Exploring Behavioral Tendencies on Social Media: A Perspective Through Claim Check-Worthiness

Zeyu Zhang, Zhengyuan Zhu, Haiqi Zhang, and Chengkai Li

University of Texas at Arlington, United States
{zeyu.zhang, zhengyuan.zhu, haiqi.zhang}@mavs.uta.edu, cli@uta.edu

Abstract. This study examines how factual claims of different significance influence and reflect social media users' behavioral patterns. Leveraging "check-worthiness" as a measure of the factual significance of claims, we analyze the connection between factual claims and user behaviors on Twitter. Through a series of experiments using statistical methods such as correlation analysis and hypothesis testing, we provide insights into a few pivotal inquiries: (1) whether differences exist between users' tweeting tendencies toward check-worthiness, (2) the underlying reasons for such differences, (3) whether users tend to create, share, and endorse content with check-worthiness levels similar to their own tweets, and (4) whether users with similar tendencies toward check-worthiness exhibit heightened engagement. The experiments were conducted across three datasets, comprising over 48.5 million tweets and involving 15,000 users, spanning several domains and yielding statistically significant findings. Previous studies have primarily centered on examining the effectiveness and strategies of fact-checks rather than understanding people's behavioral tendencies toward factual claims. Our research pioneers understanding in this area, offering valuable insights for behavioral modeling and social sciences.

Keywords: check-worthiness · fact-checking · social media analytics.

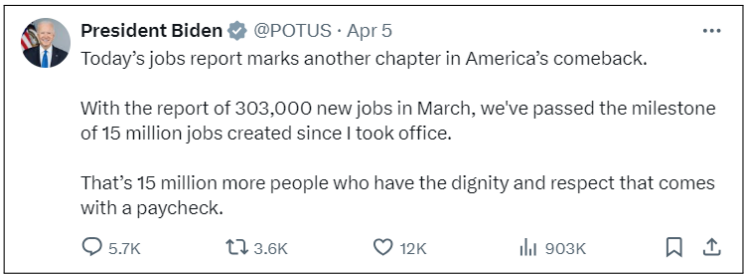
1 Introduction

In our present era, an unprecedented surge of falsehoods and partial truths has taken root within our society, posing a grave threat to national security, democratic principles, and public health. The digital worlds, notably prominent platforms such as Twitter (now called X), have become a breeding ground for fabricated news, a phenomenon that may have even cast its shadow over the presidential elections [2, 3]. In response to the epidemic of misinformation, we have witnessed a significant proliferation of fact-checking endeavors globally. Numerous researchers and experts are currently engaged in diverse works concerning factual claims, which encompass statements based on verifiable information. These efforts span activities such as detecting [13], tracking [19], and evaluating factual claims [23]. The practice of fact-checking has evolved into a pivotal interdisciplinary field, commanding attention across a spectrum of areas such as computer science, journalism, and communication.

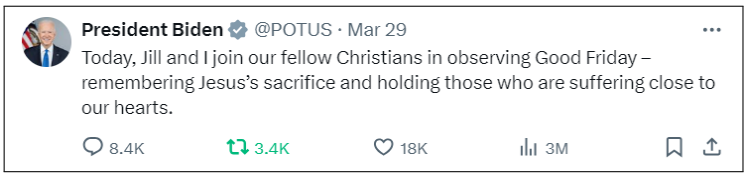
While many researchers have delved into the realm of fact-checking and factual claims, a notable knowledge gap persists in our understanding of how factual claims, each possessing varying degrees of strength or importance, wield influence over people's interactions on social media. Observations and explanations of people's behaviors pertaining to factual claims are needed in order to fill this gap. This would entail answering many crucial questions, including whether individuals exhibit behavioral tendencies toward specific factual claims and the underlying factors that drive and differentiate these tendencies. Moreover, can we apply the age-old adage "Birds of a feather flock together" to denizens of social media, particularly concerning their responsiveness to factual claims? These unknowns offer us a new perspective to study social media. The answers to these unknowns may facilitate studies such as behavioral modeling and recommendation systems by providing noteworthy new features. Moreover, it may foster research in fields such as psychology and sociology by introducing new human behavioral patterns on social media.

To study this subject, a suitable instrument is necessary to measure the strength or importance of a claim. In the field of automated fact-checking, restricting claims to those that are objectively fact-checkable makes the task more amenable to automation while reducing the volume of content needing manual fact-checking. Researchers have forged invaluable tools to detect check-worthy factual claims [14,28]. These efforts provide a yardstick for evaluating the importance of a claim to be fact-checked—denoted as "check-worthiness." As stated in [12], the initial work defined the concept of check-worthiness, check-worthy claims are those of which the general public would be interested in knowing the veracity—whether the claims are true or false. For instance, as displayed in Figure 1, (a) depicts a claim of significant check-worthiness, as it is highly probable that the public is interested in its veracity. In contrast, (b) conveys relatively low check-worthiness, as the public's interest in verifying the claim is limited. Finally, (c) exhibits the lowest check-worthiness, as there is no factual claim in the statement. By harnessing the concept of check-worthiness, a pathway emerges for investigations into how factual claims of varying importance influence and reflect people's behaviors on social media such as Twitter.

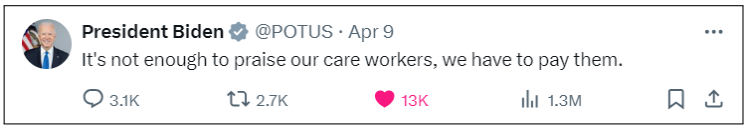
Individuals' behaviors on social media mainly consist of posting, commenting, sharing, liking, and following. They make up the communication and information diffusion on social media. Hence, many studies explored factors that influence these behaviors. For example, Comarela et al. [8] found several factors influencing user response or retweet probability, including previous responses to the same user, the user's posting rate, age, and tweet content. Firdaus et al. [9] discovered that a user's emotion towards a topic is a useful feature in modeling their retweet decision. Hopcroft et al. [15] observed that geographic distance, common friends, social status overlap, and interactions between two users (e.g., retweeting and replying) are correlated with two-way following relationships. Although a lot of studies explored the factors correlating with posting, sharing, and following behaviors on social media, none of them has looked into the impact of check-worthiness. A few slightly related studies mainly focused on the strategies and



(a) Tweet with High Check-worthiness



(b) Tweet with Relatively Low Check-worthiness



(c) Tweet with Low Check-worthiness

Fig. 1: Tweets with Different Check-worthiness Levels

effectiveness of fact-checks [6, 17, 21] rather than people's behavioral tendencies related to check-worthiness.

