Russian Trolls & Emoxicon

Tara L. Valladares1 & Hudson Golino1

1 University of Virginia

Author note

Correspondence concerning this article should be addressed to Tara L. Valladares, Postal address. E-mail: [tls8vx@virginia.edu](mailto:tls8vx@virginia.edu)

Abstract

This will be the abstract.

Russian Trolls & Emoxicon

# Introduction

State sponsored disinformation campaigns are not a new phenomenon. Purposeful efforts to spread false information and "fake news" have been documented since ancient Egypt and Rome. With the rise of the internet and social media, such campaigns can be conducted at a level and size that was previously impossible. The Russian state commonly employs disinformation campaigns in an effort to influence foreign politics. These campaigns utilize a variety of strategies to sow discord, including online media manipulation, cyber-espionage, and falsely inciting protests in the real world (Linvill, Boatwright, Grant, & Warren, 2019; Richey, 2018).

During the 2016 U.S. presidential election, Russian troll farms posed as U.S. natives on Twitter in an attempt to influence political discourse and opinion. This eventually lead to the United States Justice Department inditing 13 Russian individuals and three Russian entities on charges relating to interference in the presidential election (Barrett, Horwitz, & Heiderman, 2018). The Internet Research Agency (IRA), a company based in St. Petersburg, was among the companies charged and has been a primary focus of investigations into state-sponsored Russian interference in U.S. politics (Permanent Select Committee on Intelligence, 2018).

Extensive research has focused on how, why, and with whom the IRA-linked accounts interacted with on Twitter. Understanding the role the accounts played is critical in identifying future accounts and limiting the impact of malicious actors in shaping public opinon. ((something about studying emotion, more citations here))

The actions of Russian trolls outside of the 2016 election suggest that their main purpose is sowing discord (Broniatowski et al., 2018).

### The Role of Twitter and Emotion in Politics.

Twitter and other social media sites play a substainal role in shaping public opinion. Social media use has been shown to predict political participation, protest activity, and voting [Valenzuela (2013); skoric2016social]. But as a whole, social media users tend to follow and interact with accounts with similar beliefs, forming “echo chambers” online (Del Vicario et al., 2016; Garrett, 2009; Törnberg, 2018). More active social media users tend to be more polarized and to follow fewer sources of information (Bessi et al., 2015; Schmidt et al., 2017). Politically active twitter users rarely retweet content from accounts or news sources of different political parties; in Boutet, Kim, and Yoneki (2012) the majority of all interactions occurred between users with homologous polticial views. Though users largely retweet homologous opinions, user-to-user mentions (“@”) frequently cross political lines (Boutet et al., 2012; Conover et al., 2011). But exposure to different political views may not reliably prevent polarization, especially if users are interacting with oppositing, polarizing content from either real or troll accounts (Karlsen, Steen-Johnsen, Wollebæk, & Enjolras, 2017; Yardi & Boyd, 2010).

Emotion plays a significant role in driving public opinion and online political discourse (González-Bailón, Banchs, & Kaltenbrunner, 2012; MacKuen, Wolak, Keele, & Marcus, 2010; Peng, Sun, & Wu, 2017). Shifts in emotional arousal on a political discourse forum were linked to shifts in public opinion (González-Bailón et al., 2012). Influential Twitter users are more likely to tweet more negatively, and more emotional tweets are more likely to spread rapidly and be shared widely (Brady, Wills, Jost, Tucker, & Van Bavel, 2017; Dang-Xuan, Stieglitz, Wladarsch, & Neuberger, 2013). Supporting this, politicians with extreme ideological positions also have more Twitter followers (Hong & Kim, 2016). The use of moral-emotional language increases message spread among political leaders to their followers (Brady, Wills, Burkart, Jost, & Van Bavel, 2019). The use of anger and disgust may be especially impactful among conservative leaders (Brady et al., 2019). This may be driven by either reader characteristics or the topic. For example in Brady et al. (2017), anger words increased sharing for climate change tweets, while positive emotions increased sharing for same-sex marriage. The impact of emotion on the spread of ideas is not limited to Twitter; one study using data from The New York Times found that articles that elicited stronger emotional arousal (i.e., awe, anger, and anxiety) were more likely to be shared (Berger & Milkman, 2012). Further, moral-emotional language may have more of an impact in spreading ideas within-groups than between-groups, contributing to polarization and the “echo chamber” effect (Brady et al., 2017). Anger, specifically, may increase partisan beliefs and susceptibility to misinformation (MacKuen et al., 2010; Suhay & Erisen, 2018; Wollebæk, Karlsen, Steen-Johnsen, & Enjolras, 2019).

It is clear that emotion plays a critical role in shaping political thought online. Importantly, current research on the activies of Russian IRA accounts suggests that they leveraged both the role of emotion and current echo chamber landscape to spread political misinformation. Zannettou et al. (2019) showed that the Russian troll tweets are more negative and less positive than tweets from random subset of twitter users. In fact, some of the most popular IRA tweets were also the most racist and offensive (Freelon & Lokot, 2020)In addition, Twitter users with more homologous social networks were more likely to interact with troll accounts (Bail et al., 2020) and troll accounts were largely interacted with users already within their political cluster (Stewart, Arif, & Starbird, 2018, p. @freelon2020russian). The IRA troll accounts were also retweeted more by conservatives than liberals (Badawy, Ferrara, & Lerman, 2018).

We do not have a thourgouh understanding of the IRA accounts’ use of emotion…

### Paragraph about methods of studying emotion in text.

Because both typical human communication and disinformation campaigns commonly occur on mediums characterized by brief segments of texts (e.g., tweets), there is a great need for flexible, low-computation methods of emotion detection that reliably work on relatively short documents.

