Beaten Up on Twitter? Exploring Fake News and Satirical Responses During the *Black Panther* Movie Event

Matthew Babcock(✉), David M. Beskow, and Kathleen M. Carley

Carnegie Mellon University, Pittsburgh, PA 15213, USA
{mbabcock,dbeskow}@andrew.cmu.edu, kathleen.carley@cs.cmu.edu

Abstract. The use of user-generated online satire, itself a form of fake news, may be one strategy used to highlight and shame fake news stories and promoters. Here, we begin to explore the differences between non-satire and satire fake posts by looking at Twitter data related to false stories of racially-motivated attacks during the *Black Panther* movie opening. Overall, we found that very few fake tweets of either type had high levels of replies or retweets. We found some evidence that the satire responses were supported and shared to a greater extent than the original non-satire tweets, which leaves open the possibility that satire may have been helpful in calling out the fake attack posts. We also found some evidence that the satire responses fooled some users into believing them to be real stories.

Keywords: Fake news · Satire · Twitter

1 Introduction

The intentional and unintentional spread of false information on the internet has been the subject of continual and increasing public discussions, policy debates, and academic research. Twitter in particular has been studied as a medium in which "fake news" stories and campaigns can find footing and flourish [1–3].

With the growing amount of research and public debate has come an increased interest in whether and in which way policy makers, institutions, and the public should respond to the spread of false information. A recent Policy Forum article in the March 9, 2018 issue of Science summarizes possible interventions into two types: empowering individuals and platform-based detection and intervention [4].

A type of intervention that does not fall neatly into either of the two categories involves community-based intervention and correction. Such community-based interventions involve going beyond empowering individuals to correctly evaluate false information. It additionally involves having those individuals act to call out, mitigate, or otherwise attempt to control the spread of false information, whether on their own or in concert with others. Community-led efforts can help highlight and correct, and thus perhaps control the spread of false or misleading information [5, 6]. If independent, non-government and non-platform-directed communities are involved in the calling out, halting, and/or correcting of fake news cascades, it will be important to describe the advantages and disadvantages of different types of responses communities can engage in.

R. Thomson et al. (Eds.): SBP-BRiMS 2018, LNCS 10899, pp. 97–103, 2018.
https://doi.org/10.1007/978-3-319-93372-6_12

One possible community-based intervention involves the use of satire. Satire itself is a form of disinformation that seeks to expose and/or ridicule its target. Satire can be produced and has been studied at both the level of professional mass media and at the level of user-generated content in a specific community or social media ecosystem. Much of the recent research on satire has focused on professional mass media satire such as *The Daily Show*¸ *The Colbert Report*, and *The Onion* and how such satire impacts knowledge and perceptions of individual issues (for example see [7]).

The use of user-generated satire to specifically constrain the spread of other disinformation online has not been as thoroughly studied. The use of satire may assist in the control of the spread of other disinformation by either increasing the number of people who are exposed to the false information within the context of ridiculing it or by shaming those that spread disinformation to halt or constrain their activities. Being false information by design, satire may also be at a disadvantage as a tool to combat the spread of fake news. In fact, a recent study by Horne and Adali suggests that fake news articles are more closely related in complexity and style to satire articles than to "true" news articles [8]. If satire aimed at fake news is itself considered fake news it may only serve to spread additional false information or drain resources from attempts to control "real" fake news. User-generated satire may be susceptible in different ways from professional satire to this issue.

The preliminary research presented in this paper is therefore focused on exploring the differences and relationships between non-satirical "fake news" and satirical responses on Twitter. Using Twitter data related to the release of the Marvel comic book movie *Black Panther*, we specifically explored the retweeting and reply activity related to both types of fake posts, the presence of bots who tweet fake posts, and the network created between Twitter users responsible for such posts.

1.1 Event Background

Marvel Studios' *Black Panther* movie opened to on February 16, 2018 and tells the story of the Marvel Comics superhero of the same name, who becomes the king and protector of the hidden and technologically-advanced fictional African nation of Wakanda. *Black Panther* was the first movie in the Marvel Cinematic Universe series (and first superhero comic book movie in general) to have a predominately African and African-American cast and creative team, a fact promoted both by Marvel's parent company, Disney (who intentionally released the movie during Black History Month in the United States), and on social media prior to and during the release.

