Effects of Political Bias and Reliability on Temporal User Engagement with News Articles Shared on Facebook

Alireza Mohammadinodooshan and Niklas Carlsson

Linköping University, Sweden
{alireza.mohammadinodooshan, niklas.carlsson}@liu.se

Abstract. The reliability and political bias differ substantially between news articles published on the Internet. Recent research has examined how these two variables impact user engagement on Facebook, reflected by measures like the volume of shares, likes, and other interactions. However, most of this research is based on the ratings of publishers (not news articles), considers only bias or reliability (not combined), focuses on a limited set of user interactions, and ignores the users' engagement dynamics over time. To address these shortcomings, this paper presents a temporal study of user interactions with a large set of labeled news articles capturing the temporal user engagement dynamics, bias, and reliability ratings of each news article. For the analysis, we use the public Facebook posts sharing these articles and all user interactions observed over time for those posts. Using a broad range of bias/reliability categories, we then study how the bias and reliability of news articles impact users' engagement and how it changes as posts become older. Our findings show that the temporal interaction level is best captured when bias, reliability, time, and interaction type are evaluated jointly. We highlight many statistically significant disparities in the temporal engagement patterns (as seen across several interaction types) for different bias-reliability categories. The shared insights into engagement dynamics can benefit both publishers (to augment their temporal interaction prediction models) and moderators (to adjust efforts to post category and lifecycle stage).

Keywords: User interactions · Bias · Reliability · Temporal dynamics
.

1 Introduction

Despite 74% of all Americans believing that the propagation of online misinformation is a big problem [9], a very large fraction of users today obtains their news via social media [16]. In this environment, news articles are often propagated based on other users' interactions with the news (e.g., through likes, comments, and sharing of posts linked to various news articles). Indeed, users' interactions (and their engagement) with different news are becoming the big

driver for which news are most likely to be viewed by others, and hence also which news are given the best chance to impact other users' views of the world, including their opinions and thoughts on various current issues.

With increasing (political) polarization [11] and news articles often having vastly different reliability levels, it is therefore important to measure and understand whether there are fundamental disparities in the users' interaction dynamics with news articles that have different levels of reliability and political bias. In this paper, we provide a rigorous temporal analysis in which we identify cases of statistically significant disparity in the user interaction dynamics with different classes of news articles. Our findings provide insights into how and when to better protect against and/or slow down the spread of misinformation.

Combined impacts, granularity levels, and per-article-based news classification: While reliability represents the degree of factual reporting, bias refers to the tendency for journalists to favor one political side or another in their reporting, sometimes even without being aware that they are doing so. Prior research has established a link between the bias and reliability of news articles and how people engage with and distribute them. For example, by focusing on the reliability factor, Vosoughi et al. [20] showed that false information spreads substantially farther, faster, deeper, and more widely than the truth. Examining the bias parameter, Wischnewski et al. [22] discovered that users are more inclined to share hyperpartisan news pieces that coincide with their own political views. Limited works like [3] have considered both these parameters but studied them independently. There are even fewer studies that consider both parameters in combination. The primary exception is the work by Edelson [4], which findings indicate, among other things, that while misinformation generates less engagement than non-misinformation, it can nonetheless account for a significant percentage of the overall engagement (e.g., 37.7% on the far left).

Regardless of the bias or reliability parameter, there are also big differences in the granularity that each parameter has been classified and whether all news articles of a news outlet have been classified the same or individually. Both these aspects impact the applicability of the results. First, while a few works used several levels for the studied factors (e.g., [13]), most previous studies analyzed data at the binary level, including the only other work that considers both bias and reliability in combination. In their work, they label news as either reliable or not reliable [4]. Second, while most prior works (including the work by Edelson [4]) give the same bias/reliability score for all news articles published by a publisher, only a few papers have used the ratings of individual news articles. We argue that this is of significant importance for the generalization of the result. Otherwise, for example, political, sports, or science news published by Fox News would all receive the same reliability and political bias classification. In practice, two news articles from the same publisher or even by the same author can have significantly different ratings.

