information-sharing patterns between social media and false news websites. It was, for example, used by the EU Disinfo Lab to analyse the ecosystem of Twitter accounts involved in disinformation during the 2018 Italian elections.[105]

**Natural language processing** (NLP) has been applied to detect misinformation and disinformation, as well as to identify bots. Automated detection of false information and fake social media accounts significantly decreases the human hours necessary to check and verify such data. In NLP, when analysing the relationships between sentences, rhetorical approaches (e.g. Rhetorical Structure Theory) are used to analyse the coherence of the story. This is done by defining functional relations between text units (e.g. circumstance, evidence and purpose). Machine learning models (such as non-neural network models) have also been applied to NLP. However, the automatic detection of false information online remains a significant challenge for NLP. This is due to limited access to and availability of raw data that can be used to further develop the method specifically to counteract the speed and amounts of false information spread online, as well as the need for greater clarity of linguistic differences between real and false information.[106]

**Sentiment analysis** (SA) may be considered a form of NLP and is frequently used in marketing and customer service. Sentiment analysis is a form of data mining that 'measures the inclination of people's opinions through NLP, computational linguistics and text analysis' against specific ideas, people or products.[107] SA is a useful tool to understand potential polarities of existing population opinions and 'classify and understand the users' feelings about a topic of interest'.[108] As such, this method may be of specific interest when analysing the relationship between human behaviour and social media networks, allowing researchers to detect polarity and subjectivity, as well as the sentimental strength of the data. This type of analysis has been used in studies of Twitter discussion during the 2012 elections in Germany, the US presidential elections in 2015 and the 2016 Austrian presidential elections. However, application to real-life case studies remains challenging due to the complexity of social media networks and the differences, for example, between the literal and intended meaning of statements on social media.[109]

103 Technopedia (2019a).
104 Williams et al. (2018).
105 Alaphilippe et al. (2018).
106 Oshikawa et al. (2018).
107 Technopedia (2019b).
108 Saif et al. (2013).
109 Monkey Learn (2019).

**Stance & emotion analysis** (SEA) is an expansion of SA. Where SA is restricted to polar measures, SEA uses more granular, expanded taxonomies to measure attitudes inscribed in text. **Emotion analysis** seeks to quantify audiences' emotional engagement.[110] Artificial intelligence communities have recently drawn their attention to Man–Machine Interaction (MMI) systems that use multimodal information about their users' current emotional state.[111] *Stance* goes further, with extended categories for sociocultural information in text, including temporal, epistemic, moral, social, spatial, material and cognitive aspects of language.[112] Examples of stance categories include *future* and *past*, *certainty* and *uncertainty, social goods* and *ills, concrete objects* and *properties*, and *abstract concepts*.

**Cognitive psychology concepts** have also been applied to social media analysis to identify deception and the spread of false information. Cognitive analysis as applied to social media has predominantly focused on the analysis of the 'consistency of message, coherency of message, credibility of source and general acceptability of message'.[113] For example, the University of Edinburgh's NRLab monitors the cognitive and sentiment aspects of the online debate regarding the UK's membership of the EU by analysing hashtags.[114] Expert System in cooperation with the University of Aberdeen has used cognitive text analysis and NLP to analyse the EU referendum in the UK.'[115]

**Data mining** aims to 'extract the information from a large number of data sets and convert it into a reasonable structure for further use'.[116] This approach, therefore, is focused on finding new and actionable data hidden in the 'noise'. It is especially useful as a baseline methodology for social media analysis, and to mitigate the challenge of a large quantity of raw data. Several data mining techniques may be used to structure data, such as regression, clustering, classification and sequential pattern mining. For example, classification, generally applied to data that is already labelled, is one of the most commonly used approaches in social media analysis, while clustering is often used to mine data that is new and the contents of which is unknown.[117]

**Information retrieval** is a computer science methodology used to find data 'of an unstructured nature (usually text and one that is not structured for easy use by computers) that satisfied an information need from within large collections'.[118] It explores the relevance of the information found in respect to the information requested. This method, which otherwise could be called 'information search', is the dominant form of information gathering for research as well as everyday purposes (e.g. searching for information online).

110 Allouch (2018).
111 Balomenos et al. (2005).
112 Helberg et al. (2018); Kaufer et al. (2006).
113 Kumar & Geethakumari (2014).
114 University of Edinburgh (2019).
115 Expert System (2017).
116 Elangovan et al. (2018).
117 Barbier & Liu (2011).
118 Manning et al. (2009).

**Machine learning and deep learning to build artificial intelligence** are seen as potential solutions for detecting and countering disinformation online. Due to the character of social media (i.e. large and unstructured data and its velocity[119]), forms of automated data analytics are already used by the UK and its allies. For example, NATO's Alliance Open Source System (AOSS) applies artificial intelligence and big data analytics to access, analyse and visualise open source data. Automated and semi-automated approaches to detecting and analysing malign use of information on social media are largely focused on the identification of false information, bots and social networking analysis, as well as detection and tracking of the dissemination of false information (specifically links to false news websites) across social media and fact checking. However, some researchers consider the current level of development of automated tools to be insufficient for thorough social media analysis, specifically of its semantic contents. Furthermore, the large amount of data that would need to be processed to achieve real-time analysis renders the process financially prohibitive.[120]

## 1.8. Methodological challenges and analytical gaps

The methodologies mentioned above may also be used in such aspects of social media analysis as:

- Analysis of peer users
- Linguistic and semantic analysis of tweets
- Qualitative surveys aimed at pre-qualified participant panels
- Use of algorithms to detect bots
- Measuring the usage of false news outlets
- Content analysis
- Analysis of posting frequency to identify bots and automation
- Analysis of heterophily[121]
- Field experiments and optimised search policies to identify extremist users.

Recently there has been an increased interest in understanding the malign use of social media, and specifically how it is used to influence behaviours. This has been triggered by recent controversial political events, such as the presidential elections in the US and the Brexit referendum in the UK. While the methods outlined above have already been applied to social media analysis, application has been largely fragmented, episodic and lacking in the versatility that would be required for deployment across a variety of social media platforms, themes and languages. Therefore, there is a requirement for further tests, maturing of methodologies and theory-based analytical techniques.

Furthermore, although some aspects of social media use and characteristics have been examined in existing research, other areas have been neglected. For text-based social media, more exploration is needed to understand the impacts of exposure to malign information and the effects of laws and regulations on the spread and impact of misuse of social media. A comparative analysis of existing

119 Velocity refers to the relative growth of the big data and how quickly that data reaches sourcing users, applications and systems; Technopedia (2019c).

120 Kumar & Geethakumari (2014).

121 Heterophily is the tendency of people to maintain a higher proportion of relations with members of groups other than their own; Lozares et al. (2014)

research into bot detection and the spread of malign information online is also lacking. Moreover, existing research is often hampered by the character of social media information, availability of data (i.e. due to high costs) and training data for algorithms.[122] For example, the level of anonymity available on Twitter and Reddit (compared to Facebook, for example) makes it more difficult to base the analysis in robust contextual clues. There is also a lack of standard theory and methods that could be applied to different contexts and social media/online platforms. Most existing theories on deception and disinformation online (e.g. 'management obfuscation hypothesis', 'information manipulation theory', or 'interpersonal deception theory') focus on so-called 'leakage cues' that may reveal dishonest intent, requiring a detailed understanding of verbal clues on different social media platforms. Moreover, there is no one theory that would apply to a wide variety of themes, user groups or even social media sites. While NLP combined with similarity analysis may lead to 68.8 per cent accuracy in identifying deceptive identity, the variations between different types of online media prohibit this approach from being applied elsewhere.[123]

## 1.9. Structure of the report

In addition to this contextual chapter, this report contains two further chapters:

- **Chapter 2** describes the methodological approach and findings resulting from proof-of-concept machine detection using an established troll database, and the application of tradecraft analysis of Russian malign information operations targeting left- and right-wing publics.
- **Chapter 3** examines the different ways and environments in which the tool developed as part of this study can be practically applied, as well as potential strategies for building resilience among targeted populations to malign information.

Supporting the findings presented in the main report, **Annex A** includes a broader discussion of resilience to malign information among vulnerable publics.

122 Tucker et al. (2018).

123 Tsikerdekis & Zeadally (2014).

# 2 Findings and analysis

This chapter presents the findings that emerged as a result of conducting proof-of-concept machine detection using a known troll database (Task B) and applying tradecraft analysis of Russian malign information operations against left- and right-wing publics (Task C).

## 2.1. Introduction

The proof-of-concept machine detection application was developed to meet the following primary objectives:

1. Use theory-driven analytics to characterise differences between authentic political supporters and trolls.
2. Use machine learning to identify and distinguish malign accounts on Twitter (Russian 'trolls') from non-malign ('authentic') accounts on Twitter.
3. Suggest effective responses to counter trolls.

Our results are promising and can contribute to combatting malign information campaigns by Russia and other actors in a significant way. Our basic pilot had high accuracy detecting Russian trolls (71 per cent on average). This pilot used a highly interpretable shallow algorithm, in contrast to powerful (but black box) deep algorithms. This interpretability allowed us to understand the tradecraft employed by trolls to sow dissent online, and our shallow model can now be engineered to inform deployment using deep algorithms. We also pioneered the use of a particular kind of sociolinguistic feature set – *stance* – that is topic-agnostic, and thus has the power to work in diverse contexts.

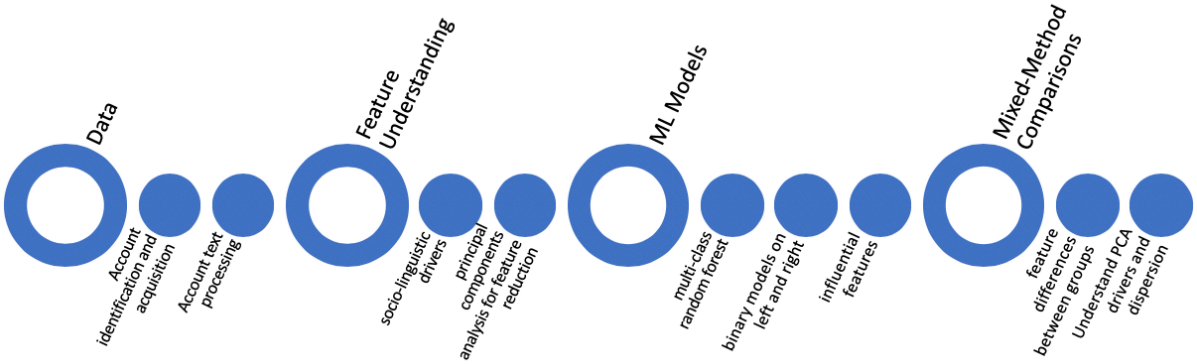
### 2.1.1. Building a theoretical and interpretable machine learning model

Starting with a dataset of known Russian malign accounts (called 'trolls'), we then worked backwards to find the online argument they were targeting within the US 2016 election. This involved transforming the data for machine learning, and then conducting quantitative statistical and qualitative text analyses.

Following this data transformation, we used an existing taxonomy of sociolinguistic features, developed at Carnegie Mellon University, adapted for use as features for training a machine learning model to discern between Russian troll accounts and authentic political supports on Twitter. We then used data reduction methods, descriptive statistics and qualitative interpretation of text mining to better understand the rhetorical strategies and tradecraft of Russian trolls.

As noted above, our analysis used a taxonomy of linguistic stance that was developed at Carnegie Mellon University. While – like all models – the CMU stance taxonomy is flawed, it has proven to be useful across a wide range

**Figure 2.1. Workflow for analysis and findings**




*Source: RAND analysis.*

of text analysis and machine learning tasks, including linguistic forensics,[124] automatic text classification tasks,[125] cross-cultural English as a second language (ESL) instruction,[126] consumer sentiment analysis,[127] and strategic communication analysis.[128] Further, as this report details, stance features allowed us to build a machine learning model that is highly accurate in distinguishing Russian trolls from authentic Twitter users. So while we cannot speak to how valid this stance technology is, we do point out from a pragmatic perspective how effective it is.

Our methodological choices reflect both our theoretical commitments and the datasets. We used network analysis to create sub-sets of data, and linguistic stance in response to the social and rhetorical nature of public sphere argumentation. We used scalable machine learning approaches, and in particular shallow algorithms, because of the vast size of our dataset, and because we needed to "look under the hood" of the analysis in order to interpret results and discover tradecraft employed by Russian trolls.

## 2.2. Scalable network analysis of the argument space

The following sub-sections elaborate in detail on the workflow and methods used by the project team.

### 2.2.1. A Twitter dataset of known Russian trolls

We started with an existing, publicly available dataset of Russian left- and right-wing troll accounts active on Twitter during the 2016 US presidential election.[129] These malign accounts were run by the Internet Research Agency, or

124 Airoldi et al. (2006); Airoldi et al. (2007).
125 Collins (2003); Hope & Witmore (2010).
126 Hu et al (2011).
127 Bai (2011).
128 Marcellino (2015).
129 Github (2018).

