Are Mutated Misinformation More Contagious? A Case Study of COVID-19 Misinformation on Twitter

Muheng Yan
muheng.yan@pitt.edu
University of Pittsburgh
Pittsburgh, United States of America

Yu-Ru Lin
yurulin@pitt.edu
University of Pittsburgh
Pittsburgh, United States of America

Wen-Ting Chung
wenting.chung@gmail.com
University of Pittsburgh
Pittsburgh, United States of America

ABSTRACT

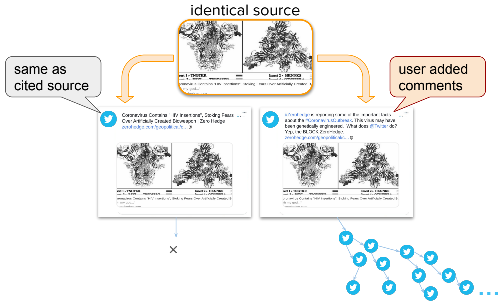
The spread of online misinformation has become a major global risk. Understanding how misinformation propagates on social media is vital. While prior studies suggest that the content factors, such as emotion and topic in texts, are closely related to the dissemination of misinformation, the effect of users' commentary on misinformation during its spreading on social media has been long overlooked. In this paper, we identify the patterns of "misinformation mutation" which captures ways misinformation is commented and shared by social media users. Our study focus on misinformation originated from digital news outlets and shared on Twitter. Through an analysis of over 240 thousand tweets capturing how users share COVID-19 pandemic-related misinformation news over a five-month period, we study the prevalence and factors of the misinformation mutation. We examine the different kinds of mutation in terms of how the article was cited from the news source, and how the content was edited, compared with its original text, and test the relationship between misinformation's mutation and its spread on Twitter. Our results indicate a positive relationship between information mutation and spreading outcome – and such a relationship is stronger for news articles shared from non-credible outlets than those from credible ones. This study provides the first quantitative evidence of how misinformation propagation may be exacerbated by users' commentary. Our study contributes to the understanding of misinformation spreading on social media and has implications for countering misinformation.

KEYWORDS

Misinformation, Social Media, Misinformation Propagation, Misinformation Mutation, COVID-19

ACM Reference Format:
Muheng Yan, Yu-Ru Lin, and Wen-Ting Chung. 2022. Are Mutated Misinformation More Contagious? A Case Study of COVID-19 Misinformation on Twitter. In *14th ACM Web Science Conference 2022 (WebSci '22), June 26–29, 2022, Barcelona, Spain.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3501247.3531562

WebSci '22, June 26–29, 2022, Barcelona, Spain
© 2022 Association for Computing Machinery.
ACM ISBN 978-1-4503-9191-7/22/06…$15.00
https://doi.org/10.1145/3501247.3531562



Coronavirus Contains "HIV Insertions", Stoking Fears Over Artificially Created Bioweapon | Zero Hedge
zerohedge.com/geopolitical/c…

#Zerohedge is reporting some of the important facts about the #CoronavirusOutbreak. This virus may have been genetically engineered. What does @Twitter do? Yep, the BLOCK ZeroHedge.
zerohedge.com/geopolitical/c…

(a) A tweet from a user having **3K** followers received **0** Retweets & **0** Likes.

(b) A tweet from a user having **3K** followers received **144** Retweets & **170** Likes.

Figure 1: Misinformation mutation. The misinformation from the same source (i.e., the same news article with an identical URL) may have different spreading outcomes. The tweet without any user-added content (a) received significantly lower Retweets and Likes, compared with the tweet with user-added content (b). (The figure was modified from the screenshots of users' tweets to highlight the content differences.)

1 INTRODUCTION

As the Internet communication platforms proliferate, the media has transformed from traditional journalism with a few limited channels serving as information gatekeepers to a broader range of online news outlets [30]. This in part gives rise to the creation of misinformation from a variety of non-credible outlets [31]. Social media has become the primary space where online misinformation propagates [30, 45]. Why does some misinformation propagate more widely than others? Existing literature has identified factors such as emotional signals or topics in content [23, 44], or information sharers' network properties (e.g., the number of followers or connectivity of social media users) [21, 23, 44]. These known factors, however, fail to explain the differential spreading outcomes when misinformation carries signals from the same sources over similar network structures. Fig 1 illustrates such a phenomenon: the two tweets distributed identical unreliable news articles reached rather different levels of propagation, even though the two users who posted the tweets have roughly the same numbers of followers.

This work aims to examine how the source misinformation reaches a wider audience through the contribution of information sharers. Here, "*misinformation mutation*" is referred to as the various degrees of work social media users contributed in a post when cited a piece of misinformation – from simply copy-and-paste to adding their own sentiments or opinions.

We target on COVID-19 misinformation on Twitter in this work to interrogate the nature of misinformation mutation. Since the onset of the pandemic, the propagation of online misinformation on social media has led to harmful behaviors and a threat to the disease control [10, 34], referred by WTO as "Infodemic." We focus on social media posts that incorporate one or more online news URL(s) from digital news outlets (namely, the *source information* or *misinformation news*). The tweets that distribute misinformation news are *misinformation tweets*. Our research can be structured into the following questions:

- **RQ1** **(Mutation)** *In what way a piece of COVID-19 relevant information can mutate, i.e., altered by its information sharers on social media? Does the information originated from less credible outlets mutate greater, or more often, compared with that from reliable outlets?*
- **RQ2** **(Spreading)** *Will mutant information, compared with non-mutant information, spread wider on social media? Will the mutant information originate from credible and non-credible outlets spread differently?* What are the factors associated with misinformation mutation and its propagation outcome?

The key challenges in answering these research questions lie in: (1) how to quantitatively characterize the types and degrees of misinformation mutation, and (2) how to simultaneously compare different factors that can also contribute to misinformation propagation outcomes. In this work, we propose an analytical approach to tackle both challenges, which results in the following contributions:

1. We develop a computational approach to differentiate various ways a piece of COVID-19 relevant information can mutate, from simply sharing URL, to alter its title and content, to change its content by adding personal input. We leverage the state-of-the-art ALBERT model to measure the *semantic mutation* from tweets with a 82% accuracy.
2. Using the new measurements of mutation, we are able to qualitatively and quantitatively capture the degree of mutation in shared information and compare that across different source credibility. We show that information originated from non-credible outlets mutate roughly 16% less often than that from reliable outlets.
3. We identify the effect of mutation on the information spreading outcome through statistical comparisons with possible confounding factors, including information topics, sharers' audience size, and use of emotional-laden languages. We incorporate generalized linear models and pathway analysis to triangulate the analysis. Our results indicate that, controlling for possible confounders, mutant information spread 1.28 times more than non-mutant information, suggesting that information becomes more viral with users' additional commentary.

This study provides the first empirical evidence on the degree and prevalence of misinformation mutation, and how the misinformation propagation differs across misinformation mutates through users' sharing on social media. Our findings identify mutation due to information shares' commentary as a significant factor where misinformation likely to spread more, which has not been well understood. Our findings shed light on on new pathways where misinformation can turn viral, which has an implication for how social media platforms may prioritize resources to counter the spread of misinformation.

2 RELATED WORKS

2.1 Social Media Misinformation Propagation

Many research works have studied why certain misinformation spreads more widely than others in social networks [17, 23, 26, 44, 45]. Except for recent work on multimodal characterization [32], most studies focused primarily on linguistic features that corresponded to the social media spreading [21, 23, 44, 45]. In these works, the social media misinformation was characterized by the socio-psychological linguistic features, including emotional words [44], hashtags, mentions [21], and writing styles defined by combinations of the above features [45]. Heimbach and colleagues provide a comprehensive collection of the linguistic factors and demonstrate that the negative emotions and the use of hashtags and mention functions are positively related to the information spreading across different platforms [23]. Chou et al. demonstrate that the linguistic factors, combined with social network characteristics (i.e., neighborhood users' socio-cultural identity and value) can to a great extent decide the receptivity of health-related misinformation [14]. Notably, among these works, the audience size measured by the number of followers on social media is considered a significant baseline factor to spreading [21, 23, 44], or naturally controlled or canceled by the study design [14, 45].

As digital news is being shared on social media, users' comments become part of the spreading messages themselves. Thus, the original signals carried in misinformation may be altered while being spread in users' network. To our best knowledge, such evolving content in misinformation spreading has not been explored.

2.2 Mutation of Information/Misinformation

Two streams of research trying to understand misinformation mutation: topic shifts of misinformation and meme tracking. The works in uncovering topic shifts have focused on exploring change in the topic of rumors and fake news in certain domains (e.g., the U.S. presidential election, vaccines, and the recent COVID-19 pandemic [5, 24, 42]), at a long-term time range (weeks to months). The studies in tracking memes – a form of information online containing images and texts that is fast-spreading and mutates rapidly [43] – model large families of memes and does not focus on heterogeneous contents in individual memes [1, 15, 16, 22, 27, 38, 39, 43]. As illustrated in Fig 1, even a slight change within a family of same-topic posts can result in a radically different spreading outcome, which differences cannot be understood by previous meme studies without sufficient content granularity. Our study makes efforts to fill this gap by unfolding how users comment on the shared news.

