Partisan Responses to Fact-Checking in Online News Platforms: Evidence from a Political Rumor about the North Korean Leader

TaeYoung Kang,[1,2] **Jaeung Sim***[3]
[1]Underscore, Seoul, South Korea
[2]MyMusicTaste, Seoul, South Korea
[3]KAIST College of Business, Seoul, South Korea
minvv23@underscore.kr, jaeung@kaist.ac.kr

Abstract

To correct misinformation and mitigate the social costs of political rumors and fake news, news providers, politicians, and researchers have exerted significant efforts on fact-checking and rumor debunking. This study examined how individuals will respond when a political rumor is debunked by large-scale fact-checking. To explore this question, we leveraged a quasi-experimental setting where the North Korean leader's reappearance in the public event suddenly rebutted a political rumor about his death. Collecting 2.6 million comments from the largest online news portal in South Korea, we employed a difference-in-differences approach comparing differences in commenting behaviors between liberals and conservatives before and after this event. The results show that a political side empowered by the fact-checking coverage became more vocal and hostile. However, their explicit support level for the rumor did not change significantly compared to their partisan counterparts. Besides, we found that news outlets rebutted by fact-checking attracted more user comments than supported news outlets. Swearing comments of the supported political side mostly drove this difference, suggesting that partisans tend to utilize favorable fact-checking to empower their political side through blaming the other side. Our research stresses the importance of capturing the silence of partisans in considering the effectiveness of fact-checking and provides an alternative explanation on why fact-checking evokes hostile communication in online media.

Introduction

Although rumors are unproven and potentially misleading, people make critical political decisions based on them (Stieglitz and Dang-Xuan 2013) and share these rumors even under a military security threat (Kwon and Rao 2017). While rumors might occur naturally in the early stages of a social crisis (Shibutani 1966), some rumors, especially for political ones, are deliberately fabricated and shared on social media as a form of news (i.e., fake news) (Bovet and Makse 2019).

Unverified rumors induce confusion in society, and if they are false or inaccurate, they may lead to potentially devastating consequences. For instance, erroneous beliefs about the vaccine have inhibited vaccination and national health outcomes (Berinsky 2017). Fabricated stories might alter voters' perception of candidates (Allcott and Gentzkow 2017) and erode trust in institutions (Ciampaglia et al. 2018). Therefore, it is crucial to correct misinformation transmitted through rumors and fake news.

To compete with the misinformation in online media, news providers, politicians, and researchers have paid massive attention to fact-checking. The main objective of fact-checking is to correct people's belief in a rumor and reduce social costs from misinformation (Allcott and Gentzkow 2017; Thorson 2016). Numerous studies have focused on the fact-checking message's diffusion and its impact on attitudes toward the rumor (e.g., Shin et al. 2017), but the effects of such messages have not been significant in practice due to the public's little attention on rumor debunking (Shin et al. 2017) as well as remaining doubts on fact-checking entities and results (Thorson 2016; Walter et al. 2020). Furthermore, some studies suggested the "backfire" effect of fact-checking, implying that users alter their beliefs toward the ideology-consistent and wrong direction (Nyhan and Reifler 2010) or behave more aggressively (Jiang and Wilson 2018). For these reasons, we need to understand the motivations behind seemingly undesirable responses to estimate and improve the real impact of fact-checking properly.

Our study aims to understand how and why partisans respond to fact-checking, focusing on user comments on

* Corresponding author.

online news platforms. In particular, we analyze the interactive effects of fact-checking and political slant of news outlets, which the extant literature has not considered. To overcome the challenge of little attention and trustworthiness of fact-checking, we leveraged a large-scale political event in North Korea. In mid-April 2020, several media raised a question on whether the North Korean leader Kim Jong-un was alive since he did not appear in the celebration of the first supreme leader of North Korea, Kim Il-sung's birthday. Numerous media (e.g., CNN, New York Times, and Washington Post) covered relevant rumors that argued Kim had heart surgery and died due to the surgery's failure. Although the South Korean government officially denied this rumor (Perrett, 2020), the media and some local conservative politicians continued to state the possibility that Kim got heart surgery and died after its failure. Kim's reappearance after two weeks refuted the rumor, and most of the media immediately covered that the leader was alive (Choe 2020; Barnes and Almasy 2020).

This event provides us a quasi-experimental setting to explore the consequences of a well-known and convincing rebuttal of a political rumor as follows. First, this event was sudden and unexpected. Second, the fact-checking result was immediately and widely covered by major media. Third, since the leader's appearance directly rebutted this rumor and most news providers admitted it, there is little room for doubt about this fact-checking.

Collecting 2.6 million news comments from the largest online news portal in South Korea, we examine how this fact-checking event affected users' commenting behaviors in online news websites. Specifically, we focus on how the responses differ between liberal and conservative users, given that their attitudes toward the rumor were opposite. Some conservative politicians argued the leader's death, whereas the government supported by the liberal party officially denied the rumor. This setting allows us to employ a difference-in-differences approach analyzing how the difference in commenting behaviors between the supported side and the rebutted side changed after the large-scale fact-checking.

Related Work

Political Rumors and Fake News

Political rumors are often not based on concrete facts or even false information and likely to mislead the public. Importantly, these rumors have a tangible impact on electoral decisions (Weeks and Garrett 2014). Online channels have played a significant role in spreading political rumors. Previous studies suggested that Internet use boosts exposure to both rumors and their rebuttals, Internet users were more likely to believe rumors that are emailed from their friends or family (Garrett 2011), and social media facilitated the diffusion of emotionally charged messages in political communication (Stieglitz and Dang-Xuan 2013).

Recently, fake news—defined as "news articles that are intentionally and verifiably false and could mislead readers" (Allcott and Gentzkow 2017)—has been widely shared and spread incorrect political rumors on social media. For example, a fake news website, The Denver Guardian, intentionally published a fabricated news article with the headline, "FBI agent suspected in Hillary email leaks found dead in apparent murder-suicide."

Responses to fake news substantially vary depending on partisanship. During the 2016 US presidential election, Trump supporters' activity governed the top fake news spreaders' dynamics, whereas Clinton supporters' activity was affected by the top spreaders of traditional center news outlets (Bovet and Makse 2019).

Prior studies have suggested various mechanisms for social media to decrease the negative influences of political misinformation and fake news. For instance, in hypothetical settings, including source ratings and highlighting the source in social media substantially reduce the believability of and engagement in suspicious news articles (Kim and Dennis 2019; Kim et al. 2019). Numerous fact-checking websites such as Factcheck.org and Snopes.com were established to provide correct political rumors (Shin et al. 2017).

