# Twitter Bots and the Spread of Russian State-Media Disinformation on the White Helmets in Syria

Tim Roy

March 30, 2021

Total number of words: 2897

## 1. Introduction

Robots have become increasingly relevant in politics today through their role in the spread of information online. Programmers can automate posting and interaction by bots *en masse* to create artificial interest in a story, a method knows as astroturfing, or filling hashtag feeds with similar content, a method known as flooding. In a form of cyber-warfare, governments have been involved in the sponsorship of bots and trolls for the spread of particular viewpoints and disinformation, a reality that has led to the coining of "cyber troops" (Bradshaw and Howard 2017). In this paper, I study disinformation stories shared by the Russian state news agency, RT, and how the spread of these stories was aided by Twitter bots and engaged with by users.

Controversy over the Russian government's influence over the recent Presidential elections due to bots has brought the regime's use of cyber troops for spreading disinformation into the spotlight. I theorize RT's online production and its Twitter following and engagement as a section of the Russian cyber troop. I specifically look at false stories shared about the Syrian Civil Defense, also known as the White Helmets (WH), in Syria. in which they are depicted as terrorists and frauds. I inspect astroturfing by measuring the degree to which bots interacted with RT content on WH. I also investigate flooding of hashtags by counting duplicate-posting by bots. Using a bot threshold $\geq 0.9$, I find that 40 percent of users retweeting the stories are bots. Of these bots, 18.84% appear more than once in the data. I also find that the average users that mention RT are less likely to be bots, with a difference in means in bot scores of 0.114 [95% CI = -0.037, 0.267].

## 1.1. Russia vs. the White Helmets

The White Helmets (WH), or the Syrian Civil Defense, are a humanitarian civil defense organization focused on search-and-rescue operations in dangerous areas or in response to bombings. They often wear cameras to record their missions and publish their recordings online. They have published videos exposing evidence of chemical attacks, the targeting of civilians, and other violations of human rights. As an effort to promote its position in the Syrian civil war, Russia initiated a massive disinformation campaign against them, promoting stories that claimed they were linked with Islamist terrorist organization, organ trafficking, and argued they were frauds that used dummy victims (mannequins) and actors (Palma 2016). The seemingly fringe disinformation stories were frequently shared by RT on TV and on YouTube. I investigate their salience on Twitter.

## 1.2. RT on Social Media

RT (formerly known as Russia Today) is a Russian state-sponsored news agencies directed at foreign audiences that was founded in 2005. Their catchphrase is "Question More," and they encourage audiences to "question western and mainstream narratives of world events" (Starbird et al. 2018). They have a tremendous online footprint. RT was the first TV station ever to garner a billion cumulative views on YouTube (Von Bidder 2013). Out of the top news sources on Youtube in English, Arabic and Spanish, RT ranks highest in English and Spanish and third highest in Arabic (Orttung and Nelson 2019). Rather than focusing on overtly political content RT tends to focus on "sensational [stories], [like] eye-witness reports of catastrophes and disasters" (Chatterje-Doody and Crilley 2019; Mickiewicz 2014 as cited in Crilley et al. 2020) or other miscellaneous talk show content (Orttung and Nelson 2019). Compared to its online audience, RT's TV audience is small and the channels viewership is only growing in the Middle East and North Africa, particularly in Syria and Iraq (Connect 2018, as cited in Crilley et al. 2020).

In an analysis of the entire online disinformation ecosystem surrounding WH, Starbird et al. (2018) find that `rt.com` and `sputniknews.com`[1] are in the top ten most tweeted domains of the ecosystem (RT 879 times and Sputnik 1110 times). In terms of news more generally, RT is also a major player. When looking at the success of RT on Twitter, controlling for bots, Metzger and Siegel (2019) find that RT and RT Arabic are the news sources whose Tweets on Syria "perform" the best. The accounts outperform the New York Times and BBC in English, and Al-Arabiya and Al-Jazeera in Arabic. The authors attribute their performance to the efficient use of hashtags in English, and the sheer quantity of content in Arabic.

RT's YouTube content matters for an analysis of its Twitter content because of the abundance of YouTube media consumed on Twitter that is shared in URLs and embedded in posts.