Early showings of the film began the evening of February 15. On the morning of February 16, it was reported by Buzzfeed that there had been a series of twitter posts claiming the user or their friends or family had been physically attacked attempting to see *Black Panther* [9]. Buzzfeed also reported that other Twitter users had quickly posted replies proving that images used in the original posts had been taken from other news and entertainment media. These response tweets called out the original posts as fake stories aimed at stoking racial conflict (most depicted white family members being attacked by black moviegoers, and some depicted the opposite) and tarnishing the film's reputation. Later the same day, Vox reported that in addition to posts debunking and

calling out the original false beating tweets directly, some Twitter users were also mocking the original tweets by posting their own versions using either more clearly unrelated photos or additionally unbelievable language [10]. Additional news reports mentioned that some of these satire posts were being treated as if they were examples of the original fake posts.

2 Methods and Results

2.1 Data Collection and Analysis

Online news articles discussing the non-satire and satire tweets were collected using Google search, with the search terms, "Black Panther Fake News". A preliminary set of non-satire and satire tweets were identified from these articles.

We collected all tweets containing "#BlackPanther" that were posted from February 8 to February 23. We additionally collected tweets containing phrases found in the tweets mentioned in the news articles. By searching our combined collected tweets for those that contained such phrases (e.g. "black youths", "MAGA hats", "cracker") but not response phrases (e.g. "fake", "troll", "racist") and then reviewing our search results, we identified a total of 249 distinct fake tweets (from 238 distinct screen names), 178 which we labeled as satire and 71 which we labeled as non-satire. We then additionally collected tweets that replied to or were retweets of any of the 238 "fake" tweeters.

Satire posts were distinguished from non-satire by manual review. Posts containing images from cartoons, movies, and classical art that depicted unrealistic violence or unrelated content were labeled as satire. Posts containing text describing unrealistic events (e.g. atomic bombs) and stories that started in a similar fashion to the fake beatings but ended positively (e.g. "we were approached by black youths…who then proceeded to give us high-fives") were also labeled as satire.

The combined dataset contains a total of 5,151,935 individual tweets. We created a subset of 291,111 tweets that included all fake posts (non-satire and satire), all retweets and replies to those posts, and all retweets and replies to any other posts by the same users who posted the fake stories.

2.2 Retweets and Replies of Satire and Non-satire Tweets

We counted the retweets and replies in our dataset for each of the 178 satire and 71 non-satire tweets. Due to account suspension and/or tweet deletion, we were unable to identify the number of retweets and replies for 28% of non-satire tweets and 7% of the satire tweets. The satire tweets were in total retweeted 47,512 times and replied to 709 times. The non-satire tweets were in total retweeted 1,916 times and replied to 2,983 times. Table 1 shows that the percentages of retweets of satire and non-satire tweets are relatively similar within each class, except for the case of class 1 in which a larger percentage of the satire tweets are categorized. The same is true of the percentages of replies to satire and non-satire posts. It should be noted that less than 5% of the fake tweets were either retweeted or replied to more than 100 times.

Table 1. Percentage of total retweets and replies that fall within each count class for satire (n = 178) and non-satire (n = 71) tweets. Classes are defined as 1 (0 counts), 2 (1–10), 3 (11–100), 4 (101–1,000), 5(1,001–10,000), and 6 (10,001–100,000).

	Class						
	1	2	3	4	5	6	Unknown
RTs: non-satire	25%	32%	10%	4%	0%	0%	28%
RTs: satire	40%	46%	6%	1%	0%	1%	7%
Replies: non-satire	34%	27%	7%	3%	1%	0%	28%
Replies: satire	62%	29%	1%	1%	0%	0%	7%

We plotted the cumulative sum of replies and retweets over time for individual satire and non-satire tweets focusing on tweets that were mentioned in the news media and/or were representative of classes 3 through 6. For the top non-satire tweet, Fig. 1 (left plot) shows that the replies outweigh the retweets by an order of magnitude. Other non-satire tweets show similar patterns in that the growth in replies to the tweet outpace retweets (though not always by orders of magnitude) and in that the retweets level off sooner.