Considering our limited understanding of the impact of check-worthiness on social media behaviors, it is meaningful to give an investigation into it. Therefore, in this study, we conduct a range of experiments aimed at uncovering the underlying connections between check-worthiness and user behaviors on social media, particularly focusing on how check-worthiness influences/reflects individuals' tweeting, liking, and following actions. To achieve this goal, we collected 3 datasets from Twitter, comprising approximately 48.5 million tweets and 15,000 users. These datasets encompass a range of general topic domains including literature, arts, religion, politics, etc.

Our experiments on these datasets identified several pronounced results—(1) People do express different behavioral tendencies toward factual claims; (2) People in domains such as media, politics, and technology tend to have more factual claims, while people related to literature, arts, and religion generate fewer factual claims; (3) Individuals tend to share and like posts with similar check-worthiness levels as their own posts; (4) In instances where individuals followed each other, there is a higher likelihood of similar preferences towards factual claims when compared to one-way following relationships. In summary, our contributions can be delineated as follows:

- Pioneering investigation into the interplay between check-worthiness and user behaviors within social media.
- Provision of an expansive dataset comprising around 48.5 million tweets and 15,000 users, encompassing different types of tweets and domains.
- The outcomes of our experiments yield several noteworthy revelations that have the potential to stimulate diverse studies such as behavioral modeling and recommendation systems, particularly those pertaining to population behavior patterns in the context of social media platforms.

2 Related Work

Claim detection, which involves identifying claims worth fact-checking, is a task in the workflow of fact-checking in which check-worthiness is conceptualized. It can be viewed as a binary classification task and further as a ranking task on the claims. Numerous works have been dedicated to various modeling methodologies and their evaluations for claim detection [11, 14, 18, 27, 28]. Furthermore, researchers have used it for various application contexts, including automated check-worthy claims collection [19], factual claims visualization [22], fake news detection [26], and so on.

Research on identifying factors that influence posting, sharing, liking, and following behaviors in social media is directly related to our study. There are many existing works on this topic. Comarela et al. [8] conducted an extensive characterization of a large Twitter dataset which includes all users, social relations, and messages posted from the beginning of the platform up to August 2009. They identified several factors that influence user responses or retweet probability, such as previous interactions with the same tweeter, the tweeter's posting frequency, and entities in tweets such as hashtags and mentions. Firdaus et al. [5] uncovered that a user's emotional state can influence their retweeting behavior. They demonstrated this by constructing a retweet prediction framework based on an emotion detection model and conducting experiments using the Stanford Twitter Sentiment dataset [16] and the Obama–McCain Debate dataset [24]. Hopcroft et al. [15] identified that geographic distance, common friends, social status overlap, and the interactions between two users are correlated with two-way following relationships.

While there exists an abundance of research that separately addresses check-worthiness and human behaviors on social media, to the best of our knowledge, no existing work has linked them together. There are works with a focus on analyzing the relationship between fact checks and audience behaviors. For example, Kim et al. [17] analyzed 914 news articles with fact checks in South Korea. They found that news articles triggered more audience comments when they mentioned the importance of fact-checking the claim under scrutiny and conveyed negative content. Clayton et al. [6] evaluated the effectiveness of strategies for designing fact-checks by conducting experiments among 2,994 participants recruited from Amazon Mechanical Turk. Their experiments found the "Rated false" tag is more effective than the "Disputed" tag and the effect of a general

warning is small compared to these two tags. Park et al. [21] discovered the unexpected and diminished effect of fact-checking due to cognitive biases. They found that claims labeled "Lack of Evidence" were often treated as false, revealing an uncertainty-aversion bias, and users who initially disapproved of a claim were less likely to change their views when presented with opposite fact-checking labels, indicating a disapproval bias. All these studies concentrated on scrutinizing the connections between fact-checks and their audiences' behaviors, primarily with the goal of understanding people's responses to truthfulness of claims. In contrast, our study is centered on discerning the correlation between check-worthiness of claims and people's behaviors.

3 Research Questions

Our research focuses on investigating how check-worthiness of claims impacts or reflects social media users' behaviors. There are many different angles and means to tackle this topic such as retweeting analysis and content analysis. Regardless of the approaches, the essence of this topic lies in determining whether check-worthiness can serve as an indicator to capture people's behavioral patterns in common and to differentiate among groups of individuals. In this paper, we study it by observing the tweeting, following, and liking behaviors among different groups of Twitter users since those behaviors make up most of their activities on Twitter.

To investigate the impact of check-worthiness, our initial inquiry revolves around determining whether individuals exhibit varying behavioral tendencies when confronted with factual claims of differing check-worthiness levels (i.e., showing higher engagement with claims of low/high check-worthiness). If such disparity arises, we then dig into the underlying rationales and figure out whether this disparity has an influence on individuals' behaviors or reflects unique characteristics of the respective populations. To unravel these unknowns, we introduce the following specific research questions for our study.

Q1. Do people exhibit different behavioral tendencies toward check-worthiness?
Q2. What factors and commonalities within the population might account for such behavioral tendencies?
Q3. Do people maintain similar check-worthiness levels between the tweets they post and those they favor?
Q4. Do people tend to follow others who exhibit similar behavioral tendencies toward check-worthiness?