In summary, the majority of prior research is based on the ratings of publishers (not news articles), considers only bias or reliability (not both combined),

use a limited news article classification (e.g., binary), focuses on a limited set of user interactions, and ignores the users' engagement dynamics over time.

Main contribution: This paper addresses the above shortcomings of the current literature by presenting the first temporal analysis of the user interaction dynamics with news articles of varying degrees of (political) bias and reliability. We consider a spectrum of user interactions and study the impact of bias and reliability in combination. In contrast to prior works studying the interaction dynamics as part of the political conversations (in online social networks) during elections and other events [7, 18, 8], our focus here is instead on the roles that the bias and reliability play in the dynamics. Another novel aspect of our temporal analysis is that we compare the temporal dynamics seen using different classes of interactions with the news, including likes and shares of posts linking the news articles. Only a few works have considered all types of user interactions (e.g., Edelson et al. [4]) but none of them consider the relative dynamics or the impact of bias and reliability on the dynamics. Finally, we examine the predictability of the total amount of user engagement that news articles of different classes may receive based on the interactions it has received thus far.

Temporal dynamics and research questions: To study the temporal dynamics at the granularity and scale needed to address the above limitations of prior works, we obtain and study temporal traces of all types of user interactions for around 18K news articles that have been individually scored based on their bias and reliability. For the news article labeling, we use data from Ad Fontes Media, and we use CrowdTangle to obtain temporal data for all classes of user interactions with all Facebook posts discussing or linking the labeled news articles. Using several carefully designed prepossessing steps, we then study the observed temporal dynamics and address the following research questions:

RQ1 How do the bias and reliability of a post affect the temporal dynamics of a user's engagement with it?

RQ2 Using its intermediate interactions as a predictive criterion, how does the bias and reliability of a post affect the prediction of the total engagement it will receive?

RQ3 How do the temporal dynamics of user engagement differ across different interaction types, and how does this variation relate to the bias and the reliability of the post?

Empirical example findings: Our analysis uncovers several interesting observations. In comparison to left-leaning posts, right-leaning posts receive more interactions per post. Considering the reliability, the "Most unreliable" and "Most reliable" news receives the maximum number of posting per article. Considering the temporal dynamics of interactions, our findings show that the temporal interaction level is best captured when bias and reliability are evaluated jointly. We highlight different joint bias-reliability classes that deviate from the temporal dynamics of the bias or reliability classes they belong to. In terms of interaction changes over time, the "most reliable" posts and the "most unreliable" posts exhibit opposite trends. Here, the "most reliable" news is experiencing a faster

decrease (than average) in the interaction rates, whereas the "most unreliable" news experience a faster than average increase in the interaction rates, as seen over time.

We find that when examining just the number of likes that a post receives within the first hour of publication, the reliability of the post is positively associated with the normalized (over the total number of interactions) number of likes received. In other words, during this period of time, the posts that are considered "most reliable" receive the highest number of likes. Finally, when considering the outlet-specific analysis, we find that despite Fox News and the New York Times having different political biases, in both cases, relatively unbiased posts receive greater interaction rates during the initial stages compared to their biased posts.

Example beneficiaries: Various stakeholders can benefit from our contributions. Researchers will benefit from our quantitative analysis of the temporal dynamics of user engagement with various types of news, including our use of statistical tests to back up example findings captured in the different stages of our time-series analysis. We share the code used to produce the results,[1] allowing others to reproduce and expand on our findings. Among practitioners, Facebook content moderators may use knowledge about the statistical disparities highlighted here between the user interactions with reliable and unreliable news to better focus their resources during the different stages of a post's lifetime. Furthermore, news content providers may incorporate the mentioned temporal patterns into their engagement prediction models.