## Aims

The purpose of the current study is to introduce a novel method of emotion detection, termed Emoxicon, and detail its application to a dataset of tweets sent by Russian Trolls. *Emoxicon* combines lexicon scoring with Rasch modeling to understand the distribution and probability of emotion categories.

### Russian Trolls Dataset.

The IRA twitter data was made accessible by FiveThirtyEight and researcher’s Linvill and Warren (2019). The topical content of the Russian troll tweets ranges widely. Linvill and Warren (2018) consolidated the data and documented four major categories of English-language accounts that were active before the 2016 election: right trolls, left trolls, news feeds, and hashtag gamers. In the current analysis, we focused on only tweets from right-wing and left-wing troll accounts.

The troll accounts were actively involved in the creation and dissemination of polticized content and propaganda (Zannettou et al., 2019). Beyond following opposite political leanings, the content of the left-wing and right-wing troll tweets focused on separate topics. Political tweets from right-wing accounts largely supported Donald Trump as the presidential nominee and attacked Hilary Clinton. Other popular right-wing topics included anti-Islam sentiment and terrorism, liberal media bias, gun rights, and supporting the police (Golino, Christensen, Moulder, Kim, & Boker, 2020).

In contrast, while political tweets from left-wing accounts largely attacked Trump, they less frequently either attacked or supported Clinton (Linvill & Warren, 2018). They also largely tweeted from a pro “Black Lives Matter” perspective, focusing on topics such as police brutality, racism in america, current events involving the deaths of black americans (e.g., Trayvon Martin, Ferguson and Michael Brown), and popular black musicians and media (Golino et al., 2020)). It is notable that the majority of the Russian left-wing troll tweets were specifically from a “black perspective” instead of the many other diverse and divisive liberal perspectives, such as LGBTQ+, muslim, or feminist voices (Freelon et al., 2020).

The majority of tweets from all accounts were deemed innocuous camouflage tweets; that is, the tweets were not political, but were instead contained mundane interactions with followers or references to popular culture (Linvill et al., 2019). There were apparent differences in how the account categories tweeted, though. For example, the right-wing trolls had less camouflage tweets than the left-wing trolls; over half of tweets from left-wing troll accounts were camouflage tweets, compared to less than 15% of tweets from right-wing troll accounts (Linvill et al., 2019). This suggests that left-wing and right-wing troll accounts were not only talking about different subjects but also employing different strategies in their attempts at political influence.

# Emoxicon

*Emoxicon* is a novel method of analyzing the emotional content of text. By combining lexicon-based scoring with Rasch modeling, emoxicon provides a low-computation analysis of the probability that different emotions categorioes appear within text. Lexicon-based analysis methods are commonly used to identify emotion within text (González-Bailón et al., 2012, pp. @peng2017interplay, @brady2017emotion). Our contribution lies in the use of Rasch models to quantify the distribution and likelihood of the co-occurance of emotions and the “emotional voice” of the author.

### Lexicon Scoring.

We used the DepecheMood++ lexicon created by (Araque, Gatti, Staiano, & Guerini, 2018). Lexicons form the basis of most methods of emotion detection. Lexicons are dictionaries that contain features (such as words or punctuation), the category they belong to, and their designated score. DepecheMood++ is an emotion lexicon that contains roughly 37 thousand terms and their associated probability weights for 8 emotion categories. The lexicon was created by scraping data from the news website Rappler which features a native "*Mood Meter*" widget on each article. Readers of the site are encourage to select one of eight reactions (Afraid, Amused, Angry, Annoyed, Don’t Care, Happy, Inspired, or Sad) to the article. DepecheMood++ has several features that make it particularly useful. First, DepecheMood++ was naively crowdsourced; it did not rely on hand-annotation by knowledgeable participants. Second, DepecheMood++ is much larger than other commonly used lexicons. Lexicon-based analyses are limited by the lexicon’s size and complexity. In general, the larger the lexicon, the more robust the scoring. Smaller lexicons may be poor matches to a target text if there is little overlap in words and significant bias can occur when a target text is poorly matched to the lexicon. Previous work has shown that the Depechemood++ lexicon performs as well or better than other emotion lexicons (Araque et al., 2018; Staiano & Guerini, 2014) in text analysis.

A simple bag-of-words method is applied to score the emotional content of a given document. Though relatively crude, bag-of-words can produce compelling results on its own or in combination with machine learning techniques. Bag-of-words combined with standard machine learning algorithms can produce accuracy rates in tweet sentiment classification around 70-80%, which improves with the addition of sentiment lexicons (Da Silva, Hruschka, & Hruschka Jr, 2014). In the original Depechemood++ lexicon, each word is assigned a weight for each of the eight dimensions. Here, the lexicon is condensed; each word in the lexicon becomes associated with the emotion category with the highest probability weight. Then, each tweet is assigned a sum score of how many words are most highly associated with each emotion category. This aids in the scoring of the text and reduces noise in the Rasch model.

### Rasch Model Analysis.

The Rasch model is used to evaluate the relative probability of the an emotion weight appearing in a given tweet. By using the Rasch model, we are able to analyze the fit of the emotion weights to the document. The Rasch model originated in the field of psychometrics to relate scales and questionnaires to the underlying latent trait (Rasch, 1960). The Rasch model is a logistic model that places persons and items on the same scale to calculate the probability of a given person endorsing a given item. Fundamentally, the Rasch model can be used to generate latent trait scores for entities that produces a set of responses determined by an underlying latent trait. The basic form of the Rasch model for dichotomous items, as adapted from (Rasch, 1960), is:

Where represents the trait level of person and represents the difficulty or location of item . In Emoxicon, tweets take the place of persons and emotional scores are represented as items. Each tweet is scored based on how many words are present from each emotion category. A mean split was performed on the emotion word counts within each emotion category to dummy code each tweet as 1 (high) or 0 (low) for each category. For example, a tweet with a relatively high amount of "Happy" words would receive a 1 for the "Happy" item. The dichotomized scores are run through the Rasch model to produce trait scores for the tweets and ‘item’ locations for the emotion categories.