'IRA', commonly referred to as the Russian Troll Factory. The dataset contains 2,973,371 tweets from 2,848 IRA-operated handles, identified in a 2018 report released by the US House Intelligence Committee.[130] Researchers at Clemson University subsequently analysed and classified the data, including two specific types of troll accounts attempting to influence the election: *right-wing trolls* and *left-wing trolls*.[131] These trolls were extremely prolific, tweeting out barrages of both support and hostility towards the candidates. They also put out messages stoking conservative fears over immigration and Islamic terrorism, and liberal concerns over fascism and police violence towards African-American citizens.

We understand this Russian malign effort to be aimed at increasing social tension between both progressive and conservative *publics* – groups of people with a shared stake in public debate on a contentious issue (the 2016 presidential election in this case), and shared ways of talking about that issue.[132] Based on our working assumption, examining how left- and right-wing trolls targeted this public debate will carry over to other left/right political arguments across the globe likely targeted by Russia, including Brexit.

### 2.2.2. Text mining for authentic account identification and acquisition

We used text mining to harvest key terms from each troll group, using those as the seed for a data pull of English language tweets from 2015 to 2017, matching the time period that the Russian trolls had been active. Our assumption was that the trolls were actively engaged in malign influence activities, and we could backtrack to the conversations they were part of from their key words. In essence, *we used the trolls to trace the publics they were trying to influence.*

## 2.3. Labelling the different groups: left-/right-wing trolls vs authentic liberals/conservatives

The existing dataset of left- and right-wing trolls was labelled by researchers at Clemson University using human, qualitative coding. We kept the same naming convention for the purposes of continuity and clarity with other research efforts using the same public dataset. We then needed a naming convention for the authentic speakers those trolls were hiding among. Our network analysis showed two distinct meta-communities, and subsequent text analysis of the Twitter content showed that each group posted content in support of either the Democratic or Republican candidate and policies associated with their ideology. We wanted labels that were analogous enough with left and right wing to show correspondence, but distinct enough not to cause confusion between the trolls and authentic speakers. Because the data related to a US election, we chose the US word pairing 'liberal' and 'conservative'. While 'liberal' does have some other associations, such as 'classical liberalism', we felt the pairing with 'conservative' would disambiguate. We were guided by usage as well, in that while 'progressive' is an increasingly favoured self-description, 'liberal' is much more common in everyday usage. For example, in the Corpus of Contemporary American English (COCA), a large, 560 million-word corpus spanning 1990 to 2017, and a common reference for general

130 US House of Representatives Permanent Select Committee on Intelligence (2018).

131 Boatwright et al. (2018).

132 Hauser (2002).

usage, 'liberal' appears 25,604 times, while 'progressive' appears only 9,828 times.[133]

### 2.3.1. Mapping the 2016 US presidential election argument using network analysis

Using this data frame, we harvested tweets from approximately 1.9 million user accounts. We did this through RAND's contract with Gnip, a social media data API (Application Programming Interface) aggregation company for Twitter's full Firehouse.[134] Our hypothesis was that within this large collection of social media data we would find authentic liberals and conservatives: two large *publics* engaged in argument over the election. To determine if this was the case, we adopted a network analysis algorithm[135] that uses frequency of interaction to infer communities: people who interact with each other on social media in a consistent manner. As expected, we identified two large meta-communities, each consisting of sub-communities. Subsequent text analysis showed that one meta-community represented the supporters of Donald Trump and the other supporters of Hillary Clinton. These meta-communities contained approximately 1.8 million of the total 1.9 million users we collected – 94 per cent of the tweets we harvested were engaged in the online argument we hoped to find.

Figure 2.2 visualises the Twitter argument space around the 2016 election, and is important because it illustrates 'publics', the unit of work in public sphere argumentation. In local private matters we argue as individual persons – for example members of a family all making their own case for where to go for the family holiday. But in public sphere matters – elections, immigration law, civil rights – we argue as groups, using common language that reflects the values and ideology of the group, in ways that rhetorically match the group's advocacy goals. The argument network map below makes visible what Russian trolls targeted: a liberal public (pink) that supported Hillary Clinton, and a conservative public (green) that supported Donald Trump. Furthermore, the network shows that these large publics are composed of smaller sub-publics, allied but distinct.

What our algorithm labelled 'community 60083', we might instead call '#Black Lives Matter Advocates', because text analysis shows that the distinctive language in that community includes words like '#BlackLivesMatter', 'Hip hop', 'white supremacy', 'police brutality', and so on. In contrast, within the conservative meta-community we see 'community 64045', a community marked by negative language around immigration, Islam and 'illegals'. This is the rhetorical battlefield the Russian trolls entered, including the specific, most extreme publics they targeted.

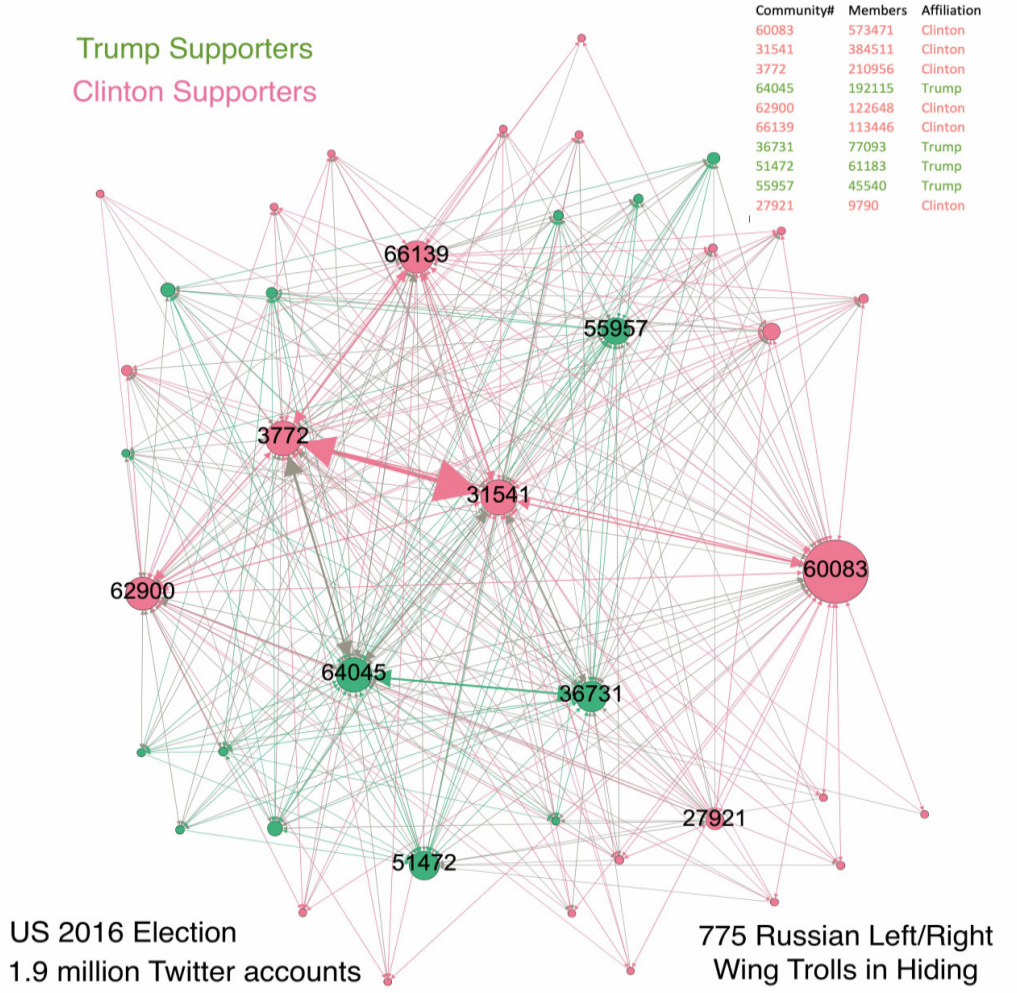
Each dot (node) is a community of Twitter users, sized by the number of accounts in the community from hundreds to hundreds of thousands of users. The colour of the nodes shows membership within a higher order 'meta-community'. We ran the network algorithm first to identify individual communities, and then a second time to discover the two larger 'meta-communities' or communities-of-communities. The pink-coloured communities supported Hillary

133 Davies (2008).

134 Because this step used public data from real people, RAND Europe's Human Subjects Protection committee reviewed our plan, determining that while our research did not constitute human subject research, we should still take steps to protect against harm. Therefore, although tweets are publicly available data, we do not quote or give examples of the tweets used in the study.

135 Clauset et al. (2004).

**Figure 2.2. Network diagram of Trump and Clinton supporters on Twitter, 2015–2017**




*Source: RAND analysis.*

Clinton, and the green ones supported Donald Trump.

Each line (edge) indicates interactions between communities, and the thicker (higher-weighted) the edge, the more interactions there are. Each edge is an arrow, but most are so small and tenuous that the point of the arrow is invisible. However, for the largest and most central communities, the interactions are so dense that the point of the arrow is visible, showing directionality.[136] The relative placement of the communities is an algorithmic function showing the position of communities by mutual connection, in a way that is visually

136 For example, the edge between the Soccer Fans community and the Pro-Russia community is bidirectional, and relatively thick, with weights of 14,874 and 16,084 in each direction, respectively. By contrast, the Pro-Ukraine to Soccer Fans edge is also bidirectional, but thinner, with weights of 5,395 and 6,625, respectively.

clear. In essence, the network diagram is a map of the 2016 US presidential election argument space, in which the Russian trolls mentioned above were broadcasting messages meant to increase dissent and social division.

As described above, starting with an existing ground truth set of 775 Russian troll accounts masquerading as liberal and conservative supporters of Hillary Clinton and Donald Trump, we then identified and acquired approximately 1.9 million authentic liberal and conservative supporters, as opposed to state-sponsored malign accounts. Now with both sets, we could proceed to training a machine learning model that could successfully distinguish between the four account categories: Russian left-wing trolls, Russian right-wing trolls, authentic liberal supporters and authentic conservative supporters.

## 2.4. A theory-based machine learning model that successfully detects Russian trolls

We used theory-guided linguistic features as inputs to machine learning (ML) models to successfully distinguish between the four categories of accounts in our study.[137] The specific feature-set was an expert dictionary and taxonomy of rhetorical functions. The most common feature set for ML classification of text documents is words (terms), although features such as punctuation, sentence length and tagged parts of speech can be used as well.[138] An advantage of using stance taxonomy is the ability to code for rhetorical effects regardless of the topic of discussion. A term-based model would see 'illegal immigration' and 'murder' as completely different, but in our taxonomy, both are considered 'Public Vices' – unwelcome phenomena in the public sphere. This allowed us to use stance as a machine learning feature, increasing the accuracy of our model, but also enabling the understanding of the rhetorical, persuasive tradecraft being employed by Russian trolls.

Another noteworthy feature of our model-building efforts was the use of simple algorithms. Current state of the art machine learning work often emphasises deep-learning algorithms that use neural networks. These algorithms are powerful and have great practical value, especially in terms of predictive performance. However, deep approaches have limitations: they depend on data abundance, they can be brittle and they may not be transferable to new problems and datasets. Furthermore, because of their dependence on hidden layers, they are considered to be black box solutions. In this sense, deep approaches are empirical rather than theoretical. Generally, pouring vast amounts of data into a powerful deep-learning algorithm produces powerful, useful results, but does not necessarily show why the algorithm works, and thus does not contribute much to the ability to derive generalisable insights.

Simple algorithms are commonly much less powerful but allow for interpretability: we were able to unpack the model to determine how it worked, and why it was better at classifying conservatives and right-wing trolls than liberals and left-wing trolls. This was an important feature given that the project pilots the use of stance as a sociocultural feature set in machine learning.

137 Our classes were driven by the network analysis: based on social interactions, we can empirically show a reason to split the Twitter dataset into two classes: liberals and conservatives. Further research could be done to determine if there were meaningful sub-classes within the larger publics (the more varied conservative one in particular), but this was out of scope for this study.

138 Maitra et al. (2016).

As such, we used a random forest algorithm,[139] in part because it performed well in our trials, but also because it allows for sorting by relative significance of features. This is an important factor in trying to understand how Russians successfully (or unsuccessfully) imitate authentic political supporters. At this stage of the project, our focus was interpretability: an effective model that not only distinguishes between types of trolls and authentic supporters, but one that we could understand. This would allow us to identify possible defensive measures and build a more precise and accurate model in the future.

To transform the dataset for machine learning modelling, we consolidated all tweets by account,[140] removed all URLs and handles (as noise), and split each account's total output into approximately 1,000-word sequences (referred to as a word 'chunk'). We chose that number because rhetorical and corpus linguistics research shows that *factors* – the co-varying language features that typify genres and pragmatic attempts to persuade – have minimum sizes. On the one hand, we wanted to maximise the *number of chunks* for training our model. On the other hand, empirical testing on chunk size shows that linguistic features are fairly stable at 1,000 words, degrade at 500, and have low gains when chunk size is increased to between 1,500 and 2,000 words.[141] Therefore, 1,000-word chunks by account were optimal for this effort.