Roadmap: The remainder of the paper is organized as follows. Section 2 describes how we collect and analyze the data. Here, we also provide detailed definitions of the normalized metrics computed and used. Section 3 presents our results for the complete dataset, as well as the outlet specific results. In Section 4, we explore the extent to which we can predict the maximum interaction volume based on the intermediate number of interactions. Sections 5 and 6 discuss related works and limitations, respectively. Ethical considerations are discussed in Section 7, before we conclude the paper in Section 8.

2 Methodology and Dataset

We first describe our methodology and dataset. Section 2.1 describes the news article selection and the labeling of articles. Section 2.2 describes the filtering we applied to have a clean dataset. Section 2.3 describes how we collected the Facebook posts sharing each news article, as well as temporal data of the user interactions associated with each such post. In Section 2.4, we determine time thresholds (based on the number of total interactions between consecutive time thresholds) that together define a sequence of time buckets with equalized (total) number of interactions per time bucket. For our (later) temporal analysis, we use these bucketized time sequences of the interactions associated with different subsets of news articles (where each subset contains the articles with a specific

[1] https://github.com/alireza-mon/pam2023

Table 1: Bias classes and their intervals.

Bias class	Far left	Skews left	Balanced Bias	Skews right	Far right
Bias range	[-42, -18]	(-18, -6]	(-6, 6)	[6, 18)	[18, 42]

bias/reliability label). In Section 2.5, we explain the normalization process we use to provide a fair head-to-head comparison between different subsets. Section 2.6 provides a summary of the final dataset.

2.1 News Article Selection and Bias/reliability Labeling

There exist several independent evaluation efforts to asses the bias and/or reliability of individual news articles and/or news sources. Examples include Media Bias Fact Check [2], Ad Fontes Media [3], AllSides [4], and NewsGuard [5]. Of these, we selected to use data from Ad Fontes Media for the following primary reasons: (1) they evaluate individual news articles, (2) each evaluated article is scored with regard to both bias and reliability, (3) the dataset contains over 30K articles covering over more than 1,500 sources, and finally (4) they provide a transparent strategy, published and explained in a white paper [14].

For each news source, Ad Fontes Media selects sample articles that are prominently featured on each source's website over multiple news cycles. To prioritize popular news sources, they rank the news sources and organize them into tiers that are given different sample frequencies. Specifically, they label approximately 15 articles per month for the top-15 sources, seven articles per month are labeled for the next 15 sources, the rest of the top-200 sources are assigned approximately five new labeled articles per quarter, and the following 200 articles (ranks 201-400) are updated approximately five times per six months. As mentioned in their white paper [14], they attempt to strike a balance between rating new sources and updating current ones with more recent samples. As a result, the dataset consists of news articles spanning both a broad range of news sources and capturing many samples from popular new sources seen over time.

Each article in the Ad Fontes Media dataset is evaluated with regard to both bias and reliability by at least three human analysts with a balance of right, left, and center self-reported political perspectives. The bias scores reported by Ad Fontes Media range from -42 to +42, with greater negative values indicating a more leftward bias and positive values leaning toward the right party. For the reliability scores, they use grades from 0 to 64, with 64 being the most reliable news. Note that 42 and 64 (not usual numbers used for scales) are arbitrarily selected by Ad Fontes Media, as described in [15]. For the analysis presented here, we binned the bias scores into belonging to one of five bias ranges and we binned the reliability scores into four reliability ranges. Ranges and assigned

[2] https://mediabiasfactcheck.com
[3] https://adfontesmedia.com
[4] https://www.allsides.com
[5] https://www.newsguardtech.com

Table 2: Reliability classes and their intervals.

Reliability class	Most Unreliable	Unreliable	Reliable	Most Reliable
Reliability range	[0, 16)	[16, 32)	[32, 48)	[48, 64]

labels are provided in Tables 1 and 2, respectively. Due to the smaller sample size of extremely biased news articles (both to the left and right) we used larger bin sizes for articles labeled as “Far left" [-42, -18] or “Far right" [18, 42].