Responses to Fact-Checking

Prior studies have suggested how difficult correcting political rumors and changing partisans' attitudes are in practice. Jiang et al. (2020) provided field evidence on responses to misinformation and fact-checking on users' belief in rumors. Based on a rich dataset of fact-checked tweets and their comments, they found that belief (disbelief) decreases (increases) in the veracity of claims, but the difference was relatively small. They also found that users' belief gradually approached the truth, but at a plodding pace. In addition, they showed that the belief in false information decreases by 3.4% after it was fact-checked by Snopes or PolitiFact, still leaving a significant belief in misinformation. Thorson (2015) also suggested that political misinformation affects an individual's attitude even when effectively discredited. The limited influences are attributable to a lack of social connections with fact-checkers (Margolin et al. 2018), attitudinal congruency of political misinformation (Hameleers and van der Meer 2020), pre-existing solid beliefs, ideology, and knowledge (Walter et al. 2020).

Some studies even suggested backfire effects of fact-checking on users' beliefs and behaviors. For instance, Nyhan and Reifler (2010) suggested that corrections of false information frequently fail to reduce and sometimes even increase misperceptions among partisan and ideological

groups, while Wood and Porter (2018) demonstrated the opposite results from different samples. Jiang and Wilson (2018) found that social media users increased swearing word usage after fact-checking.

Fact-checking messages against false rumors are rarely shared, and a significant portion of users endorse the debunked rumors and cast doubt on the objectivity of the fact-checking websites (Shin et al. 2017). Moreover, partisans selectively share fact-checking messages that are favorable (hostile) to their own (the opposite) party (Shin and Thorson 2017). In this way, fact-checking often fails to correct and stop circulating false information in the real world due to a lack of saliency as well as doubts on fact-checking entities.

In this vein, our research makes important contributions to the rich literature on rumors, misinformation, and fact-checking as follows. First of all, this research provides a unique empirical attempt to demonstrate how online commenting behaviors change after a political rumor is rebutted by a large-scale fact-checking event that occurred by the rumor's target and covered by most major media. Interestingly, despite a high saliency and trustworthiness of fact-checking in our setting, the debunked side's explicit support for the rumor did not change significantly, unlike Jiang et al. (2020). Instead, we see that such users relatively reduced their commenting behaviors. These results may suggest that rebutted rumor-supporters may seek consonant information that counters the fact-checking and recovers the consistency between their belief and the fact (Festinger 1962, Marikyan et al. 2020), at least in a very short term. Also, participation in the discussion *per se* may better reflect how confident people are about the rumor, instead of explicitly blaming or supporting, particularly when individuals are reluctant to reveal their explicit belief.

Second, by leveraging the politically polarized rumor and fact-checking event and tracking individual users over time, we identify the potential supporters (or believers) of the rumor and their responses even when they did not explicitly reveal their stance. Our findings suggest that although the vast majority of conservative users had not explicitly supported the rumor, they commented substantially less on relevant news articles than liberal users, showing how the rebutted side becomes relatively silent than the other side. Further, we showed that liberal users increased their comments with swearing words after the fact-checking, compared with conservative users. This result also partially explains why the increased use of swearing words follows fact-checking (Jiang and Wilson 2018).

Third, to our knowledge, this paper is the first empirical evidence on how fact-checking affects the selection of space for discussion. Our user-level panel data allow us to demonstrate users' strategic choices of news outlets. We found that the relative amount of liberal users' comments—particularly hostile ones—drastically increased after the fact-checking in conservative media. This result may imply that partisans are highly motivated to undermine the opposite side's legitimacy. Due to this motivation, the discussant's composition on news articles could be less polarized, while the communication may be more hostile.

Data Collection and Processing

User Partisanship Classification

We collected information on online news users and their comments from Naver News, the largest online news portal in South Korea. Specifically, among 82 thousand users who wrote comments on at least one of daily 30 most read political news articles from April 30th to May 4th, 2020 (150 articles in total), we randomly selected 1.6 thousand users (2%) and collected their 2.6 million comments from their commenting history pages.

Using the commenting history data, we classified users' political orientation in the following procedure. We first classified the political orientation of comments. To do so, we first labeled 3K randomly-sampled comments from our dataset on users' commenting history with three independent annotators. The Krippendorff's alpha coefficient, a metric to evaluate the data annotation quality, scored 0.93, showing high reliability of the labeling quality. We adopted majority voting (or two or more annotators agreed), resulting in 83.06% data preservation.

Then, we combined this dataset with Han et al (2019)'s 35K political slant data based on Korean news comments from the same online news portal (i.e., Naver News) and use them with a train/test ratio of 7:3. For training this combined dataset, we use a more recently developed model, KcBERT (Lee 2020). It is a pre-trained model based on the same data source and the data period overlapping our research period, instead of KorBERT used in Han et al. (2019).[1] The results show that our fine-tuned version of KcBERT demonstrates an F1 score of 0.89, which is higher than Han et al. (2019)'s performance on their dataset (0.83) and that on our human-labeled dataset (0.78 for the majority voting). Therefore, we adopt this new model for the political slant of user comments, which we used to estimate users' partisanship. We applied the classifier to comments with 20 or more characters written in the politics section.

Then, we assigned the average score of the predicted political slant to each user. We categorized users whose scores

[1] While KorBERT was trained based on various data sources, such as Wikipedia as well as online news comments, KcBERT used only commenting data from Naver News between January 2019 and June 2020, thoroughly covering our research period.
See https://github.com/Beomi/KcBERT for further details of KcBERT.

exceed 0.5 as conservative users and those with scores lower than 0.5 as liberal users. During this procedure, 61% of users were classified as conservatives, and 38% were classified as liberals. We discarded 1% of users who did not contain sufficient information for this procedure. As a result, we obtain 1,617 politically slanted users.

Classifying Comment Profanity

We consider abusive language, given that swearing can inhibit constructive discussion (Cho and Kwon 2015). In our setting, identifying abusive language involves several empirical challenges. First, the web news portal in this study automatically blocks a set of swearing words in its online commenting section. Hence, users who wanted to express abusive language made typos deliberately or use neologisms that implies them. For instance, a user who intends to express 'fuck' may type 'fxxk', 'f*ck', or 'fvck'. Second, Hangul, the Korean alphabet, has a sub-character architecture

(a) Choseong (b) Jungseong (c) Jongseong

Figure 1: Sub-character Example of Korean characters.