### 2.4.1. Rhetorical and sociolinguistic theories reveal Russian troll tradecraft

Our modelling effort produced three important findings:

- The model[142] we use to demonstrate the classification task was highly effective at the top-level binary task of distinguishing trolls (of any type)[143] from authentic supporters, with an F1 score (average of recall and precision, taking into account both false negatives and false positives) of 87 per cent. Given that this is a first cut at classification, using all possible stance features, we are highly encouraged by this level of performance. With additional feature engineering to use only important features, combined with potentially better performance using deep algorithms, it seems likely we could effectively identify troll accounts.
- We also attempted the more challenging, granular task of distinguishing between authentic posters and trolls by political type, a multi-class task. In this four-way model, the F1 performance score dropped to 71 per cent.[144] An interesting finding in this multi-class task was that the model was more effective at classifying right-wing trolls and authentic conservatives (74 per cent correctly classified for both), than left-wing trolls and authentic liberals (69 per cent and 70 per cent correctly classified, respectively). In essence, the model was

139 Liaw & Wiener (2002). Random forest is an ensemble method: multiple models using random subsets of data and features, classifying by majority vote. This ensemble voting helps prevent 'overfitting' of the data, where idiosyncrasies in the training data produce models that work well on the training data, but not well on new data.

140 This is a forensic effort around online personae, so the account was the appropriate unit of classification, rather than individual tweets.

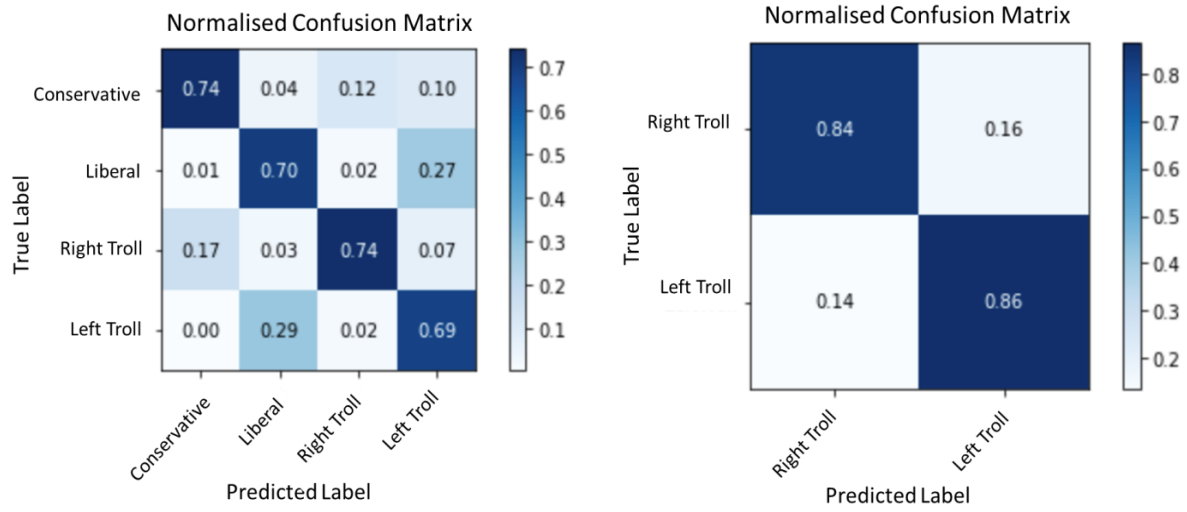
141 Biber & Finegan (1991).

142 This is a random forest classifier.

143 A multi-class model with four classes: (1) right troll, (2) right authentic, (3) left troll and (4) left authentic.

144 The Matthews Correlation Coefficient (MCC) is better than random (identified by zero) at 0.71.

**Figure 2.3. Classification of Russian trolls using only stance**




*Source: RAND analysis.*

*Note: The left-hand matrix represents multi-class trolls and authentic supporters; the right-hand matrix represents right-wing and left-wing troll accounts.*

4–5 per cent less effective when applied to the liberal half of this four-way problem (see the left-hand matrix in Figure 2.3). As an additional step we looked at only trolls (the right-hand matrix in Figure 2.3), and found the model was slightly (2 per cent) more effective in the opposite direction: 84 per cent for right-wing trolls, but 86 per cent for left-wing trolls. This helped us see that the drop in efficacy was not about trolls, so much as about the heterogeneity of authentic conservative talk, explained in more detail below.

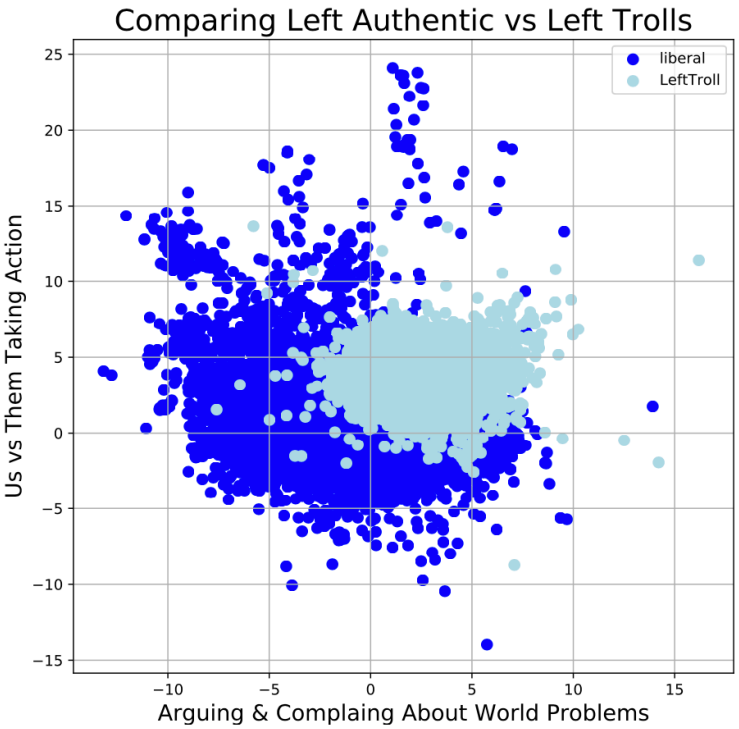
- The model allowed us to conduct a descriptive, statistical analysis of left- and right-wing trolls to understand the tradecraft Russian trolls use in trying to influence and disrupt publics.

Figure 2.3 shows the confusion matrices for our model's performance. The matrices should be read along the dark blue diagonal: 'True label' represents ground truth, 'Predicted label' is what the model classified, and the cells represent the proportion within the class. If the model was to predict behaviour perfectly, then all the values along the main diagonal would be exactly 1.00. Thus, the measure of confusion is how much of the model prediction lies off the main diagonal.

The four-way classification in Figure 2.3 has two important implications:

- The model is better at detecting left-wing trolls than right-wing trolls. As we discuss below, this is likely because liberal talk is more homogenous than conservative talk. The kind of urgent, social complaint about world problems (e.g. racial oppression and violence) that left-wing trolls performed maps well onto liberal discourse. Conversely, the rhetorically similar complaints of right-wing trolls on different topics (in this case, illegal immigration and Islam) maps onto only a portion of conservative discourse.

**Figure 2.4. PCA visualisation of left-wing trolls and authentic liberals**




*Source: RAND analysis.*

- When the model does misidentify authentic conservatives, it does this almost equally between the two types of trolls. That is, *some conservatives talk like trolls in general*. However, while liberals are more often misidentified, it is almost always as a left-wing troll, never as a right-wing troll. Again, this appears to reflect the fact that liberal speakers are more coherent in their complaints over problems, whereas only some conservatives speak in a similar way.

While we were satisfied with the performance of the model using only stance features (e.g. use of abstract concept vs concrete objects language, or talking about social problems vs social goods), we were curious about the difference in performance between left-wing/liberal classification versus right-wing/conservative classification. We turned to principle components analysis (PCA)[145] to visualise the following classification question: are Russian trolls better at imitating liberals or conservatives, or is there a modelling issue?

The analysis showed that liberal supporters of Hillary Clinton tended to write in a relatively homogenous way, while conservatives were much more varied in their linguistic stance, and thus more difficult to imitate. Figure 2.4 shows how liberal discourse was a relatively tight, roughly centre-target for imitation.

145 PCA reduces a large dataset to the most important variables, allowing us to plot them along two axes. This allows us to understand better the important variables in explaining variance between two classes, but also to visualise their difference, as in Figure 2.3 and Figure 2.4.

The horizontal axis shows the first principle component explaining the most variation in the data. This first principal component emphasises stance variables related to emotional negativity, public vice, social/personal roles and immediacy. In text, these often appeared as complaints about problems in the world, such as illegal immigration or police shootings of unarmed citizens. We can generally interpret the accounts that are more positive (or on the right) of this axis to emphasise these aforementioned stance variables. The vertical axis shows the second principle component, which emphasises variables such as social distance and closeness, and reporting events, something we describe as 'us taking action against them'. While separate PCA analysis was performed for the left and right accounts, the first two components of each PCA represented these two ideas – 'complaining about world problems' and 'us vs them taking action' – based upon the emphasis of stance features within the components.

While left-wing trolls could not perfectly cover the centre of liberal discourse around the election, it is a relatively close overlap. This means that left-wing trolls focused on a portion of the authentic user population. Whether that was a limit of their ability to imitate, or a choice to be selective in their imitation, is an open question. The overlap in this figure is based on combinations of stance features that represent two dimensions and demonstrates that the portion of overlap has more weighting on negative, uncertain and dismissive discourse (x-axis) and first-person descriptive discourse (y-axis). This rather compact region in Figure 2.4 can be contrasted with the more diverse language of conservatives in this discussion (the increased dispersion is shown in Figure 2.5). The interpretation of the components in this figure is similar to those in the above Figure 2.4, with complaints and anger over different topics (e.g. illegal immigration and Islamic terrorism).

Russian trolls imitated only a selection of conservative discourse. This may be because of the difficulty in imitating the wider variety of linguistic stances conservatives used, or possibly a tactical choice regarding which part to imitate. Of this selected conservative discourse, the portion the trolls focus on represents one end of the spectrum based upon the first principal component, the genre of complaining online about bad things happening in the world. Based on this, we concluded that disparity in classification is likely a result of the nature of liberal and conservative talk on Twitter around the 2016 US presidential election: liberal talk was more homogenous, and thus easier to imitate.

We also conducted a mixed method, human/machine analysis[146] to better understand Russian tradecraft, and to turn our quantitative description into interpretable insight. The first step was a descriptive statistical analysis[147] to quantify the meaningful stance differences between the classes. We then looked at many sample-rich features to understand what those statistical differences mean in actual speech.

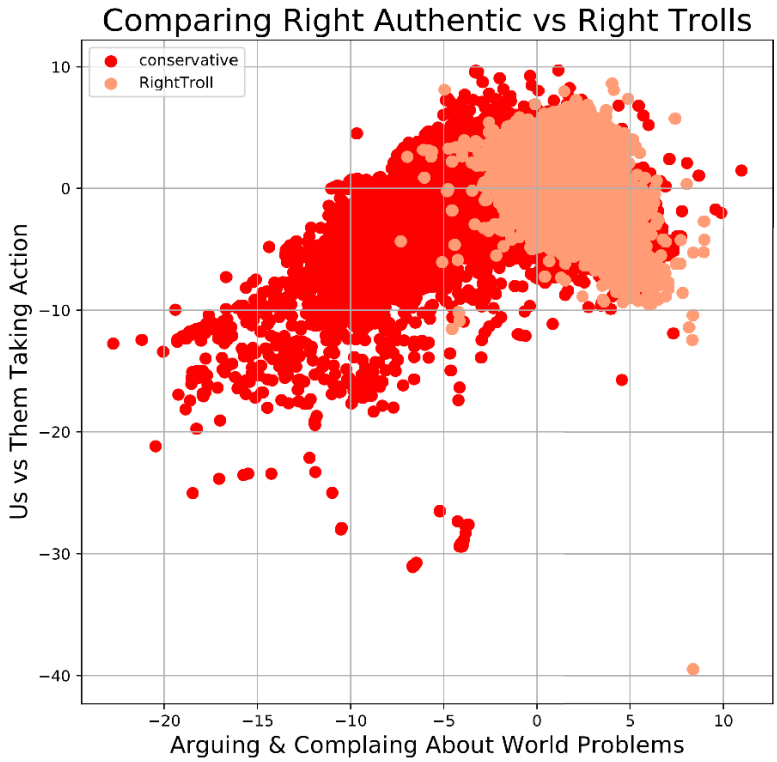
We found the following features frequent among left-wing trolls:

- ***1st person:*** I/me/my language
- ***Immediacy:*** words and phrases like 'this moment', 'right now' language
- ***Social problems:*** words and phrases like 'murder', 'injustice' or 'racism'
- ***Personal roles:*** social categories in the world, e.g. 'black man' or 'white cops'.