[1] Sputnik is another Russian state-sponsored news outlet.

Horawalavithana et al. (2020) show that Twitter acts as a megaphone for for cross-platform messaging about WH conspiracies. They find that out of all web domains sharing WH disinformation in their sample, `RT.com` co-apeared with YouTube URLs in the same Tweet the most, co-appearing 78 times. The second highest co-domain was `clarityofsignal.com`, a personal blog hosting conspiratorial articles whose catchphrase on their home page is "Exposing Geopolitical Madness." It co-appeared 60 times.

From an analysis of a data-set of 70,220 video titles of RT's YouTube media spanning from 2015 to 2017, Orttung and Nelson (2019) inferred that RT has a

> three-prong strategy that focuses on strategic groups outside the West, including Arabic, Russian, and Spanish speakers, circumvents local media in target countries to promote Kremlin aims, and spreads a positive image of Russian accomplishments in Syria.

They find that RT under-performs among Arabic speakers, compared to Russian, English, and Spanish audiences.

An overarching analysis of RT content by Crilley and Chatterje-Doody (2020) shows that comedy and satire are key pieces in RT strategic goals and part of RT's success. They show that engagement on RT content often involves praising the media for its humour. Taken together with cyber troop strategy, this finding helps explain RT's success online, where the "meme" economy dominates, compared to traditional media. Often, the line between satire and disinformation is blurred as is the case for some stories on the White Helmets.[2]

[2] see reporting by RT of the WH "Mannequin Challenge" https://www.rt.com/viral/367775-white-helmets-mannequin-challenge/

### 1.2.1. RT's Audience

Crilley et al. (2020) perform an analysis of RT's Twitter audience, rather than their content. They find that followers are more likely to be males (0.75 compared to 0.62) and are likely to be slightly older (their probability of being 30-40 is 0.2 compared to 0.15 for a random control group of Twitter uses and 0.27 compared to 0.22 for being over 40). They find that RT Twitter followers rarely engage with RT content, arguing therefore that claims that RT has a large audience supporting "anti-Western" worldview are misguided. They are exposed to RT content, but do not appear to endorse that content on Twitter often. They also find that followers do not appear to be "a niche audience of activists" (Orttung and Nelson 2019). Rather, most people who follow RT do so alongside other major international news sources. Followers are fragmented mainly along national, linguistic and cultural lines instead of political identities or ideologically extreme views.

### 1.2.2. RT and Twitter Bots

Internet bots are automated software applications, running any range of tasks and doing so repetitively. Internet bots are so widespread that in 2016, bots were estimated to make up half of all online traffic. It is estimated that between 9 and 15 percent of Twitter accounts are bot accounts. In the literature on Twitter bots, Twitter users are generally classified as human, bot, or cyborg accounts, each distinguishing itself from the other by level of automation. Cyborgs are accounts that mix automated and non-automated tweets. Twitter bots can perform all the tasks available to users (i.e. Tweet, retweet, like, follow) and Twitter does not mind the use of bots for these purposes as long as they do not break the Terms of Service through "spamming" and "misleading." Bots that are actively engaging with Twitter users (not just passive followers) to promote a particular viewpoint are social bots. Some bots are just made to follow an account to inflate the number of that account's followers and increase its influence and reach (Efthimion et al. 2018).

Different methods have been used to detect Twitter bots in user samples. Forelle et al. (2015) use a particularly simple method. They identify the platform used to Tweet from and subset their sample based on these platforms. They assume that some platforms are more bot-oriented and bot-exclusive, therefore users tweeting from them are bots. Other researchers have employed more sophisticated methods like bot-detection algorithms that were trained using machine learning. The advantage of the platform-based detection strategy over the algorithm-based strategy is in mainly in it transparency. Although less transparent, bot-detecting algorithms are useful in that they output statistics measuring the likelihood of being a bot, and researchers can declare their threshold of choice, as well as using other thresholds for robustness checks. Some of these algorithms include `R’s TweetBotOrNot` (AKA `BotOrNot`)[3] or `Python`’s `Botometer`[4].