An underlying assumption of the Rasch model is unidimensionality, which is the requirement that there is only one underlying latent trait driving all responses. In standard applications of the Rasch model, this is a trait such as math ability. Here, we propose that the dimension that the text is scored on is *emotional voice* of the author.

Under the Rasch model, items receive estimates of their relative difficulty on the latent dimension. These difficulties have meaningful order. An item with a lower difficulty is more likely to be endorsed than an item with a higher difficulty. If a high difficulty item is endorsed, it is also likely that the lower diffculty items will also be endorsed. Deviation from these assumptions will produce poor model fit. Thus, the item difficulties are also estimates of conditional probability. That is, if an emotion label recieves a high difficulty, this implies that the emotion label is both rare and generally appears alongside the lower difficulty items. This conditional fit underlies the intepretation of the latent dimension as the emotional voice of the author.

# Methods

## Data Pre-processing

The purpose of this analysis is to examine differences in tweets written by the IRA among the left-wing and the right-wing troll accounts. We will refer to such tweets as *authored* content. By removing non-authored tweets, we ensure that all the tweets support the goals of the accounts and do not belong to accounts with other “voices.” To this end, we included only English-language troll accounts labeled as Left-wing ("Left") or Right-wing ("Right") by Linvill and Warren (2018). URLs, Retweets, manual reposts (i.e., tweets that begin with "rt"), and duplicate tweets were removed.

Because the Depechmood++ lexicon was sourced from news articles, it contains emotion associations for a number of politically relevant words, such as "Hillary," "Clinton," and "Trump". We removed a small set of topically relevant terms that we anitcipated had different affective associations between the left-wing and right-wing accounts (see the appendix for a full list of removed words). That is, left-wing and right-wing trolls may use the word "Trump" just as frequently, but that word does not have the same relationship to the tweets true emotional content for each group. As these words were used often, they also would artificially inflate their respective category "easiness". Similarly, multiword hashtags (e.g., “#blacklivesmatter”) were not split up and scored to prevent misleading category inflation as the emotional valence of the whole hashtag was likely not a sum of it’s parts.

## Analysis Plan

We ran two separate but complimentary applications of the *Emoxicon* method to the Russian Trolls dataset. First, we ran one overarching Rasch model including all cleaned tweets from both left-wing and right-wing trolls. In this strategy, we are able to identify differences between left-wing and right-wing trolls in the majority of the authored tweets. However, these broad group-level differences obscur the actions of less productive accounts and are largely dominated by a handful of mega-accounts with thousands of self-authored tweets. To compliment the overall model, we also ran individual Rasch models for Twitter accounts. This stratgey allows us to examine differences between accounts regardless of their productivity. For the individual models, only accounts with at least 30 tweets were included to sufficiently fit each Rasch model (Linacre, 1994). Each method used the same data scored with the Depechemood++ lexicon (Araque et al., 2018). All models and fit statistics were produced using the R package eRm (Mair, Hatzinger, & Maier, 2020).

The main variable of interest was the estimated item parameters produced by the Rasch model, specfically the item difficulties. In the overarching model, we used Andersen’s likelihood ratio test in the eRm pacakge to identify differences in item difficulties between left-wing and right-wing trolls.

Because the mean of either the person or item parameters of the Rasch model must be set to zero before estimation, item parameters are not directly comparable across individual Rasch models. Therefore to compare the relative difficulties between models, item parameters from the individual models were transformed into ranks from 1 (easiest) to 8 (hardest) within each model. Ties in rank were broken randomly. The distribution of the ranks could then be examined for consistency within and differences between Right and Left trolls. If an individual account did not use any words from a specific emotion category, that emotion category was given the rank of 8 (hardest to endorse).

To evalute the ranks of the emotion categories in the individual models, we examined their distribution using Kolmogorov-Smirnov distribution and Welch t-tests. In addition, we evaluated the consistency of the item difficulties within the left-wing and right-wing troll groups using multiinformation (total correlation) from the R packages infotheo (Meyer, 2008), and the diversity of the emotion labels using the gini coefficient (Ong, Benson, Zautra, & Ram, 2018).

Our final sample consisted of 420447 tweets from 622 right-wing accounts ( = 347489) and 227 left-wing accounts ( = 72958). Among twitter handles with at least 30 tweets, there were 429 right-wing accounts ( = 345489) and 163 left-wing accounts ( = 72162). Tweets dated from 1/1/2013 to 9/9/2017. See Table 1 for more information on tweet distribution of the final sample.

Table 1:

*Distribution of Tweets in Accounts*

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| x | Accounts, n | Total Tweets | Minimum | 2nd Quartile | Median | 3rd Quartile | Maximum |
| Left-wing, all accounts | 227 | 72958 | 1.00 | 26.50 | 106.00 | 302.50 | 7315 |
| Left-wing, 30+ tweets | 163 | 72162 | 30.00 | 92.50 | 199.00 | 376.00 | 7315 |
| Right-wing, all accounts | 622 | 347489 | 1.00 | 18.00 | 76.00 | 450.00 | 35082 |
| Right-wing, 30+ tweets | 429 | 345489 | 30.00 | 72.00 | 191.00 | 654.00 | 35082 |

*Note.* Tweets removed during pre-processing are not included.