The second group of works focus on the long-term topical mutation of online misinformation [5, 24, 28, 42]. In a case study, Shin et al. illustrate that misinformation online recurs through time, and its topics mutate at each recurrence [42]. Johnson et al. and Boberg et al. give more examples of how misinformation topic shifts at the time scales of weeks and months, in their case studies of misinformation related to the topics of "COVID-19" and "quarantine" [5, 24]. These existing works only consider the mutation of misinformation at the very coarse level, while in this work, we take a step further to pin down the impact of mutant misinformation more precisely by examining the mutation variants associated with a single information source.

3 METHODS

This section describes our dataset, the method of measuring the misinformation mutation, and the statistical analyses that test the relationships between misinformation mutation and its spreading.

3.1 Dataset

Tweet Collection. We used the published tweet collection of COVID-19 related tweets [13], dated from January to May 2020. The dataset was first constructed by tracking a set of COVID-19 related keywords such as "coronavirus," "covid," "ncov-19," etc. and later included additional keywords that emerged from the pandemic course, such as "lockdown." We gather the IDs of the tracked tweets [1], and re-collect the metadata and content of all available tweets from Twitter's official API. This resulted in more than 80M tweets with complete information, including texts, the number of retweets and likes received, embedded URLs, and the tweet senders' information. We also collect the digital news article contents of the embedded URLs. There are in total 59K news articles collected.

Misinformation from Non-Credible Sources. In this work, we adopt the misinformation definition proposed by Lazer et al. [30], which used the credibility of the news domains as a proxy to label misinformation. Specifically, misinformation refers to information from outlets that "lack the news media's editorial norms and processes for ensuring the accuracy and credibility of information." The labeling of misinformation, in our study, is then at the level of its source rather than that of the individual news articles or tweets sharing an article. The main advantage of considering the misinformation at the outlet level as opposed to relying on fack-checked news articles is to avoid the bias toward selecting a limited set of "famous" misinformation stories that are "worth fact-checking" [30]. We employ the list of digital outlets with human-curated credibility labels [20], where the digital news outlets are classified into credibility categories of green (accountable journalism, 17 websites), yellow (low-quality journalism, 24 websites), orange (negligent or deceptive, 46 websites), and red (little regard for the truth, 61 websites). We matched the domains from the list with the URLs in the COVID-9 Tweet collection, and shortened URLs were recovered to original ones. After the matching, we found news articles from 16 green, 22 yellow, 18 orange, and 22 red domains shared in our tweet dataset. In the following analysis, we refer domains in "orange" and "red" lists as misinformation outlets. Domains in the "yellow" list, although of low quality, publish fact-based articles as reported in the analysis of Grinberg et al. [20]. For this reason, the "yellow" domains are treated as credible outlets along with those in the "green" list. Filtering the Twitter collection with the news domain list, we have the dataset (consists of 4 subsets corresponding to the green, yellow, orange, and red outlets) containing 240 thousand tweets from 33 thousand unique users. Table 1 summarizes the basic statistics of the dataset and the subsets, including the number of domains, the number of tweets, and the number of users.

	All	Green	Yellow	Orange	Red
Domain count	78	16	22	18	22
Article count	59021	23546	21615	9663	4196
Tweet count	238936	77302	80000	48349	33285
User count	33253	18806	20235	10130	6042

Table 1: The Statistics of the Dataset. We report the number of domains, the number of tweets sharing news from the domains, and the number of users corresponding to the tweets. Note that one user can share news from multiple domains, so the number of users in the "all" set does not equal to the sum of the others.

3.2 Characterizing Type and Degree of Mutation

We categorize the mutation of information in two ways. (1) Qualitatively, mutation can be distinguished by how social media users made an effort into sharing a news article – for instance, by only sharing links to the news articles or write personal comments appended to the shared news. We refer to this mutation categorization as *User Originality.* (2) For tweets including user commentary, we quantitatively distinguish the extent to which the semantic content of user commentary deviates from that of the shared news article. We refer to this as *Semantic Mutation*, measured as degree of mutation.

3.2.1 (1) Identifying User Originality in Misinformation Mutation. Typically, there are four types of actions users would take sharing news articles on Twitter. The simplest way is to share only the URL link to the articles. Other than the URL links (note that all news-sharing posts discussed here contain the link to the digital news article), one can (a) use the "sharing" functions from the news sources, which results in tweets containing the title of the shared articles; (b) add comments from the "sharing" function, which results in tweets containing both the titles and the users' added comments; and (c) compose new content in a tweet with an URL link directing to the source news. While these different ways of sharing news articles in a tweet reflect different amount of work made by the information sharers, the variations may not look distinctly by the audience, especially when the "preview window" of a cited news article also appears in a tweet. Therefore, to better distinguish the mutation types, we categorize User Originality into three discrete levels (see Table 2 for example tweets):

- **Copy-pasting Sharing (CP)**. The tweet contains no texts other than the URL directed to the shared news, or copy-pasted titles from the "sharing" functions. We consider tweets in this level with the lowest level of User Originality.
- **Commented Title (CT)**. Beyond the URLs linked to the source article, the tweet contains (1) full or part of the title of the shared news, and (2) the user's created content that may change or add new meanings besides the shared article.
- **User Originated Contents (UOC)**. Beyond the URLs, the tweet contains little or no text segments from the news title. Instead, the tweet often consists of the user's own created content. We consider such content has the highest level of originality.

[1] https://github.com/echen102/COVID-19-TweetIDs

Operationally, the tweets in CP can be directly categorized by string matching; nevertheless, the boundary between CT and UOC is less clear. Thus, we compare the tweet contents with the titles of articles on Levenshtein distance [47], and decide if the number of shared word tokens (order and deletion respected) is more than half the length of the news article titles. If so, the tweet will be categorized as CT. The copy-pasted text segments are marked as copy-pasted title contents and are excluded from user-originated texts in the later analysis. With the above definitions, we quantify the User Originality of tweets sharing news articles by how the tweets' content overlap with the article titles.

Table 2 summarizes the prevalence of each mutation type in our dataset. As shown in the table, CP is the most common type that accounts for 62.86% of the tweets, followed by UOC, which accounts for 21.85% of the tweets. We use numeric labels 0, 1 and 2 where 0 is for CP, 1 is for CT, and 2 is for UOC. Higher numeric values correspond to higher levels of User Originality.

3.2.2 (2) Measuring the Semantic Mutation. In the comments appended to the shared news articles, users may either elaborate (supporting the news article with supplemental contents or posting implications assuming the article is true), re-organize (re-stating the news article's main idea in a different language), distract (neither elaborating or denying, but changing the emphasis of the news contents), deny (negating or stating semantically opposite information), or compose contents that are fully unrelated to the shared news in semantics. Table 2 shows examples of tweets mutating the shared misinformation news articles with different semantics. We characterize the *degree of semantic mutation* by measuring the semantic differences between users' commentary appended and the main idea of the digital news.

We employ one of the state-of-the-art language models – ALBERT [29] – to inspect whether users' commentary in CT and UOC are semantically similar to the digital they share. The model is a deep neural network model pre-trained on an extensive general English corpus and fine-tuned on a "paraphrasing detection" task. In this paraphrasing detection task, a pair of texts are input to the model, and the model will judge whether the two texts are "semantically similar". Compared to other state-of-the-art models in the paraphrasing detection task, ALBERT achieves the comparable performance with significantly reduced model size [29].The model takes a pair of inputs and computes the semantic difference between the inputs. In our work, a pair consists of a given tweet and the main text from its cited news article. As suggested by prior research [9], the main idea of a news article is typically reflected by its title and leading sentences. We thus use the three leading sentences of each news article and the article titles as the article's main text. The output, the degree of semantic difference is estimated as a numeric score ranging from 0 (semantically similar) o 1 (semantically different). For each tweet (paired with title and three sentences), we take the minimum semantic difference score (most semantically similar) out of the four candidate pairs to reflect the maximum likelihood that the tweet content is semantically similar to the main idea of the cited articles.