(see Stratos 2017 for details), making identifying the deviations of original words more challenging. To overcome such obstacles, we employed a revised sub-character decomposition approach of Korean language modeling.

Figure 1 illustrates a sub-character example of Korean characters. Each character comprises two or three sub-characters called choseong, jungseong, and (optional) jongseong. We decomposed a word into sub-components while keeping the order of characters and sub-characters as suggested by Park et al. (2018). For instance, the word '정치' is decomposed into the sequence of sub-characters: {ㅈ, ㅕ, ㅇ, ㅊ, ㅣ}.

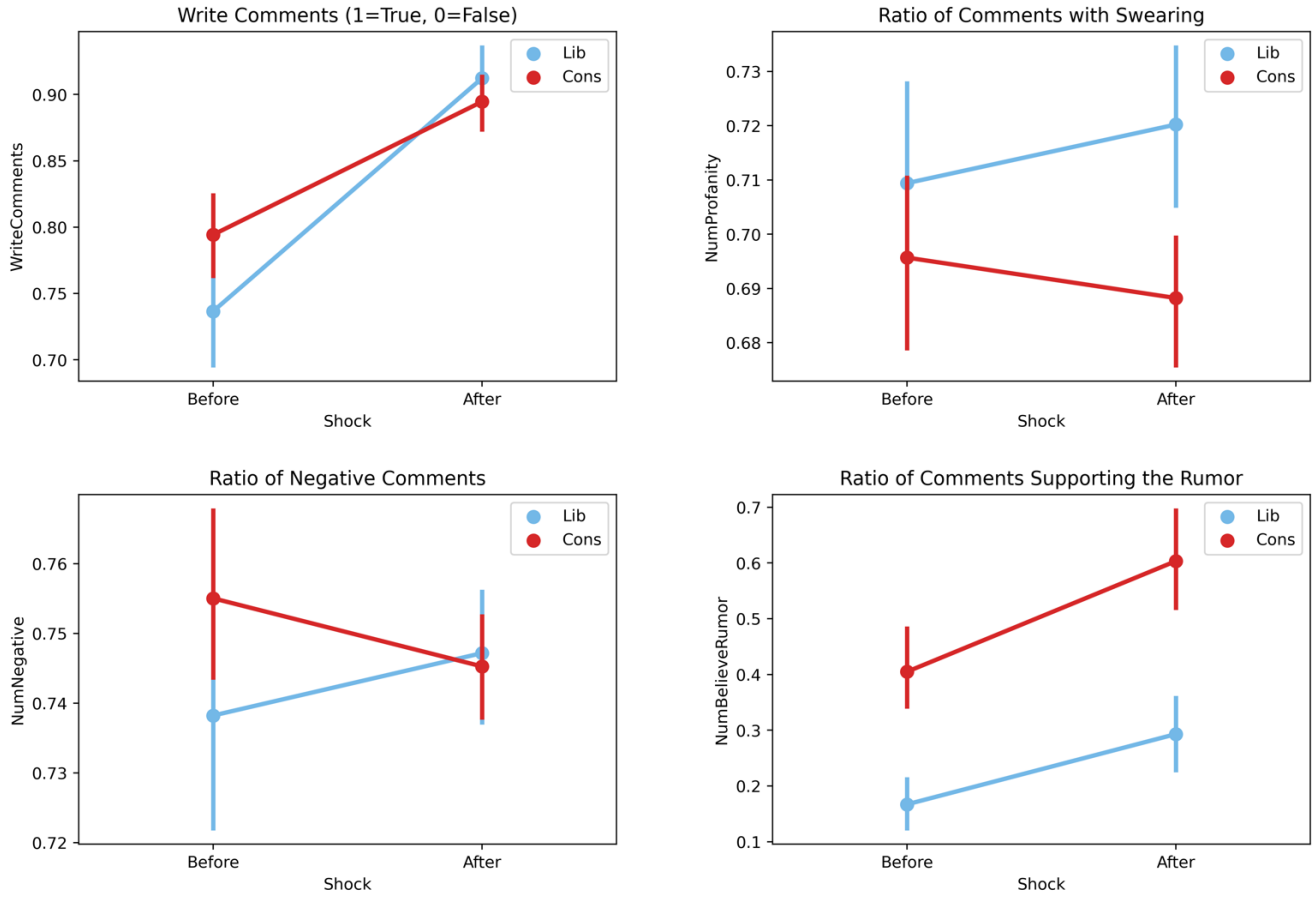

Figure 2: Behavioral changes after the fact-checking event by political slant. 'Before' denotes one week before the fact-checking event, while 'After' denotes one week since the event. Blue dots and lines denote liberal users, while red dots and lines indicate conservative users. Vertical bars indicate 95% confidence intervals.

In addition to the previous approach, we decomposed consonants into smaller units, allowing the model to consider relationships among similar-looking consonants. Specifically, we split tensed or aspirated consonants (types of Korean consonants) into special tokens and basic consonants. For example, the previous word '정치' includes an aspirated consonant, 'ㅊ.' Thus, it can be split further into the sequence of sub-characters: {ㅈ, ㅕ, ㅇ, <G>, ㅈ, ㅣ}. Considering that Korean swear words often include tensed or aspirated consonants to express high-arousal emotions, this approach can absorb substantial deviations of abusive language.

We combined two datasets on abusive online comments to test the profanity classifier. The first dataset consists of 9.4 thousand manually labeled comments on entertainment news for identifying Korean toxic speech (Moon et al, 2020). Since we focused on the narrower term of profanity, we selected only 2.3 thousand comments independent of gender-related biases as a subsample of data. Second, since our study deals with political comments only, we added a dataset of 9.3 thousand labeled comments written between 2017 and 2019 from the politics section of the web news portal. As a result, our final dataset consists of 11.7 thousand comments with profanity labels, and about 70% of these comments were swearing, possibly due to the hostile environment for political communication in Korean news outlets.

Finally, we applied the sub-character byte pair encoding (BPE) approach with FastText embedding and 1D-CNN to the collected dataset. We set aside 30% of labels as the test set and use 70% to train the model. In combining the sub-character byte pair encoding (BPE) with FastText embedding and 1D-CNN, our hyper-parameters are set to 50 epochs, batch size 16, Adamax optimizer with a learning rate of 10e-3. The trained classifier showed an accuracy of 0.91, an F1 score of 0.91, and a binary F1 score on positives of 0.94, indicating that our model's high accuracy is unlikely to be attributable only to the imbalance of our dataset.