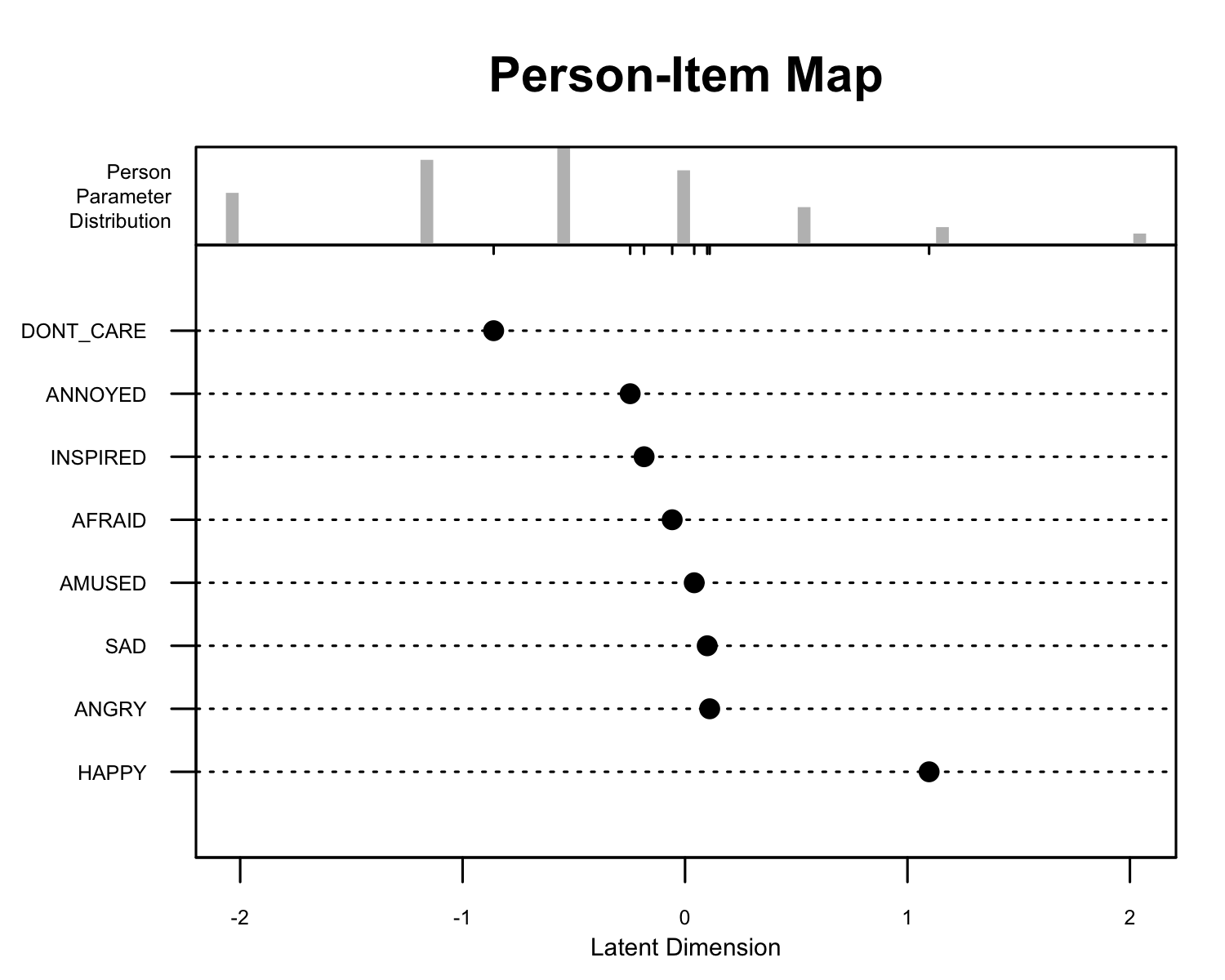
# Results

## Overarching Rasch Model

### Model Fit.

Overall model fit for the full Rasch model was appropriate. Only 1.17% of tweets showed significant misfit using a Chi-square based Z-value cutoff of 1.96. The items *Amused* and *Inspired* showed misfit based on chi-squared item fit statistics ( (399040) = 401,893.55, p = .001 and (rifit$i.df["INSPIRED"]) = 418,125.15, p < .001 , respectively). Due to the large sample size, and the results in the following sections, the misfit in these items is not of great concern. To check for multidimensionality, we performed a principle components analysis on the Rasch model residuals. The largest eigenvalue was 1.28, indicating that there was not a signficiant impact of mulitdimensionality on the model fit (Linacre, 1998).