Evaluation with human-annotated ground-truth. We use an ALBERT-xlarge (chosen among ALBERT-base, ALBERT-large, ALBERT-xlarge, and ALBERT-xxlarge, based on performance on the task) model with 60 million parameters pretrained for 125,000 steps on 16Gb raw English texts [29], and then fine-tuned on the MRPC dataset [18] for the semantic similarity detection task. After the fine-tuning, the model achieved 91% accuracy in the benchmark MPRC test. To evaluate the model's reliability on our dataset, we use stratified sampling to take 500 tweet-article text pairs based on the degree of semantic difference with ten buckets, for human annotation to have the ground truth. The human coders are instructed to judge whether the pair of tweet text and article sentences are semantically similar. Our annotation scheme extends the MRPC annotation [18], with additional annotation guideline to deal with social media languages (e.g., sarcasm). We annotate the sentence pairs with two human coders. They have agreed on 411 (82.2%) samples with the inter-coder reliability $\kappa = 0.783, k = 2, n = 500$. A third coder is introduced to resolve the disagreements. We then compare the model outputs with the human-rated ground-truth labels. The model achieved $accuracy = 0.820$, $F1 = 0.819$ and Area under Receiver Operating Characteristic Curve (AUC) = 0.856 on the 500 samples.

After the validation, we apply the trained ALBERT model on all tweets with users' comments, and infer their semantic mutation degree score. Each tweet has a score ranging from 0.0 (no semantic mutation) to 1.0 (high semantic mutation). Those tweets without comments are assigned with 0.0 as they are considered to carry the same semantic signals as the cited news articles.

3.3 Variables and Statistical Analysis

We use statistical models to estimate the effect of tweet mutation on the outcome of information spreading. Below, we describe the variable construction and the statistical models.

3.3.1 Variables. Table 3 summarizes the variables used in our statistical analyses, with a brief description of variable definition and operationalization. We offer the rationale for variable selection and construction.

The two key variables that capture the degree of mutation in a tweet are *degree of originality* and *degree of semantic mutation*. They serve both as an independent variable and a dependent variable in different statistical models. When using as an independent variable, the model seeks to test the effect of mutation on the spreading outcome. As a dependent variable, the model seeks to test the relationship between other covariates and the mutation of content in a tweet.

The key outcome variable is the *popularity* of a tweet, which captures the extent of spreading of a piece of information. It serves as a dependent variable in a statistic model. We consider both the number of retweets and the number of likes received by a tweet as indicators of the tweet's popularity. Empirically, these quantities follow a long-tail distribution. In our dataset, the sum of both quantities reaches the maximum value of 115,495, but 48.8% of tweets (116,902) do not have any likes and retweets. Therefore, to increase the robustness of the measurement, we create a binary indicator for *popularity*, with 1 indicating a tweet received at least one retweet or like and 0 otherwise. The distribution of this binarized variable is roughly balance, with 51.2% tweets having positive values (51.4%, 53.1%, 51.6%, and 45.7% in Green, Yellow, Orange, and 45.7% in Red subsets.

Type & Frequency	Example	Sem.Mut
CP 150183 (62.86%)	"*HORROR! Unconfirmed Videos From Wuhan Show People Dropping in the Street, Dropping and Dying in Hospital [URL]*"	0.00
CT 36551 (15.30%)	**[Elaboration Comment]** "*[News title] [URL] I HAVE BEEN SAYING THE SAME THING FOR OVER A MONTH..IT JUST MAKES TOO MUCH SENSE..CHINA WANTED TO HURT THE WORLD ECONOMY SPECIALLY THE USA..I AGREE WITH TED*"	0.32
	[Distraction Comment] "*Probably made in two different labs : [News title] [URL]*"	0.81
	[Denial Comment] "*[News title] [URL] Not true. Noon Sunday. I just bought 8 masks $4 ea at a local Ace hardware store. Hurry*"	0.95
UOC 52211 (21.85%)	**[Elaboration Comment]** "*Is the coronavirus getting into the US unnoticed thru luggage handling at airports? [URL]*"	0.04
	[Re-organization Comment] "*Current data shows that this virus is much less deadly that even the common flu from the 2019-2020 season... [URL]*"	0.06
	[Distraction Comment] "*Wonder WHY US is making their testing kits? [URL]*"	0.95
	[Un-related Comment] "*This is #China now Epidemic and eats animals and other toxic all [URL]*"	0.96

Table 2: Misinformation Mutation Examples. We show the number of tweets by User Originality and by semantic mutation types, followed by examples belongs to each category. We show multiple examples that are mutated differently in semantic, for the categories of CT and UOC. In general we consider *elaboration* and *re-organization* to be characterized as "semantic similar" while the *distraction*, *denial* and *un-related contents* to be the opposite as the semantic of the tweets deviated from that in the shared news. The "[News title]" token refers to the copy-pasted titles. The last column indicates the degree of semantic mutation derived from the ALBERT model.

Variable	Abbr.	Description
		Dependent Variable: the outcome of information spreading
popularity	Popularity	The popularity of a tweet that captures the extend of spreading of a piece of information. It is measured based on the sum of number of retweets and the number of likes received by a tweet and is cast to a binary variable.
		Variables of Mutation: as an independent and a dependent variable in different models
degree of originality	Originality	The degree to which the content in a user's tweet is original compared with the content of the article cited in the tweet. It is defined based on the three distinctive levels: CP, CT, and UOC.
degree of semantic mutation	Sem.Mut	The degree to which the content in a user's tweet is semantically different from the content of the article cited in the tweet. It is measured through the paraphrasing score inferred using the ALBERT model, with a continuous score between 0.0 and 1.0, ranging from most to least semantically similar.
		Other Covariates: other potential factors likely to impact the outcome
anger (article) *anxiety* (article) *sadness* (article) *positive emotion* (article)	Art.Anger Art.Anx Art.Sad Art.Posemo	The expression of emotions in an article. It is measured based on the LIWC [36] lexicon, which include words for three types of negative emotions (anger, anxiety, and sadness) and positive emotion.
anger (tweet) *anxiety* (tweet) *sadness* (tweet) *positive emotion* (tweet)	Twt.Anger Twt.Anx Twt.Sad Twt.Posemo	The expression of emotions in a tweet. It is measured similarly as the expression of emotions in an article.
audience size	Aud.Size	The expected size of audience for a tweet, measured based on the number of followers of the tweet's author.
article length	Art.Length	The length of the article shared in a tweet, measured based on the number of words in the article.
number of hashtags	#Hashtags	The number of hashtags used in a tweet.
number of mentions	#Mentions	The number of entities (people or accounts) mentioned via @-mention in a tweet.
topic	Topic	The topics of the article shared in a tweet, identified through a topic modeling approach (LDA).

Table 3: List of variables used in statistical models

We include a list of covariates that capture other potential factors that may also impact the spreading outcome or the mutation of content in a tweet. In a statistical model, the effect of tweet mutation is then adjusted by controlling for these covariates. Many of these covariates have been identified to be relevant to the outcome of information spreading, including the expression of emotions in content, either in a tweet or in the article cited by the tweet, the expected size of the audience of a tweet, the length of articles, and specific information contented in a tweet such as the use of hashtag or @-mention [17, 21, 23, 44].

Additionally, the spreading outcome may differ by topic – certain topics tend to be more popular than others. To control for the effect of different topics, we use the Latent Dirichlet Allocation (LDA) model [4] to determine the topic(s) of an article, where each article is represented using the text from its title and the first three sentences (see Sec. 3.2.2). We identify 19 topics out of the 59021 articles in our dataset. Fig. 2 shows the percentage of shared news articles over topics within each of the four credibility categories. It can be observed that the topic distributions are similar across the four credibility categories, except that in the Red subset, the percentages of Topic 1 (Scientific Myth of COVID) and Topic 9 (Trump's Public Response) are higher, while the percentages of Topic 3 (COVID's Impact on Daily Life), 5 (Travel and Quarantine), and 7 (COVID's Impact on Public Events) are lower. The Orange subset has more Topic 3 (COVID's Impact on Daily Life) and Topic 11 (COVID Timeline) articles, and the Yellow subset has more Topic 3 (COVID's Impact on Daily Life) and 4 (Public Order and Security) articles.

3.3.2 *Statistical Models.*

We employ three types of statistical analyses, including (1) logistic regression: to estimate the effect of tweet mutation on the binary spreading outcome, *popularity*, (2) ordinary least squares (OLS) regression: to estimate the relationship between the covariates and tweet mutation, measured in terms of *degree of*

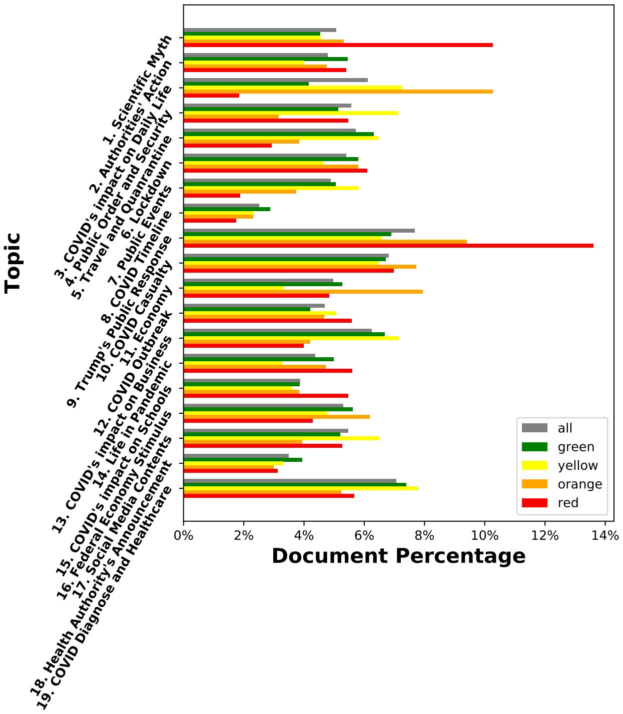


Figure 2: Article Count Percentage of Topics by Source of Different Credibility. For each topic and subset, we plot the percentage of the number of articles belonging to the topic (judged by the highest topic probability for each article) divided by the number of articles in the subset.

originality and *degree of semantic mutation*, and (3) causal analysis based on a mixed-LiNGAM model [40]: to measure the strength of the causal relationship between variable pairs.

