Tell me the Truth: Separating Fact from Fiction in Social Media Following Extreme Events

Katie Byrd
University of Southern California
ksippel@usc.edu

Richard John
University of Southern California
richardj@usc.edu

Abstract

With increased reliance on social media to spread important information during extreme events, users' reported inability to distinguish fact from fiction is a growing concern. This experiment (N=398) tests whether feedback training improves performance in identifying true and false social media content during extreme events. Respondents completed two sets of 16 binary classification judgments (true or false) of actual social media posts following either natural disasters or soft-target terror attacks. Respondents randomly assigned to the feedback training condition received feedback after each of the 16 training judgments, while those assigned to the control condition did not receive any feedback following the training judgments. Feedback training did not increase social media content classification performance for either natural disasters or soft-target terror events. Individuals' performance for correctly identifying false content was negatively related to self-identified political conservatism and was positively related to a measure of cognitive reflection.

1. Introduction

As of January 2020, there were over 3.8 billion social media users worldwide, with 321 million of those users joining since January 2019 [1]. On average, each of those social media users spends 142 minutes on social media daily [2], with 55% of U.S adults in 2019 reporting that they receive news from social media either "often" or "sometimes," an 8% increase from the previous year [3]. While the use of social media tends to be focused on younger generations, social media use is not restricted to one particular age group. As of February 2019, 90% of people in the US between the ages of 18-29, 82% of people between the ages of 30-49, 69% of people between the ages of 50-64, and 40% of people 65 and older used at least one form of social media [4].

Thus, social media plays a pivotal role in information distribution. While it is important to have accurate information in all settings, in certain contexts, such as extreme events (e.g. natural disasters or soft-target terror attacks), having accurate information can be even more vital. Social media has a unique advantage in that it allows for instant information to be shared from essentially anyone or anywhere.

Thus, providing an avenue for quick communication to a potentially large number of people. In the case of extreme events this can be particularly useful for government agencies or first response entities who are either trying to share information and resources to the population, or gather information from individuals in need [5,6,7].

Unfortunately, for the same reasons that social media can be beneficial in extreme events, it can also be detrimental. Due to the capability of instant communication that can be easily shared, false information can be spread rapidly, and without the opportunity to confirm accuracy. While the spread of false information in certain contexts might be relatively harmless, false information can be detrimental to individuals in need or to first responders. For example, false information about a bomb location or a second gunman could be used to lure individuals from one location to another, or cause unnecessary panic that could lead to injuries and deaths.

Thus, during extreme events, it is important that individuals be able to distinguish between true and false information on social media. Considering the wide-spread use of social media, and recent findings that individuals perform poorly at distinguishing between true and false information on social media [8], any method shown to improve performance would have immediate application to all on social media.

1.1 Current study

Previous research has found that respondents do not perform much better than chance (rarely exceeding 60%) at correctly identifying true and false information on social media during extreme events based on the content of the posts [8], a finding that is consistent with other deception detection research. In

URI: https://hdl.handle.net/10125/70946
978-0-9981331-4-0

addition, research has found certain individual characteristics that are related to deception detection performance, both in political contexts [9,10] and in the context of extreme events [8].

Thus, it is important to investigate whether there are ways to improve performance. While there is research related to improving deception detection on social media, this research is mostly directed toward developing computer algorithms to classify social media posts as true or false.

Existing research uses large datasets to train computer programs to identify and filter false information, mostly by identifying IP addresses, source information, sharing patterns, and other non-content aspects of the post. These programs typically ignore the content of the social media posts themselves and rely heavily on other external factors.

While these computer programs are becoming quite efficient at identifying false information, all of the information is not identified or not identified immediately, which in the case of extreme events could be detrimental due to the need for instant communication. Thus, having a way to improve users' abilities to identify false information could be highly beneficial during extreme events, yet little research has been done.

This study seeks to search for ways to increase individuals' ability to correctly identify information in social media during extreme events, through the use of feedback training. This study will determine not only whether overall ability to identify information can be improved, but also whether the ability to identify true information or the ability to identify false information can be improved through feedback learning. During active extreme event situations individuals might not have the time to wait for a computer program to identify false information or the time to research the source.

Therefore, it is vital that the individual better judge the information based on the content of the social media post. Additionally, this study aims to further explore whether certain cognitive and individual characteristics are predictive of classification performance by determining whether previously supported theories of potentially vulnerable subgroups in the population are found in this study.

By determining whether the same characteristics found in previous studies are predictive in this study as well, it could confirm that these subgroups are consistently vulnerable across social media domains. Such a finding would indicate a need for further research to identify new cognitive strategies to enhance truth detection by these vulnerable subgroups.

1.2 Previous false news research

In recent years there has been in increase in research related to false news across various domains. One domain of false news research is centered on developing computer algorithms that can detect and filter false information in social media [11, 12, 13, 14, 15], with one having shown an accuracy of 88% within 24 hours [16]. Other research is focused on how to classify false information [17], how false information spreads through social media [18, 19], factors that affect social media post credibility [20], and sharing patterns related to how users interact with false information [21, 22, 23].