2.2 Preprocessing of News Articles

We first and on August 2, 2022 received resources evaluated by Ad Fontes Media and their corresponding bias and reliability values. Second, we used Ad Fontes Media’s search functionality to prune the dataset to include only news articles. After removing television shows and podcasts, the dataset contained 27,547 articles. Third, through manual examination, we identified and removed several videos and television shows from the remaining results (e.g., some shows from https://www.rushlimbaugh.com). Fourth, to reduce the effects of potential long-term trends/biases, we restricted the final dataset to articles published after 2018. To determine the publication date of each news article, we calculated the minimum of the following four values:

- The news article’s earliest archived date on web.archive.org.
- The publication date of the article is extracted from the article page using the htmldate python package, which applies heuristics on HTML code and linguistic patterns to derive a page’s publishing date.
- The minimum post date of all Facebook posts sharing the URL.
- The minimum post date of all tweets sharing the URL.

Finally, we excluded news pages that did not refer to news articles but rather pages reporting on an event over a period of time[6]. Following these steps, the dataset contained 27,329 articles from 986 domains.

Before trying to identify social media posts pointing to a news article, it is important to note that not all links to an article will look the same. To ensure that we find as many posts referencing the identified articles as possible while avoiding false positives, we next calculated the canonical form of the URL of each news article. By canonical form, we mean the minimum form of the URL that uniquely identifies any shared version of the URL. As an example, we identified several campaign query parameters used to augment numerous URLs that we could remove. Appendix A.1 explains our procedure to compute the canonical form of the URLs.

[6] e.g., reuters.com/subjects/myanmar-reporters

2.3 Temporal User Engagement of Related Facebook Posts

We next used the CrowdTangle API to collect (1) all their Facebook posts including one of the news article URLs, as well as (2) the temporal data of users' interaction with these posts. The CrowdTangle platform [6], which is owned by Facebook, indexes the posts and engagement data for around 7 million pages, including "more than 50K likes pages, all public Facebook groups with 95k+ members, all US-based public groups with 2k+ members, and all verified profiles" [6], as well as any pages added to a CrowdTangle list by those with access to it. For collecting the Facebook posts, we opted to use the "/Links" endpoint of the CrowdTangle API. This ensured us that all shortened versions of the URLs were also collected. To collect the maximum number of posts related to each article, we passed the canonical form of the URL to this end point. In addition, we strived to account for instances in which query strings were included in the URLs' canonical form.

The data collection was done on or after Sept. 1 (2022) for all posts published before Sept. 1. By including only articles published before Aug. 2, our methodology ensures that at least 4 weeks had passed since the publication date of any articles included in our dataset. Since most posts sharing news articles occur soon after an article is posted, the 4-week gap (between the collection of articles and posts) allows us to collect (the 21 days) temporal interaction data for all posts associated with the studied news articles. Similarly, the 4-week threshold also ensures that we can catch most of the posts linking an article. In this study, we removed any articles that did not have any published posts. After this filtering, the dataset included 21,872 labeled articles for which we extract the temporal interaction data.

Using CrowdTangle, we compile temporal user interaction data for the number of likes, shares, comments, and emoji-based interactions such as Like(s), Wow(s), Sad(s), Angry(s), Love(s), and Haha(s). For each of the above metrics, as well as for the total interactions (across all actions allowed by users), CrowdTangle breaks the first (approximately) 21 days after the post is published into 74 roughly exponentially increasing time steps and provides the number of user interactions for each of the user interactions at each of these time steps. The increasingly sparse sample rate used by CrowdTangle is most likely motivated by most posts being short-lived and the interaction rates quickly reducing over time. We illustrate this in Fig. 1, where we show the cumulative fraction of all interactions that have taken place after some time since the posting time of each studied post in our dataset (with time on log scale).

For most posts, we have temporal data for the full 21-day period (the maximum age at the final data point for any posts observed in our dataset was 23 days). In addition to this temporal data, we also extract other post-related data from CrowdTangle, including the date that the post was published.