4 Datasets

The smallest research unit in this study is a tweet, which is a social media post published by a Twitter user. For the convenience of wording, we categorize tweets in our datasets into 3 types: *original-tweet*—a tweet that is initially created by a Twitter user; *retweet*—a tweet that is reposted by a Twitter user from an

original-tweet; and *liked-tweet*—a tweet that is liked by a Twitter user. As shown in Figure 1, (a) is an example of an original-tweet, (b) is an example of a retweet, and (c) is an example of a liked-tweet. We collected 3 datasets to support our study, which are accessible at *Zenodo*. [1] Due to Twitter's content redistribution policy, the datasets only include tweet IDs and user IDs instead of tweet content and user profile details. Below are detailed descriptions of the datasets.

Randomly Sampled Users Dataset (RSU) This dataset consists of 11,173 users collected through Twitter's APIs. We collected 10,000 random English tweets in February 2023 using Twitter's Volume Stream API. The tweets were posted by around 3,000 users. For each user, we collected up to 100 of its most recent followees using Twitter's Following API. Through the Timeline and Liking APIs, for each user, we collected their most recent tweets (up to 3,200 tweets due to Twitter's limit) and liked-tweets (up to 3,200 too). We then filtered out users that have insufficient tweets (less than 100 original-tweets or less than 80 retweets/liked-tweets) to ensure that the sample sizes are statistically significant in our analyses. Finally, we have 11,173 users along with 40,405,150 tweets. To prevent potential sampling biases in the collected data, we randomly selected 50 users and examined their profiles and their most recent 30 tweets to detect any biases (e.g., concentration in their backgrounds and topics). The results showed no significant concentration in the users' backgrounds or topics.

Politics Dataset (POL) This dataset contains all tweets from selected U.S. news media and U.S. politicians including *Senators*, *House Members*, *US Governors*, *US Secretaries of State*, *US Cabinet*, and *US Election Officials* at collection time. We used Twitter's Timeline API to collect the most recent tweets (up to 3,200) of the target accounts. The dataset was collected in May 2023, with 8,153,745 tweets and 3,784 Twitter accounts.

Humanities Dataset (HUM) This dataset contains 341,285 tweets and 498 Twitter accounts from selected Twitter lists including *Book Author*, *Christianity*, *Artists*, *Buddhism*, *Musician*, and *Philosophers*. We use Twitter's List and Timeline APIs to collect the accounts and their most recent tweets (up to 1,000). The dataset was collected in January 2024.

5 Methodology

Each tweet in our datasets is associated with a corresponding check-worthiness score to indicate its check-worthy level. We employed the ClaimBuster [14] API [2] to obtain check-worthiness scores. Given a tweet, the API returns a score ranging from 0 to 1, corresponding to how likely the tweet contains a check-worthy factual claim.

[1] https://zenodo.org/records/11081026
[2] https://idir.uta.edu/claimbuster/api

ClaimBuster has been used by researchers and fact-checkers in various contexts. For instance, the Duke Reporters' Lab [3] used ClaimBuster to create daily email alerts to professional fact-checkers with the most check-worthy claims from TV program transcripts and social media. These alerts have led to at least 33 claims featured in 30 different articles by fact-checking outlets, including one from The Washington Post that was discussed in a news report [10]. It has been applied in real-time for the live coverage of all primary election and general election debates of the U.S. presidential elections since 2016. Post-hoc analysis of the claims checked by professional fact-checkers at *CNN*, *PolitiFact.com*, and *FactCheck.org* reveals a highly positive correlation between ClaimBuster and fact-checkers in deciding which claims to check [13].

Although ClaimBuster has been widely applied in presidential debates, political speeches, and interviews, it is worth assessing its effectiveness on tweets, which are less formal and noisier. We did not use public datasets such as CLEF CheckThat! [4] for evaluation because their data includes multimodal features (e.g., images) and does not align perfectly with our evaluation criteria. Our labels differ from theirs by 20% in a random sample of 100 tweets from their dataset. To this end, we conducted a human evaluation on a random sample of 200 tweets selected from our datasets. Each tweet was annotated by 3 annotators who labeled them as either check-worthy or non-check-worthy. All of the 3 annotators possess the concept and experience of check-worthiness evaluation as they all have contributed to factual claims detection tasks. The final label of each tweet was decided by majority vote. We used a check-worthiness score threshold of 0.5 to classify the tweets: If a tweet received a ClaimBuster score above 0.5, it would be classified as check-worthy; otherwise, non-check-worthy. This simple classifier has an accuracy of 0.84, indicating that ClaimBuster is effective in identifying check-worthy tweets and thus it can be used as a reliable tool for analyses in our study.

Our study frequently utilizes correlation analysis and hypothesis testing in the experiments as they are simple and useful tools for identifying and verifying underlying connections between variables. In this study, we use a scatter plot to visualize the relationship between two variables and use the Pearson correlation coefficient [7] to measure the direction and strength of a linear relationship between two variables.

Hypothesis testing is widely used in verifying statistical conjectures by examining data samples. We use it to validate our presumption about individuals' behavioral tendencies toward check-worthiness. We refer to H_0 as the null hypothesis, for which we test whether to accept it. If we reject it, we will accept the alternative hypothesis H_a. In this study, we primarily use hypothesis testing to assess the equality of check-worthiness distributions across thousands of user sets, aiming to ascertain the similarities or differences between individuals' tweeting behaviors. When conducting the same hypothesis test many times using different data, one may observe some statistically significant results just by

[3] http://reporterslab.org/tech-and-check
[4] https://checkthat.gitlab.io/clef2024/task1

chance, even if there is no true effect. As we are performing some hypothesis tests thousands of times in the experiments in Section 6, the false discovery rate (FDR) using the Benjamini-Hochberg procedure [1] was applied to control false significant results by adjusting the p-values.