An additional domain of false news research examines whether certain individual characteristics are associated with false news detection performance. For example, several studies have found political orientation to be a predictor of false information detection performance [24, 9, 8]. Particularly, in false news related to politics, self-identified conservatives have been shown to perform worse at correctly identifying false information in social media when compared to self-identified liberals [9].

In the context of extreme events, namely natural disasters and soft-target terror attacks, self-identified political orientation was found to have a moderating effect within extreme event contexts [8]. Specifically, as self-identified political conservatism increased, the ability to correctly identify false social media posts related to soft-target terror events increased, while the ability to correctly identify false social media posts related to natural disasters decreased.

Additionally, cognitive measures have been found to be related to the ability to correctly identify false information in social media. In particular, delusion-prone individuals were found to be more likely to believe false news, likely due to analytic cognitive style [25], while cognitive reflection (CRT) was found to be positively related to false information identification [9].

Previous research also found that the perceived accuracy of the information in a post can be increased through just a single prior exposure to the post [10]. This was found to be especially true in individuals with relatively low cognitive ability [26]. Previous research has not found any consistent individual differences (e.g. age, sex, race) related to the ability to correctly identify true information in social media.

This study seeks to confirm whether there are similar relationships between cognitive functioning, self-identified political ideology, or other individual characteristics with ability to correctly identification information as true or false in the context of social media posts related to extreme events.

1.3 Previous deception detection research

Previous research related to deception detection has shown that individuals' ability to detect deception during in-person communication is only slightly better than chance, with performance rarely exceeding 60% [27, 28], even for professionals with deception detection experience [29, 30]. Recent research has extended this finding to include social media deception detection performance in the context of extreme events [8]. Additionally, previous research has found confidence and accuracy to be mostly unrelated in deception detection [31, 32].

In response to low observed deception detection performance, recent research has explored different training techniques in an attempt to improve performance. Overall, deception detection training research has shown mixed results, with some studies finding training to be effective [33, 34, 35], while others finding a minimal to zero effect [36]. Various researchers have given suggestions on how to test training accuracy, as well as the best training trainings to use. Levine et al. [37] suggests that researchers include an additional training group that does not receive accurate deception training to test for the placebo effect of training. While some researchers suggest that deception training should last at least one hour to be effective [34], other researchers worry that too much time and training can lead to lower accuracy due to boredom [38].

In addition, different studies take different approaches to deception training. Some studies aggregate common deception cues and provide them to individuals before they make their judgments [37, 39, 40], some provide feedback on the correct response [41, 42], and others combine both training types [33, 43]. A meta-analysis by Driskell [35] found an overall effect of type of training. Studies that used both feedback training and information training were found to be the most effective, followed by feedback only training, then information only training. The combined training approach and feedback only training approach were found to have similar effects, and both found to be significantly better than the information only training approach.

Since the results of studies using a feedback only training approach are similar to the results of studies using a combined training approach, the current study utilizes a feedback only approach in order to minimize training time in an attempt to reduce effects of boredom and fatigue. This study aims to expand previous deception detection training research to the realm of social media in the context of extreme events.

2. Hypotheses

In this study respondents were asked to make binary judgments on whether a social media post is true or false and then report their confidence rating on that judgment. Before beginning, respondents were explicitly informed that the base rate of true social media posts and false social media posts that they are judging is approximately equal. Likewise, they were informed that the error penalties of a false positive (i.e. judging a post to be true when it is actually false) and a false negative (i.e. judging a post to be false when it's actually true) are equal. By specifying the base rate (proportion of true/proportion of false =1) and error penalty (false-positive error/false-negative error =1) an optimal threshold, in terms of maximizing expected value, was specified allowing us to utilize a Signal Detection Theory (SDT) framework.

By using SDT the classification task can be assessed independent of the base-rated and error-penalties either embedded in the task, or assumed by the participant if the base rated and error penalties are not specified. Utilizing the SDT framework, the corresponding Area Under the Curve (AUC) values can be calculated to assess performance. AUC values are independent of base-rates and assumed error penalties, thus providing a more accurate measure of performance than typical performance measures. A variety of previous deception detection research has utilized SDT [44, 45], including research detecting phishing attacks [46, 47], and deception in social media [8].

This study seeks to address a number of new research questions and hypotheses that have not yet been studied in the context of social media and extreme events, in addition to confirming and extending results found in recent studies. With the exception of [8], other similar false news research, mainly in the realm of political false news, has neglected the effects that base-rates and error penalties have, by either not informing respondents of the base rate and error penalties, not assessing perceived base rate and error penalties, or not accounting for base-rates and error penalties in the evaluation of performance. The current study addresses the following research questions.

1. Does outcome feedback remove noise from the truth signal? That is, does providing feedback training improve performance? This will be evaluated in this study by addressing the following questions.