Here, it should be noted that a user sharing a post essentially pushes the post to the timeline of their friends and followers, and their statistics do not include the shares of a post (on the original shares of a post). For comments

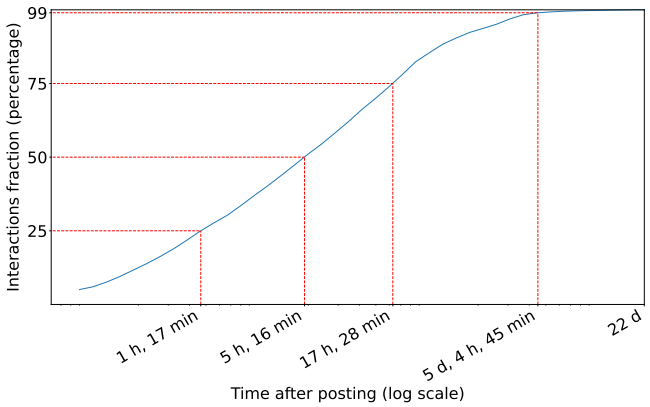


Fig. 1: Cumulative fraction of interactions (min: minutes, h: hours, and d: days).

statistics, the API counts all comments on the post and all first-level replies to those comments.

2.4 Time Partitioning

For studying the temporal dynamics of the posts, we break up the 21-day time period into smaller time buckets and then study the dynamics of user interactions over each of these time-bucket sequences. For the analysis presented here, we used four time buckets and selected the time thresholds used to define the bucket sizes so that each bucket had roughly the same total number of interactions. More specifically, we selected the time thresholds so that they represent the points where 25%, 50%, and 75% of the overall interactions (sum of over all interaction types) have been observed by CrowdTangle (and apply linear interpolation when thresholds fall between sample points). The determined threshold values are shown and highlighted (using red lines) in Fig. 1. As expected, the decreasing interaction rates, result in increasing time bucket sizes.

While we observe approximately straight-line behavior for part of the parameter range, we note that the above selection process does not require any assumptions about the actual probability distribution. This selection also helps provide fair head-to-head comparisons (using statistical tests) between the interaction differences observed during the four different stages, effectively maximizing the information gains from comparing the interaction dynamics of the users across the four phases.

2.5 Capturing Engagement Dynamics

By picking time buckets of equivalent size in terms of interaction volume, we can better compare the number of user interactions of each type in each time bucket. For our primary comparisons, we first define a metric called the Total

Table 3: Dataset summary statistics.

Articles#	Domains#	Posts#	Total interactions#	Bias mean	Reliability mean
17,966	953	106,325	81,891,888	-1.08 (std:10.11)	40.47 (std:8.59)

Interactions Covered Ratio (TICR), defined as the percentage of total interactions that a post receives which are covered within a specific period of time. For example, if a post receives a maximum of 600 interactions over the full timeline and 200 interactions between hours 1 and 5 (following its publication), then the TICR is 33% for this period.

After computing the TICR values for all the times that the current post has been probed on, we calculate the average observed over the successful probes done within each time bucket. This procedure is repeated for all time buckets and posts.

At this stage, we removed any post that was not probed at least once during each of the four buckets or that received fewer than ten interactions in total (including the ones with zero interactions). This helps remove noise from non-popular posts and improves the stability of the results.

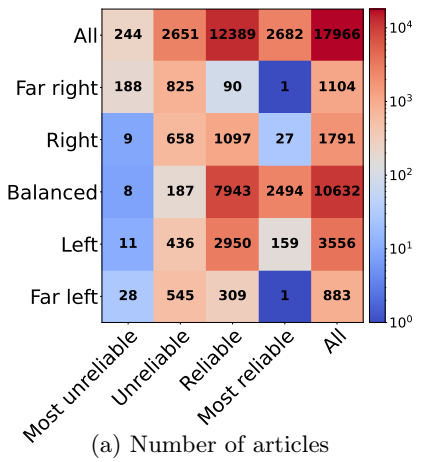
Finally, after the above per-post filtering, we removed any article without any remaining posts. Table 3 provides summary statistics for the final dataset.