146 Kavanagh et al. (2019).

147 Abdi & Williams (2010).

**Figure 2.5. PCA visualisation of right-wing trolls and authentic conservatives**




*Source: RAND analysis.*

In practice, this often looks like expressions of outrage over contested issues, for example urgent complaints over sexual assault cases, or the death of a black citizen at the hands of a white police officer. Since this genre of 'outrage stories' was a large part of liberal discourse, Russian left-wing trolls did a better job mimicking liberals.

As Figure 2.5 shows, conservative discourse was more diverse and spread out. Russian trolls were able to imitate one part of that discourse, marked by the following stance language:

- ***General negative emotion:*** words and phrases such as 'terrorism', 'destruction' or 'weak'
- ***Generic events:*** the prosaic part of our lives, language like 'the way we live', 'go to school' or 'daycare'
- ***Concrete objects:*** physical objects in the real world such 'trucks', 'the railroad' or 'guns'.

In practice, this looked like complaints about perceived threats to the US, such as illegal immigration or Islamic terrorism. Russian trolls were able to imitate this part of conservative discourse, but did not imitate other types, for example more libertarian talk that was rich in abstract concepts and public authority language ('the Constitution', 'the courts'). We note the *heterogeneity* of the conservative talk – religious and social conservatives talk very differently than libertarians, for example,

even on the same issue, presenting a difficult discourse to imitate.

Our analysis indicates that it is likely that Russian trolls imitated the most divisive, confrontational elements of liberal and conservative US publics: a 'far-left' liberal discourse that constructs the US as fundamentally oppressive, and a 'far-right' conservative discourse that constructs the US as under attack by outsiders.[148] This would be consonant with the wider Russian practice of trying to stoke discord and division in other populations, but also points to possible defensive measures. Our analysis of the specific rhetorical tactics used opens the door not only to identifying trolls in real time, but also to showing the targets of these tactics and how they are being manipulated.

We can conclude that stance alone, without any topical content from raw text such as word terms and frequencies, differentiates groups and provides valuable insights about them. While trolls are able to replicate authentic speakers, they only replicate a small portion of the variability exhibited by them. Left-wing trolls more consistently represent authentic liberal discourse whereas right-wing trolls primarily represent a sub-section of conservatives, failing to replicate the broad range of conservative discourse in the argument network we examined. These models serve as a baseline, where the addition of the topical data should improve the predictive power of classifying malign ('troll') accounts. In addition, potential improvements could include:

- Increasing the feature set and reducing misclassifications by adding term frequencies in addition to stance features.
- Making the feature inputs more interpretable by adding derived features that represent an idea (where the idea consists of combined features).
- Using word 'embeddings' to help link similar terms in the vocabulary.
- Reducing misclassifications and improving predictive power by augmenting the machine learning architecture (moving from a shallow to a deep model or an ensemble model).
- Adding cutting-edge models and pertained word 'embeddings' from other institutions (e.g. BERT[149] and Grover[150]) to distinguish between troll and authentic accounts.

148 This is evidenced in the PCA analysis.

149 BERT refers to Bidirectional Encoder Representation from Transformers, a technique developed by Google AI Language for Natural Language Processing (NLP) applying bidirectional training of Transformer to language modelling. For more information, see Horev (2018).

150 Grover represents a model directed at 'neural fake news' – fake news techniques that apply AI to mimic the tone and language patterns of particular publics on social media. For more information, see Zellers et al. (2019).

# 3 Considerations and future opportunities

In this final chapter we explore the different ways and environments in which the tool can be practically applied. This chapter also discusses potential plausible strategies for building resilience to malign information in targeted populations.

## 3.1. Concluding remarks

We have shown that combining social network and stance analysis, guided by rhetorical and publics theory, is an effective way to detect malign troll accounts hiding in the social media sphere. With a simple algorithm, and only using stance features, we were able to detect troll and authentic accounts in a four-way classification problem. This preliminary effort yielded high levels of accuracy, and through feature engineering and the use of deep algorithms, we can reasonably expect even better performance. Through the use of a shallow algorithm for our first pass, we were also able to improve our understanding of Russian tradecraft, which opens up the possibility of robust defence against their tactics. More importantly, our model architecture should be transferable to other contexts.

As such, our project has four main takeaways that could be operationalised by UK government entities concerned with foreign malign information campaigns:

1. Publics theory and network analytics can help map out the rhetorical battlefields that Russia (and others) attempt to influence in malign ways. Publics-based analytics that make visible socio-political aggregate structures such as meta-communities should be widely adopted by government entities seeking to counter malign foreign efforts.
2. Russian trolls work both sides of a controversy, and to stoke social discord they highlight hot button issues for each side using a repeated rhetorical pattern. Government entities should seek to make this visible to members of targeted publics.
3. Using less power but shallow interpretable models is an important preliminary step: interpretability allows for insight on tradecraft (and thus defence), but also informs subsequent efforts to engineer a more powerful deep algorithm.
4. Stance features (sociocultural features) can detect influence/rhetorical patterns across different topics and can powerfully enhance machine learning classification that seeks to counter foreign malign information efforts.

## 3.2. Potential plausible strategies for building resilience to malign information in targeted populations

While Section 1.6 outlines a multitude of existing technical, regulatory and educational approaches to countering bots and other types of malign actors, an effective resilience-building framework ought to take a broader approach to reduce the vulnerability of relevant publics to social media manipulation. Although resilience can apply to many different contexts,[151] we focus on exploring ways of achieving societal resilience[152] as a way to integrate the developed tools to detect the tactics of social media manipulation into a more systemic and proactive approach. This section builds on a comprehensive review of relevant literature on resilience-building in the context of social media manipulation, misinformation and disinformation. While this section focuses on outlining key aspects of the proposed approach, Annex A includes a broader discussion of the resilience-building framework based on the literature review.

Existing research on hostile social manipulation demonstrates that most methods utilising social media become effective through the exploitation of existing vulnerabilities, rather than the creation of new ones.[153] Therefore, resilience-building ought to focus on reducing the specific **vulnerabilities** of relevant stakeholders to malign information, as well as restoring and strengthening relevant **capacities** to respond to and counter the effects of malign information. These capacities can include the ability to cope with an immediate threat in the short term, adapting to new circumstances through learning and improving existing processes, as well as creating societal robustness through the empowerment of relevant stakeholders and the improvement of inter-stakeholder collaboration.[154]

A review of existing approaches and initiatives[155] indicates that in this context, the creation of a misinformation or social media manipulation detection tool serves as a crucial enabling factor for resilience-building. Detection tools based on machine learning methods are particularly advantageous in the resilience-building context as various potential biases limit the credibility of human fact-checking processes.[156] At the same time, strengthening resilience solely on the basis of exposing misinformation is challenging because institutions often lack credibility, and users have pre-existing beliefs that are difficult to challenge only by correcting the information at hand.[157]

This indicates that detection tools should be integrated within wider approaches to address specific stakeholder capacities, such

151 Early notions of resilience have focused on the resistance and stability of ecological systems, with perspectives slowly expanding into systems behaviour in sociology, economy and engineering. For more background on the concept of resilience, see Rehak & Hromada (2018).

152 Reflecting a systems behaviour perspective, our understanding of societal resilience is that of 'the ability of a system, community or society exposed to hazards to resist, absorb, accommodate, adapt to, transform and recover from the effects of a hazard in a timely and efficient manner, including through the preservation and restoration of its essential basic structures and functions through risk management.' For more information, see UNDRR (2017).

153 Mazarr et al. (2019).

154 Keck & Sakdapolrak (2013).

155 See Section 1.6 and Annex A.

156 See Annex A for more discussion on this issue.

157 Pamment et al. (2018).

as user-friendly educational online tools and platforms. Any such strategic framework on building social resilience would therefore benefit from considering the following elements:

1. Fostering **critical thinking** and driving **digital literacy.**
2. Identifying **relevant stakeholders** and strengthening institutional and **stakeholder capacity** to enhance their ability to address social media manipulation.[158]
3. Integrating the developed **social media manipulation detection tool** into user-friendly misinformation resilience platforms to advance the development of media literacy and other relevant skills among vulnerable publics. Beyond detection and understanding, the tool could be used to proactively focus on accounts at risk for malign propagation and malign influence (via the social network analysis). The language and key markers for trolls could be used to proactively identify malign information in other social settings (e.g. forums, blogs).
4. Ensuring continued **collaboration** across the stakeholder coalition to foster **inter-stakeholder trust and community ownership**.
5. Creating avenues for transparent and tailored **strategic communications and dissemination** of the resilience platform, ensuring that it reaches vulnerable publics and fosters user engagement.

Figure 3.1 demonstrates the building blocks of the proposed framework, integrating malign user detection into a wider proactive collaborative approach serving to strengthen resilience through fostering media and digital literacy; empowering users though incentivising user engagement; and creating inter-stakeholder trust and collaboration.

A relevant example of the integration of misinformation detection methods into media literacy-focused platforms is **Tanbih**, a recent initiative from Hamad Bin Khalifa University's Qatar Computing Research Institute (QCRI) that includes the training of users to recognise social media manipulation techniques.[159] This allows stakeholders to address effectively the threat of malign information in a preventative manner, tackling online social manipulation before information is published online. Similarly, the **Co-Inform** initiative (see Box 3) included the development of two tools integrating online fact-checking techniques: a browser plug-in serving as an awareness-raising tool for users to recognise misinformation, and a platform for journalists and policy-makers to monitor the nature and effects of misinformation techniques. These two cases also highlight the issue of media literacy, which creates the foundation of societal resilience to malign information (see Box 2 for more information on media literacy in Ukraine, and Annex A for discussion of the centrality of media literacy in resilience-building more broadly).

158 Though our proposed approach focuses predominantly on fostering bottom-up societal resilience, as Annex A discusses, strengthening the capacity of key institutions is also significant where existing institutional bodies are ill-equipped to address social media manipulation. However, policies should avoid adding an additional layer of complexity to response efforts through the immediate creation of new bodies rather than strengthening/revising the mandates and resources of existing ones.

159 Tanbih (2019).

**Figure 3.1. Integrating the malign user detection tool to build resilience among vulnerable publics**


*Source: RAND analysis.*

**Box 2. Learn to Discern (L2D): Improving media literacy in Ukraine**

Between 2015 and 2016, the International Research and Exchanges Board (IREX) conducted a media literacy programme in Ukraine called Learn to Discern (L2D). The initiative aimed at addressing various structural features of disinformation vulnerability in Ukraine: though almost three-quarters of Ukrainians consume news frequently, less than a quarter have trust in the media, and media literacy skills are generally under-developed.

IREX's evaluation determined that through intensive skills-building seminars, the programme reached an estimated 2.5 million Ukrainians through a combination of direct participation, indirect engagement and publicity activities. Evaluation results show that in comparison to a control group, L2D seminar participants were:160

- 28% more likely to demonstrate sophisticated knowledge of the news media industry,
- 25% more likely to self-report checking multiple news sources
- 13% more likely to correctly identify and critically analyse a fake news story
- 4% more likely to express a sense of agency over what news sources they can access.

*Source: Murrock et al. (2018).*

160 Murrock et al. (2018, 4).

**Fostering critical thinking** and creating the skills to recognise online malign information, however, only represents one aspect of addressing the vulnerabilities commonly exploited by the techniques of social media manipulation. The case of Co-Inform, launched by a consortium of organisations from various EU countries to foster critical thinking and digital literacy (see Box 3), exemplifies this through the so-called Co-Creation method. Co-Creation allows for the integration of various stakeholder collaboration methods in the process of developing tools and platforms to strengthen the stated goals of critical thinking and digital literacy. An equally important, though less easily conceptualised, problem for strengthening resilience is identified as **strengthening trust between relevant stakeholders**. This creates somewhat paradoxical imperatives for resilience-building through media literacy, whereby users are encouraged to think critically while building more trust towards the media and democratic institutions.[161] To address this caveat, resilience-building initiatives ought to incorporate processes and techniques of **inter-stakeholder collaboration**. The identification of relevant stakeholders in this context could include civil society organisations, media stakeholders including journalists, industry and private sector actors, as well as key policy-makers and government agencies.

Apart from enhancing inter-stakeholder trust, this can serve to address an additional potential challenge, namely that media literacy initiatives are more likely to be effective if designed in a format and context that users and participants are familiar with.[162] Similarly to strategic communications, media literacy requires a **tailored outreach approach** that takes into account the experiences and circumstances of individual communities and stakeholders.[163] By incentivising participation from all relevant stakeholders, initiatives can foster a sense of collective community ownership. Additionally, collaboration methods can contribute to more tailored strategic messaging and dissemination of the platform to relevant vulnerable publics.

More holistically, engaging a broad coalition of stakeholders can ensure sufficient **buy-in from key policy-makers** as well as civil society organisations (CSOs), which are equally crucial for the success of resilience-building.[164] Annex A discusses in more detail how CSO engagement as well as buy-in from policy leaders strengthens the overall framework through a bottom-up resilience-building dynamic.