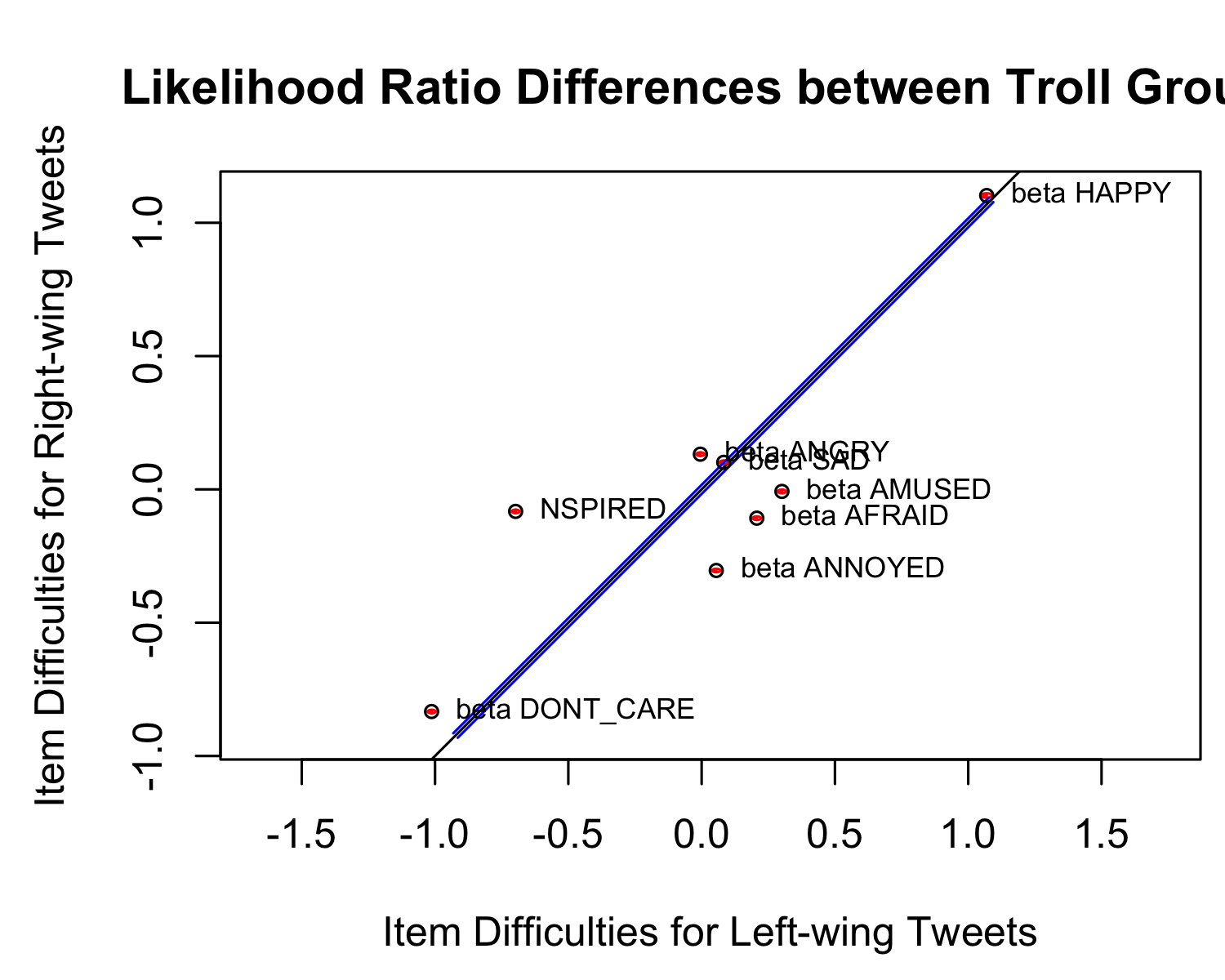
Items were generally well-matched to tweet scores (Figure 1). Right-wing tweets had slightly higher raw scores (M = 2.85, SD = 1.38) than Left-wing tweets M = 2.71, SD = 1.59; *t*(97,338.97) = -21.34, *p* < .001 ), though the difference was empirically small. In the Rasch model, *Don’t Care* was the easiest item with the lowest difficulty, while *Happy* was the item with the highest difficulty. The difficulties of other items were estimated to be relatively close to one another, indicating that the categories were used at relatively similar rates when both Left-wing and Right-wing trolls were combined.



*Figure* *1:*. Note that 5.09% of tweets had full zero/complete response patterns and are not included in the pictured person distribution.

### Left-wing and Right-wing Differences.

The Andersen’s likelihood ratio test showed significant differences between the Left-wing and Right-wing troll tweets ((7) = 9,768.27, p < .001; Figure 2. The items *Amused*, *Afraid*, and *Annoyed* were more difficult among Left-wing Tweets. *Inspired* was more difficult for Right-wing tweets. *Don’t Care* and *Angry* were also more difficult for Right-wing tweets, but to a much smaller degree.The items *Happy* and *Sad* were similarly used between both Left-wing and Right-wing trolls.



*Figure* *2:*. Items closer to the black identity line were estimated more similarly between Left-wing and Right-wing tweets. The 95% confidence intervals are shown in blue for the identity line and in red for the item difficulties.

## Individual Rasch Models

### Model Fit.

We fit 592 Rasch models for each of the individual twitter handles with at least 30 authored tweets. Among these models, there were 10 models with full zero responses for the category *Happy* and 1 model with full zero responses for the category *Afraid*. The average percentage of misfitting persons was 0.73% (SD = 0.82) for Left-wing accounts and 0.99% (SD = 0.91) for Right-wing accounts. Item misfit was also generally low (Table 2). The majority of items misfit in less than 6% of models, except for *Amused*, *Happy*, and *Inspired*. Overall, the measures of model fit were appropriate.