Generally speaking, when determining if two samples originate from the same distribution, our preference would be *Z-test* or *T-test* in instances where we have equally sized samples and can make the assumption that the underlying populations adhere to normal distributions with known variances. Nonetheless, the check-worthiness of a Twitter user's posts hardly conforms to a normal distribution. We substantiated this claim by performing Shapiro–Wilk tests [25] on the randomly sampled users dataset RSU (Section 4). Shapiro–Wilk test is one of the most popular hypothesis tests for examining how close the sample data fit to a normal distribution by ordering and standardizing the sample. Given each user in the RSU dataset, we performed Shapiro–Wilk test with significance level $\alpha = 0.05$ on the check-worthiness scores of the user's original-tweets, retweets, and liked-tweets, respectively. The result, as displayed in Table 1, shows that only a few of the null hypotheses were accepted across all users and tweet types. This suggests a very low probability that the check-worthiness scores of a user's original-tweets, retweets, or liked-tweets follow a normal distribution.

Table 1: Normality Test on Check-worthiness Distributions

H_0 ($\alpha = 0.05$)	Accept	Reject
The check-worthiness scores of a user's original-tweets are normally distributed	37	11136
The check-worthiness scores of a user's retweets are normally distributed	259	10914
The check-worthiness scores of a user's liked-tweets are normally distributed	58	11115

Given that it is highly unlikely the check-worthiness scores follow normal distributions, *Z-test* and *T-test* become less applicable. Hence, we select two non-parametric tests that are applicable under less rigorous conditions—Brunner Munzel test [4] and Kolmogorov-Smirnov test [20]. Both tests possess the capability to assess the stochastic equality of two random variables—whether one is "larger" than another—without rigorous assumptions such as identical distribution type and equal variances. The Kolmogorov-Smirnov test is more strict since it tests whether two samples are from the same distribution, while the Brunner Munzel test only examines the stochastic equality of two samples. We articulate the formal null hypotheses of these two tests as follows.

- **H_0 of Brunner Munzel (BM) test**: For randomly selected values X and Y from two populations, the probability of X being greater than Y is equal to the probability of Y being greater than X.
- **H_0 of Kolmogorov-Smirnov (KS) test**: Two sets of samples are drawn from the same (but unknown) probability distribution.

6 Experiments

6.1 Q1: Individuals' Behavioral Tendencies Toward Check-Worthiness

The very first question we want to answer is whether people exhibit different behavioral tendencies toward check-worthiness. The RSU dataset contains a large number of random users and their corresponding tweets, making it a suitable dataset for investigating this query.

The most straightforward way of checking an individual's behavioral tendency toward check-worthiness is the overall check-worthiness of their posts. Hence, for each user in the RSU dataset, we computed the median check-worthiness score of their tweets, denoted as *individual check-worthiness*. We chose the median because check-worthiness scores are typically not normally distributed and tend to be skewed. We present in a histogram (Figure 2) the distribution of individual check-worthiness of all users in the RSU dataset. It shows that, although individual check-worthiness mostly concentrates between 0.3 and 0.4, there are people who exhibit a particular tendency toward higher or lower check-worthiness. That motivates us to explore more about the underlying rationales and behavioral consequences of those preferences.

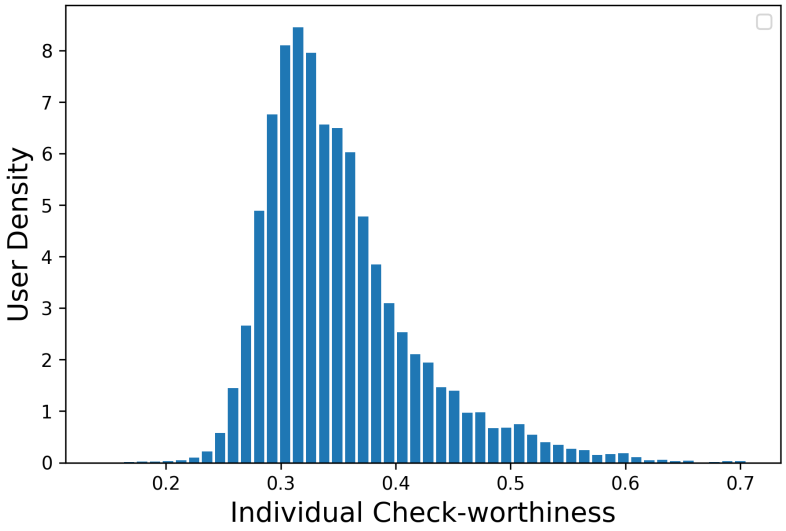


Fig. 2: Individual Check-worthiness Distribution

6.2 Q2: Causes of Different Behavioral Tendencies Toward Check-Worthiness

Knowing the difference between people's behavioral tendencies toward check-worthiness prompts us to speculate whether there are some common attributes

correlated with those preferences. To investigate this question, we conducted correlation analyses on various features of users in the RSU dataset.

First of all, we analyzed the numeric features—posts count, favorites count, followers count, followees count, listed count (number of lists containing the user), and media count (number of posts containing images or videos). These features primarily reflect a user's popularity and activity level. We performed a univariate correlation analysis by calculating the Pearson correlation coefficients between individual check-worthiness and log-transformed feature values. The results, as Figure 3 shows, are all weak correlations along with most p-values less than 0.005. In addition, a multivariate regression analysis on these features yielded both tiny coefficients and an R^2 value of 0.25, indicating a weak correlation between individual check-worthiness and these features. Based on the results, we cannot conclude that features related to popularity and activity level are indicators of individuals' behavioral tendencies toward check-worthiness.