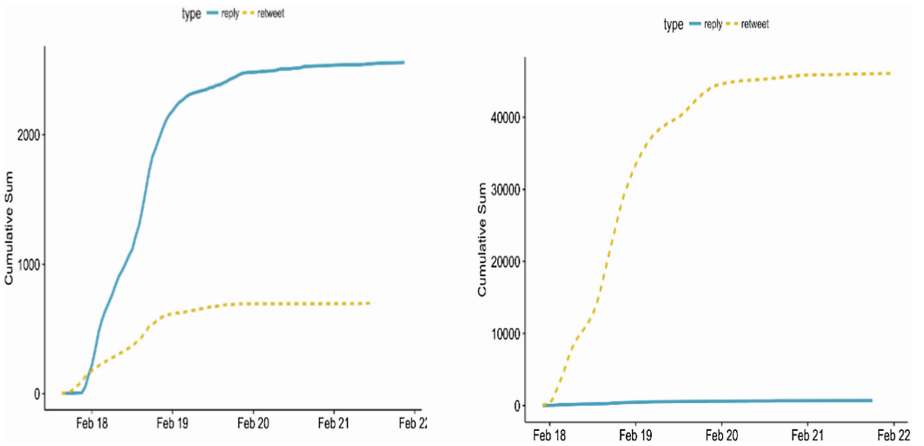

Fig. 1. Cumulative sum of retweets and replies to top non-satire tweet (left plot) and top satire tweet (right plot).

For the top satire tweet, Fig. 1 (right plot) shows a different relationship, where the retweets outweigh the replies by two orders of magnitude. Other top satire tweets show a similar pattern, with the growth in retweets outpacing replies and the replies leveling off sooner. We manually read through the response thread to the top non-satire tweet and a large majority of the replies were calling out the non-satire user for posting "fake" information. This appeared to be the case with other non-satire tweet threads that we looked at. We also read through the responses to the top satire tweet and found the replies to be a mix of supporters acknowledging the satire and some responses that attacked the satire tweet as if it was one of the non-satire tweets.

2.3 Bot Detection

We used CMU Bothunter, an integrated ML approach, to categorize the 238 individual user accounts who posted a fake tweet as bots or not. Due to accounts being deleted and/or suspended prior to our running the bot detection algorithms, only 187 user accounts could be categorized. A total of 12 (6.7%) of the satire accounts and 2 (2.3%) of the non-satire accounts were categorized as bots. Within classes 3–6 from Table 1 there was only one satire account and no non-satire accounts that were categorized as bots.

2.4 Network of Satire and Non-satire Posters Over Time

We created a dynamic network with the 238 fake posting accounts as the nodes and where an edge exists between nodes if either of the users retweeted, replied, or otherwise mentioned the other. Figure 2 shows the cumulative progression of the network at each of 4 days.

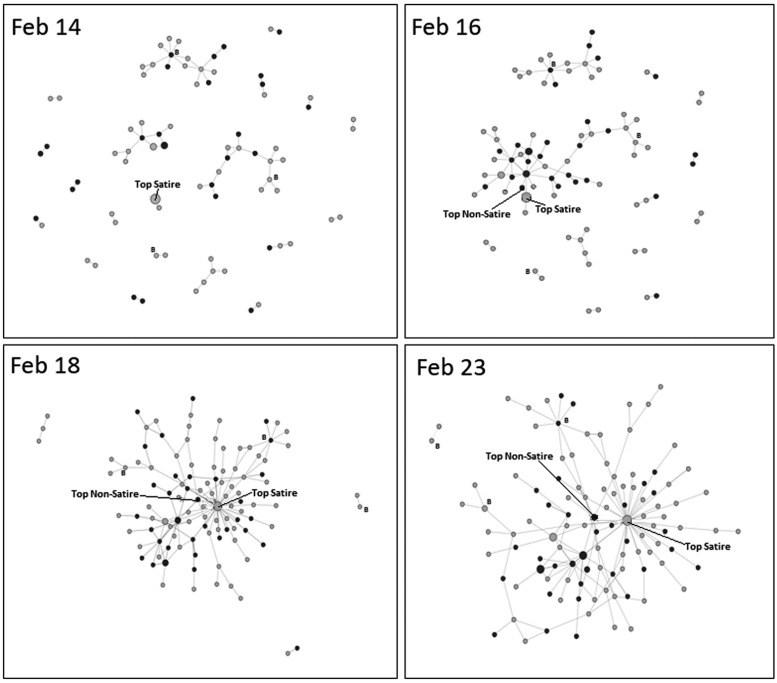

Fig. 2. Network of users who posted satire (grey) and non-satire (black) tweets. Isolates have been removed for clarity. "B" represents the three users in this network that were classified as bots (other bots were isolates). On February 14th, prior to the posting of any fake attack posts, there were three main components, and the top satire user existed in the network but was only connected to one other user. On February 16th, opening day and the day of the first news media reports, two of the main components were connected to each other and to both the top satire and top non-satire user. By February 18th, after the top satire tweet was posted, that user has become the center of the main component of the network. Between February 18th and the 23rd, only 12 additional network connections were made.