Table 2:

*Percent of Models with Misfitting Items*

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Group | Afraid | Amused | Angry | Annoyed | Don’t Care | Happy | Inspired | Sad |
| Left-wing | 2% | 7% | 2% | 4% | 2% | 9% | 11% | 6% |
| Right-wing | 2% | 13% | 0% | 6% | 1% | 7% | 4% | 3% |

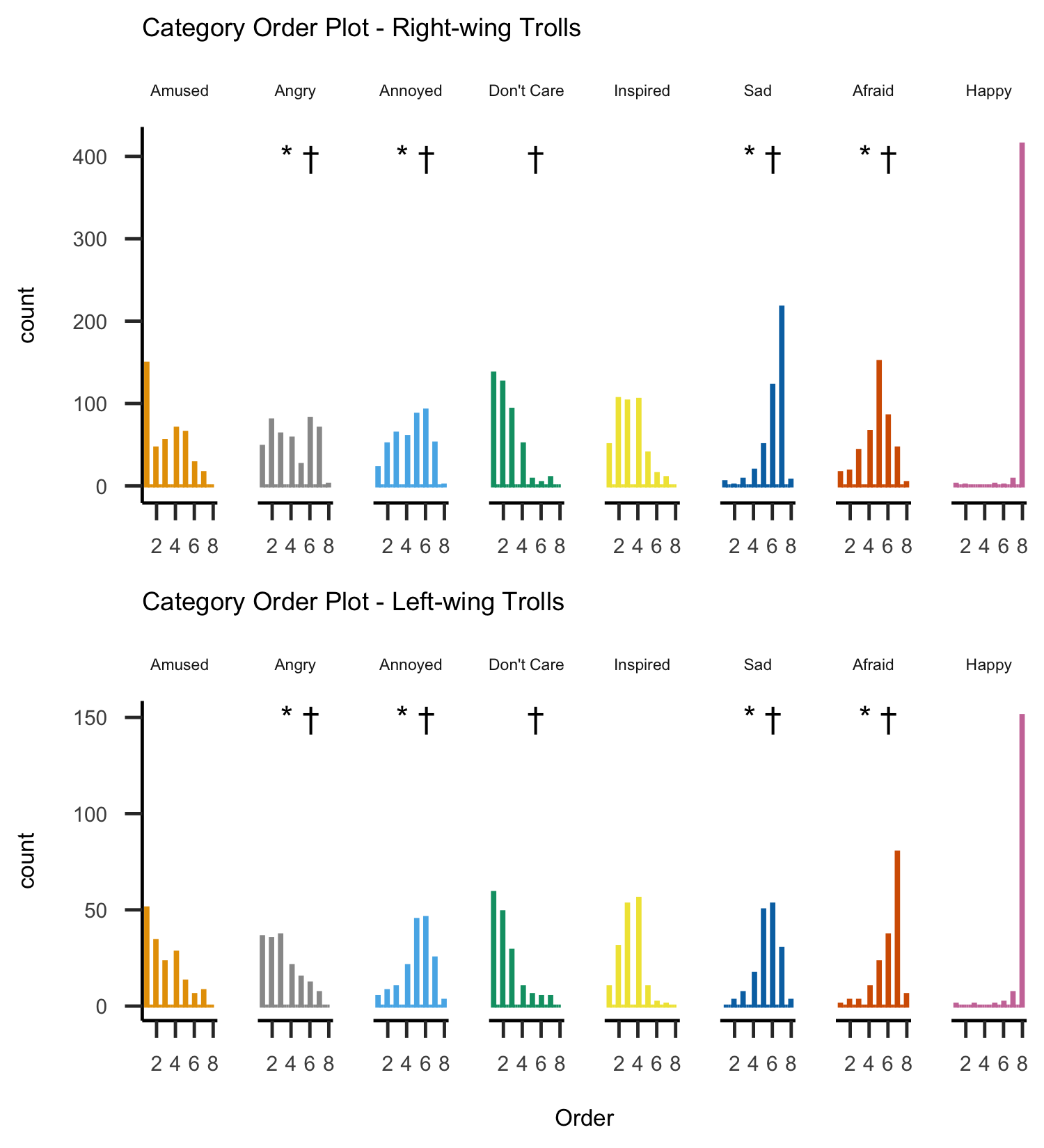
### Category Ordering.

Figure 3 displays the distributions of the order of item difficulties for both left-wing and right-wing troll accounts. Across both groups, the hardest category was *Happy* for 95.61% of accounts. There were significant mean and distributional differences between left-wing and right-wing trolls for the categories *Angry* (*t*(348.54) = 6.24, *p* < .001; D = 0.24, *p* = < .001), *Annoyed* (*t*(330.42) = -4.48, *p* < .001; D = 0.19, *p* < .001), *Sad* (*t*(282.93) = 6.50, *p* < .001; D = 0.32, *p* < .001), and *Afraid* (*t*(338.12) = -10.81, *p* < .001; D = 0.44, *p* < .001; see Table 3 for means and standard deviations). The category *Don’t Care* showed distributional differences (D = 0.05, *p* = 0.049), but not mean differences. There were no differences in the item difficulty orders for the categorys *Amused*, *Inspired*, or *Happy*.

Table 3:

*Means and Standard Deviations of Emotion Category Orders*

|  |  |  |
| --- | --- | --- |
| Category | Mean (SD) - Right-Wing | Mean (SD) - Left-Wing |
| Amused | 3.01 (1.86) | 2.80 (1.75) |
| Angry | 4.10 (2.07) | 3.05 (1.72) |
| Annoyed | 4.46 (1.75) | 5.12 (1.54) |
| Don’t Care | 2.34 (1.34) | 2.33 (1.51) |
| Inspired | 3.15 (1.41) | 3.21 (1.09) |
| Sad | 6.21 (1.15) | 5.50 (1.19) |
| Afraid | 4.81 (1.49) | 6.14 (1.28) |
| Happy | 7.92 (0.61) | 7.84 (0.76) |



*Figure* *3:*. The distributions of emotion categories within accounts when ordered from 1 (easiest) to 8 (hardest). indicates significant differences between left-wing and right-wing accounts based on *t* tests. indicates significant differences based on Kolmogorov-Smirnov distribution tests.