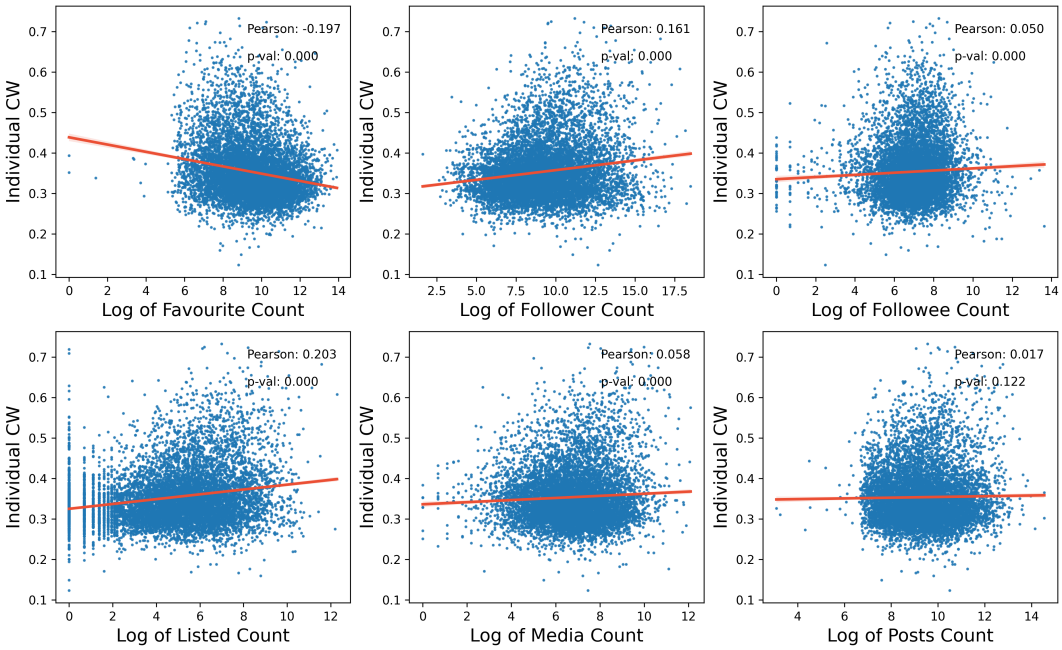


Fig. 3: Correlation between Individual CW and Popularity/Activity Features

Besides those numerical features representing popularity and activity levels, there are other more complex features that could affect/indicate the tendencies. Such features may include occupation, political spectrum, and educational

background. Since these features are either hidden or challenging to identify using automatic methods, we decided to conduct a content analysis to get some insights.

Firstly, among the 11,173 users in the RSU dataset, we selected all users with individual check-worthiness scores less than 0.25 or greater than 0.55 as they encompass two tails of the individual check-worthiness distribution, and thereby represent two small groups with weak and strong behavioral tendencies toward check-worthiness. The group denoted as U_0 comprises 146 users with low individual check-worthiness, while the group denoted as U_1 consists of 169 users with high individual check-worthiness. For both groups, we gathered all the user profile descriptions and tweets from the users, and conducted word frequency analysis. More specifically, for all the user profile descriptions and tweets respectively, we tokenized, removed stopwords, lemmatized, and counted word frequencies. The result is interesting, as Table 2 shows. The top frequent words in tweets from U_0 are general and irrelevant to specific people/events/affairs (e.g., love, life, like, god, good), while the top frequent words in tweets from U_1 are more concrete and highly related to trending topics/events (e.g., russian, ukraine, cannabis). The analysis of the user profile descriptions further enhances this observation. The top frequent words in user profile descriptions from U_0 are more related to literature/art, life/entertainment, and religion, while the top frequent words in user profile descriptions from U_1 are more related to journalism, politics, and technology.

Table 2: Top Words in Tweets/Profiles from U_0 and U_1

U_0's Tweets	U_1's Tweets	U_0's Profiles	U_1's Profiles
people (13112)	new (9591)	author (15)	news (23)
love (12462)	russian (6841)	podcast (7)	reporter (12)
life (12419)	ukraine (5135)	com (7)	newsletter (9)
one (10759)	people (4152)	book (7)	com (9)
like (9956)	energy (3906)	producer (6)	world (8)
god (9138)	report (3667)	life (5)	public (7)
us (8654)	said (3576)	people (5)	tech (7)
time (8249)	russia (3485)	views (5)	government (7)
get (7506)	cannabis (3215)	buddhist (5)	research (6)
good (7368)	us (3135)	writer (5)	policy (6)

The results from the word frequency analyses appear to suggest that individuals' professions, backgrounds, and interests are potentially related to their behavioral tendencies toward check-worthiness. To confirm this conjecture, we randomly selected 100 users from U_0 and U_1 respectively, and then we annotated each user account based on their backgrounds and interests. The results, as shown in Table 3, reveal that a majority of users in U_0 lack explicit backgrounds, though a considerable portion comprises writers and influencers. Their primary focus lies in sharing their ideologies and daily lives. On the other hand, users in U_1 are prominently associated with the media, with over half of the selected 100

users actively doing media-related jobs. Additionally, users with backgrounds in research and politics are also notably present. The dominant interests within U_1 encompass politics, general news, public interests, and technology/science, validating our initial conjecture.

Table 3: Top-ranked Backgrounds/Interests in U_0 and U_1

U_0's BGs	U_1's BGs	U_0's Interests	U_1's Interests
unknown (36)	media (33)	ideology (40)	politics (25)
writer (19)	reporter (13)	daily life (39)	general news (14)
influencer (12)	research (9)	religion (7)	public good (12)
pastor (3)	politician (6)	entertainment (3)	tech&science (12)
speaker (3)	analyst (5)	photography (2	climate (9
singer (2)	journalist (4)	writing (2)	energy (8)
photographer (2)	unknown (4)	general (2)	security (7)
consultant (2)	writer (4)		business (6)
student (2)	advocate (4)		war (3)
teacher (2)	editor (3)		economics (2)

To further solidify this conclusion, we also compared the individual check-worthiness distributions of three specific groups of Twitter users using the aforementioned datasets. The first group contains all users from the HUM dataset, which encompasses individuals related to humanities such as literature, arts and religion. The second group consists of all users from the POL dataset, which represents individuals related to politics and journalism. The third group consists of all users from the RSU dataset which are randomly sampled user accounts. Figure 4 depicts the individual check-worthiness distributions of these three groups of users. The figure shows a left-skewed distribution for the HUM dataset, a right-skewed distribution for the POL dataset, with the RSU in the middle. This finding suggests that users in the HUM dataset generally possess lower individual check-worthiness, whereas those in POL tend to exhibit higher levels of check-worthiness.