A. Does feedback increase AUC values? Previous studies have reported mixed results on how effective feedback training is at improving deception detection; previous research suggests that respondents do not perform very well at correctly identifying information in social media following extreme events in terms of AUC. Since this is an exploratory study, we make no prediction regarding the effect feedback will have on AUC values.

B. Does feedback increase the ability to correctly identify true posts? In previous studies, sensitivity (i.e. the probability of saying true when the post is true) has been found to be relatively stable across individuals; thus, unlike specificity (i.e. the probability of saying false when the post is false), there does not appear to be many factors that predict sensitivity. In addition, people tend to be better at identifying true information compared to false information. Hence, we do not expect sensitivity will increase due to feedback.

C. Does feedback increase the ability to correctly identify false posts? Previous research has demonstrated that specificity varies considerably across individuals, with certain demographic and cognitive measures being predictive of performance. Additionally, performance on correctly identifying false information is relatively low. Thus, we expect that if feedback training were to be effective, it would most likely be manifest through improved specificity performance.

3. Are there any cognitive variables related to specificity or sensitivity? Considering results from previous false news research we hypothesize that the cognitive measures (CRT, skepticism, and conscientiousness) will be positively related to specificity.

4. Are there any individual characteristics (e.g. age, sex, political ideology) related to specificity or sensitivity? Considering the results from previous false news research which found political ideology to be related to specificity, we hypothesize that a relationship between self-identified political ideology and specificity will be present.

3. Methods

Two separate US based samples (N=204 for soft-target terror attacks, N=194 for natural disasters) were recruited through Amazon Mechanical Turk for a total N=398. Respondents each made 32 binary judgments on the accuracy of actual social media posts, and reported confidence in their judgment. The social media posts were taken from four scales constructed prior to this study, related to either soft-target terror attacks or natural disasters that took place in the US between 2016 and 2019.

Respondents were randomly assigned to either the feedback condition or the control condition. Respondents in both conditions made 16 training judgments and 16 trial judgments. For the 16 training judgments, respondents assigned to the feedback condition were given feedback after each judgment as to whether their judgment was correct or incorrect. Respondents in the control condition did not receive any feedback during the 16 training judgments.

In each set of 16 judgments there were 8 true posts and 8 false posts. To control for order effects and potential differences between sets of social media posts, both conditions (feedback/control) were partitioned into two groups, and the sets of social media posts for feedback vs. control groups were counterbalanced.

After completing all 32 judgments, respondents completed three psychometric measures and 7 demographic questions. Responses were recorded using Qualtrics.com (http://www.qualtrics.com). Summary statistics, search engine keywords, and Item Characteristic Curves (ICC) are available in the supplementary materials.

3.1 Social Media Post Selection

Using Snopes.com (http://www.snopes.com) and Twitter's advanced search interface (http://www.twitter.com), a collection of 80 posts, 20 true and 20 false for both US natural disasters and soft-target terror attacks, was constructed. To identify social media posts related to recent extreme events, only natural disasters and soft-target terror attacks that occurred in the US between 2016-2019 were considered (e.g. Hurricane Harvey, Orlando nightclub shooting). When searching Snopes.com and Twitter's advanced search interface, keywords related to each event were used.

Using Twitter's advanced search interface, posts containing non-verifiable facts or non-pertinent information (e.g. opinions, prayers, personal remarks, sympathies), as well as reposts and retweets, were excluded. True social media posts selected through Twitter's advanced database were taken from credible media sources (e.g. Government agencies, local police, NBC, CNN). Social media posts selected from Snopes.com were limited to only posts verified as either "True" or "False". Posts that Snopes.com identified as being something other than "True" or "False" (e.g. "Mostly True," "Mostly False," "Mixture," or "Unproven") were not considered for this study since the respondents were asked to make binary judgments. A full list of Snopes.com ratings and definitions can be found at www.snopes.com/fact-check-ratings/.

Prior to this study, two separate groups of human respondents, N=205 for natural disasters and N=197 for soft-target terror attacks, were recruited through Amazon Mechanical Turk (Mturk), for a total sample of N=402, to construct performance metrics measuring correct identification of true information and correct identification of false information, separately for soft-target terror attacks and natural disasters. Using Item Response Theory (IRT), four scales were constructed with 16 social media content items each, exhibiting a broad range of difficulty item parameters and generally moderate to high discriminability item parameters.

The four scales were used to measure truth and deception ability in this study. Each of the four scales (true/false, soft-target terror/natural disasters) has 16 social media posts for a total of 64 social media posts utilized. Each of the four groups of 16 was then separated into two sets of 8 each, one set of 8 for the training judgments and one set of 8 for the trial judgments, referred to later as set A and set B. These sets were determined using comparable difficulty and discriminability values.

3.2 Respondents

Recruited through Mturk, 398 adult US respondents participated in the study. In addition to the task, respondents self-reported their age, race, sex, education, income, and political orientation, as well as three psychometric measures of conscientiousness, skepticism, and cognitive reflection.