Regression analyses. To test the relationship between tweet mutation and the information spreading outcome, we use logistic regression, where the outcome is measured as a binary variable *popularity*, and tweet mutation is measured in terms of both *degree of originality* and *degree of semantic mutation*. The estimated effect is adjusted by controlling for other potential factors such as *audience size*, *topic*, and the expression of emotions either in a tweet or in the article cited by the tweet (see the list of covariates in Table 3). By construction, `Sem.Mut` and `Originality` have linear correspondence – for example, the `CP` category corresponds to a zero value of `Sem.Mut`. Therefore, we separately estimate the effect of the mutation in terms of `Sem.Mut` and `Originality` using two Logistic Regression models with the same set of additional covariates.

These controlled factors in the digital news articles may also influence not only the spreading outcome but also the level of tweet mutation. Prior work in psychological studies found that social media users' commenting patterns may be explained by motivation theories [2, 11]. For example, the commentary patterns may be affected by the emotions expressed in the content being shared [3, 7], suggesting the negative emotion-laden languages in content, such as the expressions of anger and anxiety, are related to users' sharing and commenting behaviors. These prior observations motivate us to further examine the effect of the emotion-laden languages in the context of misinformation spreading on tweet mutation. We use the OLS regression, to examine the relationship between the different expressions of emotions and the tweet mutation, where the outcome variable is either the discrete levels of *degree of originality* or the of continuous measurement *degree of semantic mutation*. The estimated effect is adjusted by controlling for other potential factors such as *audience size* and *article length*, which has been found to be linked to users' response to the information being shared [23].

Causality Inference Model. In a regression model, the effect of a variable is typically adjusted by controlling for other potentially confounding factors – that is, by adding these factors as covariates in the model. Nevertheless, in many cases, the confounding factors may not be observable, which may lead to an over-estimation of the effect of observed factors. For example, social media users' trust in the information outlets, which is not readily observable, may increase both users' engagement to comment and wiliness to share information from the outlets [2]. Thus, we employ a causality inference model to further examine the causal relationship between variable pairs of interest, as discussed in the regression analyses, by considering potentially unobserved confounders.

While Structural Equation Modelling (SEM) [6] has been widely used in social science studies to discover potential quasi-causal relationships from observational data, this method heavily relies on two assumptions that cannot be guaranteed to hold in our case: the modeled variables should follow Gaussian distribution, and there exist no latent (unobserved) confounding variables [40, 41]. Therefore, we choose to use a recently developed Mixed-LiNGAM model [40], which relaxes the assumptions of both variable distribution and unobserved variables. Mixed-LiNGAM allows us for estimating the quasi-causal coefficient as the strength of the causal relationship between two variables, v_1 and v_2, which can have two possible directions: $v_1 \rightarrow v_2$ or $v_2 \rightarrow v_1$. The direction with a significant and greater quasi-causal strength then is considered to be more likely. Specifically, for any two models $\mathcal{M}_1$ (for $v_1 \rightarrow v_2$) and $\mathcal{M}_2$ (for $v_2 \rightarrow v_1$), we can compare the model likelihood $P(\mathcal{D}|\mathcal{M}_1)$ and $P(\mathcal{D}|\mathcal{M}_2)$ to decide whether the a quasi-causal direction is more likely to exist given the observed data [40]. To ensure the robustness of the estimation, we use a bootstrap sampling method to test each possible relationship for each direction 500 times.

4 RESULTS

We first answer **RQ1** by showing statistical distributions of mutation variation. We then answer **RQ2** by analyzing results from the statistical models. Finally, we explore factors associated with different information mutation and their propagation outcome.

4.1 Mutation and Mutation's Difference by Source Credibility

One-fifth of COVID-19 tweets mutated, and tweets cited low-credibility sources mutated less often. Overall, we found that 21.85% of the tweets sharing COVID-realted news mutated through users' comments. Note that those categories are mutually exclusive for a tweet, but a user might be involved in one or more categories if one created multiple tweets in different categories. As shown in the table, more than half of the tweets (60%, 59%, 63%, and

		CP	CT	UOC
Green	#Tweets	46531 (60.20%)	**12378 (16.01%)**	**18393 (23.79%)**
	#Users	11919 (45.56%)	5666 (21.66%)	8578 (32.79%)
Yellow	#Tweets	47592 (59.49%)	**12525 (15.66%)**	**19883 (24.85%)**
	#Users	13276 (47.50%)	5684 (20.34%)	8989 (32.16%)
Orange	#Tweets	**30598 (63.29%)**	7154 (14.80%)	10597 (21.92%)
	#Users	6685 (48.55%)	2735 (19.86%)	4349 (31.59%)
Red	#Tweets	**25453 (76.47%)**	4494 (13.50%)	3338 (10.03%)
	#Users	4576 (57.54%)	1723 (21.66%)	1654 (20.80%)

Table 4: Prevalence of mutation categories. The table shows the number of tweets/users fall into each mutation categories, grouped by the credibility of the cited source. The percentages in parenthesis show the proportion of tweets/users within each credibility group.

76% of the tweets sharing news from Green, Yellow, Orange, and Red sources, respectively) share COVID-19 related information by simply forwarding the URLs or using the "share" functions, resulting tweets of CP; among the four groups, tweets sharing information from the least credible outlets (Red) have the highest proportion in CP. So tweets from credible outlets (Green and Yellow groups) mutate more often: roughly 40% of the tweets either add comments to copy-pasted texts from the news or compose new content in both Green and Yellow, compared to 36% in Orange and 24% in Red. Table 4 also reports the number of users corresponding to the tweets by User Originality and credibility group. Red group has much larger user proportion in CP (58% Red vs. around 47% for *others*), and smaller user proportion posting user-originated content (21% Red vs. around 32% for *others*). Overall, the Red group has both a smaller proportion of users compose user-originated contents when sharing news, and a smaller proportion of tweets containing user-originated content.

Tweets sharing news from low-credibility sources have lower semantic mutation degrees. We ran the non-parametric Mann-Whitney U tests to compare, across groups, regarding their degrees of semantic difference (i.e., how different the tweet content commenting on a news is far from the content of the original news article). For CP (Copy-pasting Sharing), no such analysis is needed. For CT mutation (Commended Title), all the pairwise comparisons are significant, Orange < Red < Yellow < Green ($p < 0.001$). And for UOC (User Originated Contents), only the comparison, Red < Green, is significant ($p = 0.008$). Overall, tweets in the least credible group (Red) tend to mutate to a lower degree in semantics upon the shared articles, compared to those in the most credible group (Green).