### Gini Index (Emodiveristy).

Following Ong et al. (2018)’s emodiversity method, we used the Gini coefficient to estimate the diversity in category usage between left-wing and right-wing trolls. High gini coefficient values indicate more diversity (more inequality) among emotion categories. We used the raw emotion scores rather than the Rasch model dichotomized scores. Left-wing trolls showed more inequality among emotion categories than right-wing trolls when the gini coefficient was calculated either across all tweets (left-wing = 0.64, right-wing = 0.60; (*t*(103,073.64) = 65.18, *p* < .001) or within authors (left-wing = 0.46, right-wing = 0.42; (*t*(413.76) = 7.02, *p* < .001).

### Multiinformation.

We calculated multiinformation (also known as total correlation) with bootstrapped confidence intervals for the category orders for the left-wing and right-wing accounts to understand the consistency in the orders within groups. Higher values of multiinformation signify higher entropy, that is, lower relationships between variables. For left-wing accounts, multiinformation was 7.05 (95% CI [6.34, 7.26], bias = 0.25); for right-wing accounts, multiinformation was 6.71 (95% CI [6.22, 6.62], bias = 0.29). We also bootstrapped the average multiinformation among a random ordering of the numbers 1 thorugh 8, resulting in a maximum value of multiinformation value of 10.58 (95% CI [10.02,10.16]). Both left-wing and right-wing trolls showed more cohesion in their category orders than would be expected by chance, but they did not differ from one another.

# Discussion

*Happy* was the hardest emotion category for both left-wing or right-wing trolls when examining the total number of tweets and the individual models.

Of note, the emotion categories with that showed misfit most often (*Amused*, *Happy*, and *Inspired*) were also the categories that generally did not show differences between Left-wing and Right-wing trolls.

## Interpretation

The differences in emotion expression may be linked to different methods of eliciting emotion. Different purposes, topics eliciting different emotions

It is critical to interpret the results of this study in context of its methods. Emotion is not a universal, context-free phenomenon; statements that are neutral in one context may elicit powerful reactions in another. The emotion categories delinated in the Depechemood++ lexicon were sourced from a specific time and group membership. When a tweet is rated highly on *Happy*, it does not imply that all readers of the tweet would have the same reaction. This may be even more true when dealing with highly politically and emotionally charged content that was written to elicit strong emotion reactions.

We propose that differences in emotion expression in this study should rightly be interpreted as differences in how one, perhaps neutral, subset of readers may react. We do not believe that these differences are not true or present in reality - rather that it is an overstep to assume that every person who reads a left-wing troll tweet is more likely to react with anger than if they read a right-wing troll tweet.

## Future Directions

## Limitations

## Conclusions

# Words removed from Lexicon

hilary hillary russia russian trump donald bernie sanders clinton rt obama barack america president black white racist

Araque, O., Gatti, L., Staiano, J., & Guerini, M. (2018). DepecheMood++: A bilingual emotion lexicon built through simple yet powerful techniques. Retrieved from [https://arxiv.org/abs/1810.03660](%20https://arxiv.org/abs/1810.03660)

Badawy, A., Ferrara, E., & Lerman, K. (2018). Analyzing the digital traces of political manipulation: The 2016 russian interference twitter campaign. In *2018 ieee/acm international conference on advances in social networks analysis and mining (asonam)* (pp. 258–265). IEEE.

Bail, C. A., Guay, B., Maloney, E., Combs, A., Hillygus, D. S., Merhout, F., … Volfovsky, A. (2020). Assessing the russian internet research agency’s impact on the political attitudes and behaviors of american twitter users in late 2017. *Proceedings of the National Academy of Sciences*, *117*(1), 243–250.

Barrett, D., Horwitz, S., & Heiderman, R. (2018). Russian troll farm, 13 suspects indicted in 2016 election interference. *Washington Post*.

Berger, J., & Milkman, K. L. (2012). What makes online content viral? *Journal of Marketing Research*, *49*(2), 192–205.

Bessi, A., Zollo, F., Del Vicario, M., Scala, A., Caldarelli, G., & Quattrociocchi, W. (2015). Trend of narratives in the age of misinformation. *PloS One*, *10*(8).

Boutet, A., Kim, H., & Yoneki, E. (2012). What’s in your tweets? I know who you supported in the [uk] 2010 general election. In *Sixth international aaai conference on weblogs and social media* (pp. 411–414).

Brady, W. J., Wills, J. A., Burkart, D., Jost, J. T., & Van Bavel, J. J. (2019). An ideological asymmetry in the diffusion of moralized content on social media among political leaders. *Journal of Experimental Psychology: General*, *148*(10), 1802–1813.

Brady, W. J., Wills, J. A., Jost, J. T., Tucker, J. A., & Van Bavel, J. J. (2017). Emotion shapes the diffusion of moralized content in social networks. *Proceedings of the National Academy of Sciences*, *114*(28), 7313–7318.

Broniatowski, D. A., Jamison, A. M., Qi, S., AlKulaib, L., Chen, T., Benton, A., … Dredze, M. (2018). Weaponized health communication: Twitter bots and russian trolls amplify the vaccine debate. *American Journal of Public Health*, *108*(10), 1378–1384.

Conover, M. D., Ratkiewicz, J., Francisco, M., Gonçalves, B., Menczer, F., & Flammini, A. (2011). Political polarization on twitter. In *Fifth international [aaai] conference on weblogs and social media* (pp. 89–96).

Dang-Xuan, L., Stieglitz, S., Wladarsch, J., & Neuberger, C. (2013). An investigation of influentials and the role of sentiment in political communication on twitter during election periods. *Information, Communication & Society*, *16*(5), 795–825.

Da Silva, N. F., Hruschka, E. R., & Hruschka Jr, E. R. (2014). Tweet sentiment analysis with classifier ensembles. *Decision Support Systems*, *66*, 170–179.

