City Research Online

City, University of London Institutional Repository

Citation: Farkas, J. & Bastos, M. T. (2018). IRA Propaganda on Twitter: Stoking Antagonism and Tweeting Local News. Proceedings of the 9th International Conference on Social Media & Society, pp. 281-285. doi: 10.1145/3217804.3217929

This is the accepted version of the paper.

This version of the publication may differ from the final published version.

Permanent repository link: https://openaccess.city.ac.uk/id/eprint/19401/

Link to published version: https://doi.org/10.1145/3217804.3217929

Copyright: City Research Online aims to make research outputs of City, University of London available to a wider audience. Copyright and Moral Rights remain with the author(s) and/or copyright holders. URLs from City Research Online may be freely distributed and linked to.

Reuse: Copies of full items can be used for personal research or study, educational, or not-for-profit purposes without prior permission or charge. Provided that the authors, title and full bibliographic details are credited, a hyperlink and/or URL is given for the original metadata page and the content is not changed in any way.

City Research Online: http://openaccess.city.ac.uk/ publications@city.ac.uk

IRA Propaganda on Twitter: Stoking Antagonism and Tweeting Local News

Johan Farkas
Malmö University
Nordenskiöldsgatan 1, 211 19
Malmö, Sweden
+46 765830854
johan.farkas@mau.se

Marco Bastos
City, University of London
Northampton Square, Clerkenwell, London
EC1V 0HB, United Kingdom
+44 2070400469
marco.bastos@city.ac.uk

ABSTRACT

This paper presents preliminary findings of a content analysis of tweets posted by false accounts operated by the Internet Research Agency (IRA) in St Petersburg. We relied on a historical database of tweets to retrieve 4539 tweets posted by IRA-linked accounts in 2012-2017 and coded 2501 tweets manually. The messages cover US newsworthy events, the Charlie Hebdo terrorist attack in 2015, and the Brexit referendum in 2016. Tweets were annotated using 19 control variables to investigate whether IRA operations on social media are consistent with classic propaganda models. The results show that the IRA operates a composite of user accounts tailored to perform specific tasks, with the lion's share of their work focusing on US daily news activity and the diffusion of polarized news across different national contexts.

CCS CONCEPTS

• **Social media propaganda → IRA propaganda**; Social network sites → Manipulation -> Disinformation

KEYWORDS

Social media, Propaganda, Internet Research Agency, Russia, Disinformation, Twitter, Information warfare

ACM Reference format:
Johan Farkas and Marco Bastos. 2018. IRA Propaganda on Twitter: Stoking Antagonism and Tweeting Local News. In Proceedings of the *International Conference on Social Media & Society*, Copenhagen, Denmark (SMSociety).

1 INTRODUCTION

In this article, we present preliminary findings of a research investigation into, what the social media company Twitter defines as, "a propaganda effort by a Russian government-linked organization known as the Internet Research Agency" [19]. We analyze 2501 messages posted on Twitter between 2012 and 2017 by accounts with false identities operated by the Internet Research Agency (IRA). According to Twitter, these accounts were part of "Russian efforts to influence the 2016 [US] election through automation, coordinated activity, and advertising" [7]. Drawing on theoretical concepts from propaganda studies, we investigate dominant themes and discourses produced by the IRA through false accounts claiming to represent US citizens as well as news channels and organizations.

The study relies on a list of 2752 deleted Twitter accounts that was handed over to the US Congress by Twitter on 31 October 2017 as part of investigations into Russia's meddling in the 2016 US elections [7]. In his testimony before Congress, Twitter's acting General Counsel, Sean Edgett, stated that the company identified a total of 36,746 "Russian-linked accounts", which produced about 1.4 million tweets in connection to the US elections [7]. Twitter also identified 2572 "Human-Coordinated Russian-Linked Accounts" operated by the IRA [7]. As Twitter handed over the list of IRA accounts, their names became public. The company, however, has yet to share the corpus of deleted tweets posted by these accounts [11].

This article examines 2501 tweets posted by IRA accounts found in connection to US daily news, Brazilian and Ukrainian protests in 2013-2014, the Charlie Hebdo terrorist attack in 2015, and the Brexit referendum in 2016. As data was collected using event-specific hashtags and keywords, the resulting dataset is not representative of the activity of the IRA. Yet, the tweets offers a unique glimpse into the workings of IRA's subversive propaganda strategies, which remain largely underexamined.

There are important epistemological issues that need to be taken into consideration within this line of inquiry and the specifics of the data being analyzed [10]. This is particularly the case of information potentially designed to induce a state of *psychological warfare* [15]. In the following section, we briefly present an overview of scholarly contributions to social media propaganda, outline the theoretical framework underpinning this study, and present the research questions deriving from propaganda theory.