4.2 Mutation's relationship with Spreading

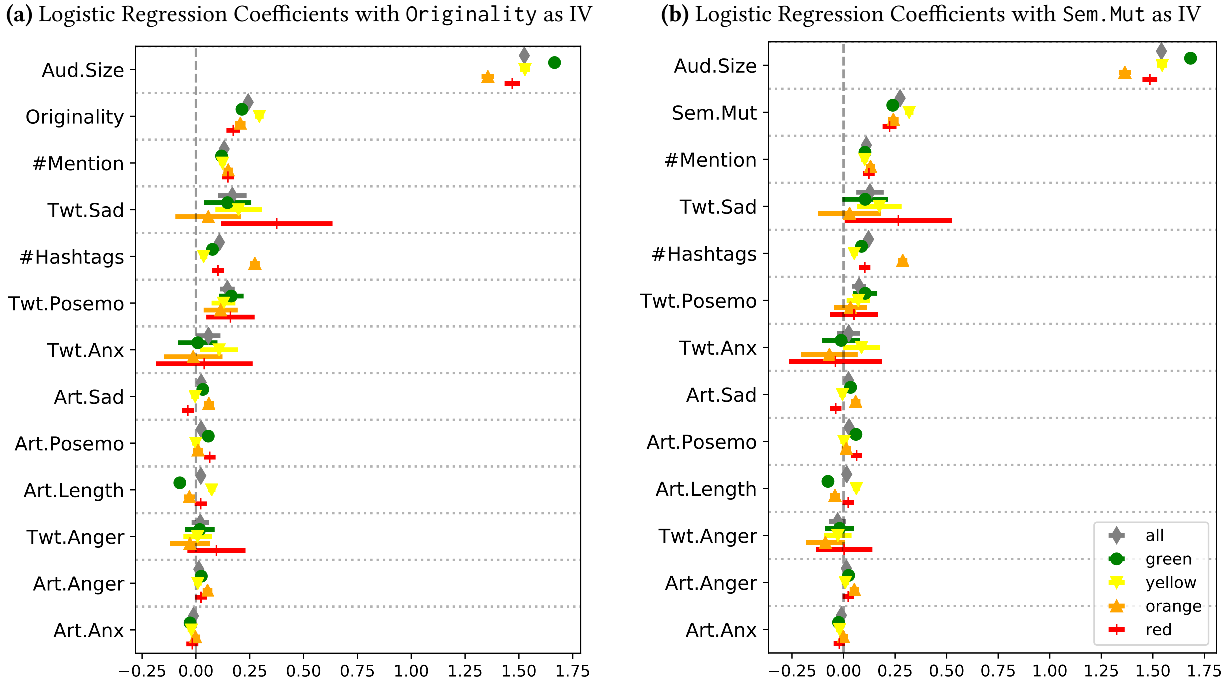
Mutation is positively associated with spreading. Fig 3 summarizes the estimated effect of mutation (measured as Originality, or Sem.Mut) on the spreading outcome (measured as Popularity). Fig 3 (a) and (b) show the results of Logistic Regression models that predict Popularity from Originality and Sem.Mut, respectively. The effects of variables, including covariates, are shown as standardized regression coefficients with a 95% confidence interval (CI); the colors indicate the four credibility groups of the news sources. Both Originality and Sem.Mut, as shown in Fig 3, are significant predictors to Popularity with baseline variables controlled. Specifically, the more additional content users put in a tweet (Originality, in Fig 3(a)), the higher its Popularity. Furthermore, the more users' commentary was semantically different from the shared news source (Sem.Mut, in Fig 3(b)), the more the tweet was reshared. Other factors, such as Aud.Size, #Hashtags, #Mentions, are also significant predictors, as expected from the prior studies [17, 23]. The effect of Originality (0.243, 0.214, 0.295, 0.206, and 0.174 for *all*, Green, Yellow, Orange, and Red sets, respectively) is stronger than other baseline variables except for Aud.Size. The effect of Sem.Mut (0.274, 0.239, 0.318, 0.241, and 0.223, respectively) are also greater than other variables. Among all the linguistic cue factors, only #Hashtags and #Mentions are consistently significant among all sets. Twt.Sad has a relatively large effect size but is not always significant across different levels of credibility.

The estimated effects of Originality and Sem.Mut on Popularity appear to be consistent across different credibility groups, suggesting that the relationship between mutation and the spreading outcome is robust over the different levels of credibility of the shared news sources.

Mutation is significantly associated with spreading while controlling for unobserved confounders. Table 5 shows the mixed-LiNGAM model results revealing the directed relationship from information mutation to spreading. We report the mean likelihoods (along with the standard deviation) of the quasi-causality path from the bootstrap experiments. The casual-relation of variable pairs follows chronological order – for example, mutation may lead to spreading outcome, but not vice versa. Relational pairs that violate the chronological order are removed from our causal inference results. The causal effects (coefficients) are estimated by using Markov Chain Monte Carlo (MCMC) to sample the causal coefficients from the identified model parameter posteriors. The last column of Table 5 reports the estimated coefficients for the corresponding causal relational pairs with a 95% confidence interval (CI).

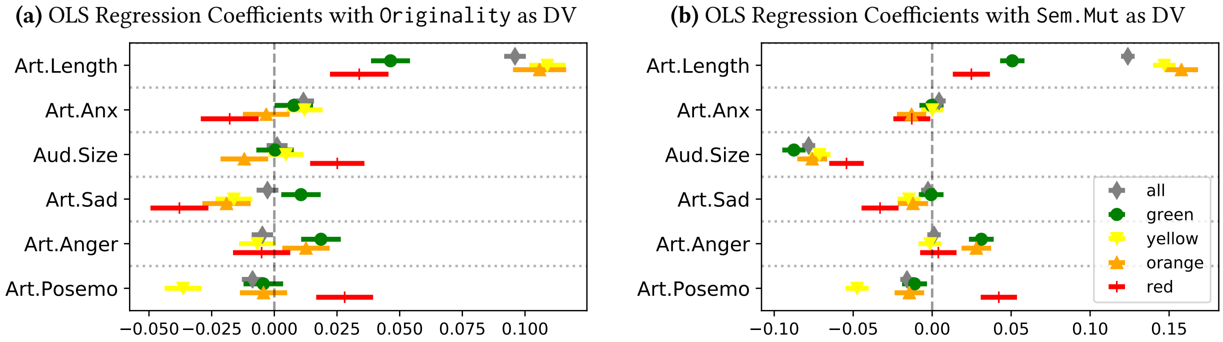
Table 5 indicates that with potential latent confounders controlled, mutation significantly relates to spreading. As shown in the column of estimated effect size, in the Originality → Popularity relationship, the model detects effects in the *all*, Green, Orange, and Red sets, and the Orange sets have significant larger coefficients compared to the Green set with non-overlapping CIs. In the Sem.Mut → Popularity relationship, we similarly detect quasi-causation effects in the *all*, Green, Orange, and Red sets, and Red set has greater coefficients than Orange and Green sets. The results show that when potential unobserved confounders controlled: (a) resonating the Logistic Regression analysis, both the Originality and Sem.Mut have a positive effect on Popularity, and (b) such positive effect is stronger on the news from less credible sources of Orange and Red, compared to those from Green and Yellow.

The results of causal estimation for the mutation-popularity relationship indicate that both causal links, Originality → Popularity and Sem.Mut → Popularity, are invalid in the Yellow subset. These results help triangulate the results from the regression model. They suggested the causal strength of these links but also offer additional insights – the causal inference model assumption could be potentially violated in a heterogeneous sample set like the Yellow, and additional steps should be taken to deal the heterogeneity in such group of samples.


(a) Logistic Regression Coefficients with `Originality` as IV

(b) Logistic Regression Coefficients with `Sem.Mut` as IV

Figure 3: Logistic Regression Coefficients with 95% confidence interval. In each sub-figure, the variables are sorted by the estimated coefficients of the *all* sub-dataset. The coefficients of the topic variables are omitted in the figure for readability. (a) Variable Coefficients of the `Originality` Predicting the `Popularity` in Logistic Regression. Among all variables besides the audience size, the `Originality` has the highest coefficient values in most data subsets (except for the `Red` set, where the estimated coefficient value of sadness in comment (Twt.Sad) is greater). (b) Variable Coefficients of the `Sem.Mut` Predicting the `Popularity`. The `Sem.Mut` have greater estimated coefficients than most of the other variables including the emotional cues in article texts and emotional cues in the users' comments (except for sadness).


(a) OLS Regression Coefficients with `Originality` as DV

(b) OLS Regression Coefficients with `Sem.Mut` as DV

Figure 4: OLS Regression Coefficients with 95% confidence interval. In each sub-figure, the variables are sorted by the estimated coefficients of the *all* sub-dataset. The coefficients of the topic variables are omitted in the figure for readability. (a) Variable Coefficients of the OLS Regression Predicting the `Originality`. No variable other than article length have consistent signs on the estimated coefficients among different sub-datasets. (b) Variable Coefficients of the OLS Regression Predicting the `Sem.Mut`. Similar to the model for User Originality, no variables other than article length have consistent coefficient signs among different sub-datasets.