161 Huguet et al. (2019).

162 Huguet et al. (2019).

163 EPRS (2019).

164 Jankowitz (2019).

**Box 3. Co-Inform: Co-Creation of misinformation resilience platforms**

Co-Inform is a resilience-building initiative that embeds the 'Co-Creation' method for 'achieving the right equilibrium between actors and types of solutions against misinformation' and fostering interaction between academics, journalists, the private and non-profit sectors, as well as civil society 'with minimal intervention'.[165] With a view to fostering digital literacy and critical thinking, the method is directed at the development of 'Co-Inform Tools' consisting of:

- A browser plugin as an awareness-raising tool, which highlights misinformation content for users.
- A Dashboard for journalists, fact-checkers and policy-makers showing the origin of identified misinformation as well as its current and potential extension and relevance for public opinion.

The project's timeline revolves around a consultative 'tools development' stage in 2019, release of the 'Co-Inform tools' in 2020, and release of three pilot studies in 2021. Given this timeline, no official programme evaluation is available as of yet. However, the project design serves as a positive example of the engagement of a broader coalition of stakeholders not only in the process of detecting malign information, but also in developing relevant tools and platforms, fostering a broader sense of community ownership.

*Source: Co-Inform (2019).*

## 3.3. Future opportunities and next steps

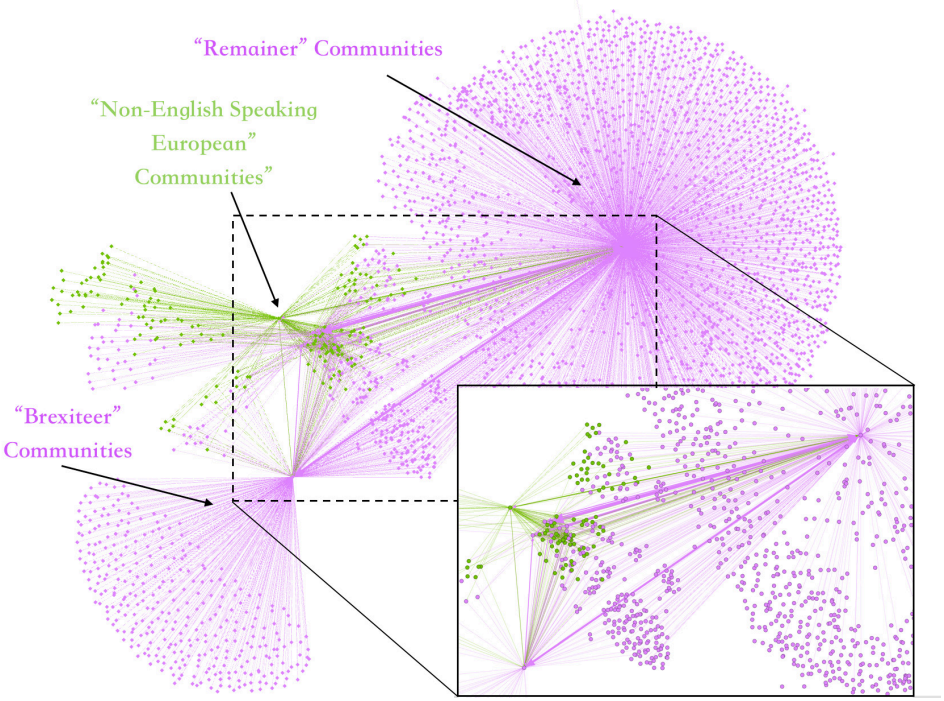
Based on the findings of this study, we can identify several possibilities for future work on this issue:

- Build on the current findings and extend the modelling capabilities to leverage other state-of-the art deep learning and artificial intelligence approaches. These include combining the salient features from the sociolinguistic approach with neural detection – including, but not limited to, comparisons to Allen Institute's ELMo, University of Washington's Grover model or Open AI's GPT-2.
- Leverage the sociolinguistic features and derived features as inputs into partner models.
- Extend and generalise findings to other datasets and platforms.

Beyond the specific problem of detecting and resisting malign information efforts, the component technologies we piloted have wider applications. Specifically, during the Behavioural Analytics Showcase, prospective UK MOD customers understood and showed interest in what we had done, but asked how adaptable our technology was for more general social media monitoring applications. Based on our discussions with those potential users, the community detection, text analysis, machine learning and visualisation components could be re-configured to create a robust general purpose social media monitoring tool. Unlike monitoring tools ported from the commercial world, we could provide a system that accounts for aggregated structures like publics, and for

165 Co-Inform (2019).

**Figure 3.2. Brexit online debate, January to March 2019**




*Source: RAND analysis.*

the socio-cultural purpose of social media talk. These are both gaps in current off-the-shelf solutions, and would be of immediate use to a number of Dstl customers.

### 3.3.1. Exploring the model's portability through deployment in a new context

A feasible next stage in our effort could be aimed at testing our model for detecting malign troll accounts on social media in a new context in order to trial the model's portability.

A relevant choice for this could be the online debate over Brexit, in anticipation of the UK's departure from the EU. As a preliminary step – for proposed further analysis – we collected Brexit Twitter data from 1 January 2019 to 31 March 2019. In order to explore the portability of the model, we conducted an initial network analysis on this Brexit dataset in which we identified two large publics, which we labelled 'Brexiteers' and 'Remainers'. We were also surprised to find a third meta-community of Continental non-English speaker accounts involved in the discourse, engaged primarily with the 'Remainer' meta-community, as shown in Figure 3.2.[166]

As in Figure 2.2, each dot (node) is a community of Twitter users, sized by the number of accounts in the community. The colour of the nodes shows membership within a higher-order 'meta-community': the

166 These international Twitter users retweeted and tweeted at Remainers, with little response back, likely because these were generally non-English tweets or with few English words. This international community does not appear to have any traction within the internal UK debate.

purple-coloured communities are engaged in an internal UK debate over Brexit, and the green ones are non-English speakers that we characterised as part of the broader European argument over Brexit. Each line (edge) indicates interactions between communities, and the thicker (higher-weighted) the edge, the more interactions there are. Each edge is an arrow, but most are so small and tenuous that the point of the arrow is invisible. However, for the largest and most central communities, the interactions are so dense that the point of the arrow is visible, showing directionality.

Given Russia's prior malign information efforts it is likely that, similarly to the 2016 US presidential election, Russian-sponsored trolls were active in the above argument, imitating discourse in both camps and stoking discord and enmity. Our aspiration would be to prove the robustness and utility of our model by applying it to the online Brexit argument, building on the preliminary work described above. In this way, as a next step we hope to further trial the portability of the model developed as part of this project through the lens of a new context such as the online Brexit debate.

# Annex A. Contextualising resilience to malign information among vulnerable publics

Resilience to social media manipulation can connote a wide array of policy responses and instruments employed with the same objective of reducing vulnerability and strengthening capacity to mitigate risk. Though no comprehensive taxonomy of misinformation resilience efforts exists, it is possible to differentiate between those that address the 'supply-side' and 'demand-side' of the information ecosystem. While an intervention to address the 'supply-side' would focus on institutions which *supply* information to the public, such as the media, 'demand-side' initiatives seek to foster resilience among consumers from the bottom up.[167] Both logics play a substantial role in determining the efficacy of responses following the detection of social manipulation. This annex provides additional context for developing resilience-building strategies that integrate social manipulation detection mechanisms.

## A.1. Strengthening institutional capacity for engagement with at-risk communities

Strengthening institutional capacity to engage and communicate with at-risk publics is often determined as a key top-down response to disinformation and social media manipulation. Shortages in institutional capacity are generally easier to identify and define than more structural societal vulnerabilities, which often leads to their prioritisation in this context. Unfortunately, efforts to enhance institutional capacity to provide solutions to social media manipulation have frequently resulted in the so-called 'Band-Aid Effect', with the creation of new bodies and institutional entities adding a layer of complexity and thus preventing a more coherent and coordinated response. While this represents a challenge notably in the European context, some existing initiatives reveal a range of best practices to avoid the 'Band-Aid' scenario.[168] This includes ensuring sufficient vertical and horizontal buy-in from a range of stakeholders including civil society and high-level officials to key policies and objectives, such as through widening the scope of membership and strengthening engagement from key audiences. As organisations also frequently 'suffer from a gap in understanding or expectations from the public or the media', active and transparent outreach should be prioritised to ensure public awareness and

167 Claesson (2019).

168 Jankowitz (2019).

enhance the level of trust between institutions and their audiences.[169]

The case of the EU's newly established EEAS East StratCom Task Force, directed to take on the disinformation portfolio, illustrates several other challenges to resilience-building at the institutional level.[170] Despite wide-ranging support for more comprehensive disinformation efforts throughout the EU, the Task Force suffered staff and resource shortages. A lack of transparency in the Task Force's fact checking methodology also contributed to legal challenges, with several media platforms accusing the Task Force of including legitimate news stories in its disinformation database.[171]

A positive aspect of the EU's response has been a phased understanding of strategic communications, wherein a difference is drawn between 'making sense' of the information at hand (i.e. fact checking), and 'creating meaning' for it through strategic communication and 'persuasive messaging', allowing the publics to understand and assess the narrative at hand.[172] In a similar way, Swedish responses have focused on 'informing and educating [citizens] on the methods used, rather than by passing legislation' to counter disinformation.[173] This highlights that identifying malign information and social media manipulation only constitutes a part of the response, and efforts to strengthen institutional capacity must take into account the need to 'create meaning' for the detected information.

## A.2. Engaging civil society for bottom-up resilience-building

The effectiveness of measures such as strategic communications depends not only on the capacity of actors to issue such communications, but also on the response from the public to the initiative. The success of most responses thus depends not only on the quality of the institutional input, but also on the structural factors determining the response from vulnerable publics. The Latvian experience of efforts to curtail the spread of Russian-sponsored disinformation has highlighted this aspect, with the country's lack of success in top-down regulatory measures directed against pro-Kremlin media. The failure to provide viable alternatives to Russian-speaking communities has meant that restrictions on some media channels merely pushed audiences to other channels following similar narratives.[174] Similarly, top-down institutional responses in Ukraine, including new regulations on the national media and communications infrastructure and the establishment of the Ministry of Information Policy, had to be complemented by bottom-up resilience-building efforts mainly led by the civil society space. [175]

Strengthening civil society capacity and engagement features prominently in many of the identified 'blueprint' approaches to countering malign information. The 'civil society approach' to disinformation broadly emphasises the empowerment of civil society

169 Jankowitz (2019, 2).
170 Scheidt (2019).
171 Jankowitz (2019).
172 Scheidt (2019, 14).
173 Robinson et al. (2019, 4).
174 Robinson el al. (2019).
175 Teperik et al. (2018).

actors to resist malign information activities due to potential biases imposing limits on the credibility of state-sponsored initiatives.[176] Similarly, the 'collaborative approach' emphasises the integration of civil society actors in a wider national and international network of stakeholders. This is intended to 'jointly increase [capacity] to counter information influence activities by, for example, supporting information and experience sharing.'[177]

In this space, users and civil society organisations (CSOs) are generally identified as the key actors in strengthening bottom-up societal resilience. The engagement of 'Elves', online volunteers 'fighting the Russian trolls', has for example demonstrated the potential of proactive civic engagement in the fight against social media manipulation. Other citizen-led monitoring and fact-checking initiatives such as InformNapalm in Ukraine have also been seen as effective early warning systems against disinformation.[178] Due to the above-mentioned credibility challenges associated with government or state-sponsored narratives, user empowerment and civil society engagement is also crucial in the context of strategies such as 'naming and shaming'. 'Naming and shaming' relies on the public release of information regarding the activities of a malign actor as well as their methods. While critics argue that this is little more effective than 'yapping in the wind',[179] it can impose significant costs for sponsors of social media manipulation as well as unify the response through public identification and attribution.[180] This necessarily depends on the robustness of the response, in which the capacity of civil society stakeholders plays a crucial role.