Del Vicario, M., Bessi, A., Zollo, F., Petroni, F., Scala, A., Caldarelli, G., … Quattrociocchi, W. (2016). The spreading of misinformation online. *Proceedings of the National Academy of Sciences*, *113*(3), 554–559.

Freelon, D., Bossetta, M., Wells, C., Lukito, J., Xia, Y., & Adams, K. (2020). Black trolls matter: Racial and ideological asymmetries in social media disinformation. *Social Science Computer Review*, 1–19.

Freelon, D., & Lokot, T. (2020). Russian twitter disinformation campaigns reach across the american political spectrum. *Harvard Kennedy School Misinformation Review*.

Garrett, R. K. (2009). Echo chambers online?: Politically motivated selective exposure among internet news users. *Journal of Computer-Mediated Communication*, *14*(2), 265–285.

Golino, H., Christensen, A. P., Moulder, R. G., Kim, S., & Boker, S. M. (2020). Modeling latent topics in social media using dynamic exploratory graph analysis: The case of the right-wing and left-wing trolls in the 2016 us elections. Preprint at https://psyarxiv.com/tfs7c.

González-Bailón, S., Banchs, R. E., & Kaltenbrunner, A. (2012). Emotions, public opinion, and us presidential approval rates: A 5-year analysis of online political discussions. *Human Communication Research*, *38*(2), 121–143.

Hong, S., & Kim, S. H. (2016). Political polarization on twitter: Implications for the use of social media in digital governments. *Government Information Quarterly*, *33*(4), 777–782.

Karlsen, R., Steen-Johnsen, K., Wollebæk, D., & Enjolras, B. (2017). Echo chamber and trench warfare dynamics in online debates. *European Journal of Communication*, *32*(3), 257–273.

Linacre, J. M. (1994). Sample size and item calibration stability. *Rasch Measurement Transactions*, *7*, 328.

Linacre, J. M. (1998). Detecting multidimensionality: Which residual data-type works best? *Journal of Outcome Measurement*, *2*, 266–283.

Linvill, D. L., Boatwright, B. C., Grant, W. J., & Warren, P. L. (2019). “THE RUSSIANS ARE HACKING MY BRAIN!” Investigating russia’s internet research agency twitter tactics during the 2016 united states presidential campaign. *Computers in Human Behavior*.

Linvill, D. L., & Warren, P. L. (2018). Troll factories: Manufacturing specialized disinformation on twitter. Retrieved from <http://pwarren.people.clemson.edu/Troll_Factories_v2_Linvill_Warren.pdf>

MacKuen, M., Wolak, J., Keele, L., & Marcus, G. E. (2010). Civic engagements: Resolute partisanship or reflective deliberation. *American Journal of Political Science*, *54*(2), 440–458.

Mair, P., Hatzinger, R., & Maier, M. J. (2020). *eRm: Extended Rasch Modeling*.

Meyer, P. E. (2008). *Information-theoretic variable selection and network inference from microarray data* (PhD thesis). Universite Libre de Bruxelles.

Ong, A. D., Benson, L., Zautra, A. J., & Ram, N. (2018). Emodiversity and biomarkers of inflammation. *Emotion*, *18*(1), 3–14.

Peng, T.-Q., Sun, G., & Wu, Y. (2017). Interplay between public attention and public emotion toward multiple social issues on twitter. *PloS One*, *12*(1).

Permanent Select Committee on Intelligence. (2018). Schiff statement on release of twitter ads, accounts and data. Retrieved from <https://democrats-intelligence.house.gov/news/documentsingle.aspx?DocumentID=396>

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Oxford, England: Nielsen & Lydiche.

Richey, M. (2018). Contemporary Russian revisionism: Understanding the Kremlin’s hybrid warfare and the strategic and tactical deployment of disinformation. *Asia Europe Journal*, *16*(1), 101–113.

Schmidt, A. L., Zollo, F., Del Vicario, M., Bessi, A., Scala, A., Caldarelli, G., … Quattrociocchi, W. (2017). Anatomy of news consumption on facebook. *Proceedings of the National Academy of Sciences*, *114*(12), 3035–3039.

Staiano, J., & Guerini, M. (2014). Depechemood: A lexicon for emotion analysis from crowd-annotated news. *arXiv Preprint arXiv:1405.1605*.

Stewart, L. G., Arif, A., & Starbird, K. (2018). Examining trolls and polarization with a retweet network. In *Proceedings of wsdm workshop on misinformation and misbehavior mining on the web*.

Suhay, E., & Erisen, C. (2018). The role of anger in the biased assimilation of political information. *Political Psychology*, *39*(4), 793–810.

Törnberg, P. (2018). Echo chambers and viral misinformation: Modeling fake news as complex contagion. *PLoS One*, *13*(9).

Valenzuela, S. (2013). Unpacking the use of social media for protest behavior: The roles of information, opinion expression, and activism. *American Behavioral Scientist*, *57*(7), 920–942.

Wollebæk, D., Karlsen, R., Steen-Johnsen, K., & Enjolras, B. (2019). Anger, fear, and echo chambers: The emotional basis for online behavior. *Social Media+ Society*, *5*(2), 257–273.

Yardi, S., & Boyd, D. (2010). Dynamic debates: An analysis of group polarization over time on twitter. *Bulletin of Science, Technology & Society*, *30*(5), 316–327.

Zannettou, S., Caulfield, T., De Cristofaro, E., Sirivianos, M., Stringhini, G., & Blackburn, J. (2019). Disinformation warfare: Understanding state-sponsored trolls on twitter and their influence on the web. In *Companion proceedings of the 2019 world wide web conference* (pp. 218–226).