Classifying Comment Stance

We measure the stance on the rumor or whether a comment explicitly supports it, as it is one of the most widely explored variables in the extant literature (e.g., Jiang et al. 2020). Given that support for a specific issue can hardly be inferred from pre-trained models or datasets on other issues, unlike relatively general features such as swearing or sentiment, we try to construct a set of true labels relying on human annotators as much as possible. To be specific, we randomly sample 3K comments from 9.3K comments on political news about North Korea and Kim Jong-un. Then, three independent and knowledgeable annotators classified all of the sampled comments as either 'supports the rumor' or 'doubts the rumor / neutral' with three annotators just as the political slant data stated above. Krippendorff's alpha coefficient was 0.89, showing sufficiently high reliability. To confirm the labels, we adopt majority voting (or two or more annotators agreed).

We split the labeled dataset into 70% as a training set and 30% as a test set. The trained 1D textCNN model showed an F1 score of 0.61. This limited performance could be attributable to a small sample size and class imbalance. However, it is also worth noting that one-third of all comments in our dataset are manually labeled, and due to the high reliability inferred from Krippendorff's alpha coefficient, we can virtually consider them as true values. We expect this will substantially alleviate measurement errors, leading to downward-biased effect estimates (Scharkow and Bachl 2017).

Classifying Comment Sentiment

We measure sentiment, indicating the positivity of comments. Since more positivity often implies more favorable attitudes or comfortable states, non-positive comments may work as another proxy of hostility or aggressiveness.

To measure this variable, we adopt a pre-trained classifier, KoBERT, which was developed to overcome the limited performance of Google's BERT base multilingual cased in the Korean language.[2] This Korean sentiment classifier is based on the NSMC dataset with 200K comments from the movie review corpus of Naver, one of the most widely used text classification datasets in the Korean NLP community. This model demonstrates an accuracy of 0.90.

It is worth noting that Naver, the data source of KoBERT, also serves the online news portal in our setting, implying that users' composition and language patterns might be similar to our sample. Even if they are not similar, sentiment is a generally accepted concept that is relatively insensitive to time and contexts, compared to swearing words that often denote specific political figures.

Empirical Analysis

Research Setting

In South Korea, the reappearance was first reported at 6 A.M. on May 2nd, 2020. Therefore, we postulate that commenting behaviors after this moment were affected by the fact-checking event. Our analysis considers online comments written between April 25th (a week before the event)

[2] See https://github.com/SKTBrain/KoBERT for more details.

and May 8th, 2020 (a week since the event). We are interested in the following outcome variables at the user level.

Write a Comment. This variable indicates whether a user wrote at least one comment in the rumor-related news articles in the given week. For statistical analyses, it is operationalized as one if a user wrote a comment, 0 otherwise. It can be interpreted as whether a user participated in the discussion or not.

Write a Rumor-Supporting Comment. It indicates whether a user wrote at least one comment supporting the rumor in the given week. It is operationalized as one if a user wrote a comment, 0 otherwise, for each week.

Number of Comments. It is defined as the number of comments that a user wrote in the rumor-related news articles in the week. It can be interpreted as the intensity of participation in the discussion.

Number of Rumor-Supporting Comments. It denotes the number of rumor-supporting comments that a user wrote in the week.

Number of Non-Supporting Comments. It is the number of comments not including explicit support for the rumor in the week.

Number of Comments with Swearing. It is the number of comments that include a swearing word that a user wrote in the rumor-related news articles in the given week. This variable represents the amount of hostile communication.

Number of Comments without Swearing. It denotes the number of comments which do not include swearing words. This variable presents the amount of civil communication.

Number of Positive Comments. It is the number of comments showing positive sentiment in the rumor-related news articles in the given week. This variable proxies the amount of less hostile or relatively civil communication.

Number of Non-Positive Comments. It indicates the number of comments not showing positive sentiment. As a flip side of positive comments, these comments represent the higher likelihood of hostile communication.

Figure 2 presents how commenting behaviors of partisans changed after the fact-checking event. We observed that conservative users were more likely to write comments on the rumor-related articles than liberal users before the event. However, the relationship became reversed after the event, suggesting that the fact-checking event encouraged liberals to be vocal more than conservatives. In addition, we found that the proportion of comments with swearing and negative comments relatively increased (decreased) among liberals (conservatives). Notably, we did not see that conservatives decreased the proportion of rumor-supporting comments.

To understand where this difference came from, we divided the comments into comments with swearing and those without swearing. Interestingly, we found that liberals presented a much higher increase in swearing comments than conservatives, suggesting that liberals or users supported by fact-checking were highly motivated to behave aggressively. In contrast, the two groups of users demonstrated a similar increase in non-swearing comments.

DID Estimates of Eq. (1) Dependent Variables	Coeff. (γ)	Robust Std. Err.
Write a Comment	-0.0950***	0.0320
Write a Rumor-Supporting Comm.	-0.0127	0.0216
No. of Comments		
All Comments	-0.153***	0.0428
Rumor-Supporting Comments	-0.0169	0.0182
Non-Supporting Comments	-0.152***	0.0423

Table 1: DID estimates of the effects of fact-checking on commenting amount and stance toward the rumor. Estimated γ's of Equation (1) are reported for each dependent variable. *$p < 0.1$; **$p < 0.05$; ***$p < 0.01$. 1,617 users x 2 (before vs. after) = 3,234 observations are used in this analysis.

DID Estimates of Eq. (1) Dependent Variables	Coeff. (γ)	Robust Std. Err.
No. of Comments		
With Swearing	-0.148***	0.0393
Without Swearing	-0.0271	0.0318
Non-Positive Sentiment	-0.143***	0.0399
Positive Sentiment	-0.0379	0.0307

Table 2: DID estimates of the effects of fact-checking on commenting hostility. Estimated γ's of Equation (1) are reported for each dependent variable. *$p < 0.1$; **$p < 0.05$; ***$p < 0.01$. 1,617 users x 2 (before vs. after) = 3,234 observations are used in this analysis.