The possibility of mustering a robust response from a broad coalition of civil society and institutional stakeholders at a structural level depends on the cohesion between those groups of stakeholders.[181] Despite the concept's sparse definitional specificity, lack of public trust in the stability and responsiveness of core democratic institutions is often placed at the core of the problem of social media manipulation. Countries such as Belarus are also often found to have a 'weak national identity' which exposes additional vulnerabilities to narratives of historical revisionism frequently incorporated in the use of malign information on social media.[182] Some initiatives have addressed this issue through fostering knowledge of national history as a way of improving social cohesion and strengthening the connection citizens perceive with the state. For example, the Canada History Fund, part of an extensive programme for citizen-focused disinformation resilience initiatives, has focused on '[encouraging] Canadians to improve their knowledge about Canada's history, civics and public policy'.[183] Similarly, in Belarus, efforts to strengthen resilience have turned to a 'soft Belarusianisation' approach 'seeking

176 Pamment et al. (2018).

177 Pamment et al. (2018, 83).

178 Marin (2018, 3); Teperik et al. (2018).

179 Jasper (2018).

180 Gressel (2019).

181 Damarad & Yeliseyeu (2018).

182 Filipec (2019).

183 Government of Canada (2019).

to promote national characteristics, using affirmative action in support of Belarusian culture and language' despite continued repression from the regime.[184]

## A.3. Improving media literacy

The concept most widely associated with disinformation resilience is media literacy. As described in Section 3.2, incorporating media literacy into resilience-building efforts increases the likelihood of users recognising social media manipulation techniques, though the isolated effect of media literacy training is also contested by some experts who point to the need for strengthening the quality of education and critical thinking skills more widely. To the extent that media literacy is promoted across Europe, it has been formalised and institutionalised to varying degrees. In Latvia, for example, media literacy is not institutionally integrated into formal education programmes.[185] In the Czech Republic, recent analysis shows that close to 75 per cent of teachers dealing with media education have never completed media education training. Additionally, close to 10 per cent of teaching staff use disinformation media sources for information in their teaching, with close to 30 per cent generally trusting such sources.[186]

A crucial component of media literacy is therefore fostering the ability of students, educators and all citizens to recognise disinformation and manipulation on social media. Existing evaluations indicate that the efficiency of media literacy efforts is amplified particularly when initiatives are targeted at a broader spectrum of civil society and are not limited to students and youth.[187] In Moldova, for example, initiatives focused on raising awareness about Russian disinformation have integrated 'Western-based diaspora as a spearhead for change' and 'a critical amplifier for progressive thinking'.[188] In countries in which Russian-sponsored disinformation is a particularly strong threat, such efforts are also part of broader initiatives to enhance civil cohesion through the ability of citizens to recognise Russian influence and interference in public institutions.[189]

The media, as well as social media platforms, plays an important role in this with its ability to influence or amplify the response to any given narrative. As observed in various contexts, the challenge is to avoid the media becoming 'unwitting multipliers of misleading information'.[190] Initiatives focused on media literacy should therefore seek to incentivise engagement with journalists and other stakeholders from within the media ecosystem, which serves the above-mentioned goal of creating broader and stronger stakeholder coalitions through vertical and horizontal engagement.

## A.4. Considerations

The detection and identification of social media manipulation remains an important enabling factor for countering the effects of malign

184 Boulègue et al. (2018, 3).
185 Damarad & Yeliseyeu (2018).
186 Filipec (2019).
187 Claesson (2019).
188 Boulègue et al. (2018, 3, 35).
189 Boulègue et al. (2018).
190 Helmus et al. (2018, 80).

information. It is safe to assume that the more concretely a threat is defined, the more able a society is to cope, adapt and transform in order to become more resilient. Denying malign actors the ability to remain undetected and act without responsibility furthermore serves to undermine the principle of plausible deniability as a key determinant of the effectiveness of disinformation.[191] However, it is also clear that the process of strengthening such capacities necessitates not only effective detection, but also proactive engagement with at-risk publics and the creation of holistic resilience-building frameworks.

The examination of existing approaches and relevant case studies reveals a number of best practices that should be incorporated into a resilience-building strategy. Firstly, institutional actors should be able to map relevant vulnerabilities that social media manipulation techniques can exploit among at-risk communities. Identification tools should therefore be integrated with corresponding responses to such factors in a holistic and proactive manner. Secondly, the robustness of resilience-building strategies depends on the vertical and horizontal integration of a wide coalition of relevant stakeholders. While user empowerment and civil society engagement is crucial, a lack of support from policy-makers and political elites playing key roles in the national discourse can have an equally detrimental effect. Any initiative should thus start with the identification of relevant stakeholders. Overall, resilience-building approaches should also be tailored to the needs and experiences of individual at-risk communities. This can be reflected not only in the nature of the identified response or initiative, but also in the manner in which specific responses and initiatives are formulated, ensuring sufficient community buy-in. Lastly, all stages of resilience-building should be characterised by transparent and open communications between relevant stakeholders to prevent and mitigate credibility challenges and advance the process of trust-building to reduce corresponding vulnerabilities from malign social media manipulation.

191 Bjola & Pamment (2016).

# References


Abdi, Hervé, & Lynne J, Williams. 2010. 'Tukey's honestly significant difference (HSD) test.' In *Encyclopedia of Research Design*, edited by Neil Salkind. Thousand Oaks, CA: Sage.

Airoldi, Edoardo, Annelise Anderson, Stephen Fienberg, & Kiron Skinner. 2006. 'Who wrote Ronald Reagan's radio addresses?' *Bayesian Analysis,* 2, 289-320.

---. 2007. 'Whose ideas? Whose words? Authorship of the Ronald Reagan radio addresses', *Political Science & Politics,* 40, 501–506.

Alaphilippe, Alexandre, Chiara Ceccarelli, Léa Charlet & Martin Mycielski. 2018. 'Developing a disinformation detection system and sourcing it live. The case study of the 2018 Italian elections.' EU Disinfo Lab, 17 May. As of 9 June 2019: https://www.disinfo.eu/wp-content/uploads/2018/05/20180517_Scientific_paper.pdf

Allouch, Nada. 2018. 'Sentiment and Emotional Analysis: The Absolute Difference.' Emojics Blog, 21 May. As of 28 October 2019: http://blog.emojics.com/emotional-analysis-vs-sentiment-analysis/

Bai, Xue. 2011. 'Predicting consumer sentiments from online text', *Decision Support Systems* 50: 732–742.

Balomenos, Themis, Amaryllis Raouzaiou, Spiros Ioannou, Athanasios Drosopoulos, Kostas Karpouzis, & Stefanos Kollias. 2005. 'Emotion Analysis in Man-Machine Interaction Systems.' In *Machine Learning for Multimodal Interaction*, edited by Samy Bengio & Hervé Bourlard, 318–28. Berlin: Springer-Verlag.

Barbier, Geoffrey, & Huan Liu. 2011. 'Data mining in social media.' In *Social Network Data Analytics*, edited by Charu C. Aggarwal, 327–52. Springer Science+Business Media. As of 9 June 2019: https://pdfs.semanticscholar.org/8a60/b082aa758c317e9677beed7e7776acde5e4c.pdf

Bay, Sebastian, & Nora Biteniece. 2019. *The Current Digital Arena and its Risks to Serving Military Personnel*. NATO Stratcom. As of 19 June 2019: https://stratcomcoe.org/current-digital-arena-and-its-risks-serving-military-personnel

Biber, Douglas, & Edward Finegan. 1991. 'On the exploitation of computerized corpora in variation studies.' In *English Corpus Linguistics: Studies in Honour of Jan Svartvik*, edited by Karin Aijmer & Bengt Altenberg, 204–20. Harlow: Longman.

Bienkov, Adam. 2012. 'Astroturfing: what is it and why does it matter?' *Guardian*, 8 February. As of 19 June 2019: https://www.theguardian.com/commentisfree/2012/feb/08/what-is-astroturfing

Bjola, Corneliu, & James Pamment. 2016. 'Digital Containment: Revising Containment Strategy in the Digital Age.' *Global Affairs* 2(2):131–42. As of 17 October 2019: https://www.tandfonline.com/doi/abs/10.1080/23340460.2016.1182244?journalCode=rgaf20

Bloom, Mia, Hicham Tiflati & John Horgan. 2017. 'Navigating ISIS's Preferred Platform: Telegram.' *Terrorism and Political Violence* 31(6):1242–54. As of 3 January 2018: http://www.tandfonline.com/doi/pdf/10.1080/09546553.2017.1339695?needAccess=true

Boatwright, Brandon C., Darren L. Linvill & Patrick L. Warren. 2018. 'Troll factories: The internet research agency and state-sponsored agenda building.' Resource Centre on Media Freedom in Europe. As of 22 November 2019: https://www.rcmediafreedom.eu/Publications/Academic-sources/Troll-Factories-The-Internet-Research-Agency-and-State-Sponsored-Agenda-Building

Boulègue, Mathieu, Orysia Lutsevych & Anais Marin. 2018. *Civil Society Under Russia's Threat: Building Resilience in Ukraine, Belarus, and Moldova*. London: Chatham House. As of 17 October 2019: https://www.chathamhouse.org/publication/civil-society-under-russias-threat-building-resilience-ukraine-belarus-and-moldova

Boyd, Ryan L., Alexander Spangher, Adam Fourney, Besmira Nushi, Gireeja Ranade, James Pennebaker & Eric Horvitz. 2018. 'Characterizing the Internet Research Agency's Social Media Operations During the 2016 US Presidential Election using Linguistic Analyses.' As of 9 June 2019: https://www.davidpuente.it/blog/wp-content/uploads/2018/08/Linvill_Warren_TrollFactory.pdf

Bradshaw, Samantha, & Philip N. Howard. 2018. *Challenging Truth and Trust: A Global Inventory of Organized Social Media Manipulation*. Oxford: Oxford Internet Institute. As of 19 June 2019: https://blogs.oii.ox.ac.uk/wp-content/uploads/sites/93/2018/07/ct2018.pdf

Bradshaw, Samantha, Lisa-Maria Neudert & Philip N. Howard. 2018. *Government Responses to Malicious Use of Social Media*. NATO Stratcom. As of 19 June 2019: https://www.stratcomcoe.org/government-responses-malicious-use-social-media

Brooking, Emerson. T., & P.W. Singer. 2016. 'War Goes Viral.' The Atlantic, November. As of 6 June 2019: https://www.theatlantic.com/magazine/archive/2016/11/war-goes-viral/501125/

Chen, Yimin, Niall J. Conroy & Victoria L. Rubin. 2015. 'Misleading Online Content: Recognizing Clickbait as "False News".' WMDD '15: Proceedings of the 2015 ACM on Workshop on Multimodal Deception Detection. As of 28 November 2019: https://dl.acm.org/citation.cfm?doid=2823465.2823467

CITS (Center for Information Technology & Society). 2019. 'A citizen's guide to fake news.' Cits.usb.edu. As of 28 November 2019: https://www.cits.ucsb.edu/fake-news/what-is-fake-news

Claesson, Annina. 2019. *Coming together to fight fake news: Lessons from the European Union approach to disinformation*. London: CSIS Europe Program. As of 17 October 2019: https://www.csis.org/coming-together-fight-fake-news-lessons-european-approach-disinformation

Clauset, Aaron, M. E.J. Newman & Cristopher Moore. 2004. 'Finding Community Structure in Very Large Networks.' *Physical Review E* 70(6):1–6. As of 22 November 2019: http://ece-research.unm.edu/ifis/papers/community-moore.pdf

Co-Inform (homepage). 2019. As of 17 October 2019: https://coinform.eu/about/the-project/

Collins, J. 2003. *Variations in Written English*. As of 20 June 2019: http://www.dtic.mil/cgi-bin/GetTRDoc?AD=ADA416543

Crimson Hexagon. 2019. 'Twitter: About Crimson Hexagon.' As of 20 June 2019: https://twitter.com/crimsonhexagon?lang=en

Cox, Kate, William Marcellino, Jacopo Bellasio, Antonia Ward, Katerina Galai, Sofia Meranto & Giacomo Persi Paoli. 2018. *Social Media in Africa: A double-edged sword for security and development*. Santa Monica, Calif.: RAND Corporation. As of 19 June 2019: http://www.africa.undp.org/content/dam/rba/docs/Reports/UNDP-RAND-Social-Media-Africa-Research-Report_final_3%20Oct.pdf

Daily Beast. 2017. 'Study: Russian Propaganda Targeted Engaged U.S. Vets, Active-Duty Troops on Social Media.' Daily Beast, 9 October. As of 7 June 2019: https://www.thedailybeast.com/study-russian-propaganda-targeted-engaged-us-vets-active-duty-troops-on-social-media

Damarad, Volha. 2017. Building governmental resilience to information threats: The case of Ukraine. *Polish Quarterly of International Relations* 26(3). As of 17 October 2019: https://www.questia.com/library/journal/1P4-1999053303/building-governmental-resilience-to-information-threats

Damarad, Volha, & Andrei Yeliseyeu. 2018. *Disinformation Resilience in Central and Eastern Europe*. Kyiv: Eurasian States in Transition Research Centre. As of 17 October 2019: http://prismua.org/en/dri-cee/

Davies, Mark. 2008. *The Corpus of Contemporary American English (COCA): 560 million words, 1990–present*. As of 17 December 2019: https://www.english-corpora.org/coca/

Davis, Susan. 2018. *Russian Meddling in Elections and Referenda in the Alliance: General Report.* Brussels: NATO Parliamentary Assembly. As of 6 June 2019: https://www.nato-pa.int/download-file?filename=sites/default/files/2018-11/181%20STC%2018%20E%20fin%20-%20RUSSIAN%20MEDDLING%20-%20DAVIS%20REPORT.pdf

Digital Forensic Research Lab. 2016. 'Human, Bot or Cyborg? Three clues that can tell you if a Twitter user is fake.' Medium, 23 December. As of 19 June 2019: https://medium.com/@DFRLab/human-bot-or-cyborg-41273cdb1e17

EISC (European Integration Studies Centre). 2018. *Lithuanian-Swedish Roundtable Expert Discussions on Societal Resilience and Psychological Defence*. EISC Policy Brief, September 2018. As of 17 October 2019: http://www.eisc.lt/uploads/documents/files/EISC_policy%20brief(1).pdf

Elangovan, D., V. Subedha, R. Sathishkumar & V.D. Ambeth kumar. 2018. 'A Survey: data mining techniques for social media analysis.' *Advances in Engineering Research* 142:109–15. As of 19 June 2019: https://download.atlantis-press.com/article/25893606.pdf

el Hjouji, Zakaria, David Scott Hunter, Nicolas Guenon des Mesnards & Tauhid Zaman. 2018. 'The Impact of Bots on Opinions in Social Networks.' As of 19 June 2019: https://arxiv.org/pdf/1810.12398.pdf

EPRS (European Parliament Research Service). 2019. *Automated Tackling of Disinformation*. PE 624.278. Brussels: EPRS. As of 17 October 2019: http://www.europarl.europa.eu/RegData/etudes/STUD/2019/624278/EPRS_STU(2019)624278_EN.pdf

European Commission. 2016. 'European Commission and IT Companies Announce Code of Conduct on Illegal Online Hate Speech.' Press release, 31 May. As of 19 June 2019: http://europa.eu/rapid/press-release_IP-16-1937_en.htm

Expert System. 2017. 'Tweeting Brexit: What cognitive software revealed in analysis of Brexit tweets.' Expert System, 6 April. As of 10 June 2019: https://www.expertsystem.com/tweeting-brexit-cognitive-software-revealed-analysis-brexit-tweets/

FactBar EDU. 2018. *Elections approach – are you ready? Fact-checking for educators and future voters*. As of 19 June 2019: https://www.faktabaari.fi/assets/FactBar_EDU_Fact-checking_for_educators_and_future_voters_13112018.pdf

Ferrara, Emilio. 2015. 'Manipulation and abuse on social media.' ACM SIGWEB Newsletter, Spring.