There is already significant evidence that RT engagement is driven heavily by bots. A significant amount of RT followers are bots and these bots tend to fall into specific categories of Twitter users. After running the botometer algorithm over a sample of RT followers, (Crilley et al. 2020) find that 39 percent are bots (i.e. accounts with a score $\geq 0.8$). When they applied a Louvain algorithm to identify different segments of users and estimated the proportion of bots in each segment, they found that the proportion of bots in each segment ranged from 65 percent to 17 percent, suggesting that some bots engaging with RT operate as coordinated networks.

In an analysis of Twitter networks behind WH disinformation, Pacheco et al. (2020) find evidence of coordinated groups by using network analysis. They identify a “rapid retweet network” and a “similar tweet network” to prove that coordinated cyborg or bot efforts existed surrounding WH stories on Twitter. They find that most tweets replicating content are created a few seconds after the original and they a large number of pairs with matching text using a text similarity index.

[3] https://github.com/mkearney/tweetbotornot

[4] https://botometer.osome.iu.edu/

In an analysis of YouTube data on White Helmet videos and bots, Choudhury et al. (2020) find evidence of strategic operations by bots to increase the popularity of certain videos through astroturfing. Specifically, they discovered the same messages ("with slight variations in emojis or other embellishments") posted repeatedly on the same or different videos and also find instances where different user accounts post the same message. They find that a group of 12 main commenters on RT videos always have higher likes on average than the rest of the commenters.

## 2. Data & Methods

I use `Python`'s `twint` library[5] to scrape all RT tweets and tweets mentioning RT between 2015 and 2017. I then use R to subset all tweets for those that were hash-tagged "#whitehelmets." Next, I get the 100 most recent retweets for each WH-related RT tweet that we scraped using the `rtweet R` package[6] because this method of searching for retweets is not yet supported by `twint`. Mentions were when users tagged "@RT_com" whereas retweets were when users retweeted RT statuses. I follow by using the `tweetbotornot R` package to predict bot probabilities for users that mentioned RT and then for users that retweeted RT for the subsets of WH-related tweets.[7] `tweetbotornot`'s default model is 93.33 percent accurate when classifying bots.[8] Its likelihood estimates range from 0 to 1 with zero meaning no likelihood of being a bot. When I refer to bots from my study I use a conservative threshold held at $\geq 0.9$. The data from our sample of WH-related RT tweets, retweets, and mentions are summarized in Table 1.

I test whether there is evidence of astroturfing and flooding. Specifically, I calculate the proportion of retweets and mentions that are done by bots to check for astroturfing and

[5] https://github.com/twintproject/twint

[6] https://github.com/ropensci/rtweet

[7] All replication files are publicly available for review at https://github.com/timroy/white-helmets and data is hosted at https://drive.google.com/drive/folders/1pRiIkeQHrrO8NVMFGOdg7SkTChXQCtCB?usp=sharing

[8] https://www.rdocumentation.org/packages/tweetbotornot/versions/0.1.0

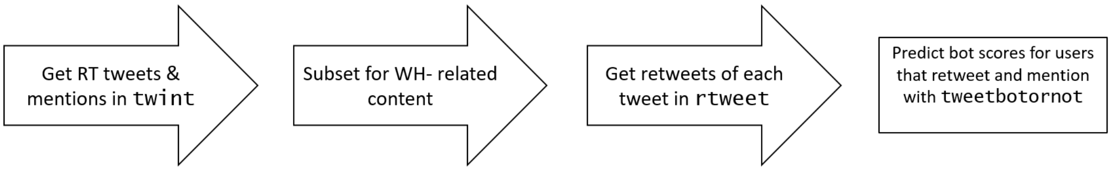

Get RT tweets & mentions in twint
Subset for WH- related content
Get retweets of each tweet in rtweet
Predict bot scores for users that retweet and mention with tweetbotornot

Figure 1: Workflow Diagram for Extracting and Predicting

calculate the amount of duplication in the data to check for flooding. As an added finding, I test to see if there is a significant difference between mention and retweet bot probabilities.