3.3. Measures

In this study Cognitive Reflection (CRT) scores were calculated by summing the total number of correct item responses on a 7-item CRT scale. This same measure of Cognitive Reflection has been used in similar false news research and has been found to be a predictor of false information identification accuracy [9,10]. The following is an example item, "In a lake, there is a patch of lily pads. Every day, the patch doubles in size. If it takes 48 days for the patch to cover the entire lake, how long would it take for the patch to cover half of the lake?" while the fast incorrect "intuitive" response is 24 days, the correct answer is 47 days.

In this study skepticism was measured using the 30-item Professional Skepticism Scale [48]. Each participant rated the extent to which they agreed with each of the 30 item statements using a 6-point Likert-scale. In this 30-item scale there are items related to search for knowledge, interpersonal understanding, suspension of judgment, a questioning mind, and self-confidence. Skepticism scores were then calculated by averaging the responses across all 30 items; higher scores indicate higher levels of skepticism.

In this study conscientiousness was measured using a 9-item conscientiousness subscale of the Big Five Inventory [49]. Using a 5-point Likert scale, respondents indicated level of agreement with each of nine statements. Conscientiousness scores were then calculated by averaging responses across all 9 items. In this study higher scores indicate higher levels of conscientiousness.

Additionally, sensitivity and specificity values were also calculated for each participant. For this study sensitivity refers to the participant's ability to correctly identify true information as true. This measure was calculated as the proportion of true social media posts correctly identified as true by the participant. Specificity in this case represents the participant's ability to correctly identify false information as false. Specificity scores were calculated as the proportion of false social media posts that the individual correctly identified as false. For each participant, three separate sets of sensitivity and specificity values were calculated: a training sensitivity and specificity, a trial sensitivity and specificity, and a total sensitivity and specificity. Mean scores for each of the cognitive variables is shown in Table 1.

Table 1. Mean psychometric scores. (C=Control, F=Feedback)

Psychometric Scale	Soft Target Terror Attacks		Natural Disasters	
	C	F	C	F
Cognitive Reflection (CRT)	3.50	3.80	3.87	3.35
Skepticism	4.45	4.40	4.51	4.36
Conscientiousness	3.93	3.92	3.93	4.01

3.4. Procedure

Respondents were randomly assigned to either the feedback condition or the control condition. Respondents in both conditions were shown 16 social media posts as training judgments, followed by 16 social media posts as trial judgments, and asked to make a binary judgment of true or false for each post followed by their confidence rating. Confidence ratings were reported on a 5-point rating scale from 1 Not at all confident to 5 Extremely confident. Both the feedback group and the control group were randomly partitioned into two groups.

One group received set A for the training judgments and set B for the trial judgments, while the other group received set B for the training judgments and set A for the trial judgments. Respondents were informed that approximately half of the social media posts were true and half false.

Respondents were informed they would receive $0.50 for their participation, regardless of their performance. Additionally, they were incentivized with an additional bonus of up to $8.00 based on their performance on the 16 trial judgments. Respondents were informed they would receive an additional $0.50 for each correct response, and be penalized $0.50 for each incorrect response. The total bonus awarded was determined by summing the amount from the last 16 (trial) judgments only.

To minimize cheating and thoughtless responses, each social media post presented was displayed for a minimum of 10 seconds and a maximum of 30 seconds. After completing all 32 judgments respondents completed the CRT, skepticism, and conscientiousness items and were asked to self-report their age, sex, ethnicity, education, income, and political orientation.

4. Results

Results for the study are presented in three sections: (1) Evaluating the effect of feedback training on performance using Signal Detection Theory (SDT) metrics; (2) Evaluating the effect of feedback training on performance in relation to sensitivity/specificity using Bayesian regression models; and (3) Examining relationships between cognitive/ demographic variables and specificity/sensitivity using Bayesian regression models. All analyses were completed separately for terror events and natural disasters.

4.1. Signal Detection Theory Analysis

To estimate an ROC curve a stimulus strength is needed. To create a continuous range of information strength on the horizontal axis, each of the confidence ratings for the 32 judgments was unfolded. While confidence ratings were reported on an ordinal scale, Byrd & John [8] found that an ordinal confidence rating and a continuous probability of correct judgment converged for this task. Each of the confidence ratings was coded 1 (Extremely confident) to 5 (Not at all confident) for the social media posts that they judged to be false, and 6 (Not at all confident) to 10 (Extremely confident) for the social media posts that they judged to be true.

Using the information strength values, Receiver Operating Characteristic (ROC) curves were constructed separately for soft-target terror events and natural disasters using the ROC procedure in SPSS version 25. The ROC curves, sensitivity values, and specificity values were similar within each condition across the counterbalanced social media posts order; thus, for the remainder of the analyses the counterbalanced groups are aggregated. Using the ROC curves, AUC values were calculated. Table 2 displays the AUC estimates for the training judgments and the trial judgments separated by the control condition and the feedback condition for both the soft-target terror attacks and natural disasters.