Econometric Analysis

To formally examine the differences in the effects of fact-checking on online commenting behaviors between liberal users and conservative users, we employed a difference-in-differences (DID) approach that compares the change in the outcomes for a treated group with that of a control group (Meyer 1995). The DID approach has been widely adopted to estimate causal inferences in recent social science studies (e.g., Antecol et al. 2018; Datta et al. 2018; Larcom et al. 2019). In the current setting, we used liberal users (i.e., the supported group) as a control group of conservative users (i.e., the rebutted group) and estimated the following regression model:

$$y_{it} = \alpha_i + \beta \cdot After_t + \gamma \cdot After_t \cdot Conservative_i + \varepsilon_{it}, \quad (1)$$

where i indexes users; t indexes week; y is a dependent variable; α_i is a set of user fixed effects; $After_t$ is a binary variable that indicates 1 if week t is after the reappearance of the North Korean leader, and 0 otherwise; $Conservative_i$ indicates 1 if user i is politically conservative, and 0 otherwise; and ε_{it} is an error term. Counting variables, such as Number of Comments, are highly skewed, so they are log-transformed after adding one to address zero values.

In this specification, any time-invariant differences among users are captured by the fixed effects, omitting any other user-specific controls (Datta et al. 2018). To correct the statistical inefficiency of heteroscedasticity, we used the Huber-White robust standard errors. In this model, our main interest is the DID estimator γ which captures the relative change of conservative users compared with that of liberal users after the fact-checking event.

The DID estimate of Equation (1) for each dependent variable is reported in Table 1. The complete estimation results, including other coefficients and goodness-of-fit, are provided in the Appendix. A positive (negative) coefficient of γ can be interpreted as the increase in conservatives' (or the rebutted group) commenting activities was larger (smaller) than that of liberals (or the supported group). We see that conservative users showed a smaller increase in the likelihood of writing comments on the rumor-related articles ('Write Comments') by 9.5 percentage points, and the difference was statistically significant at the 1% level. We also observe negative coefficients for the number of comments in line with our observations in the model-free analysis. These findings suggest that liberal or supported users became relatively more vocal after fact-checking compared with conservative or rebutted users. Insignificant results of rumor-supporting comments might be attributable to the small amount of rumor-supporting comments before the fact-checking event or self-censorship behaviors (see the next section for more details).

Table 2 shows the DID estimates of commenting hostility. We found that users on the rebutted side substantially reduced swearing and non-positive sentiments compared with the supported side. Unlike hostile comments, the coefficients for comments without swearing or positive sentiment were statistically insignificant. In other words, these results suggest that users supported by fact-checking increased their commenting hostility.

Heterogeneous Responses to Fact-Checking by Political Orientation of News Outlets

To dive into this phenomenon more deeply, we need to explore why commenters changed their behaviors. We propose that users supported by fact-checking intended to strengthen their legitimacy and power through blaming the opposite side (Boin et al. 2010). Since major news outlets often support a particular political party, they are likely to be attacked by supported partisans. Conversely, rebutted partisans are unlikely more hostile than they were before, as their legitimacy has dropped. For these reasons, we expect that conservative news outlets were more likely to be blamed, particularly by liberal users, after the fact-checking event.

DDD Estimates of Eq. (2) Dependent Variables	Coeff. (γ)	Robust Std. Err.
Write a Comment	-0.112***	0.0346
Write a Rumor-Supporting Comm.	-0.0379***	0.0133
No. of Comments		
All Comments	-0.0985***	0.0334
Rumor-Supporting Comments	-0.0309***	0.00999
Non-Supporting Comments	-0.0793**	0.0329

Table 3: DDD estimates of heterogeneous effects of fact-checking on commenting amount and stance toward the rumor by users' political orientation (or prior support for the rumor) and news outlets. Estimated γ's of Equations (2) are reported for each dependent variable. *$p < 0.1$; **$p < 0.05$; ***$p < 0.01$. 1,617 users x 2 (before vs. after) x 2 news outlet groups (conservative vs. liberal) = 6,468 observations are used in this analysis.

DDD Estimates of Eq. (2) Dependent Variables	Coeff. (γ)	Robust Std. Err.
No. of Comments		
With Swearing	-0.0853***	0.0284
Without Swearing	-0.00966	0.0206
Non-Positive Sentiment	-0.0829***	0.0301
Positive Sentiment	-0.0213	0.0186

Table 4: DDD estimates of heterogeneous effects of fact-checking on commenting hostility by users' political orientation (or prior support for the rumor) and news outlets. Estimated γ's of Equation (2) are reported for each dependent variable. *$p < 0.1$; **$p < 0.05$; ***$p < 0.01$. 1,617 users x 2 (before vs. after) x 2 news outlet groups (conservative vs. liberal) = 6,468 observations are used in this analysis.

To test this proposition, we analyzed how user responses to fact-checking differ across news outlets' political slants. We selected major liberal and conservative media, following Han et al. (2019). To estimate the heterogeneous effects across groups and media, we employ a difference-in-difference-in-differences (DDD) model (Gilje 2019):

$$y_{ijt} = \alpha_{ij} + \beta_1 \cdot After_t + \beta_2 \cdot After_t \cdot Conservative_i + \beta_3 \cdot After_t \cdot Conservative_j + \gamma \cdot After_t \cdot Conservative_i \cdot Conservative_j + \varepsilon_{ijt}, \quad (2)$$

where j indexes news outlets; α_{ij} is a set of interactions between user fixed effects and political slant of news outlets;

$Conservative_j$ indicates 1 if outlet j is politically conservative, and 0 otherwise; other variables are identically defined as in Equation (1). Our interest is the three-way interaction term γ which picks up the difference in user responses to fact-checking depending on news outlets' political slant.

Table 3 shows the DDD estimates of Equation (2). A positive (negative) coefficient of γ can be interpreted as conservative users, or rumor supporters, became relatively vocal (silent) than liberals or non-supporters in conservative news outlets, compared to liberal news outlets. The results indicate that the adverse effects on conservatives' commenting behaviors were more prominent in conservative outlets. These results suggest that the voice of rumor supporters became relatively quiet, as conservative outlets—which had provided the favorable environment for conservative users—were rebutted and blamed by the opponents.

The DDD estimates for commenting hostility in Table 4 support this argument. The results show that the decrease in conservative users' comments in conservative outlets was significant only for swearing or non-positive comments, suggesting that conservative users became relatively less hostile than liberal users. In other words, liberal users became more vocal and hostile in conservative news media to blame conservative media and rumor believers.[3] These results support our proposition that individuals supported by fact-checking (liberal users) aimed to strengthen their legitimacy and power through blaming news outlets on the opposite side (conservative news media) for spreading wrong information.

Discussion and Conclusion

Summary of Findings

In this study, we aimed to examine how and why partisans respond to the fact-checking event, focusing on user comments on online news platforms. Moreover, we analyze the interactive effects of fact-checking and political slant of news outlets, which the existing studies have not considered. To tackle challenges pointed out in the extant literature, we leveraged a large-scale natural experiment wherein the North Korean leader Kim Jong-un's reappearance suddenly rebutted a political rumor about his death.