Table 1: Summary of White Helmet Related Tweets, Mentions, and Retweets

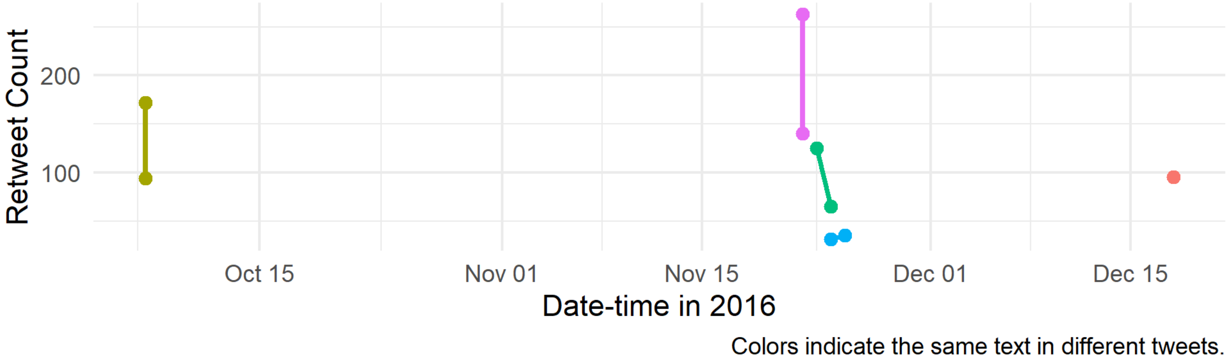
| | RT.Tweets | Mentions | Retweets | Total.Engagement |
|---|---|---|---|---|
| N. of Tweets | 9 | 19 | 569 | 588 |
| N. of Users | | 17 | 529 | 546 |
| Mean bot score | | 0.590 | 0.704 | 0.705 |
| Bot score $\geq$ 0.95 | | 0.294 | 0.296 | 0.312 |
| Bot score $\geq$ 0.9 | | 0.353 | 0.392 | 0.462 |
| Bot score $\geq$ 0.75 | | 0.471 | 0.580 | 0.588 |
| Bot score $\geq$ 0.5 | | 0.647 | 0.760 | 0.744 |

## 3. Results

RT only hashtagged “#whitehelmets” 9 times in two years.[9] RT tended to post two tweets with the same text within short intervals of each other (see Figure 2). For those 9 tweets, I was only able to collect retweets from 8 of them due to limitations in `rtweet`. There ended up being a total of 569 retweets and 529 unique accounts in the sample. In the sample of mentions, RT was only mentioned 19 times in tweets that hash-tagged “#whitehelmets”. Out of that sample, only one user (5.9%) appeared more than once and that user was a bot.