Filipec, Ondrej. 2019. 'Towards a disinformation resilient society? The experience of the Czech Republic.' *Cosmopolitan Civil Societies* 11(1). As of 22 November 2019: https://epress.lib.uts.edu.au/journals/index.php/mcs/article/view/6065/7126

Funke, Daniel & Daniela Flamini. 2019. 'A guide to anti-misinformation actions around the world.' Poynter. As of 21 June 2019: https://www.poynter.org/ifcn/anti-misinformation-actions/

Gallacher, John D., Vlad Barash, Phillip N. Howard & John Kelley. 2017. 'Junk News on Military Affairs and National Security: Social Media Disinformation Campaigns Against US Military Personnel and Veterans.' COMPROP Data Memo 2017.9. As of 6 June 2019: http://blogs.oii.ox.ac.uk/comprop/wp-content/uploads/sites/93/2017/10/Junk-News-on-Military-Affairs-and-National-Security-1.pdf

Github. 2018. 'fivethirtyeight / russian-troll-tweets.' As of 19 June 2019: https://github.com/fivethirtyeight/russian-troll-tweets

Government of Canada. 2019. 'Helping Citizens Critically Assess and Become Resilience Against Harmful Online Disinformation.' Canadian Heritage, Canada.ca, 2 July. As of 17 October 2019: https://www.canada.ca/en/canadian-heritage/news/2019/07/helping-citizens-critically-assess-and-become-resilient-against-harmful-online-disinformation.html

Gressel, Gustav. 2019. 'Protecting Europe Against Hybrid Threats.' European Council on Foreign Relations, 25 June. As of 17 October 2019: https://www.ecfr.eu/publications/summary/protecting_europe_against_hybrid_threats

Hauser, Gerard A. 2002. *Introduction to rhetorical theory*. Long Grove: Waveland Press.

Helberg, Alex, Maria Poznahovska, Suguru Ishizaki, David Kaufer, Necia Werner & Danielle Wetzel. 2018. 'Teaching textual awareness with DocuScope: Using corpus-driven tools and reflection to support students' written decision-making.' *Assessing Writing* 38(1):40–45. As of 22 November 2019: https://www.sciencedirect.com/science/article/pii/S107529351830062X

Helmus, Todd C., Elizabeth Bodine-Baron, Andrew Radin, Madeline Magnuson, Joshua Mendelsohn, William Marcellino, Andriy Bega & Zev Winkelman. 2018. *Russian Social Media Influence: Understanding Russian Propaganda in Eastern Europe*. Santa Monica, Calif.: RAND Corporation. RR-2237-OSD. As of 19 June 2019: https://www.rand.org/pubs/research_reports/RR2237.html

Hope, Jonathan & Michael Witmore. 2010. 'The hundredth psalm to the tune of "Green Sleeves": Digital approaches Shakespeare's language of genre', *Shakespeare Quarterly* 61.

Horev, Rani. 2018. 'BERT Explained: State of the art language model for NLP.' Towarddatascience.com, 10 November. As of 31 October 2019: https://towardsdatascience.com/bert-explained-state-of-the-art-language-model-for-nlp-f8b21a9b6270

Howard, Philip N., & Bence Kollanyi. 2016. *Bots, #Strongerin, and #Brexit: Computational Propaganda during the UK-EU Referendum*. Working Paper 2016.1. Oxford, UK: Project on Computational Propaganda. As of 22 November 2019: https://comprop.oii.ox.ac.uk/research/working-papers/bots-strongerin-and-brexit-computational-propaganda-during-the-uk-eu-referendum/

Hu, Yongmei, David Kaufer & Suguru Ishizaki. 2011. 'Genre and instinct.' In *Computing with instinct*, pp. 58-81. Springer Berlin Heidelberg.

Huguet, Alice, Jennifer Kavanagh, Garrett Baker & Marjory S. Blumenthal. 2019. *Exploring Media Literacy Education as a Tool for Mitigating Truth Decay*. Santa Monica, Calif.: RAND Corporation. RR-3050-RC. As of 17 October 2019: https://www.rand.org/pubs/research_reports/RR3050.html

Hybrid CoE (homepage). 2019. As of 19 June 2019: https://www.hybridcoe.fi/

IATE. 2019. 'Misinformation.' As of 14 June 2019: https://iate.europa.eu/search/standard/result/1560521282914/1

IHS Markit. 2018. 'Fact of fallacy? Fake news and the military.' 28 November. As of 6 June 2019: https://ihsmarkit.com/research-analysis/fake-news-and-military.html

Irving, Doug. 2017. 'Big Data, Big Questions.' RAND Blog, 16 October. As of 16 June 2019: https://www.rand.org/blog/rand-review/2017/10/big-data-big-questions.html

Jankowicz, Nina. 2019. *Avoiding the Band-Aid Effect in Institutional Responses to Disinformation and Hybrid Threats*. The German Marshall Fund of the United States, August 2019. As of 17 October 2019: http://www.gmfus.org/publications/avoiding-band-aid-effect-institutional-responses-disinformation-and-hybrid-threats

Jasper, Scott. 2018. *U.S. Strategic Cyber Deterrence Options*. Dissertation, Doctor of Philosophy in Politics, University of Reading. As of 17 October 2019: http://centaur.reading.ac.uk/79976/1/22839264_Jasper_thesis.pdf

Jigsaw Research. 2018. *News Consumption in the UK: 2018*. As of 3 June 2019: https://www.ofcom.org.uk/__data/assets/pdf_file/0024/116529/news-consumption-2018.pdf

Kaufer, David S., & Brian S. Butler. 2010. *Rhetoric and the Arts of Design*. London: Routledge.

Kaufer, David, Cheryl Geisler, Pantelis Vlachos & Suguru Ishizaki. 2006. 'Mining textual knowledge for writing education and research: The DocuScope project.' In *Writing and Digital Media,* edited by Luuk van Waes, Mariëlle Leijten & Christophe Neuwirth, 115–29. Amsterdam: Elsevier.

Kavanagh, Jennifer, William Marcellino, Jonathan S. Blake, Shawn Smith, Steven Davenport & Mahlet G. Tebeka. 2019. *News in a Digital Age: Comparing the Presentation of News Information over Time and Across Media Platforms*. Santa Monica, Calif.: RAND Corporation, RR-2960-RC. As of 27 October 2019: https://www.rand.org/pubs/research_reports/RR2960.html

Keck, Markus & Patrick Sakdapolrak. 2013. 'What is Societal Resilience? Lessons Learned and Ways Forward.' *Erdkunde* 67(1):5–19. As of 17 October 2019: http://transre.uni-bonn.de/files/9414/1449/3358/Keck_Etzold_-_2013_-_What_is_social_resilience_Lessons_learned_and_ways_forward.pdf

Klausen, Jytte, Christopher E. Marks & Tauhid Zaman. 2018. 'Finding Extremists in Online Social Networks.' *Operations Research* 66(4). As of 22 November 2019: https://pubsonline.informs.org/doi/10.1287/opre.2018.1719

Krelim Watch. 2019. 'Finland.' As of 19 June 2019: https://www.kremlinwatch.eu/countries-compared-states/finland/

Kumar, K.P., & Gopalan Geethakumari. 2014. 'Detecting misinformation in online social networks using cognitive psychology.' *Human-centric Computing and Information Sciences* 4(14). As of 9 June 2019: https://hcis-journal.springeropen.com/articles/10.1186/s13673-014-0014-x

Liaw, Andy, & Matthew Wiener. 2002. 'Classification and regression by randomForest.' *R news* 2(3):18–22. As of 22 November 2019: https://www.r-project.org/doc/Rnews/Rnews_2002-3.pdf

Lozares, Carlos, Joan Miquel Verd, Irene Cruz & Oriol Barranco. 2014. 'Homophily and heterophily in personal networks. From mutual acquaintance to relationship intensity.' *Quality & Quantity* 48(5):2657–70. As of 29 October 2019: https://link.springer.com/article/10.1007/s11135-013-9915-4

Mackintosh, Eliza. 2019. 'Finland is winning the war on fake news. What it's learned may be crucial to Western democracy.' CNN. As of 19 June 2019: https://edition.cnn.com/interactive/2019/05/europe/finland-fake-news-intl/

Maitra, Promita, Souvick Ghosh & Dipankar Das. 2016. 'Authorship Verification-An Approach based on Random Forest.' As of 22 November 2019: https://www.researchgate.net/publication/305735928_Authorship_Verification_-_An_Approach_based_on_Random_Forest

Manning, Christopher, Prabhakar Raghavan & Hinrich Schutze. 2009. 'Boolean retrieval'. In *An Introduction to Information Retrieval*. Cambridge: Cambridge University Press. As of 20 June 2019: https://nlp.stanford.edu/IR-book/pdf/01bool.pdf

Marcellino, William M. 2014. 'Talk like a Marine: USMC linguistic acculturation and civil–military argument.' *Discourse Studies* 16(3):385–405. As of 22 November 2019: https://journals.sagepub.com/doi/full/10.1177/1461445613508895

Marcellino, William M. 2015. 'Revisioning Strategic Communication Through Rhetoric & Discourse Analysis.' *Joint Forces Quarterly* 76.

Marchal, Nahema, Lisa-Maria Neudert, Bence Kollanyi & Philip N. Howard. 2018. 'Polarization, Partisanship and Junk News Consumption on Social Media During the 2018 US Midterm Elections.' COMPROP Data Memo 2018.5. As of 19 June 2019: https://comprop.oii.ox.ac.uk/research/midterms2018/

Marin, Daru. 2018. Cultivating information resilience in Moldova's media sector. *Media Forward Policy Brief* 4, April. As of 17 October 2019: https://freedomhouse.org/sites/default/files/inline_images/Moldova_Policy_Brief_Information_Resilience_ENG_2.pdf

Marwick, Alice, & Rebecca Lewis. 2017. 'Media Manipulation and Disinformation Online.' New York: Data & Society Research Institute. As of 19 June 2019: https://datasociety.net/pubs/oh/DataAndSociety_MediaManipulationAndDisinformationOnline.pdf

Mazarr, Michael J., Abigail Casey, Alyssa Demus, Scott W. Harold, Luke J. Matthews, Nathan Beauchamp-Mustafaga, James Sladden. 2019. *Hostile Social Manipulation: Present Social Realities and Emerging Trends*. Santa Monica, Calif.: RAND Corporation. RR-2713-OSD. As of 17 October 2019: https://www.rand.org/pubs/research_reports/RR2713.html

McAfee. 2013. 'Social Media Manipulation is for Real, Some Call it as Crowd-Turfing!' As of 19 June 2019: https://securingtomorrow.mcafee.com/consumer/identity-protection/social-media-manipulation-is-for-real-some-call-it-as-crowd-turfing/

Merriam Webster. 2019. 'Clickbait.' As of 29 October 2019: https://www.merriam-webster.com/dictionary/clickbait

Mihaylov, Todor, Georgi D. Georgiev & Preslav Nakov. 2015. 'Finding Opinion Manipulation Trolls in News Community Forums.' Proceedings of the 19th Conference on Computational Language Learning, Beijing, China, 30–31 July, 310–14.