To analyze partisan responses to this fact-checking event, we collected user comments on rumor-related news articles from the largest online news portal in South Korea. Focusing on relative differences in responses to the fact-checking event between liberals and conservatives, we employed a difference-in-differences approach, which is widely adopted for causal inferences in social science (e.g., Antecol et al. 2018; Datta et al. 2018; Larcom et al. 2019). To infer users' political orientation, we improved a pre-trained political orientation classifier and observed a higher F1 score of 0.89 in our dataset. In classifying profanity of comments, we introduced a novel extension of Park et al. (2018)'s sub-character decomposition approach and obtained an F1 score of 0.91 for our classifier. Besides, based on intensive human-labeled observations, and a newly-trained or pre-trained model, we considered the stance on the rumor and sentiment of comments.

We show that liberal users became relatively vocal than conservative users after the fact-checking event. This difference was mainly attributable to hostile comments rather than neutral or civil comments. These results suggest that fact-checking tends to encourage the supported side to raise their voice but increases hostile communication. To further understand this phenomenon, we explored the underlying motivation of partisan responses by analyzing how such responses were heterogeneous across political slants of news outlets. We found that fact-checking had a more significant effect on conservative news outlets (on the rebutted side) than liberal outlets (on the supported side) and liberal users' swearing comments mostly drove this gap. These findings imply that individuals supported by fact-checking intend to strengthen their legitimacy and power by blaming the opposite-side media for spreading wrong information.

Discussions

Concerning rumor-supporting comments, we found that behavioral responses between conservatives and liberals were not distinguishable. This result could be associated with the small portion of individuals who explicitly expressed their support for the rumor before the fact-checking (234 / 1,641 = 14.3%). However, this may not be plausible given that the number of explicit supporters among our sample increased after the event (from 234 to 257). Another possibility is that the cognitive dissonance that arose from disconfirming rumor supporters' beliefs might have led the supporters to seek consonant information (Festinger 1962, Marikyan et al. 2020). To be specific, individuals feel guilt when they find their decision was based on their false belief, motivating them to seek information consistent with their prior belief to overcome this feeling (Marikyan et al. 2020). Throughout processing and reviewing rumor-supporting comments after the fact-checking event, we observed several comments showing consonant-information-seeking motivation, such as suspecting whether Kim Jong-un in the picture is the true person or a body double. In this regard, our findings might suggest that sudden fact-checking motivates firm believers to overcome the cognitive dissonance by rebutting the fact-

[3] We found positive and significant estimates of β_3 for comments with swearing and those with the non-positive sentiment, does not for other comments. These results imply that liberal users become relatively vocal in conservative media than liberal media. The complete estimates are available in the Appendix.

check at least in a very short term, while it silences relatively weak believers.

We observed that the effects on hostile comments were highly asymmetrical between liberals and conservatives. Specifically, we found evidence that liberals who were supported by fact-checking became more vocal and hostile, particularly in conservative media. Such responses might be closely related to efforts to strengthen their legitimacy and power through blaming the opponents (Boin et al. 2010). Importantly, these findings might provide a possible explanation on why the increased use of swearing words follows fact-checking (Jiang and Wilson 2018).

Our findings suggest that fact-checking may reshape the composition of discussants in online media. According to our findings, partisans increase their commenting activities in the opposite-side outlets when fact-checking is favorable to their side. In particular, such an increase is prominent only for swearing comments, further supporting the claim that partisans seek legitimacy by attacking their opponents.

Limitations and Future Research

As the first empirical effort examining the effects of a large-scale fact-checking in the field, some limitations encumber this work. First, this research examines a political rumor only. While political rumors are widely shared in social media and have substantial societal impacts (Shin et al. 2017; Weeks and Garrett 2014), other rumor types are also worthy of attention. For instance, despite solid scientific results that reject the argument that vaccination causes autism, many people still believe this association and deny vaccination (Davidson 2017). Considering that such scientific rumors can also affect government policies, future research should examine our question in various rumoring contexts.

Second, we did not show whether our findings hold in the opposite case wherein conservatives are supported, while liberals are rebutted. Notwithstanding prior studies suggesting that both liberals and conservatives interpret factual information in ways consistent with their political orientation (Gerber and Huber 2010; Nyhan and Reifler 2010), it is impossible to determine if fact-checking that rebuts a pro-liberal rumor also induces similar responses.

Third, the current work focuses on heterogeneity of commenting volume, hostility, explicit support for the rumor, and news outlets across political orientations. Although beliefs in rumors, participation and incivility in online communication have attracted considerable attention from policymakers and researchers (Cho et al. 2012; Miškolci et al. 2020; Jiang et al. 2020), more nuanced features, such as a target of swearing, might also be considered.

Fourth, our work considers online comments only, while responses to fact-checking might have various forms. Future works may benefit from exploring different outcomes such as attitudes toward policies or politicians.

Lastly, the reappearance of a political leader is not necessarily the same as a conventional fact-checking effort. Although this event presented a fact, it might have signaled strategic aspects of the North Korean government. Future studies need to investigate fact-checking settings that can rule out this possibility.

Despite the noted shortcomings, this paper offers several important contributions. First, our paper provides the first empirical evidence on how partisans respond to a large-scale fact-checking of a political rumor in the field. Despite a high level of saliency and trustworthiness of fact-checking in our setting, we did not find substantial differences in explicit support for the rumor from the debunked political side, unlike Jiang et al. (2020). Instead, we showed that users on that side relatively reduced their commenting behaviors. Such results may imply that participation in discussion per se may reflect how confident people are about the rumor more effectively, rather than explicit blaming or supporting.

Second, our unique setting based on the politically slanted rumor and tracking individual users over time allowed us to identify the potential rumor supporters and their responses even if they did not explicitly reveal their stance. Our results show that although conservative users had not mostly expressed their stance, they substantially reduced their comments on relevant news articles than liberal users. Our method may also be applied to other rumor-debunking contexts where pro- and anti-rumor sides are distinguished.

Third, our paper provides the first empirical evidence on how fact-checking affects space selection for discussion. Our panel data tracking each user over time allow us to show users' strategic news outlet choices. We found that the relative amount of liberal users' comments—only hostile ones—substantially increased after the fact-checking in conservative media, implying that partisans are highly motivated to undermine the opponents' legitimacy. Also, our findings may suggest one of the viable mechanisms on how asymmetry in an online discussion can be formed and destructed in the absence of social networks. We showed that fact-checking makes partisans more vocal and hostile when their beliefs are supported than debunked. We also found that this increase in voice is asymmetric across political slants of news outlets. It is worth noting that such asymmetric changes in partisan voice were induced where social functions such as following and friending are absent. These findings provide insights on how online discussions are biased and contribute to polarization.