2 STATE PROPAGANDA IN DIGITAL MEDIA

While propaganda predates mass communication technologies by several centuries, 20th century state propaganda was intimately connected to the rise of mass media such as newspapers, radio, and television [12]. Mass media evolved along with increasingly complex propaganda techniques, ultimately leading to a state of globalized warfare when propaganda dissemination reached unprecedented scales [18]. Propaganda went through considerable changes [20], but the centrality of mass media remained a stable component in propaganda diffusion [8, 12].

The emergence of social network sites was greeted as a formidable challenger to the monopoly of mass media and centralized publishing systems. The decentralized nature of social networks would allow for dissenting voices to be expressed and heard [5, 6]. Social platforms were heralded, as exemplified by Boler and Nemorin, writing that "the proliferating use of social media and communication technologies for purposes of dissent from official government and/or corporate-interest propaganda offers genuine cause for hope" [5]. While mass media relies on one-to-many communication, which is difficult and at times impossible for activists to circumvent, social media enable citizens to organize and coordinate protests through distributed networks [6].

Despite early optimism around social media, recent research has shown that rather than empowering citizens and disempowering authoritarian states, social media is increasingly appropriated by state actors to enforce mass censorship and surveillance along with propaganda and disinformation campaigns [13, 14]. Technological advances in software development and machine learning enable automated detection of political dissidents, removal of political criticism, and mass dissemination of government propaganda through social media. These emerging forms of political manipulation and control constitute a difficult object of analysis due to scant and often non-existing data along with extant methodological and epistemological challenges.

While mass mediated propaganda requires extensive resources, any individual with an internet-capable device can potentially disseminate propaganda through social media [9]. Social network sites are distinctly dynamic platforms, in which social actors of all types communicate and interact. The network structure of digital environments enables citizens to produce counter-discourses to established norms, practices, and policies. The platforms' decentralized structure, however, also enables large-scale actors, such as authoritarian states, to disseminate disguised propaganda appearing to derive from within a target population. State propaganda can be further disseminated by users unaware of the manipulation. For scholars and journalists, such propaganda poses considerable challenges due to the difficulty of establishing authorship. Social media companies have so far been hesitant to provide support for such investigations, while offering extensive anonymity for content producers and handling abusive content by simply removing it [10]. This has led to a scenario in which little research has been carried out on the topic.

In the context of the 2016 British EU membership referendum, research estimates that 13,493 Twitter accounts were so-called social bots: software-driven digital agents producing and distributing social media messages [1]. Researchers identify bot-like accounts based on distinct characteristics that set them apart from regular accounts, most prominently the number and ratio of tweet to retweets, which is higher for social bots [1]. Bessi and Ferrara [3] used similar bot-detection techniques to estimate that 400,000 bots operated during the 2016 US elections. Despite these findings, literature is yet to establish the origin of such social bots, as Bessi and Ferrara [3] summarizes:

> … it is impossible to determine who operates such bots. State- and non-state actors, local and foreign governments, political parties, private organizations, and even single individuals with adequate resources… could obtain the operational capabilities and technical tools to deploy armies of social bots and affect the directions of online political conversation.

It is difficult to establish the identity of disguised social media accounts [9, 10] and their country of origin [7]. While social bots can be identified based on traces of computer automation, disguised human-driven accounts can be difficult to recognize, as they do not display features clearly associated with automation. Disguised human-driven accounts can neither easily be found nor traced back to an original source or controller. Furthermore, potential identification of accounts require collaboration with social media companies, which are reluctant to provide such support [10]. Within the scope of this study, Twitter has released a list of 2752 deleted accounts identified as operated by the IRA. Although Twitter has not shared the tweets posted by these accounts [11], it is possible to trace campaign and social media activity spearheaded by the IRA.

3 THEORETICAL FRAMEWORK & OBJECTIVES

Jowett and O'Donnell [12] define propaganda as the "deliberate, systematic attempt to shape perceptions, manipulate cognitions, and direct behavior to achieve a response that furthers the desired intent of the propagandist" (p. 7). One such agenda pursued extensively by state actors throughout the 20th century is *psychological warfare* [12, 15], which ac-

cording to Linebarger [15] encompasses "the use of propaganda against an enemy, together with such other operational measures of a military, economic, or political nature" (p. 40). Unlike propaganda targeted at a state's own population, psychological warfare is waged against foreign states. Despite its name, psychological warfare is not restricted to periods of armed warfare. Jowett and O'Donnell [12] argue that it "commences long before hostilities break out or war is declared… [and] continues long after peace treaties have been signed".