[9] See Table A1 in the Appendix for tweet wording

Figure 2: Duplicate RT WH Tweets Over Time



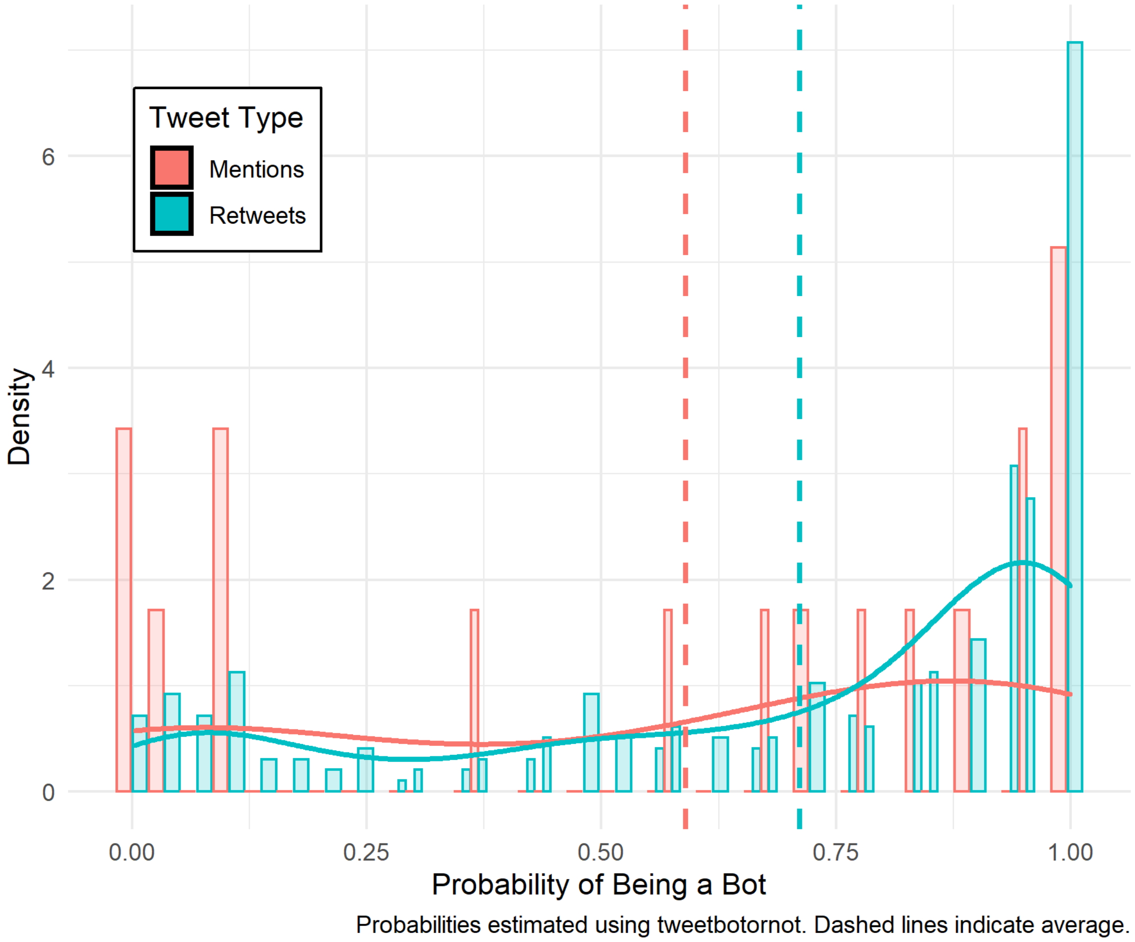
## 3.1. Astroturfing

Figure 3 plots histograms and density curves for the distribution of bot probabilities for users that retweeted or mentioned RT. The distribution for retweeters is left-skewed; the majority of observations fall above 0.5 and a significant proportion score 0.9 and above. As shown in Table 1, 29.3 percent score above 0.95 and 40.3 percent score above 0.9. I employ a conservative threshold and consider accounts above 0.9 to be bots. This finding echoes the finding in Crilley et al. (2020) that 39 percent of RT followers are bots. The distribution of users that mentioned RT is more even across probabilities, with a high proportion at either end of the distribution.

## 3.2. Flooding

There are 207 unique bot accounts (39.2%) involved in retweeting (threshold $\geq 0.9$). In the retweet sample, 39 bots (18.84% of users $\geq 0.9$) appear twice or more times. From the 39 double-posting bots, 8 retweeted duplicate tweets by RT (i.e. different statuses with the exact same text). The other bots seemed to have been programmed to avoid this situation. The same bots appearing more than once points strongly towards concerted flooding efforts.

Figure 3: Bot Probabilities for RT WH “Tweeters”



### 3.3. Mention-ers and Retweeters

Compared to retweeting, mentions involve the production of original content to create the post, which requires sophisticated bots or cyborg accounts. This explains the difference in bot probabilities between the retweet and mention distributions and the fact that the mean for retweets is 0.704 compared to 0.59 for mentions, for a difference of means of 0.114 [95% CI = -0.037, 0.267]. This estimate is not statistically significant.[10] Nonetheless, the results imply it is easier for bots to re-post and share already created content than to create original content through mentions. Bots that mention rather than retweet are potentially better at "hiding."

## 4. Limitations

I only extracted tweets that were hash-tagged "#whitehelmets." In the future, I could add "#SyrianCivilDefense" and I could search in the text of the tweet for "white helmets" or "SyrianCivilDefense". In this paper, I also only extracted English tweets. Twitter bots may be active in Arabic. In the future, I could extract Arabic tweets for RT Arabic and replicate the analysis in Arabic.