Acknowledgments

The authors thank the associate editor and reviewers for their constructive comments that significantly helped improve this paper. We also thank Neoul Shim and Hyunchae Lim for their research assistance. All errors are our own.

Appendix

Complete Results of Difference-in-Differences

Dependent Variables	Write a Comment	Write a Rumor-Supporting Comment	ln(Number of Comments + 1)	ln(Number of Rumor-Supporting Comments + 1)	ln(Number of Non-Supporting Comments + 1)
β	0.193***	0.0207	0.335***	0.0191	0.336***
	(0.0255)	(0.0160)	(0.0336)	(0.0133)	(0.0332)
γ	-0.0950***	-0.0127	-0.153***	-0.0169	-0.152***
	(0.0320)	(0.0216)	(0.0428)	(0.0182)	(0.0423)
User FE	Included	Included	Included	Included	Included
No. of Users	1,617	1,617	1,617	1,617	1,617
Observations	3,234	3,234	3,234	3,234	3,234
R-Squared	0.546	0.632	0.758	0.660	0.756

Table A1: The estimates of Equation (1) for commenting amount and stance toward the rumor. Robust standard errors are in parentheses. *$p < 0.1$; **$p < 0.05$; ***$p < 0.01$.

Dependent Variables	ln(Number of Comments with Swearing +1)	ln(Number of Comments without Swearing +1)	ln(Number of Non-Positive Comments +1)	ln(Number of Positive Comments +1)
β	0.300***	0.123***	0.305***	0.113***
	(0.0308)	(0.0249)	(0.0313)	(0.0238)
γ	-0.148***	-0.0271	-0.143***	-0.0379
	(0.0393)	(0.0318)	(0.0399)	(0.0307)
User FE	Included	Included	Included	Included
No. of Users	1,617	1,617	1,617	1,617
Observations	3,234	3,234	3,234	3,234
R-Squared	0.746	0.734	0.757	0.685

Table A2: The estimates of Equation (1) for commenting hostility. Robust standard errors are in parentheses. *$p < 0.1$; **$p < 0.05$; ***$p < 0.01$.

Dependent Variables	Write a Comment	Write a Rumor-Supporting Comment	ln(Number of Comments + 1)	ln(Number of Rumor-Supporting Comments + 1)	ln(Number of Non-Supporting Comments + 1)
β_1	0.0271*	-0.00478	0.0298**	-0.00332	0.0331**
	(0.0163)	(0.00478)	(0.0145)	(0.00332)	(0.0141)
β_2	0.0295	0.0159**	0.0298	0.0114**	0.0216
	(0.0205)	(0.00665)	(0.0188)	(0.00469)	(0.0183)
β_3	0.128***	0.0207**	0.119***	0.0137**	0.108***
	(0.0270)	(0.00886)	(0.0256)	(0.00629)	(0.0253)
γ	-0.112***	-0.0379***	-0.0985***	-0.0309***	-0.0793**
	(0.0346)	(0.0133)	(0.0334)	(0.00999)	(0.0329)
User FE x Media Slant	Included	Included	Included	Included	Included
No. of Users	1,617	1,617	1,617	1,617	1,617
Observations	6,468	6,468	6,468	6,468	6,468
R-Squared	0.635	0.553	0.690	0.558	0.680

Table A3: The estimates of Equation (2) for commenting amount and stance toward the rumor. Robust standard errors are in parentheses. *$p < 0.1$; **$p < 0.05$; ***$p < 0.01$.

Dependent Variables	ln(Number of Comments with Swearing +1)	ln(Number of Comments without Swearing +1)	ln(Number of Non-Positive Comments +1)	ln(Number of Positive Comments +1)
β_1	0.0219*	0.0123	0.0286**	0.00194
	(0.0119)	(0.00882)	(0.0134)	(0.00794)
β_2	0.0220	0.0107	0.0244	0.0109
	(0.0153)	(0.0120)	(0.0169)	(0.0103)
β_3	0.105***	0.0225	0.0958***	0.0328**
	(0.0217)	(0.0146)	(0.0231)	(0.0139)
γ	-0.0853***	-0.00966	-0.0829***	-0.0213
	(0.0284)	(0.0206)	(0.0301)	(0.0186)
User FE	Included	Included	Included	Included
No. of Users	1,617	1,617	1,617	1,617
Observations	6,468	6,468	6,468	6,468
R-Squared	0.668	0.631	0.675	0.607

Table A4: The estimates of Equation (2) for commenting hostility. Robust standard errors are in parentheses. *$p < 0.1$; **$p < 0.05$; ***$p < 0.01$.

References

Allcott, H., and Gentzkow, M. 2017. Social Media and Fake News in the 2016 Election. *Journal of Economic Perspectives* 31(2): 211-236.

Antecol, H.; Bedard, K.; and Stearns, J. 2018. Equal but Inequitable: Who Benefits from Gender-Neutral Tenure Clock Stopping Policies? *American Economic Review* 108(9): 2420-2441.

Barnes, T., and Almasy, S. May 2, 2020. Kim Jong Un seen laughing, smiling, smoking and waving to crowds, North Korea state media reports. *CNN*. https://edition.cnn.com/2020/05/01/asia/kim-jong-un-public-appearance-kcna/index.html (accessed on May 31, 2020).

Berinsky, A. J. 2017. Rumors and Health Care Reform: Experiments in Political Misinformation. British Journal of Political Science 47(2): 241-262.

Boin, A.; Hart, P. T.; McConnell, A.; and Preston, T. 2010. Leadership style, crisis response and blame management: The case of Hurricane Katrina. *Public Administration* 88(3): 706-723.

Bovet, A., and Makse, H. A. 2019. Influence of fake news in Twitter during the 2016 US presidential election. *Nature Communications* 10(7): 1-14.

Ciampagnia, G. L.; Mantzarlis, A.; Maus, G.; and Menczer, F. 2018. Research Challenges of Digital Misinformation: Toward a Trustworthy Web. *AI Magazine* 39(1): 65-74.

Cho, D.; Kim, S.; and Acquisti, A. 2012. Empirical analysis of online anonymity and user behaviors: the impact of real name policy. In *the 45th Hawaii International Conference on System Sciences (HICSS)*. Maui, HI.