The study of propaganda and psychological warfare depends on identifying the ideology, context and underlying identities of the propagandist, the latter being particularly challenging for disguised propaganda [12]. Analysts can nonetheless engage in source-identification by studying "the apparent ideology, purpose, and context of the propaganda message. The analyst can then ask, Who or what has the most to gain from this?" [12]. In relation to tweets produced by the IRA, we do not know the extent to which the Russian government was involved, but in view of the mutual military build-up and trade sanctions between the US and Russia [4], it is conceivable that Russia would benefit from supporting a Russian-friendly presidential candidate in the US. According to Twitter, the IRA accounts were part of "Russian efforts to influence the 2016 election" [7], a characterization that implies a close connection to psychological warfare on social media.

A key goal of psychological warfare throughout modern history has been to create confusion, disorder, and distrust behind enemy lines [12, 18]. Through the use of grey or black propaganda, conflicting nation states have disseminated rumors and conspiracy theories within enemy territories for "morale-sapping, confusing and disorganising purposes" [2]. Within propaganda theory, grey propaganda refers to that which has an unidentifiable or difficult to identify source, while black propaganda refers to that which claims to derive from within the enemy population [2, 12]. In this article, we use the term disguised propaganda to encompass both forms. According to Becker [2], black propaganda is particularly effective as means of psychological warfare "when there is widespread distrust of ordinary news sources" [2]. Considering the contemporary political landscape, in which only 33% of Americans, 50% of Brits, and 52% of Germans trust news sources "most of the time" [17], we hypothesize that IRA-linked Twitter accounts deploy disguised propaganda (i.e., grey and black) to spread falsehoods and conspiracy theories. In view of that, we posit the following research questions:

RQ1 *Does the IRA propaganda effort on social media rely on grey and black propaganda?*

RQ2 *Is IRA propaganda on social media centered around spreading rumors and conspiracy theories*?

The seminal work of Ellul [8] has detailed psychological warfare along a range of characteristics. Subversive psychological warfare most often comes in the form of *propaganda of agitation* [8], which refers to propaganda disseminated to stir up tension through use of "the most simple and violent sentiments… Hate is generally its most profitable resource" [8]. According to Ellul [8], propaganda of agitation not only seeks to prompt emotional responses, but also to direct behavior: "it operates inside a crisis or actually provokes the crisis itself" [8]. Drawing on these propositions, our third and fourth research questions are:

RQ3 *Is IRA propaganda on social media focused on disseminating emotional and antagonistic content?*

RQ4 *Do IRA propaganda efforts encourage antagonistic action online and offline?*

4 DATA & METHODS

The disguised IRA propaganda in our study has been sampled by trawling through millions of historical tweets and searching for messages authored by IRA accounts, as identified by Twitter [7]. One account turned out to be a false-positive, which has been excluded from our study [16]. The dataset spans six years and includes tweets with a topical focus on US news outlets, the Charlie Hebdo terrorist attack in 2015 (e.g. #CharlieHebdo, #JeSuisCharlie), and the Brexit debate in 2016 (e.g. #Brexit, #GoodbyeBritain). Upon querying the database, we found 4539 tweets posted by IRA accounts between 2012 and 2017. The available data cannot account for the totality of messages posted by these accounts nor a representative sample. Accordingly, our study cannot estimate the extent of IRA propaganda on social media nor the prevalence of other forms of propaganda tactics. The findings presented in the following section are conditional on these constraints.

Out of the 4539 tweets identified as posted by IRA accounts, a total of 1848 messages could not be annotated because they did not include text, were posted using the Cyrillic alphabet, or a combination of the above. The database is encoded in Latin-1 Supplement of the Unicode block standard, which does support Cyrillic characters, hence messages in Russian or Ukrainian were removed from the sample. The database archives only text and therefore we do not have access to images or videos embedded to tweets, except in cases of content that is still available. The remaining 2501 tweets were manually annotated along 19 variables established to explore the four research questions underpinning the study. Eighteen of these variables are deductive and one variable was found inductively based on an initial coding of a subsample of 10%

of tweets. One of the authors with previous experience coding social media propaganda coded the totality of messages. The variables listed below are not mutually-exclusive nor do they apply to all tweets in the dataset.