Overall, feedback training did not consistently improve AUC values. In the natural disaster context, the AUC value for the feedback condition increased between the training judgments and the trial judgments, while the AUC value for the control condition decreased between the training judgments and the trial judgments. However, the opposite pattern was observed for the soft-target terror attack context, where AUC values for the feedback condition decreased between the training judgments and the trial judgments, while increasing in the control condition.

In general, the AUC values for the soft-target terror attack condition were considerably higher than those of the natural disaster condition. Considering the AUC value for the training judgments for the natural disaster feedback condition was .600 (SE = .013) compared to .713 (SE=.013) for soft-target terror attacks, one possibility is that feedback did improve performance for natural disasters but not soft-target terror attacks. One reason for this possible explanation is a ceiling effect; performance for the feedback soft-target terror condition was already high to begin with. An AUC of .600 allows more room for improvement than a substantially higher AUC of .713

Table 2. Area Under the Curve (AUC) values.

Condition	Information Stimulus	Terror: AUC	Natural Disaster : AUC
Control	Training	.670***	.637***
	Trial	.697***	.629***
Feedback	Training	.713***	.600***
	Trial	.706***	.623***
Average		.697	.622

*** significantly different from 0.5, p < 0.001

4.2. Feedback Training Bayesian Regression Models

While feedback appears not to have improved overall performance in terms of AUC values, previous research has found that the ability to correctly identify true information (sensitivity) and correctly identify false information (specificity) are separate abilities. By utilizing Bayesian regression models, we can determine whether feedback training improves either sensitivity or specificity separately. Two regression models were used for both soft-target terror attacks and natural disasters to predict the effect of feedback during the training judgments on sensitivity and specificity estimates in the trial judgments. In the models, the training sensitivity or specificity value was used as a covariate in the model to account for previous individual and group differences, non-informative priors were used.

The results of the sensitivity and specificity models for natural disasters and soft-target terror events are shown in Table 3. The models show that outcome feedback did not have an effect on trial judgment sensitivity and specificity values when covarying out sensitivity and specificity values from the 16 training judgments. Thus, receiving feedback did not improve ability to better distinguish between either true information or false information in social media posts following either extreme event context.

Table 3. Bayesian Regression results for feedback on sensitivity and specificity.

Natural Disaster	Specificity	Sensitivity
Intercept	0.22*	0.34*
Specificity Training	0.63*	
Sensitivity Training		0.48*
Feedback	0.00	0.03
R^2	.37	.21

Terror	Specificity	Sensitivity
Intercept	0.14*	0.33*
Specificity Training	0.82*	
Sensitivity Training		0.55*
Feedback	0.01	-0.01
R^2	.37	.21

* Coefficient different from 0.0, p < 0.05

4.3 Individual Differences Bayesian Regression Models

Again, sensitivity and specificity are evaluated separately in the individual difference variables models for both extreme event contexts. The four regressions predict overall sensitivity and specificity and include feedback training, six demographic variables (i.e. sex, education, race, income, age, political ideology) as covariates, and three cognitive functioning measures. The results of the models for both natural disasters and soft-target terror events are shown in Table 4.

Self-identified political conservatism was found to be negatively related to specificity for both the terror and natural disaster context. In this study, political orientation is measured on a self-identified 7-point scale from 1 (extremely liberal) to 7 (extremely conservative), therefore as respondents increase on the political ideology scale towards political conservatism, specificity decreases. While this finding is different from the moderating effect extreme event context (natural disasters and soft-target terror attacks) had on political orientation's relationship with skepticism found in Byrd & John [8], this result is similar to the effect found in previous false news research in the context of politics [9,10].

Regarding cognitive functioning, CRT was found to be positively related to specificity for both terror and natural disasters, with the effect of CRT for the terror condition being double the effect of CRT for the natural disaster condition. This result is consistent with previous false news research [9,10]. Conscientiousness and skepticism were both positively related to specificity for the soft-target terror attack condition. Additionally, skepticism was also positively related to sensitivity for soft-target terror attacks.

Table 4. Individual differences regression results.

Natural Disasters	Specificity	Sensitivity
Intercept	0.29 [0.05;0.52]	0.72 [0.49; 0.94]
Feedback	0 [-0.06;0.05]	0.02 [-0.03; 0.07]
Sex	0.02 [-0.04;0.08]	-0.05 [-0.10; 0.00]
Education	-0.06* [-0.12;-0.00]	0.05* [0.00; 0.10]
Race	-0.06 [-0.12;0.00]	-0.04 [-0.10; 0.01]
Income	0 [-0.02;0.02]	0 [-0.02; 0.01]
Age	0 [-0.00;0.01]	0 [-0.00; 0.00]
Political	-0.02* [-0.04;-0.00]	0 [-0.02; 0.02]
CRT	0.02* [0.01;0.03]	0 [-0.01; 0.02]
Conscientiousness	0.01 [-0.03;0.06]	-0.02 [-0.06; 0.02]
Skepticism	0.04 [-0.02;0.09]	0.03 [-0.02; 0.08]
R^2	0.26	0.12