We next use the bias-reliability labels of the articles associated with each post to compute statistics for each bias-reliability pair and time bucket. For most of our analysis presented here, we report the mean values observed for each time bucket and interaction type, as well as perform statistical tests on the relative mean values.

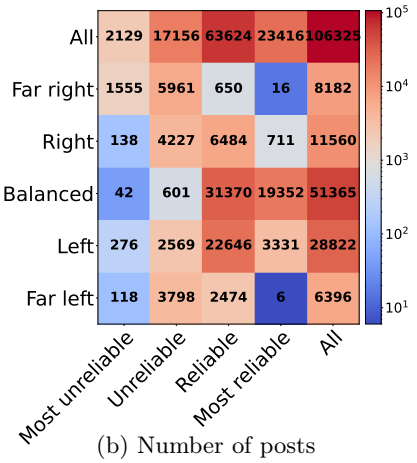
2.6 Dataset Summary

Given the above steps, for each bias-reliability class and for each interaction type, we have the bucket-based temporal sequences of the user interactions to the posts associated with news articles of that class. Figs. 2a and 2b summarize the number of articles we have in each bias and reliability class and the number of posts for which we have completed such sequences, respectively, as broken down per bias-reliability category. In addition to five categories for bias and four categories for reliability, we include one column and one row for the aggregate statistics combining all categories of reliability and bias, respectively. With this design, the overall observed articles (17,966) and posts (106,325) are shown in the top-right corners of the first two sub-plots, respectively.

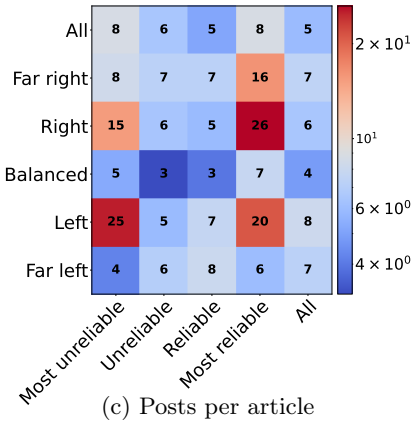
While there are some categories (primarily the “most unreliable less biased" articles and the “most reliable" but extremely biased articles) for which we only have a small number of articles with complete stats, we often have enough posts for our analysis also for these categories. In fact, most categories have a significant number of posts per article on average (Fig. 2c). In terms of the normalized number of posts that shared them, the eleven articles in the most unreliable-left



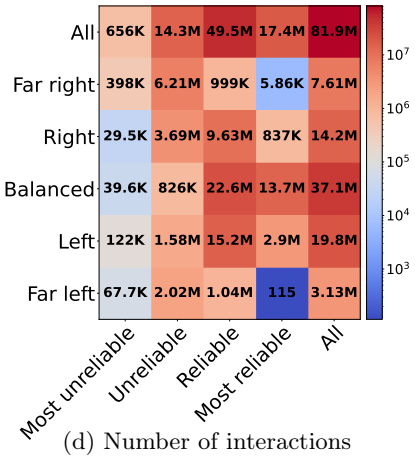
(a) Number of articles



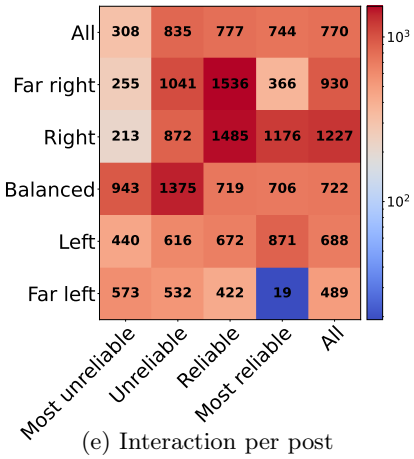
(b) Number of posts



(c) Posts per article



(d) Number of interactions



(e) Interaction per post

Fig. 2: Dataset summary statistics.

and the most reliable-right classes were the most successful, as shown in this figure. Moreover, we observe that, among all classes of reliability, the two extreme classes are shared the most.