1. National identity (based on of five attributes, including self-descriptions, language and Twitter names/handles - e.g. LAOnlineDaily).
2. National context of tweets
3. Language
4. Retweeted Twitter account
5. Mentioned or replied Twitter account
6. Mentioned person or organization (non-Twitter mentions)
7. Political party of mentioned, retweeted or replied person or account
8. Endorsement of individual, organization or cause
9. Disapproval of individual, organization or cause
10. Religion
11. Fatalities (five attributes: 'risk of fatality', 'fatality', 'fatalities', '5+ fatalities' and 'mass murder')
12. Issues (up to four attributes for each tweet based on seventeen attributes established through an inductive coding of a sub-set of 10% of tweets)
13. Encouragement of action (explicit encouragement, e.g. 'Vote for X' or 'Share this!')
14. Rumor/Conspiracy (two attributes: 'yes' and 'high', defined as the dissemination of claims with no referenced sources)
15. Aggressiveness (two attributes: 'yes' and 'high', defined as use of curse words, threats and/or capitalized sentences).
16. Antagonism (two attributes: 'yes' and 'high').
17. Emotional (two attributes: 'yes' and 'high').
18. Populism (eight attributes: 'Reference to the people', 'anti-establishment', 'anti-mainstream media', 'scapegoating', 'call for action', 'ethno-cultural antagonism', 'state of crisis/threat against society', 'the need for a strong leader')
19. Populism spectrum (two attributes: 'Low' and 'High')

5 FINDINGS

After manually annotating the tweets (N=2501), we found that most of them were written in English (n=2082), 324 in German, and 84 in Italian. The remaining tweets were written in French (8), Dutch (1), Swedish (1), and Filipino (1). Most of the tweets (n=1607) address or are situated in a US national context, 923 refer to a British context, and 272 to Germany, with the coding scheme allowing several contexts to apply to the same tweet. The most prevalent topics are local affairs (n=1453), encompassing news pieces related to specific cities or municipalities, followed by politics (n=1184), crime (n=788), economy (n=272), and entertainment (n=257). Only 5.72% of tweets cover rumors or conspiracy theories (n=148), but 11.76% include antagonisms (n=294), 10% comprise emotional statements, and 3.12% encourage online or offline antagonistic behavior (n=78).

These issues are segmented across different types of accounts, displaying distinct characteristics. This suggests that IRA propaganda efforts incorporate independent lines of action that can be assigned to a typology of user accounts. To this end, we did a preliminary classification of accounts in the sample according to prevailing features, resulting in nine primary groups:

Individuals

1. Conservative patriots (Trump/Brexit supporters; US)
2. "Ordinary" accounts (Personal experiences and sometimes conspiracy theories; US)
3. Political news disseminators (US & Italy)
4. Anti-EU Brexit supporters (Germany)
5. Pro-EU Brexit supporters (Germany)

News and Organizations

6. Local news (US)
7. War news (German, US, and unidentifiable)
8. Political commentary (US)
9. Conservative organizations (US)

The preliminary classification highlights that the IRA uses different types of accounts to support various political agendas. Many of the accounts impersonate local news outlets in the US. This includes *DailyLosAngeles, ChicagoDailyNew, DailySanFran* and *KansasDailyNews* (type 6). Upon probing into the data, we found that they operate by relaying information sourced from established news outlets in the area they operate. The tweeting pattern comprises a single headline and not always include a link to the original source.

When available, we resolved the shortened URLs embedded to tweets to identify the news source tweeted by disguised local news accounts. *LAOnlineDaily* tweeted exclusively Los Angeles Times content and *ChicagoDailyNew* follows a similar pattern having tweeted content from Chicago Tribune. As such, this cohort of news repeaters seems dedicate to replicating local news content with a potential bias towards news items in the crime section and issues surrounding public safety. The local news stories distributed by IRA accounts are dominated by negative and contentious narratives and/or amplify concerns about public security, particularly crime incidents, but also fatal accidents and natural disasters. The most prolific account in our dataset is user 2624554209 with a total of 1212 tweets. This account operated under the handle *DailyLosAngeles* in 2016, but it was also active in 2015 under the username *LAOnlineDaily*. Below is an example of the type of content relayed by

LAOnlineDaily.