Terror	Specificity	Sensitivity
Intercept	-0.15 [-0.41; 0.10]	0.57 [0.39; 0.75]
Feedback	0.03 [-0.03; 0.09]	-0.01 [-0.06; 0.03]
Sex	0.02 [-0.04; 0.08]	0 [-0.04; 0.05]
Education	-0.02 [-0.08; 0.05]	0.04 [-0.00; 0.09]
Race	-0.02 [-0.10; 0.04]	-0.03 [-0.07; 0.03]

Income	0 [-0.02; 0.02]	0 [-0.01; 0.02]
Age	0 [-0.00; 0.01]	0 [-0.00; 0.01]
Political	-0.02* [-0.05; -0.01]	0 [-0.02; 0.01]
CRT	0.04* [0.03; 0.06]	0 [-0.01; 0.01]
Conscientiousness	0.05* [0.00; 0.10]	-0.04* [-0.08; -0.00]
Skepticism	0.12* [0.06; 0.17]	0.06* [0.02; 0.10]
R^2	0.47	0.12

* Coefficient different from 0.0, p < 0.05

5. Discussion

The average AUC values in this study (AUC=.697 soft-target terror events, AUC=.622 natural disasters) are substantially higher than the AUC results reported in previous studies (AUC=.617 terror events, AUC=.521 natural disasters) [8]. This likely reflects the difference in using a set of social media posts that form a reliable measure of truth detection ability, compared to a set of posts not selected for difficulty and discriminability.

One common predictor found throughout social media false information research is political ideology. In the context of political social media posts, political conservatism has been found to be a negative predictor of false information performance [9, 10], while in the context of extreme events political ideology has been found to be a moderating variable, with political conservatism being positively related to false information performance in the context of soft-target terror attacks and negatively related in the natural disaster context [8]. This study, however, did not find the same moderating effect between extreme events. Similar to research on political false information, this study found political conservatism to be negatively related to specificity in both extreme event contexts. Respondents who identified as more conservative performed worse at correctly identifying false information for both soft-target terror attacks and natural disasters. Ability to correctly identify true information (sensitivity) was not found to be related to political ideology in either extreme event context, which is consistent with previous literature.

Consistent with previous research [9, 25] we found the cognitive reflection (CRT) to be positively related to skepticism (i.e. the ability to correctly identify false information) across both extreme event contexts, providing evidence that the more reflective one is in her cognitive thinking, the better she performs in identifying false information in social media across a variety of different contexts.

The other cognitive measures, skepticism and conscientiousness, were also found to be positively related to specificity, but only in the soft-target terror attack context. Additionally, in the soft-target terror context, we found a positive relationship between skepticism and sensitivity. The lack of these findings in the natural disaster context could imply that this relationship is context dependent. In other words, this relationship is not consistent across every extreme event, but is instead dependent on the context.

6. Limitations

The aim of this study was to determine whether feedback training increases performance in classifying true and false information in social media following extreme events based on the content of the social media post alone. During an extreme event, individuals most likely do not have the luxury of being able to fact check information or verify the source because information is either needed quickly, or cannot be fact checked because there is no credible source available. Additional information about the posts was excluded, and the time constraints included, to isolate truth detection based on content only.

Additionally, in extreme event contexts, it is more common for information to come from unfamiliar sources (e.g. ones that are close to the event) than from familiar sources. Even information from familiar sources may simply be copied from posts from unfamiliar sources. Further, previous research has found the source of social media posts to be unrelated to correct identification performance [9]. However, the inclusion of such information could potentially affect performance, so the results of this study should be interpreted with that in mind. While numerous studies have shown MTurk samples to be valid for research [50, 51] due to the nature of online samples there are potentially uncontrolled variables that could contribute to performance. Additionally, the time constraint placed to avoid cheating could contribute to performance differences.

In addition, this study was restricted to 16 feedback judgments. Therefore, this study makes no claim that more extensive feedback training would not be effective. The results of this study are related to the effect that 16 feedback training trials had on helping individuals improve their truth detection skills based on the content of the social media posts alone. Additionally, sensitivity and specificity estimates were measured using only eight items, it's possible that these measures may require more than eight judgments to reliably estimate.

7. Future Research

While previous research has developed machine-learning algorithms for social media posts, these algorithms have limitations in regard to utilizing content features of the posts. Recent research has begun utilizing critical thinking guidelines to improve false information detection in social media [52]. Future research should explore combining feedback training with a critical thinking intervention. Future research should also seek to combine both machine-learning algorithms and human judgment. Human decision makers may provide unique insight into the content of the social media posts addressing some limitations that computer algorithms alone have. In a similar way, computer algorithms could be used to evaluate non-content attributes of social media posts. By combining both machine-learning algorithms and human judgment, classification performance could be enhanced beyond that possible by either alone.

8. Conclusion

This study extends previous deception detection feedback training to the context of social media training in relation to extreme events in the US. In this study, feedback training was not found to increase ability to correctly identify true and false information in social media posts following natural disasters or soft-target terror events. Future research should investigate other training methods to improve correct information identification in social media.