3 Discussion and Conclusions

Our preliminary results show that the fake stories of racially-motivated attacks – including both satire and non-satire versions – make up a small fraction of the overall conversation surrounding the Black Panther movie. This is even though many online news and social media outlets covered the story. The sets of identified satire and non-satire tweets were found to be similar in that only a small percentage of each type had high levels of retweeting and reply activity.

The comparison of retweets and replies made in response to individual satire and non-satire tweets suggests that in general the satire tweets were supported and spread by the community while the non-satire tweets were mostly called out and the posters shamed or attacked. This may be an indication of community peer pressure successfully mitigating the spread of non-satire fake news. The fact that response posts began before the first news story ran is also indicative of community self-correction. On the other hand, the high retweeting of specific satire posts may be leading to confusion for those that don't get the pop-culture jokes at the heart of many of such posts. This may also be making the overall "fake attack" story appear larger than it is.

The fact that the top satire tweet becomes the central network node of the main discussion amongst those that post either kind of fake story is interesting in part because it connects both satire and non-satire tweeters. The three bot accounts do not appear to have played a large role in making connections in this network. Further exploring the directionality of the network and expanding the analysis to the larger one-hop network (including all additionally mentions, retweets, and replies to the 249 fake posts) will help to describe how top satire posts may be bringing parts of the networks together. Future work will also include exploring the non-satire responses to the original fake stories as such responses started earlier and have the advantage of being less likely to be confused for the story they are attacking - though it remains to be seen if they spread as fast and/or deep as the satire responses. Exploring a more detailed timeline may provide an indication as to whether the news drove additional activity or active posts gained the attention of the news media.

There is uncertainty in the total number of fake posts of both kinds in our dataset due to our use of keyword searches based on the news articles. This is somewhat mitigated by the fact that many of the false posts that are worth exploring further due to their number of retweets and mentions are ones that the news media picked up. There is also some uncertainty in the labeling of satire posts, as we could not confirm the intent of such posts. This preliminary work is additionally limited in that we only currently have access to Twitter data and therefore are missing network connections between users on other social media. Deleted tweets and suspended accounts also inhibited some of the data collection and bot categorization.

Acknowledgements. This work was supported in part by the Office of Naval Research (ONR) under grants - N000141812108 and N00014-17-1-2605/25323818CMU and the Center for Computational Analysis of Social and Organization Systems (CASOS). The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the ONR or the U.S. government.

References

1. Chamberlain, P.: Twitter as a Vector for Disinformation. School of Computer & Security Science, Edith Cowan University, Australia (2009)
2. Shao, C., Ciampaglia, G.L., Varol, O., Flammini, A., Menczer, F.: The spread of fake news by social bots. arXiv preprint arXiv:1707.07592 (2017)
3. Vosoughi, S., Roy, D., Aral, S.: The spread of true and false news online. Science **359**, 1146–1151 (2018)
4. Lazer, D.M., Baum, M.A., Benkler, Y., Berinsky, A.J., Greenhill, K.M., Menczer, F., Metzger, M.J., Nyhan, B., Pennycook, G., Rothschild, D.: The science of fake news. Science **359**, 1094–1096 (2018)
5. Arif, A., Robinson, J.J., Stanek, S.A., Fichet, E.S., Townsend, P., Worku, Z., Starbird, K.: A closer look at the self-correcting crowd: examining corrections in online rumors. Presented at the Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing (2017)
6. Dailey, D., Starbird, K.: Visible skepticism: community vetting after Hurricane Irene. In: ISCRAM (2014)
7. Day, A.: Breaking boundaries—shifting the conversation: colbert's super PAC and the measurement of satirical efficacy. Int. J. Commun. **7**, 414–429 (2013)
8. Horne, B.D., Adali, S.: This just in: fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news. arXiv preprint arXiv:1703.09398 (2017)
9. Silverman, C.: Trolls are Posting Fake Claims of Being Assaulted at Showings of "Black Panther". https://www.buzzfeed.com/craigsilverman/trolls-are-posting-fake-claims-of-being-assaulted-at
10. Romano, A.: Racist trolls are saying Black Panther fans attacked them. They're lying. https://www.vox.com/culture/2018/2/16/17020230/black-panther-movie-theater-attacks-fake-trolls