#breaking #LA Two fetuses found beside road in Fallbrook http://t.co/IIYpmtXaGCaking (LA Online Daily, Twitter, 3 January 2015)

The dataset also contains accounts impersonating American, British, German, and Italian individuals. These accounts often distribute content from established news sources (type 3), but also post content written in a personal, emotional, and antagonistic style (type 1, 4, and 5). These users also offer clear support for political actors and agendas such as Britain's withdrawal from the EU, US President Donald Trump, or the German Chancellor Angela Merkel. The following tweets exemplify such content:

Europe is killing itself. How long until there will be Belgium and French Sultanates? #StopIslam #Brexit #MAGA #MEGA (Williams_Diana, Twitter, 18 June 2016)

After #Brexit #Merkel will make Frankfurt stronger! #Merkelmussbleiben (LarsWolflars, Twitter, 21 July 2016, own translation from German)

6 CONCLUSIONS

This paper offers preliminary insights into the strategies employed by the IRA on social media. We manually annotated messages to identify the extent to which IRA's modus operandi is consistent with classic propaganda models. We found numerous and conflicting types of disguised accounts suggesting that the IRA employs different propagandistic techniques depending on the country and targeted political agenda. Contrary to our expectations, we found that most activity in the dataset was associated with accounts mimicking local news outlets. This group of accounts display a preference for news stories dominated by contentious narratives that amplifies concerns about public security.

The extent to which the Russian governments was involved in the IRA activity remains unknown and, thus, we lack a clear understanding of the strategic role played by the IRA. We nonetheless expect the investigation into Russia's meddling in the 2016 elections to shed new light on these issues. Lastly, the results reported in this study are preliminary and contingent on the limitations of our data. Further research should explore the profiles of IRA-linked accounts to reveal the extent to which a classic distinction between covert and overt propaganda remains valid in the age of social media.

REFERENCES

[1] Bastos, M.T. and Mercea, D. 2017. The Brexit Botnet and User-Generated Hyperpartisan News. *Social Science Computer Review*. (2017), 1–18. DOI:https://doi.org/10.1177/0894439317734157.

[2] Becker, H. 1949. The Nature and Consequences of Black Propaganda. *American Sociological Association*. 14, 2 (1949), 221–235.

[3] Bessi, A. and Ferrara, E. 2016. Social bots distort the 2016 us presidential election online discussion. *First Monday*. 21, 11 (2016).

[4] Birnbaum, M. 2015. 3 maps that show how Russia and NATO might accidentally escalate into war. *The Washington Post*.

[5] Boler, M. and Nemorin, S. 2013. Dissent, Truthiness, and Skepticism in the Global Media Landscape: Twenty-First Century Propaganda in Times of War. *The Oxford Handbook of Propaganda Studies*. J. Auerbach and R. Castronovo, eds. Oxford University Press. 395–417.

[6] Castells, M. 2012. *Networks of Outrage and Hope: Social Movements in the Internet Age,*. Polity Press.

[7] Edgett, S. 2017. Testimony of Sean J. Edgett. *United States Senate Committee on the Judiciary, Subcommittee on Crime and Terrorism*. (2017).

[8] Ellul, J. 1965. *Propaganda: The Formation of Men's Attitudes*. Vintage Books.

[9] Farkas, J. et al. 2017. Cloaked Facebook Pages: Exploring Fake Islamist Propaganda in Social Media. *New Media & Society*. (2017). DOI:https://doi.org/https://doi.org/10.1177/1461444817707759.

[10] Farkas, J. and Neumayer, C. 2017. "Stop fake hate profiles on Facebook": Challenges for crowdsourced activism on social media. *First Monday*. 22, 9 (2017).

[11] Hern, A. 2017. Russian troll factories: researchers damn Twitter's refusal to share data. *The Guardian*.

[12] Jowett, G.S. and O'Donnell, V. 2012. *Propaganda and Persuasion*. SAGE Publications.

[13] Khamis, S. et al. 2013. Propaganda in Egypt and Syria's "Cyberwars": Contexts, Actors, Tools, and Tactics. *The Oxford Handbook of Internet Studies*. J. Auerbach and R. Castronovo, eds. Oxford University Press. 418–438.

[14] King, G. et al. 2017. How the Chinese Government Fabricates Social Media Posts for Strategic Distraction, Not Engaged Argument. *Gking.Harvard.edu*. (2017). DOI:https://doi.org/10.1017/S0003055417000144.

[15] Linebarger, P.M.A. 1954. *Psychological Warfare*. Duell, Sloan, & Pearce.

[16] Matsakis, L. 2017. Twitter Told Congress This Random American Is a Russian Propaganda Troll. *Vice Motherboard*.

[17] Newman, N. et al. 2016. *Digital News Report 2016*.

[18] Taylor, P.M. 2003. *Munitions of the mind: A history of propaganda from the ancient world to the present era*. Manchester University Press.

[19] Update on Twitter's Review of the 2016 U.S. Election: 2018. *https://blog.twitter.com/official/en_us/topics/company/2018/2016-election-update.html*. Accessed: 2018-04-11.

[20] Welch, D. 2014. *"Opening Pandora"s Box': Propaganda, Power and Persuasion*. I.B Tauris an Co Ltd.