Emotion-laden languages from source news have no significant impact on mutation. We conduct OLS regression and causal analyses to examine how the controlled factors in the digital news articles – specifically, the emotions expressed in the content – relate to mutation. As shown in Fig 3, the emotion-laden languages are not significant predictors to `Originality` nor `Sem.Mut`. When predicting `Sem.Mut`, only article length and audience size have consistent significant effect estimation across subsets. Similarly, in mixed-LiNGAM model (Table 5), the article emotional-laden cues (anger, anxiety, and sadness) are not detected as causal factors for

Relation	Data Subset	Est. Effect [95% CI]	Likelihood (STD)
Mutation-Popularity Relationship			
`Originality` ⟶ `Popularity`	All	0.062 [0.060, 0.064]	0.111 (0.000)
	Green	0.055 [0.052, 0.058]	0.112 (0.000)
	Yellow	NA	-0.111 (0.000)
	Orange	0.184 [0.168, 0.200]	0.108 (0.000)
	Red	0.080 [0.074, 0.086]	0.128 (0.012)
`Sem.Mut` ⟶ `Popularity`	All	0.096 [0.093, 0.099]	0.111 (0.000)
	Green	0.069 [0.062, 0.069]	0.114 (0.000)
	Yellow	NA	-0.110 (0.002)
	Orange	0.066 [0.059, 0.073]	0.113 (0.000)
	Red	0.127 [0.115, 0.139]	0.111 (0.000)
Emotion-Mutation Relationship			
`Art.Anger` ⟶ `Originality`	All	-0.041 [-0.042, -0.040]	0.129 (0.124)
	Green	NA	-0.111 (0.000)
	Yellow	-0.041 [-0.042, -0.040]	0.111 (0.000)
	Orange	NA	-0.111 (0.000)
	Red	NA	-0.111 (0.000)
`Art.Anx` ⟶ `Originality`	All	NA	-0.164 (0.211)
	Green	NA	-0.111 (0.000)
	Yellow	NA	-0.111 (0.000)
	Orange	NA	-0.111 (0.000)
	Red	-0.027 [-0.030, -0.024]	0.988 (0.069)
`Art.Sad` ⟶ `Originality`	All	NA	-0.971 (0.100)
	Green	NA	-0.990 (0.040)
	Yellow	NA	-0.147 (0.174)
	Orange	NA	-0.164 (0.211)
	Red	NA	-1.000 (0.000)
`Art.Anger` ⟶ `Sem.Mut`	All	-0.089 [-0.091, -0.087]	0.111 (0.000)
	Green	NA	-0.111 (0.000)
	Yellow	-0.089 [-0.093, -0.085]	0.111 (0.000)
	Orange	NA	-0.111 (0.000)
	Red	NA	-0.998 (0.009)
`Art.Anx` ⟶ `Sem.Mut`	All	NA	-0.111 (0.000)
	Green	NA	-0.706 (0.411)
	Yellow	NA	-0.129 (0.124)
	Orange	NA	-0.128 (0.110)
	Red	NA	-0.973 (0.082)
`Art.Sad` ⟶ `Sem.Mut`	All	NA	-0.999 (0.007)
	Green	NA	-0.302 (0.358)
	Yellow	NA	-0.164 (0.211)
	Orange	NA	-0.289 (0.356)
	Red	-0.103 [-0.121, -0.085]	0.998 (0.001)

Table 5: Causal Inference Results. This table shows the relationship direction, corresponding data (sub)set, estimated quasi-causal coefficient, and likelihood of the relation of the relationship. For the likelihood column, we report the higher likelihood of the forward and backward directions, and denote forward likelihoods with positive numbers and the backward ones with negative numbers. Since the relationships are chronically order, we consider the negative likelihood (indicating reversed quasi-causality direction) as a failure of detecting the quasi-causal relationship. Only the coefficient of successfully detected quasi-causality path is meaningful and is reported in the last table column. NA represents the invalid quasi-causal relationship with negative likelihoods.

mutation – both in terms of originality and semantic mutation degree.

5 DISCUSSION

In this study, we explored how users' commentary on COVID-19 relevant information relates to information credibility, and how the commentary relates to misinformation spreading. To answer these, we quantitatively characterize "mutation" as the types and levels of user input on a piece of information. Our method includes a measure of semantic mutation with a 82% accuracy. We found that information from non-credible outlets mutates less often than that from reliable outlets. Even though it was less observed, mutant information tends to reach large online populations. This result has an implication in allocating resources for counter misinformation: while fact-checking information is expensive, only a small portion of misinformation received users' comments, and those tend to spread more. Such mutant misinformation – one that involves more active interactions in the social media communities – should be tracked more carefully.

The results from logistic regression analysis and a quasi-causality analysis show that the mutant information spread wider on social media than non-mutant information. With relevant factors controlled, the boost from the mutation on spreading is greater in the digital news from less credible sources than that from credible sources.

We note that the effect of emotional expressions in both tweets and original articles are either not significant or have minimal effects when predicting the spreading outcome with mutation variable, which may seem to contradict to some prior studies (e.g.,[23]). To clarify, in understanding the effect of emotion-laden languages, we conducted more sophisticated analyses than prior works. First, beyond simply examining the relationship between emotional expressions in text and spreading outcome, we examines how emotional expressions have an implication on mutation, and how the mutation variation is associated with spreading outcome. Second, we also distinguish emotions into several positive and negative ones, in order to explore whether different types of emotion may have distinct effects. The weak but significant associations were found between some emotion expression and mutation, but not between emotion expression and spreading outcome. Further studies will need to better understand why in some cases that information spreading is associated with emotional expression in text, and others not. Finally, the results of our sophisticated models offer an alternative explanation to reconcile the reported contradictory findings in prior studies – whether there is a positive [3] or negative [21] relationship between the tweet sentiments and the propagation [23]. We show that the mutation relationship between the original texts and users' comments can complicate the expressed sentiments, but itself can be more predictive to the spreading outcome.

These findings open up new possibilities for counter misinformation. For example, future studies may consider the mutant information due to user commentary as one of the predictive factors for widespread misinformation and allocate more resources for tracking such information. This work provides a methodological framework that allows for tracking potentially high-virality misinformation from how they mutate on social media. While some mutation of misinformation is meant to counter misinformation – such as users' adding debunking messages for misinformation – it has been found that the retraction or correction of misinformation may not be effective or even counterproductive [12, 31]. Therefore, being able to locate and prevent potentially viral misinformation from spreading further is an important step.

6 LIMITATIONS AND FUTURE DIRECTION

This study has three limitations. First, we used the credibility levels of digital news outlets as a proxy for misinformation, which may err at the individual article level. With the proxy, information shared from the same domain will be treated to have the same credibility. Nevertheless, credible news sources can sometimes publish information that is false or misleading (and likewise, low credible outlets may publish factual stories). For example, the U.S. government recommended not wearing face masks at the early stages of the COVID-19 pandemic. Compared to the fact-check labels, using the source credibility as the proxy of the information credibility, on one hand, avoids our study being biased towards a certain set of "popular false stories" or "stories that worth fact-checking", but on the other hand may make the misinformation labels less accurate. Moreover, the list of domains reported by Grinberg et al. [20] is targeted initially to study fake news in the U.S. election. This list may be biased towards the websites actively reporting political news. Future works may examine the findings in differently sampled data to address this limitation.

Second, our findings were derived from the data specific to COVID-19, so further studies are required to examine whether the mutation patterns and their associations with information spreading may vary across contexts, as well as what confounding factors will be involved. While particular contexts may elicit unique public needs for information and motives for spreading the news (fake or true), our results are expected to be more applicable to events that share similar social significance and consequences such as health issues, epidemics, or risky situations with high uncertainty. For example, compared with a significant political event that typically would lead to people's active commenting due to distinct stances and a desire to debate (i.e., mutative spreading), a health event such as COVID-19 may attract people at large to seek information that can address uncertainty and so tend to forward the news about resolutions such as cures and vaccine developments as it is (i.e., no mutation).

Third, this work assumed that users acted independently when sharing news articles, which may not hold true in some cases. Keller et al. [25] suggested that misinformation may be propagated as a "Centrally Organized Message Coordination" in which fake or biased digital news can be amplified on the social media by different methods such as maneuvered passive re-sharing or fabricated comments by opinion leaders [19]. Thus the users' interactions on social media can be different depending on whether a misinformation topic has campaigned and how it has campaigned. For example, political misinformation propaganda may leverage crowdturfing, where human agents are employed to disseminate propaganda by various actions, including repeated automatic forwarding or composing elaborative posts [33, 46]. Additionally, automated accounts, human or cyborg accounts may all participate in the campaigns [8]. This study does not consider coordinated or bot-related campaigns. Future works may extend our study by exploring misinformation mutation in other domains and validate if the discovered pattern is applicable in those domains.

Finally, future work may further improve the measurement of semantic mutation by extending the present work with more label data or semi-supervised learning approaches [35]. We leveraged the latest state-of-the-art NLP techniques to infer the semantic entailment between mutant and un-mutant contents – but the model is not perfectly predicting the entailment still. If more gold-standard labels for model training are available in the future, the noises in the analysis can be further reduced and controlled.

ACKNOWLEDGMENTS

The authors would like to acknowledge the support from NSF #2027713 and AFOSR awards. Any opinions, findings, and conclusions or recommendations expressed in this material do not necessarily reflect the views of the funding sources.

REFERENCES

[1] Lada A Adamic, Thomas M Lento, Eytan Adar, and Pauline C Ng. 2016. Information evolution in social networks. In *Proceedings of the ninth ACM international conference on web search and data mining*. 473–482.

[2] Oberiri Destiny Apuke and Bahiyah Omar. 2020. Fake news and COVID-19: modelling the predictors of fake news sharing among social media users. *Telematics and Informatics* (2020), 101475.