Two individual difference predictors were found to be consistent across both extreme event contexts, the cognitive reflection (CRT) and self-identified political ideology. Both of these predictors were found to be related to ability to correctly identify false information (specificity). As an individual's cognitive reflection increased, so did their ability to correctly identify false information. Additionally, political conservatism was associated with decreased performance in identifying false information in social media. Overall, those most challenged to detect truth in social media posts following extreme events are those with low cognitive reflection and those who self-identify as more politically conservative.

9. References

[1] S. Kemp, "Digital 2020: 3.8 Billion People Use Social Media," TheNextWeb, 2020.

[2] S. Salim, "How Much Time Do You Spend on Social Media? Research says 142 Minutes Per Day," Digital Information World, 2019.

[3] P. Suciu, "More Americans are Getting Their News from Social Media," Forbes, 2019.

[4] J. Clement, "Share of U.S. Adults Who Use Social Media 2019, by Age," Statista, 2019.

[5] J.B. Houston, J. Hawthorne, M.F. Perreault, E.H. Park, M. Goldstein Hode, M. R…, & S.A. Griffith, "Social media and disasters: a functional framework for social media use in disaster planning, response, and research," Disasters, 39(1), 2015, pp. 1-22.

[6] M. Keim, & E. Noji, "Emergent use of social media: a new age of opportunity for disaster resilience," American Journal of Disaster Medicine, 6(1), 2011, pp. 47-54.

[7] P. Jaeger, B. Shneiderman, K. Fleischmann, J. Preece, Y. Qu, P. Wu, "Community response grids: E government, social networks, and effective emergency management," Telecommunications Policy, 31(10- 11), 2007, pp. 592-604.

[8] K. Byrd, & R. S. John, "Lies, Damned Lies, and Social Media Following Extreme Events," (Unpublished manuscript under review).

[9] G. Pennycook, D. G. Rand, "Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning," Cognition, 188, 2019, pp. 39-50.

[10] G. Pennycook, T. Cannon, D. G. Rand, "Prior exposure increases perceived accuracy of fake news," Journal of Experimental Psychology: General, 147(12), 2018, pp. 1865-1880.

[11] V. L. Rubin, "Deception detection and rumor debunking for social media," In The SAGE Handbook of Social Media Research Methods, 2017, p. 342.

[12] S. Hamidian, M. Diab, "Rumor detection and classification for Twitter data," Presented at (SOTICS), 2015.

[13] R. Nieva, "Google fights fake news in search with 'fact check' tags," CNET, 2017.

[14] M. Snider, "Facebook aims to filter more fake news from news feeds," USA Today, 2017.

[15] J. Vanian, "Facebook, Twitter take new steps to combat fake news and manipulation," Fortune, 2018.

[16] K. Wu, S. Yang, K. Zhu, "False rumors detection on sina weibo by propagation structures," IEEE 31st international conference on data engineering, 2015, pp. 651-662.

[17] A. Zubiaga, A. Aker, K. Bontcheva, M. Liakata, R. Procter, "Detection and resolution of rumours in social media: A survey," ACM Computing Surveys (CSUR), 51(2), 2018, pp. 1-36.

[18] S. Vosoughi, D. Roy, S. Aral, "The spread of true and false news online," Science, 359(6380), 2018, pp. 1146-1151.

[19] Y. L. Huang, K. Starbird, M. Orand, S.A. Stanek, H.T. Pedersen, "Connected through crisis: Emotional proximity and the spread of misinformation online," ACM conference on computer supported cooperative work & social computing, 2015, pp. 969-980.

[20] M.R. Morris, S. Counts, A. Roseway, A. Hoff, J. Schwarz, "Tweeting is believing? Understanding microblog credibility perceptions," ACM conference on computer supported cooperative work, 2012, pp.

441-450.

[21] B. Wang, J. Zhuang, “Rumor response, debunking response, and decision makings of misinformed Twitter users during disasters,” Natural Hazards, 93(3), 2018, pp. 1145-1162.

[22] A. Zubiaga, M. Liakata, R. Procter, G. W. S. Hoi, P. Tolmie, “Analysing how people orient to and spread rumours in social media by looking at conversational threads,” PloS One, 11(3), 2016.

[23] K. Starbird, J. Maddock, M. Ornad, P. Achterman, R.M. Mason, “Rumors, false flags, and digital vigilantes: Misinformation on twitter after the 2013 Boston marathon Bombing,” IConference 2014 Proceedings.

[24] H. Allcott, M. Gentzkow, “Social media and fake news in the 2016 election (No. w23089),” National Bureau of Economic Research, 2017.

[25] M. Bronstein, G. Pennycook, A. Bear, D. Rand, T. Cannon, “Belief in fake news is associated with delusionality, dogmatism, religious fundamentalism, and reduced analytic thinking,” Journal of Applied Research in Memory and Cognition, 8(1), 2019, pp. 108-117.

[26] A. Roets, “‘Fake news’: Incorrect, but hard to correct. The role of cognitive ability on the impact of false information on social impressions,” Intelligence, 65, 2017, pp. 107-110.