The sample size for our mentions is substantially smaller than the sample size for retweets. That means there is high uncertainty associated with the mean bot probability estimate of 0.59 for users that mention ($se_{mentions} = 0.096$, $se_{retweets} = 0.013$).

Engagement is not limited to retweets and mentions. Replies and likes are also important aspects of content-sharing on Twitter. In the future I could possibly account for the reply and like mechanics. Replies are are a bit trickier, however, because retweets and and mentions

[10] see T.Test results in Table A2 in the Appendix

imply endorsement more than replies do, therefore the way we detect disinformation tweets in this paper through hashtags will not be adequate when going through replies.[11]

Lastly, Twitter could have deleted users which they identified as bots that broke their Terms of Service. I intended to address this criticism by using a data-set of users that were removed by Twitter to check their WH-related RT retweeting.[12] The data-set for the time period under study in this paper (accounts removed before October 2018) was unfortunately corrupted and I have yet to hear back from Twitter about it. In the following data-set of Twitter users that were removed in January 2019, of those statuses posted between 2015 and 2017, 260 are retweets of RT and 2 are retweets with "#whitehelmets," but there is no intersection of the two. The January 2019 data-set is significantly smaller than the October 2018 data-set (the file size is 10 times smaller) because many accounts posting during the time-period under study in this paper were already removed from Twitter. Given the large file size of the corrupted data-set, I suspect some tweets intersecting RT and WH are stored there.

# 5. Conclusion

This analysis has shown that twitter bots were heavily involved in the spread of RT disinformation online. Even after holding our bot threshold at a conservative level ($\geq 0.9$), I find that 40.3 percent of retweets of RT WH content were done by bots. This is sign of clear overall astroturfing effort. I additionally find that 18.84 percent of those bots appear in the sample at least twice, implying efforts at flooding hashtags with RT content by the same bots. Lastly, I find that users that mentioned "@RT_com" were more likely to be real people than were users that retweeted. This points to the limits of using bots. Content that is more

[11] https://help.twitter.com/en/using-twitter/types-of-tweets

[12] https://transparency.twitter.com/en/reports/information-operations.html

original, like content mentioning RT instead of retweeting RT, is not as easily generated by a pure bot. Nonetheless, a significant portion of bots are involved in mentions as well.

## References

Bradshaw, Samantha, and Philip N Howard. 2017. *Troops, Trolls and Troublemakers: A Global Inventory of Organized Social Media Manipulation.* Computational Propaganda Research Project 2017.12. University of Oxford.

Chatterje-Doody, Precious N., and Rhys Crilley. 2019. "Populism and Contemporary Global Media: Populist Communication Logics and the Co-construction of Transnational Identities." In *Populism and World Politics: Exploring Inter- and Transnational Dimensions,* edited by Frank A. Stengel, David B. MacDonald, and Dirk Nabers, 73–99. Global Political Sociology. Cham: Springer International Publishing.

Choudhury, Nazim, Kin Wai Ng, and Adriana Iamnitchi. 2020. "Strategic Information Operation in YouTube: The Case of the White Helmets." In *Social, Cultural, and Behavioral Modeling,* edited by Robert Thomson, Halil Bisgin, Christopher Dancy, Ayaz Hyder, and Muhammad Hussain, 318–328. Lecture Notes in Computer Science. Cham: Springer International Publishing.

Connect, Ipsos. 2018. *RT's Audiences in Middle East and North Africa. Internal 17-Country survey report for France 24 shared with research team.* Technical report. France.

Crilley, Rhys, and Precious Chatterje-Doody. 2020. "From Russia with Lols: Humour, RT, and the Legitimation of Russian Foreign Policy." *Global Society,* (Early Access).

Crilley, Rhys, Marie Gillespie, Bertie Vidgen, and Alistair Willis. 2020. "Understanding RT's Audiences: Exposure Not Endorsement for Twitter Followers of Russian State-Sponsored Media." Publisher: SAGE Publications Inc, *The International Journal of Press/Politics,* 1940161220980692.

Efthimion, Phillip George, Scott Payne, and Nicholas Proferes. 2018. "Supervised Machine Learning Bot Detection Techniques to Identify Social Twitter Bots." 1 (2): 71.