6.3 Q3: Impact of Check-Worthiness on Tweeting Behaviors

With the conclusion that people express different behavioral tendencies toward check-worthiness, one may ask how well it aligns with people's tweeting behaviors. More specifically, it would be useful to know whether people tend to share and like posts with similar check-worthiness as their own posts. To answer this question, we conducted experiments on the RSU dataset.

We define O, R, and L as random variables of the check-worthiness of a randomly picked original-tweet, retweet, and liked-tweet from a given user. Moreover, given the dataset, we define X and P as random variables of the check-worthiness of a random tweet and a random popular tweet (liked or retweeted by anyone) from that dataset. Hence, with the RSU dataset, we have 4 hypotheses defined as follows to test the stochastic equality between O and other random variables:

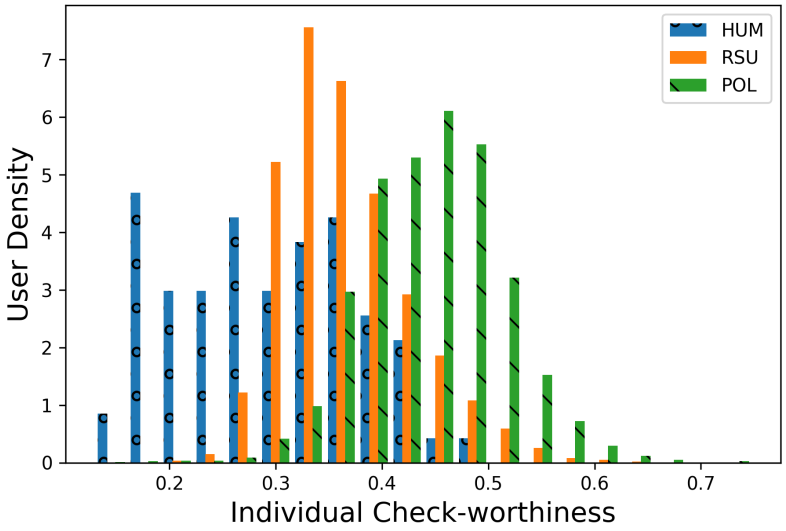


Fig. 4: Individual CW Distributions of HUM, RSU, and POL

– **Hyp1** $\begin{cases} H_0: & P(O > R) = P(O < R) \\ H_a: & P(O > R) \neq P(O < R) \end{cases}$

– **Hyp2** $\begin{cases} H_0: & P(O > L) = P(O < L) \\ H_a: & P(O > L) \neq P(O < L) \end{cases}$

– **Hyp3** $\begin{cases} H_0: & P(O > P) = P(O < P) \\ H_a: & P(O > P) \neq P(O < P) \end{cases}$

– **Hyp4** $\begin{cases} H_0: & P(O > X) = P(O < X) \\ H_a: & P(O > X) \neq P(O < X) \end{cases}$

For each user in the RSU dataset, we performed Brunner Munzel (BM) test and Kolmogorov-Smirnov (KS) test on Hyp1-4, as explained in Section 5. Table 4 shows the results of the acceptance for those hypotheses with alpha (significance level) equal to 0.05. We can see that the acceptance rates of Hyp1-2 are greater than that of Hyp3-4 for all the tests, which means the check-worthiness distribution of a user's original-tweets is more likely to have the same shape as the check-worthiness distributions of same user's retweets and liked-tweets, in comparison to random and popular tweets from arbitrary users.

Table 4: Acceptances of Hyp1-4

Test	Hyp1	Hyp2	Hyp3	Hyp4
BM Test	1797/16.1%	2896/25.9%	939/8.4%	1013/9.1%
KS Test	1126/10.1%	1831/16.4%	248/2.2%	284/2.5%

In addition, we also performed a correlation analysis on the median check-worthiness of original-tweets, retweets, and liked-tweets for all the users in the RSU dataset. As Figure 5 shows, where each point represents a user account, there exist strong correlations between the median check-worthiness scores of the users' original-tweets, retweets, and liked-tweets. To reduce the effects of content redundancy (such as when a user retweets or likes their own original-tweet, or when a retweet is also liked by the retweeter), we computed the overlap ratios among these three types of tweets for each user. The result shows that, on average, merely 1.3% of original-tweets are additionally retweeted by their authors, and less than 1% of original-tweets are also liked by the authors. Additionally, about 16% of retweets are also liked by the retweeters. Following the removal of overlapping tweets, we conducted the correlation analysis again, and the results as shown in Figure 5 remain largely unchanged. Therefore, we are able to conclude that people overall have the behavioral tendency to post, share, and favor tweets with similar check-worthiness levels.

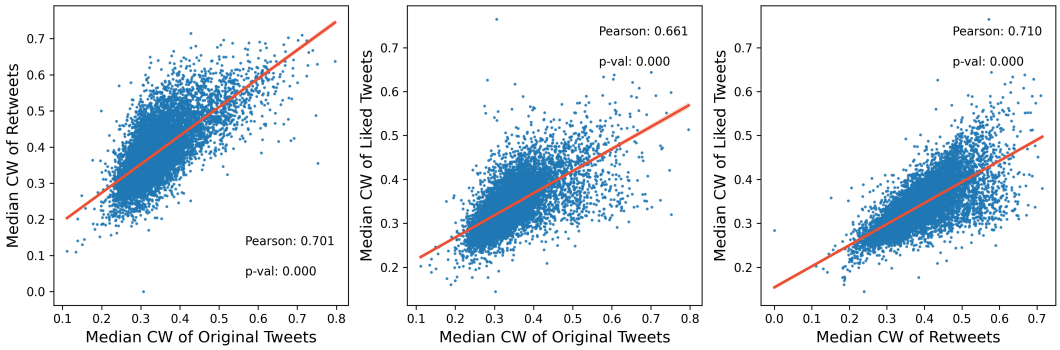


Fig. 5: Correlation in Median Check-Worthiness Among Three Types of Tweets

6.4 Q4: Impact of Check-Worthiness on Following Behaviors

Besides tweeting activities, another important activity on social media is following, which influences a large portion of the information a user receives. Therefore, it is natural and crucial to find out whether people tend to follow others with similar tendencies toward check-worthiness. More specifically, we want to examine whether the check-worthiness distribution of a user's tweets is more similar to that of its followers than other users.