[27] R. Kraut, “Humans as lie detectors,” Journal of Communication, 30(4), 1980, pp. 209-218.

[28] C. Bond, B. DePaulo, “Accuracy of deception judgments,” Personality and Social Psychology Review, 10(3), 2006, pp. 214-234.

[29] S. Porter, M. Juodis, L. Brinke, R. Klein, K. Wilson, “Evaluation of the effectiveness of a brief deception detection training program,” Journal of Forensic Psychiatry & Psychology, 21(1), 2010, pp. 66-76.

[30] G. Köhnken, “Training police officers to detect deceptive eyewitness statements: Does it work?” Social Behaviour, 2(1), 1987, pp. 1-17.

[31] B. DePaulo, K. Charlton, H. Cooper, J. Lindsay, L. Muhlenbruck, “The accuracy-confidence correlation in the detection of deception,” Personality and Social Psychology Review, 1(4), 1997, pp. 346-357

[32] A. Vrij, M. Baxter, “Accuracy and confidence in detecting truths and lies in elaborations and denials: Truth bias, lie bias and individual differences,” Expert Evidence, 7(1), 1999, pp. 25-36.

[33] M. Hartwig, P. A. Granhag, L. A. Strömwall, O. Kronkvist, “Strategic use of evidence during police interviews: When training to detect deception works,” Law and Human Behavior, 30(5), 2006, pp. 603-619.

[34] M. G. Frank, T. H. Feeley, “To catch a liar: Challenges for research in lie detection training,” Journal of Applied Communication Research, 31(1), 2003, pp. 58-75.

[35] J. E. Driskell, “Effectiveness of deception detection training: A meta-analysis,” Psychology, Crime & Law, 18(8), 2012, pp. 713-731.

[36] L. Akehurst, R. Bull, A. Vrij, G. Köhnken, “The effects of training professional groups and lay persons to use criteria-based content analysis to detect deception,” Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition, 18(7), 2004, pp. 877-891.

[37] T.R. Levine, T.H. Feeley, S.A. McCornack, M. Hughes, C.M. Harms, “Testing the effects of nonverbal behavior training on accuracy in deception detection with the inclusion of a bogus training control group,” Western Journal of Communication, 69(3), 2015, pp. 203-217.

[38] M.A. deTurck, G.R. Miller, “Training observers to detect deception: Effects of self-monitoring and rehearsal,” Human Communication Research, 16(4), 1990, pp. 603-620.

[39] A. Vrij, S. Graham, “Individual differences between liars and the ability to detect lies,” Expert Evidence, 5(4), 1997, pp. 144-148.

[40] S. Kassin, C. Fong, “"I'm innocent!": Effects of training on judgments of truth and deception in the interrogation room,” Law and Human Behavior, 23(5), 1999, pp. 499-516.

[41] M. Zuckerman, R. Koestner, A. Alton, “Learning to detect deception,” Journal of Personality and Social Psychology, 46(3), 1984, pp. 519-528.

[42] M. Zuckerman, R. Koestner, M. J. Colella, “Learning to detect deception from three communication channels,” Journal of Nonverbal Behavior, 9(3), 1986, pp. 188-194.

[43] K. Fiedler, I. Walka, “Training lie detectors to use nonverbal cues instead of global heuristics,” Human Communication Research, 20(2), 1993, pp. 199-223.

[44] C. Meissner, S. Kassin, ““He's guilty!”: Investigator bias in judgments of truth and deception,” Law and Human Behavior, 26(5), 2012, pp. 469-480.

[45] G.D. Bond, “Deception detection expertise,” Law and Human Behavior, 32(4), 2008, pp. 339-351.

[46] J. Martin, C. Dubé, M. Coovert, “Signal detection theory (SDT) is effective for modeling user behavior toward phishing and spear-phishing attacks,” Human Factors, 60(8), 2018, pp. 1179-1191.

[47] C. Canfield, B. Fischoff, A. Davis, “Quantifying phishing susceptibility for detection and behavior decisions,” Human Factors, 58(8), 2016, pp. 1158.

[48] R. Hurtt, “Development of a scale to measure professional skepticism,” Auditing: A Journal of Practice and Theory, 29(1), 2010, pp. 149-171.

[49] O. John, S. Srivastava, “The Big Five trait taxonomy: History, measurement, and theoretical perspectives,” Handbook of personality: Theory and research, 2(4), 1999, pp. 102-138.

[50] K. Mullinix, T. Leeper, J. Druckman, & Freese, J. The generalizability of survey experiments. Journal of Experimental Political Science, 2(2), 2015, 109-138.

[51] A. Coppock. “Generalizing from survey experiments conducted on Mechanical Turk: A replication approach,” Political Science Research and Methods, 7(3), 2019, pp. 613-628.

[52] L. Lutzke, C. Drummond, P. Slovic, & J. Árvai. “Priming critical thinking: Simple interventions limit the influence of fake news about climate change on Facebook,” Global Environmental Change, 58, 2019, 101964.