[3] Jonah Berger and Katherine L Milkman. 2012. What makes online content viral? *Journal of marketing research* 49, 2 (2012), 192–205.

[4] David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research* 3, Jan (2003), 993–1022.

[5] Svenja Boberg, Thorsten Quandt, Tim Schatto-Eckrodt, and Lena Frischlich. 2020. Pandemic populism: Facebook pages of alternative news media and the corona crisis–A computational content analysis. *arXiv preprint arXiv:2004.02566* (2020).

[6] Kenneth A Bollen and Rick H Hoyle. 2012. Latent variables in structural equation modeling. (2012).

[7] Margaret M Bradley and Peter J Lang. 2007. Emotion and motivation. (2007).

[8] Samantha Bradshaw and Philip N Howard. 2018. Challenging truth and trust: A global inventory of organized social media manipulation. *The Computational Propaganda Project* 1 (2018).

[9] Ronald Brandow, Karl Mitze, and Lisa F Rau. 1995. Automatic condensation of electronic publications by sentence selection. *Information Processing & Management* 31, 5 (1995), 675–685.

[10] Aengus Bridgman, Eric Merkley, Peter John Loewen, Taylor Owen, Derek Ruths, Lisa Teichmann, and Oleg Zhilin. 2020. The Causes and Consequences of COVID-19 Misperceptions: Understanding the Role of News and Social Media. (2020).

[11] Andrew Chadwick and Cristian Vaccari. 2019. News sharing on UK social media: Misinformation, disinformation, and correction. *Survey Report. Available online: https://repository. lboro. ac. uk/articles/News_sharing_on_UK_social_media_misinformation_disinformation_and_correction/9471269 Accessed on* 25 (2019).

[12] Man-pui Sally Chan, Christopher R Jones, Kathleen Hall Jamieson, and Dolores Albarracín. 2017. Debunking: A meta-analysis of the psychological efficacy of messages countering misinformation. *Psychological science* 28, 11 (2017), 1531–1546.

[13] Emily Chen, Kristina Lerman, and Emilio Ferrara. 2020. Tracking Social Media Discourse About the COVID-19 Pandemic: Development of a Public Coronavirus Twitter Data Set. *JMIR Public Health and Surveillance* 6, 2 (2020), e19273.

[14] Wen-Ying Sylvia Chou, April Oh, and William MP Klein. 2018. Addressing health-related misinformation on social media. *Jama* 320, 23 (2018), 2417–2418.

[15] Michele Coscia. 2013. Competition and success in the meme pool: a case study on quickmeme. com. *arXiv preprint arXiv:1304.1712* (2013).

[16] Constance de Saint Laurent, Vlad P Glăveanu, and Ioana Literat. 2021. Internet memes as partial stories: Identifying political narratives in coronavirus memes. *Social Media+ Society* 7, 1 (2021), 2056305121988932.

[17] Michela Del Vicario, Alessandro Bessi, Fabiana Zollo, Fabio Petroni, Antonio Scala, Guido Caldarelli, H Eugene Stanley, and Walter Quattrociocchi. 2016. The spreading of misinformation online. *Proceedings of the National Academy of Sciences* 113, 3 (2016), 554–559.

[18] William B Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*.

[19] Ryan J Gallagher, Larissa Doroshenko, Sarah Shugars, David Lazer, and Brooke Foucault Welles. 2021. Sustained online amplification of COVID-19 elites in the United States. *Social Media+ Society* 7, 2 (2021), 20563051211024957.

[20] Nir Grinberg, Kenneth Joseph, Lisa Friedland, Briony Swire-Thompson, and David Lazer. 2019. Fake news on Twitter during the 2016 US presidential election. *Science* 363, 6425 (2019), 374–378.

[21] Lars Kai Hansen, Adam Arvidsson, Finn Årup Nielsen, Elanor Colleoni, and Michael Etter. 2011. Good friends, bad news-affect and virality in twitter. In *Future information technology*. Springer, 34–43.

[22] Chip Heath, Chris Bell, and Emily Sternberg. 2001. Emotional selection in memes: the case of urban legends. *Journal of personality and social psychology* 81, 6 (2001), 1028.

[23] Irina Heimbach, Benjamin Schiller, Thorsten Strufe, and Oliver Hinz. 2015. Content virality on online social networks: Empirical evidence from Twitter, Facebook, and Google+ on German news websites. In *Proceedings of the 26th ACM Conference on Hypertext & Social Media*. 39–47.

[24] NF Johnson, N Velasquez, OK Jha, H Niyazi, R Leahy, N Johnson Restrepo, R Sear, P Manrique, Y Lupu, P Devkota, et al. 2020. Covid-19 infodemic reveals new tipping

point epidemiology and a revised *R* formula. *arXiv preprint arXiv:2008.08513* (2020).

[25] Franziska B Keller, David Schoch, Sebastian Stier, and JungHwan Yang. 2020. Political Astroturfing on Twitter: How to coordinate a disinformation Campaign. *Political Communication* 37, 2 (2020), 256–280.

[26] Jooho Kim and Makarand Hastak. 2018. Social network analysis: Characteristics of online social networks after a disaster. *International Journal of Information Management* 38, 1 (2018), 86–96.

[27] Simon Kirby, Hannah Cornish, and Kenny Smith. 2008. Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language. *Proceedings of the National Academy of Sciences* 105, 31 (2008), 10681–10686.

[28] Ania Korsunska. 2019. The Spread and Mutation of Science Misinformation. In *International Conference on Information*. Springer, 162–169.

[29] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942* (2019).

[30] David MJ Lazer, Matthew A Baum, Yochai Benkler, Adam J Berinsky, Kelly M Greenhill, Filippo Menczer, Miriam J Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, et al. 2018. The science of fake news. *Science* 359, 6380 (2018), 1094–1096.

[31] Stephan Lewandowsky, Ullrich KH Ecker, Colleen M Seifert, Norbert Schwarz, and John Cook. 2012. Misinformation and its correction: Continued influence and successful debiasing. *Psychological science in the public interest* 13, 3 (2012), 106–131.

[32] Unal Mesut Erhan, Adriana Kovashka, Wen-Ting Chung, and Yu-Ru Lin. 2022. Visual Persuasion in COVID-19 Social Media Content: A Multi-Modal Characterization. In *Companion Proceedings of the Web Conference 2022 (WWW '22 Companion)*. ACM.

[33] JD Moffitt, Catherine King, and Kathleen M Carley. 2021. Hunting Conspiracy Theories During the COVID-19 Pandemic. *Social Media+ Society* 7, 3 (2021), 20563051211043212.

[34] Katherine Ognyanova, David Lazer, Ronald E Robertson, and Christo Wilson. 2020. Misinformation in Action: Fake News Exposure Is Linked to Lower Trust in Media, Higher Trust in Government When Your Side Is In Power. *HKS Misinfo. Rev.* (2020).

[35] Aarthi Paramasivam and S Jaya Nirmala. 2021. A survey on textual entailment based question answering. *Journal of King Saud University-Computer and Information Sciences* (2021).

[36] James W Pennebaker, Martha E Francis, and Roger J Booth. 2001. Linguistic inquiry and word count: LIWC 2001. *Mahway: Lawrence Erlbaum Associates* 71, 2001 (2001), 2001.

[37] Michael Röder, Andreas Both, and Alexander Hinneburg. 2015. Exploring the space of topic coherence measures. In *Proceedings of the eighth ACM international conference on Web search and data mining*. 399–408.

[38] Jens Seiffert-Brockmann, Trevor Diehl, and Leonhard Dobusch. 2018. Memes as games: The evolution of a digital discourse online. *new media & society* 20, 8 (2018), 2862–2879.

[39] Limor Shifman and Mike Thelwall. 2009. Assessing global diffusion with Web memetics: The spread and evolution of a popular joke. *Journal of the American society for information science and technology* 60, 12 (2009), 2567–2576.

[40] Shohei Shimizu and Kenneth Bollen. 2014. Bayesian estimation of causal direction in acyclic structural equation models with individual-specific confounder variables and non-Gaussian distributions. *The Journal of Machine Learning Research* 15, 1 (2014), 2629–2652.

[41] Shohei Shimizu, Patrik O Hoyer, Aapo Hyvärinen, and Antti Kerminen. 2006. A linear non-Gaussian acyclic model for causal discovery. *Journal of Machine Learning Research* 7, Oct (2006), 2003–2030.

[42] Jieun Shin, Lian Jian, Kevin Driscoll, and François Bar. 2018. The diffusion of misinformation on social media: Temporal pattern, message, and source. *Computers in Human Behavior* 83 (2018), 278–287.

[43] Matthew P Simmons, Lada A Adamic, and Eytan Adar. 2011. Memes online: extracted, subtracted, injected, and recollected. *icwsm* 11 (2011), 17–21.