Mikros, George K. 2002. 'Quantitative parameters in corpus design: Estimating the optimum text size in Modern Greek language.' In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02),* edited by Manuel González Rodríguez & Carmen Paz Suarez Araujo, 834–38. As of 19 June 2019: https://aclweb.org/anthology/papers/L/L02/L02-1099/

Molnar, Christoph. 2019. 'Interpretable Machine Learning: A Guide for Making Black Box Models Explainable.' Christophm.github.io, 28 October 2019. As of 29 October 2019: https://christophm.github.io/interpretable-ml-book/

Monkey Learn. 2019. 'Sentiment Analysis.' As of 16 June 2019: https://monkeylearn.com/sentiment-analysis/

Mueller-Hsia, Kaylana. 2019. 'WMD disinformation campaigns: major events in the 20th and 21st century.' Medium, 3 October. As of 7 June 2019: https://medium.com/@Tech4GS/wmd-disinformation-campaigns-major-events-in-the-20th-and-21st-century-96db9e6785b4

Murrock, Erin, Joy Amulya, Mehri Druckman & Tetiana Liubyva. 2018. *Winning the war on state-sponsored propaganda*. Washington, DC: International Research and Exchanges Board (IREX). As of 17 October 2019: https://www.irex.org/sites/default/files/node/resource/impact-study-media-literacy-ukraine.pdf

NATO Stratcom. 2019. 'Robotrolling 1/2019.' As of 19 June 2019: https://stratcomcoe.org/robotrolling-20191

Neudert, Lisa-Maria. 2018. 'Future elections may be swayed by intelligent, weaponized chatbots.' MIT Technology Review, 22 August. As of 19 June 2019: https://www.technologyreview.com/s/611832/future-elections-may-be-swayed-by-intelligent-weaponized-chatbots/

Nielsen, Rasmus Kleis. 2017. 'Where do people get their news? The British media landscape in 5 charts.' Medium.com, May 30**.** As of 19 June 2019:
https://medium.com/oxford-university/where-do-people-get-their-news-8e850a0dea03

Nyabola, Nanjala. 2017. 'Media Perspectives: Social Media and New Narratives: Kenyans Tweet Back.' In *Africa's Media Image in the 21st Century: From the 'Heart of Darkness' to 'Africa Rising'*, edited by Mel Bunce, Suzanne Franks & Chris Paterson. London and New York: Routledge.

Oshikawa, Ray, Jing Qian & William Yang Wang. 2018. 'A Survey on Natural Language Processing for Fake News Detection.' As of 9 June 2019:
https://arxiv.org/pdf/1811.00770v1.pdf

Pamment, James, Noward Nothhaft, Henrik Agardt-Twetman & Alicia Fjallhed. 2018. *Countering Information Influence Activities: The State of the Art.* Stockholm: MSB (Swedish Civil Contingencies Agency) and Lund University.

Pennebaker, James W., Martha E, Francis & Roger J. Booth. 2001. *Linguistic inquiry and word count: LIWC 2001*. Mahwah: Lawrence Erlbaum Associates.

Poynter. 2019. 'Mission & Vision.' As of 19 June 2019: https://www.poynter.org/mission-vision/

Rehak, David, & Martin Hromada. 2018. 'Failures in a Critical Infrastructure System.' In *System of System Failures*, edited by Takafumi Nakamura. As of 17 October 2019:
https://www.intechopen.com/books/system-of-system-failures

Reuters. 2017. 'Russian Twitter accounts promoted Brexit ahead of EU referendum: Times newspaper.' As of 19 June 2019:
https://www.reuters.com/article/us-britain-eu-russia/russian-twitter-accounts-promoted-brexit-ahead-of-eu-referendum-times-newspaper-idUSKBN1DF0ZR

Robinson, Olga. 2018. 'Malicious Use of Social Media: Case Studies.' BBC Monitoring/NATO Stratcom. As of 19 June 2019:
https://www.stratcomcoe.org/download/file/fid/79730

Robinson, Olga, Alistair Coleman & Shayan Sardarizadeh. 2019. *A Report of Anti-Disinformation Initiatives*. Oxford: Oxford Technology & Elections Commission. As of 17 October 2019:
https://comprop.oii.ox.ac.uk/wp-content/uploads/sites/93/2019/08/A-Report-of-Anti-Disinformation-Initiatives.pdf

Ruge, T.M.S. 2013. 'How the African Diaspora is Using Social Media to Influence Development.' *Guardian*, 6 February. As of 21 June 2019:
https://www.theguardian.com/global-development-professionals-network/2013/feb/06/african-diaspora-social-media-tms-ruge

Saif, Hassan, Miriam Fernández, Yulan He & Harith Alani. 2013. 'Evaluation datasets for Twitter sentiment analysis: a survey and a new dataset, the STS-Gold.' 1st International Workshop on Emotion and Sentiment in Social and Expressive Media: Approaches and Perspectives from AI (ESSEM 2013), 3 December 2013, Turin, Italy. As of 10 June 2019:
http://oro.open.ac.uk/40660/1/paper1.pdf

Scheidt, Melanie. 2019. *The European Union versus External Disinformation Campaigns in the Midst of Information Warfare: Ready for the Battle?* Bruges: College of Europe. As of 22 November 2019:
https://www.coleurope.eu/system/files_force/research-paper/edp_1_2019_scheidt.pdf?download=1

Shao, Chengcheng, Giovanni Luca Ciampaglia, Onur Varol, Alessandro Flammini & Filippo Menczer. 2017. 'The spread of fake news by social bots.' Bloomington: Indiana University. As of 19 June 2019:
https://www.researchgate.net/publication/318671211_The_spread_of_fake_news_by_social_bots

Shu, Kai, Amy Silva, Suhang Wang, Jiliang Tang & Huan Liu. 2017. 'Fake News Detection on Social Media: A Data Mining Perspective.' *ACM SIGKDD Explorations Newsletter* 19(1):22–36. As of 22 November 2019: http://delivery.acm.org/10.1145/3140000/3137600/p22-shu.pdf?ip=130.154.51.250&id=3137600&acc=ACTIVE%20SERVICE&key=D0E502E9DB58724B%2ED0E502E9DB58724B%2E4D4702B0C3E38B35%2E4D4702B0C3E38B35&__acm__=1574437822_eefa2eba66d50edd43420a4520c64505

Stewart, James, & Maurice Dawson. 2018. 'How the Modification of Personality Traits Leave One Vulnerable to Manipulation in Social Engineering.' *Int. J. Information Privacy, Security and Integrity* 3(3):187–208.

Tanbih (homepage). 2019. As of 17 October 2019: http://tanbih.qcri.org/

Technopedia. 2019a. 'Social network analysis (SNA).' As of 10 June 2019: https://www.techopedia.com/definition/3205/social-network-analysis-sna

Technopedia. 2019b. 'Sentiment analysis.' As of 19 June 2019: https://www.techopedia.com/definition/29695/sentiment-analysis

Technopedia. 2019c. 'Velocity.' As of 29 October 2019: https://www.techopedia.com/definition/16762/velocity-big-data

Technopedia. 2019d. 'Deepfake.' As of 26 November 2019: https://www.techopedia.com/definition/33835/deepfake

Teperik, Dmitri, Tomas Jermalavicius, Grigori Senkiv, Dmytro Dubov, Yevhen Onyshchuk, Oleh Pokalchuk & Mykhailo Samus. 2018. *A route to national resilience: Building a whole-of-society security in Ukraine*. Tallinn: International Centre for Defence and Security. As of 17 October 2019: https://euagenda.eu/upload/publications/untitled-205948-ea.pdf

Thamm, Marianne. 2018. 'UK report reveals how social media is used to undermine global democracy.' Daily Maverick, 2 August. As of 19 June 2019: https://www.dailymaverick.co.za/article/2018-08-02-uk-report-reveals-how-social-media-is-used-to-undermine-global-democracy/

Theohary, Catherine A. 2018. 'Defense Primer: Information Operations.' Congressional Research Service, 18 December. As of 14 June 2019: https://fas.org/sgp/crs/natsec/IF10771.pdf

Tsikerdekis, Michail, & Sherali Zeadally. 2014. 'Online deception in social media.' *Communications of the ACT* 57(9). As of 22 November 2019: https://uknowledge.uky.edu/cgi/viewcontent.cgi?article=1013&context=slis_facpub

Tucker, Joshua A., Andrew Guess, Pablo Barberá, Cristian Vaccari, Alexandra Siegel, Sergey Sanovich, Denis Stukal & Brendan Nyhan. 2018. 'Social media, political polarization and political disinformation: a review of the scientific literature.' As of 10 June 2019: https://www.hewlett.org/wp-content/uploads/2018/03/Social-Media-Political-Polarization-and-Political-Disinformation-Literature-Review.pdf

UK Government Communication Service. 2019. 'Rapid Response Unit: A Year in Digital Trends.' Civilservice.gov.uk, 22 January. As of 14 June 2019: https://gcs.civilservice.gov.uk/rapid-response-unit-a-year-in-digital-trends/

UNDRR (United Nations Office for Disaster Risk Reduction). 2017. 'Terminology on Disaster Risk Reduction.' unisdr.org. As of 17 October 2019: https://www.unisdr.org/we/inform/terminology

University of Edinburgh. 2019. 'Neuropolitics research #ImagineEurope Twitter Demo.' As of 9 June 2019: http://www.pol.ed.ac.uk/neuropoliticsresearch/sections/remote_content?sq_content_src=%2BdXJsPWh0dHAlM0ElMkYlMkYxMjkuMjE1LjE4NC4xNTQlMkZjZ2ktYmluJTJGdHdpdHRlcl9hbmFseXNpcy5jZ2kmYWxsPTE%3D

University of Leiden. 2019. 'Disinformation as cybersecurity threat: value considerations and ethical issues.' As of 14 June 2019: https://www.universiteitleiden.nl/en/events/2019/06/disinformation-as-cybersecurity-threat-value-considerations-and-ethical-issues

University of Oxford. 2018. 'Social media manipulation rising globally, new report wars.' Phys.org, 20 July. As of 7 June 2019: https://phys.org/news/2018-07-social-media-globally.html

US House of Representatives Permanent Select Committee on Intelligence. 2018. 'Exposing Russia's Effort to Sow Discord Online: The Internet Research Agency and Advertisements.' intelligence.house.gov. As of 31 October 2019: https://intelligence.house.gov/social-media-content/

Van Oudenaren, John. 2019. 'Contested Waters: Great Power Naval Competition in the 21st Century.' National Interest, 4 February. As of 14 June 2019: https://nationalinterest.org/feature/contested-waters-great-power-naval-competition-21st-century-43247

Varol, Onur, Emilio Ferrara, Filippo Menczer & Alessandro Flammini. 2017. 'Early Detection of Promoted Campaigns on Social Media.' *EPJ Data Science* 6(13). As of 19 June 2019: https://epjdatascience.springeropen.com/articles/10.1140/epjds/s13688-017-0111-y

von Ahn, Luis, Manuel Blum, Nicholas J. Hopper & John Langford. 2003. 'Captcha: Using hard AI problems for security.' In *Advances in Cryptology – Proceedings of EUROCRYPT 2003: International Conference on the Theory and Applications of Cryptographic Techniques*, edited by Eli Biham, 294–311. New York: Springer.

Wardle, Claire, & Hossein Derakhshan. 2017. *Information Disorder: Toward an interdisciplinary framework for research and policy making*. Brussels: Council of Europe. As of 19 June 2019: https://rm.coe.int/information-disorder-toward-an-interdisciplinary-framework-for-researc/168076277c

Williams, Malcolm V., Danielle M. Varda, Ryan Andrew Brown, Courtney Armstrong, Praise O. Iyiewuare, Rachel Ross, Kayleigh Newman & Sara Sprong. 2018. *A Social Network Analysis and Qualitative Assessment of Partnerships for Million Hearts®*. Santa Monica, Calif.: RAND Corporation. RR-1875-ASPE. As of 10 June 2019: https://www.rand.org/pubs/research_reports/RR1875.html

Williams, Zoe. 2012. 'What is an internet troll?' *Guardian*, 12 June. As of 19 June 2019: https://www.theguardian.com/technology/2012/jun/12/what-is-an-internet-troll

World Economic Forum. 2018. 'Digital Wildfires.' As of 14 June 2019: http://reports.weforum.org/global-risks-2018/digital-wildfires/

Zaman, Tauhid. 2018. 'Even a few bots can shift public opinion in big ways.' The Conversation, 5 November. As of 19 June 2019: http://theconversation.com/even-a-few-bots-can-shift-public-opinion-in-big-ways-104377

Zeitzoff, Thomas. 2017. 'How Social Media is Changing Conflict.' *Journal of Conflict Resolution* 61(9). As of 22 November 2019: https://journals.sagepub.com/doi/pdf/10.1177/0022002717721392

Zellers, Rowan, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner & Yejin Choi. 2019. 'Defending Against Neural Fake News.' As of 31 October 2019: https://arxiv.org/pdf/1905.12616.pdf