Cho, D., and Kwon, K. H. 2015. The impacts of identity verification and disclosure of social cues on flaming in online user comments. *Computers in Human Behavior* 51: 363–372.

Choe, S.-H. May 2, 2020. Kim Jong-un Is Back. What Happens When He's Really Gone? *The New York Times* https://www.nytimes.com/2020/05/02/world/asia/kim-jong-un-alive.html (accessed on May 31, 2020).

Datta, H.; Knox, G.; and Bronnenberg, B. J. 2018. Changing Their Tune: How Consumers' Adoption of Online Streaming Affects Music Consumption and Discovery. *Marketing Science* 37(1): 5-21.

Davidson, M. 2017. Vaccination as a cause of autism—myths and controversies. *Dialogues in Clinical Neuroscience* 19(4): 403-407.

Festinger, L. 1962. Cognitive Dissonance. *Scientific American* 207(4): 93-106.

Garrett, R. K. 2011. Troubling Consequences of Online Political Rumoring. *Human Communication Research* 37(2): 255-274.

Gerber, A. S., & Huber, G. A. (2010). Partisanship, Political Control, and Economic Assessments. *American Journal of Political Science* 54(1): 153–173.

Gilge, E. P. 2019. Does Local Access to Finance Matter? Evidence from U.S. Oil and Natural Gas Shale Booms. *Management Science* 65(1): 1-18.

Hameleers, M.; and van der Meer T. G. L. A. 2020. Misinformation and Polarization in a High-Choice Media Environment: How Effective Are Political Fact-Checkers? *Communication Research* 47(2): 227-250.

Han, J.; Lee, Y.; Lee, J.; and Cha, M. 2019. The Fallacy of Echo Chambers: Analyzing the Political Slants of User-Generated News Comments in Korean Media. In *Proceedings of the 2019 EMNLP Workshop W-NUT: The 5th Workshop on Noisy User-generated Text*. Hong Kong.

Jiang, S.; Metzger, M.; Flanagin, A.; and Wilson, C. 2020. Modeling and Measuring Expressed (Dis)belief in (Mis)information. In *Proceedings of the 14th International AAAI Conference on Web and Social Media (ICWSM)*. Atlanta, Georgia, U.S.

Jiang, S.; and Wilson, S. 2018. Linguistic Signals under Misinformation and Fact-Checking: Evidence from User Comments on Social Media. In *Proceedings of the ACM on Human-Computer Interaction* 2(CSCW), Article 82.

Joulin, A.; Grave, E.; Bojanowski, P.; Douze, M.; Jégou, H.; and Mikolov, T. 2016. FastText.zip: Compressing text classification models. *arXiv preprint* arXiv:1612.03651.

Kwon, K. H.; and Rao, H. R. 2017. Cyber-rumor sharing under a homeland security threat in the context of government Internet surveillance: The case of South-North Korea conflict. *Government Information Quarterly* 34: 307-316.

Larcom, S.; She, P.; and van Gevelt, T. 2019. The UK summer heatwave of 2018 and public concern over energy security. *Nature Climate Change* 9(5): 370-373.

Lee, J. 2020. KcBERT: Korean Comments BERT. In *Proceedings of the 32nd Annual Conference on Human and Cognitive Language Technology*: 437-440.

Margolin, D. B.; Hannak, A.; and Weber, I. 2018. Political Fact-Checking on Twitter: When Do Corrections Have an Effect? *Political Communication* 35(2): 196-219.

Marikyan, D.; Papagiannidis, S.; and Alamanos, E. 2020. Cognitive Dissonance in Technology Adoption: A Study of Smart Home Users. *Information Systems Frontiers*, forthcoming.

Meyer, B. D. 1995. Natural and Quasi-Experiments in Economics. *Journal of Business & Economic Statistics* 13(2): 151-161.

Miškolci, J.; Kováčová, L.; and Rigová, E. 2020. Countering Hate Speech on Facebook: The Case of the Roma Minority in Slovakia. *Social Science Computer Review* 38(2): 128-146.

Moon, J.; Cho, W. I.; and Lee, J. 2020. BEEP! Korean Corpus of Online News Comments for Toxic Speech Detection. *arXiv preprint* arXiv:2005.12503.

Nyhan, B.; and Reifler, J. 2010. When Corrections Fail: The Persistence of Political Misperceptions. *Political Behavior* 32(2): 303-330.

Park, S.; Byun, J.; Baek, S.; Cho, Y.; and Oh, A. 2018. Subword-level word vector representations for Korean. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*. Melbourne, Australia.

Perrett, C. April 27, 2020. South Korean official says Kim Jong Un is 'alive and well' amid long public absence and rumors he died or was in grave condition. *Business Insider*. https://www.businessinsider.com/kim-jon-un-alive-south-korean-official-2020-4 (accessed on May 31, 2020).

Shibutani, T. 1966. *Improvised News: A Sociological Study of Rumor*. Indianapolis, IN: The Bobbs-Merrill Company Inc.

Shin, J.; Jian, L.; Driscoll, K.; and Bar, F. 2017. Political rumoring on Twitter during the 2012 US presidential election: Rumor diffusion and correction. *New Media & Society* 19(8): 1214-1235.

Shin, J., and Thorson, K. 2017. Partisan Selective Sharing: The Biased Diffusion of Fact-Checking Messages on Social Media. *Journal of Communication* 67(2): 233-255.

Stieglitz, S., and Dang-Xuan, L. 2013. Emotions and Information Diffusion in Social Media—Sentiment of Microblogs and Sharing Behavior. *Journal of Management Information Systems* 29(4): 217-248.

Stratos, K. 2017. A sub-character architecture for Korean language processing. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Copenhagen, Denmark.

Thorson, E. 2016. Belief Echoes: The Persistent Effects of Corrected Misinformation. *Political Communication* 33(3): 460-480.

Walter, N.; Cohen, J.; Holbert, R. L.; and Morag, Y. 2020. Fact-Checking: A Meta-Analysis of What Works and for Whom. *Political Communication* 37(3): 350-375.

Weeks, B. E., and Garrett, R. K. 2014. Electoral Consequences of Political Rumors: Motivated Reasoning, Candidate Rumors, and Vote Choice during the 2008 U.S. Presidential Election. *International Journal of Public Opinion Research* 26(4): 401-422.

Wood, T.; Porter, E. 2019. The Elusive Backfire Effect: Mass Attitudes' Steadfast Factual Adherence. *Political Behavior* 41(1): 135-163.