We define U, V, F, and X as random variables of the check-worthiness of a randomly picked tweet from a given user, one of its followers, one of its friends (being both follower and followee), and a random user respectively. Here we have the hypotheses defined as follows to test the stochastic equality between U and other random variables:

- **Hyp5** $\begin{cases} H_0: & P(U > V) = P(U < V) \\ H_a: & P(U > V) \neq P(U < V) \end{cases}$
- **Hyp6** $\begin{cases} H_0: & P(U > F) = P(U < F) \\ H_a: & P(U > F) \neq P(U < F) \end{cases}$
- **Hyp7** $\begin{cases} H_0: & P(U > X) = P(U < X) \\ H_a: & P(U > X) \neq P(U < X) \end{cases}$

In the RSU dataset, we have $10,402$ (follower, followee) pairs and 351 friend pairs, with a total of $9,124$ distinct accounts involved. Table 5 shows the results of the acceptance for hypotheses Hyp5-7 with alpha (significance level) equal to 0.05. We can see that the acceptance rates of Hyp5 are greater than Hyp7, meaning the check-worthiness distribution of a user's tweets is more likely to have the same shape as the check-worthiness distribution of its followers' tweets compared with a random user's tweets. However, this likelihood is not strong since the acceptance rates of Hyp5 do not exceed Hyp7 by a lot. A more substantial result comes from the acceptance rates of Hyp6, which are much higher. This indicates a higher likelihood of check-worthiness similarity between tweets from a pair of users in a two-way following relationship than in a one-way following relationship.

Table 5: Acceptances of Hyp5-7

Test	Hyp5	Hyp6	Hyp7
BM Test	1043/10%	59/16.9%	696/7.6%
KS Test	335/3.2%	50/14.3%	130/1.4%

To further verify this conclusion, we again performed a correlation analysis on the individual check-worthiness of users and their friends. As Figure 6 shows, there exists a weak correlation between the individual check-worthiness of followers and followees. However, the correlation becomes stronger when we compare the individual check-worthiness of users with a two-way following relationship. Similar to what we discussed in Section 6.3, our calculation shows that the average ratio of tweet overlap between each pair is less than 1%. This means the possibility of the result being influenced by overlapping tweets among friends is low. Therefore, the result is solid and aligns with our conjecture.

7 Limitation

While we have examined certain social media behaviors and identified several patterns, a few questions about the observations remain unanswered. For example, since most users may not consciously choose high or low check-worthy content when posting, sharing and liking tweets, the observations cannot be interpreted as definitive indicators of users' behavioral patterns. However, our

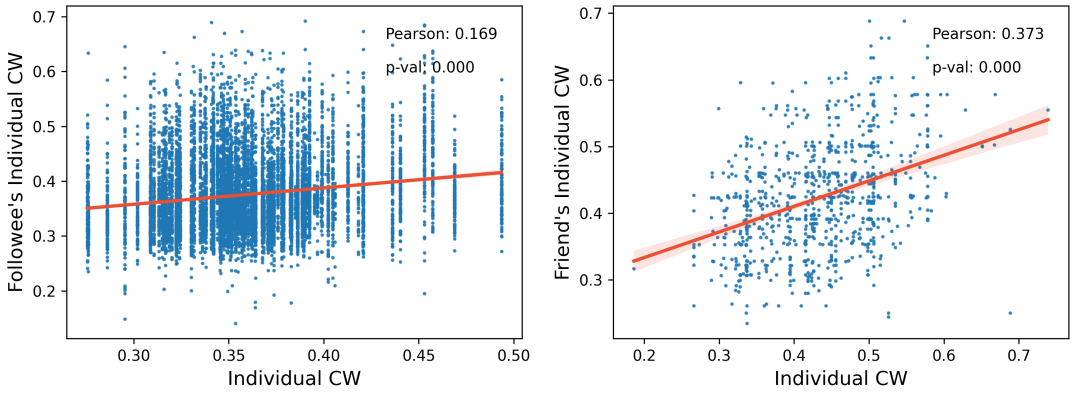


Fig. 6: Correlation between Following Parties' Individual Check-worthiness

statistics reveal that a significant portion of users demonstrate consistent behavioral patterns associated with check-worthiness, suggesting a possible subconscious adherence to such patterns. Undoubtedly, a more nuanced investigation is warranted, particularly concerning where the observed patterns come from. The inherent value of our study lies not only in the answers it provides but also in the thought-provoking questions it raises. This study possesses the potential to not only address current gaps in knowledge but also to act as a catalyst, possibly inaugurating a new line of research endeavors.

8 Conclusions

This work identified the existence of the difference between individuals' behavioral tendencies toward factual claims. In particular, the population from domains such as politics, general news, and technology is more likely to engage with check-worthy content compared to those associated with arts, literature, and religions. Through a set of experiments, the research has established a strong correlation between these tendencies and users' posting, sharing, and liking behaviors, indicating a conspicuous pattern to engage with content of similar check-worthiness. Furthermore, the findings emphasize the heightened efficacy of two-way following relationships in reflecting shared preferences towards factual claims.

The concept of check-worthiness emerges as a potent tool for understanding human behaviors within the realm of social media. Our results not only provide valuable insights into the impact and adaptability of check-worthiness but also lay the groundwork for future investigations to delve deeper into the various dimensions of its influence on social media behaviors and its potential applications across diverse domains.

Acknowledgement

This work is partially supported by the National Science Foundation award #2346261. The authors acknowledge the Texas Advanced Computing Center (TACC) at The University of Texas at Austin for providing HPC resources that have contributed to the research results reported within this paper.