[44] Stefan Stieglitz and Linh Dang-Xuan. 2013. Emotions and information diffusion in social media—sentiment of microblogs and sharing behavior. *Journal of management information systems* 29, 4 (2013), 217–248.

[45] Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science* 359, 6380 (2018), 1146–1151.

[46] Gang Wang, Christo Wilson, Xiaohan Zhao, Yibo Zhu, Manish Mohanlal, Haitao Zheng, and Ben Y Zhao. 2012. Serf and turf: crowdturfing for fun and profit. In *Proceedings of the 21st international conference on World Wide Web*. 679–688.

[47] Li Yujian and Liu Bo. 2007. A normalized Levenshtein distance metric. *IEEE transactions on pattern analysis and machine intelligence* 29, 6 (2007), 1091–1095.

A APPENDIX

List of News Outlets. From the list of news websites reported by Grinberg et al. [20], we identify the following domains that are active in the COVID-19 Twitter dataset:

- **Green Domains:** washingtontimes.com, buzzfeed.com, foxnews.com, time.com, independent.co.uk, telegraph.co.uk, dw.com, msn.com, yahoo.com, dallasnews.com, talkingpointsmemo.com, newsday.com, huffingtonpost.com, montgomeryadvertiser.com, magicvalley.com, lawofficer.com
- **Yellow Domains:** dailymail.co.uk, nypost.com, thesun.co.uk, breitbart.com, redstate.com, dailykos.com, metro.co.uk, standard.co.uk, thefederalist.com, deadstate.org, tmz.com, lifesitenews.com, lifenews.com, cosmopolitan.com, hngn.com, christiannews.net, metalsucks.net, cheezburger.com
- **Orange Domains:** zerohedge.com, express.co.uk, crooksandliars.com, dailywire.com, theconservativetreehouse.com, dailypost.ng, endoftheamericandream.com, palmerreport.com, dailycaller.com, themindunleashed.com, inquisitr.com, dennismichaellynch.com, thehornnews.com, healthnutnews.com, afa.net, ahtribune.com, medicalkidnap.com, iotwreport.com, dailyheadlines.net, concealednation.org, trueactivist.com, awarenessact.com
- **Red Domains:** thegatewaypundit.com, infowars.com, trunews.com, activistpost.com, frontpagemag.com, thelastamericanvagabond.com, dcclothesline.com, wnd.com, eutimes.net, worldtruth.tv, judicialwatch.org, 100percentfedup.com, collective-evolution.com, endtimeheadlines.org, barenakedislam.com, bipartisanreport.com, powderedwigsociety.com, conservativefiringline.com, nowtheendbegins.com, anonhq.com, wearechange.org, conservativepost.com

Sentence Pair Semantic Similarity Annotation. As reported by Dolan et al. [18] in the development of the MPRC dataset, the annotators are given pairs of texts and asked to judge whether the two texts are "semantically similar". The pair of texts are not restricted to be bi-directional entailment to each other, to be considered as "semantically similar". Different from their corpus that is extracted from only news articles, our dataset includes commentary texts appended to the other text (one-directional entailment), but the additional texts can be either elaboration, direct denial (which does not exist in the MPRC dataset), emphasis-changing texts, or context-irrelevant expressions (such as "wow" or "check this"). To solve such ambiguity for the annotators, we develop our coding instruction as follow:

In this task, you will be given a series of sentence pairs. You are asked to judge whether the two sentences are, at a high level, conveying similar meanings. To be judged as with similar meanings, the sentences need to meet the following criteria: (1) share the same context, (2) delivering similar information, and (3) in the same attitude. If these criteria are met, the pair of sentences is labeled as semantically similar to each other.

(1) Same context: the two sentences refer to the same topic, use the same schema of interpretation, emphasize the same aspect of the topic; *(2) Similar information: One sentence MAY have slightly more information. The extra information may only give supplemental information, without contradictions, make irony/sarcasm against the shared information nor change the context (specifically, do not change the topic, do not use a different schema of interpretation, and do not change the emphasized aspect)*; *(3) Same attitude: With the same context and similar information, the two sentences should NOT hold the opposite attitude (if one denying the other, they should be judged as different from each other).*

Based on these rules, please first check examples on the next page for illustration. Then please label these sentence pairs with labels 0 (different) and 1 (similar).

Misinformation Mutation Examples with Article Contents. Table **??** is the extended version of Table 2, where we present the titles and

Topic Number	**Top Keywords**	**Topic Name**	**All**	Green	Yellow	Orange	Red
1	chinese,virus,find,vaccine,scientis	Scientific Myths of COVID	2994 (5.07%)	1066 (4.53%)	983 (4.55%)	514 (5.32%)	**431 (10.27%)**
2	call,crisis,fight,nation,leader	Authorities' Action in Crisis	2829 (4.79%)	**1280 (5.44%)**	864 (4.00%)	459 (4.75%)	226 (5.39%)
3	die,late,home,family,work	COVID's impact on Daily Life	3615 (6.12%)	976 (4.15%)	1570 (7.26%)	**992 (10.27%)**	77 (1.84%)
4	man,release,woman,police,break	Public Order and Security	3287 (5.57%)	1211 (5.14%)	**1542 (7.13%)**	305 (3.16%)	229 (5.46%)
5	travel,quarantine,leave,return,passenger	Travel and Quanrantine	3378 (5.72%)	1486 (6.31%)	1398 (6.47%)	371 (3.84%)	123 (2.93%)
6	state,lockdown,order,close,measure	Lockdown	3185 (5.40%)	1366 (5.80%)	1002 (4.64%)	560 (5.80%)	**256 (6.10%)**
7	due,announce,cancel,year,hold	Public Events	2884 (4.89%)	1190 (5.05%)	**1255 (5.81%)**	360 (3.73%)	79 (1.88%)
8	week,month,day,government,early	COVID Timeline	1475 (2.50%)	**675 (2.87%)**	504 (2.33%)	223 (2.31%)	73 (1.74%)
9	trump,claim,president,medium,response	Trump's Public Response	4524 (7.67%)	1622 (6.89%)	1423 (6.58%)	908 (9.40%)	**571 (13.61%)**
10	case,death,confirm,number,report	COVID Casualty	4018 (6.81%)	1579 (6.71%)	1400 (6.48%)	**746 (7.72%)**	293 (6.98%)
11	year,global,economy,market,economic	Economy	2928 (4.96%)	1241 (5.27%)	717 (3.32%)	**767 (7.94%)**	203 (4.84%)
12	people,virus,spread,outbreak,infect	COVID Outbreak	2769 (4.69%)	988 (4.20%)	1096 (5.07%)	451 (4.67%)	**234 (5.58%)**
13	worker,staff,work,mask,food	COVID's impact on Business	3689 (6.25%)	1572 (6.68%)	**1543 (7.14%)**	407 (4.21%)	167 (3.98%)
14	time,make,pandemic,world,good	Life in Pandemic	2574 (4.36%)	1175 (4.99%)	708 (3.28%)	456 (4.72%)	**235 (5.60%)**
15	report,accord,school,send,member	COVID's impact on Schools	2286 (3.87%)	907 (3.85%)	779 (3.60%)	371 (3.84%)	**229 (5.46%)**
16	government,federal,pay,include,provide	Federal Economy Stimulus	3129 (5.30%)	1321 (5.61%)	1031 (4.77%)	**597 (6.18%)**	180 (4.29%)
17	show,video,share,post,social	Social Media Contents around COVID	3230 (5.47%)	1224 (5.20%)	**1404 (6.50%)**	381 (3.94%)	221 (5.27%)
18	health,official,public,risk,emergency	Health Authority's Announcement	2062 (3.49%)	**928 (3.94%)**	714 (3.30%)	289 (2.99%)	131 (3.12%)
19	test,patient,hospital,covid,doctor	COVID Diagnosis and Healthcare	4165 (7.06%)	1739 (7.39%)	**1682 (7.78%)**	506 (5.24%)	238 (5.67%)
Sum			59021	23546	21615	9663	4196

Table 6: Extracted Topics from the News Articles. In this table, we first show the top ten keywords characterizing each extracted topics, followed by the names we assign to each topics by summarizing the keywords. In the rest part, we report the article frequency of each topics in our data subsets, with the percentage (with respect to each subsets) in the parenthesis. The most-appeared subsets for each topic (by percentages) are marked in bold.

article contents in addition to the tweets sharing them. This gives the context for a better understanding of the examples.

Determining the Number of Topics. We evaluate the LDA models by their topic coherence score. The coherence score is calculated from the framework proposed by Roder et al. [37], where the higher the score, the extracted topics (characterized by keywords) are more coherent. We run a search on the hyper-parameter k, the number of topics from 5 to 30, and decide the optimal $k = 19$ by the elbow rule on the topic coherence score.

The top keywords and assigned topic names we manually assigned according to the top keywords are reported in Table 6. The numbers of documents in each dataset belonging to each topic are also reported in the same table.