Forelle, Michelle, Phil Howard, Andrés Monroy-Hernández, and Saiph Savage. 2015. "Political Bots and the Manipulation of Public Opinion in Venezuela." ArXiv: 1507.07109, *arXiv:1507.07109 [physics].*

Horawalavithana, Sameera, Kin Wai Ng, and Adriana Iamnitchi. 2020. "Twitter Is the Megaphone of Cross-platform Messaging on the White Helmets." In *Social, Cultural, and Behavioral Modeling,* edited by Robert Thomson, Halil Bisgin, Christopher Dancy, Ayaz Hyder, and Muhammad Hussain, 235–244. Lecture Notes in Computer Science. Cham: Springer International Publishing.

Metzger, Megan M, and Alexandra A Siegel. 2019. "When State-Sponsored Media Goes Viral: Russia's Use of RT to Shape Global Discourse on Syria," 32.

Mickiewicz, Ellen. 2014. *No Illusions: The Voices of Russia's Future Leaders.* Publication Title: No Illusions. Oxford University Press.

Orttung, Robert W., and Elizabeth Nelson. 2019. "Russia Today's strategy and effectiveness on YouTube." *Post-Soviet Affairs* 35 (2): 77–92.

Pacheco, Diogo, Alessandro Flammini, and Filippo Menczer. 2020. "Unveiling Coordinated Groups Behind White Helmets Disinformation." In *Companion Proceedings of the Web Conference 2020,* 611–616. Taipei Taiwan: ACM.

Palma, Bethania. 2016. *FACT CHECK: Are the Syrian 'White Helmets' Rescue Organization Terrorists?*

Starbird, Kate, Ahmer Arif, Tom Wilson, Katherine Van Koevering, Katya Yefimova, and Daniel Scarnecchia. 2018. "Ecosystem or echo-system? exploring content sharing across alternative media domains." In *Proceedings of the International AAAI Conference on Web and Social Media,* vol. 12. Issue: 1.

Von Bidder, Benjamin. 2013. "Putin Fights War of Images and Propaganda with Russia Today Channel." *Spiegel.*

# A. Appendix

Table A1: Text in RT WH-Related Tweets

| | tweet |
|---|---|
| 1 | 'I saw no evidence of executions in #Syria as reported by #WhiteHelmets amp; #MSM sources' (Op-Ed) https://t.co/KP68gjeGD2 https://t.co/NfjeIFOJAP |
| 2 | Error of judgement: #WhiteHelmets apologize for war zone  #Mannequin challenge challenge https://t.co/gUhvIJ55x4 |
| 3 | Error of judgement: #WhiteHelmets apologize for war zone  #Mannequin challenge challenge https://t.co/gUhvIJmGoC |
| 4 | #WhiteHelmets 'deserve an Oscar' for mannequin challenge performance in Syria war zone (Op-Edge) https://t.co/oBTkV6wcV3 |
| 5 | #WhiteHelmets 'deserve an Oscar' for #MannequinChallenge performance in Syria war zone (Op-Edge) https://t.co/oBTkV6wcV3 |
| 6 | Confusing #WhiteHelmets "#MannequinChallenge" video goes viral, leaving many questioning authenticity. DETAILS: https://t.co/aa2VYFKQtQ https://t.co/Np1S4pT2ep |
| 7 | Confusing #WhiteHelmets "#MannequinChallenge" video goes viral, leaving many questioning authenticity. DETAILS: https://t.co/aa2VYFKQtQ https://t.co/lPeSYJhjWn |
| 8 | 'Massive evidence foreign-funded #WhiteHelmets support terrorist entities in Syria' - independent researcher https://t.co/MeCHOuHOyI |
| 9 | 'Massive evidence foreign-funded #WhiteHelmets support terrorist entities in Syria' - independent researcher https://t.co/MeCHOuZpXi |

Table A2: Joint T.Test for Bot Probs on Mentions and Retweet Samples

| Retweets Mean | Mentions Mean | statistic | p.value | parameter | conf.low | conf.high |
|---|---|---|---|---|---|---|
| 0.704 | 0.59 | c(t = 1.478) | 0.14 | c(df = 582) | -0.038 | 0.267 |