References

1. Benjamini, Y., Hochberg, Y.: Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal statistical society: series B (Methodological) **57**(1), 289–300 (1995)
2. Bovet, A., Makse, H.A.: Influence of fake news in twitter during the 2016 us presidential election. Nature communications **10**(1), 7 (2019)
3. Brummette, J., DiStaso, M., Vafeiadis, M., Messner, M.: Read all about it: The politicization of "fake news" on twitter. Journalism & Mass Communication Quarterly **95**(2), 497–517 (2018)
4. Brunner, E., Munzel, U.: The nonparametric behrens-fisher problem: asymptotic theory and a small-sample approximation. Biometrical Journal: Journal of Mathematical Methods in Biosciences **42**(1), 17–25 (2000)
5. Chen, J., Liu, Y., Zou, M.: User emotion for modeling retweeting behaviors. Neural Networks **96**, 11–21 (2017)
6. Clayton, K., Blair, S., Busam, J.A., Forstner, S., Glance, J., Green, G., Kawata, A., Kovvuri, A., Martin, J., Morgan, E., et al.: Real solutions for fake news? measuring the effectiveness of general warnings and fact-check tags in reducing belief in false stories on social media. Political behavior **42**, 1073–1095 (2020)
7. Cohen, I., Huang, Y., Chen, J., Benesty, J., Benesty, J., Chen, J., Huang, Y., Cohen, I.: Pearson correlation coefficient. Noise reduction in speech processing pp. 1–4 (2009)
8. Comarela, G., Crovella, M., Almeida, V., Benevenuto, F.: Understanding factors that affect response rates in twitter. In: Proceedings of the 23rd ACM conference on Hypertext and social media. pp. 123–132 (2012)
9. Firdaus, S.N., Ding, C., Sadeghian, A.: Topic specific emotion detection for retweet prediction. International Journal of Machine Learning and Cybernetics **10**, 2071–2083 (2019)
10. Funke, D.: This washington post fact check was chosen by a bot (2018)
11. Hansen, C., Hansen, C., Alstrup, S., Grue Simonsen, J., Lioma, C.: Neural check-worthiness ranking with weak supervision: Finding sentences for fact-checking. In: Companion proceedings of the 2019 world wide web conference. pp. 994–1000 (2019)
12. Hassan, N., Li, C., Tremayne, M.: Detecting check-worthy factual claims in presidential debates. In: Proceedings of the 24th acm international on conference on information and knowledge management. pp. 1835–1838 (2015)
13. Hassan, N., Tremayne, M., Arslan, F., Li, C.: Comparing automated factual claim detection against judgments of journalism organizations. In: Computation+ journalism symposium. pp. 1–5 (2016)
14. Hassan, N., Zhang, G., Arslan, F., Caraballo, J., Jimenez, D., Gawsane, S., Hasan, S., Joseph, M., Kulkarni, A., Nayak, A.K., et al.: Claimbuster: The first-ever end-to-end fact-checking system. Proceedings of the VLDB Endowment **10**(12), 1945–1948 (2017)

15. Hopcroft, J., Lou, T., Tang, J.: Who will follow you back? reciprocal relationship prediction. In: Proceedings of the 20th ACM international conference on Information and knowledge management. pp. 1137–1146 (2011)
16. Hu, X., Tang, L., Tang, J., Liu, H.: Exploiting social relations for sentiment analysis in microblogging. In: Proceedings of the sixth ACM international conference on Web search and data mining. pp. 537–546 (2013)
17. Kim, H.S., Suh, Y.J., Kim, E.m., Chong, E., Hong, H., Song, B., Ko, Y., Choi, J.S.: Fact-checking and audience engagement: A study of content analysis and audience behavioral data of fact-checking coverage from news media. Digital journalism **10**(5), 781–800 (2022)
18. Lespagnol, C., Mothe, J., Ullah, M.Z.: Information nutritional label and word embedding to estimate information check-worthiness. In: Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 941–944 (2019)
19. Majithia, S., Arslan, F., Lubal, S., Jimenez, D., Arora, P., Caraballo, J., Li, C.: Claimportal: Integrated monitoring, searching, checking, and analytics of factual claims on twitter. In: Proceedings of the 57th annual meeting of the association for computational linguistics: system demonstrations. pp. 153–158 (2019)
20. Massey Jr, F.J.: The kolmogorov-smirnov test for goodness of fit. Journal of the American statistical Association **46**(253), 68–78 (1951)
21. Park, S., Park, J.Y., Chin, H., Kang, J.h., Cha, M.: An experimental study to understand user experience and perception bias occurred by fact-checking messages. In: Proceedings of the Web Conference 2021. pp. 2769–2780 (2021)
22. Rony, M.M.U., Hoque, E., Hassan, N.: Claimviz: Visual analytics for identifying and verifying factual claims. In: 2020 IEEE Visualization Conference (VIS). pp. 246–250. IEEE (2020)
23. Samadi, M., Talukdar, P., Veloso, M., Blum, M.: Claimeval: Integrated and flexible framework for claim evaluation using credibility of sources. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 30 (2016)
24. Shamma, D.A., Kennedy, L., Churchill, E.F.: Tweet the debates: understanding community annotation of uncollected sources. In: Proceedings of the first SIGMM workshop on Social media. pp. 3–10 (2009)
25. Shapiro, S.S., Wilk, M.B.: An analysis of variance test for normality (complete samples). Biometrika **52**(3/4), 591–611 (1965)
26. Shu, K., Cui, L., Wang, S., Lee, D., Liu, H.: defend: Explainable fake news detection. In: Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining. pp. 395–405 (2019)
27. Vasileva, S., Atanasova, P., Màrquez, L., Barrón-Cedeño, A., Nakov, P.: It takes nine to smell a rat: Neural multi-task learning for check-worthiness prediction. arXiv preprint arXiv:1908.07912 (2019)
28. Wright, D., Augenstein, I.: Claim check-worthiness detection as positive unlabelled learning. arXiv preprint arXiv:2003.02736 (2020)