We also provide summary statistics for the total number of interactions (irrespective of interaction type), calculated as the sum of all interactions.[7] As shown in Fig. 2d, 81.9 million interactions have been recorded for the posts included in our final dataset. As expected, the interactions are correlated with the number of posts. To determine objectively which class performs better in terms of interactions per post, we present the normalized number of interactions (over the number of posts) in Fig. 2e. As is noticeable, the right party (both “far right" and “right") receives more interactions. Regarding reliability, however, it is shown that the “most unreliable" news are the least engaging for users. We next present the results and analysis of the temporal sequences.

Key observations: In comparison to left-leaning posts, right-leaning posts receive more interaction per post. In terms of reliability, the “Most unreliable" news receives the minimum interaction per posting.

3 Results

3.1 High-level Analysis of the All Interactions Dynamics

Let us first consider the cases when all interactions are aggregated into one interaction metric, calculated as the sum over all interaction types. Fig. 3 shows the temporal interaction dynamics of this metric in terms of TICR. Here, we again show the five categories of the bias and an “All" category (that combines all observations regardless of bias) as rows in each sub-plot and show the four categories of the reliability plus an aggregate “All" category (that combines all observations regardless of reliability) as columns. The four sub-plots, going from left to right, show the results for the time buckets containing all sample points (as described in Sections 2.3 and 2.5) associated with the following time buckets: (1) 0 to 1 hour and 17 minutes, (2) 1 hour and 17 minutes to 5 hours and 16 minutes, (3) 5 hours and 16 minutes to 17 hours and 28 minutes, and (4) 17 hours and 28 minutes until the end of the timeline of each post we study (typically 21 days). We use a timeline with green markers to illustrate this bucketization. As expected from the definition of TICR (Section 2.5) and our selection of time bucket thresholds (Section 2.4), the TICR value for the “all news" case (i.e., the right-top-most cell) of each bucket is 25%.

In each bucket, the mean TICR value for all posts belonging to the respective class and bucket is depicted (using heatmap colors). To capture the variances of each class and thereby quantify the reliability of the mean reported for each group, the coefficient of variation of the mean (cv_{mean})(i.e., standard error of

[7] For example, if a post receives 6 likes, 2 comments, 3 shares, and no other interactions, the value of the total interactions for this post is 11.

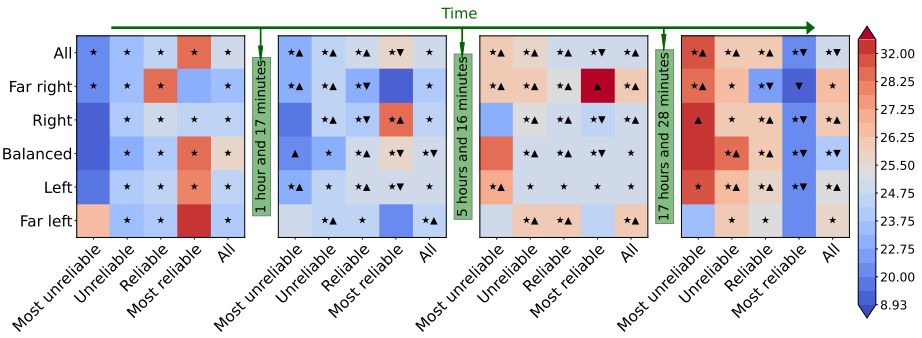


Fig. 3: Temporal dynamics of the total interactions (⋆: coefficient of variation of the mean is smaller than 4%, ▲ and ▼: has deviation from the previous time bucket with p-value <0.05).

the class divided by its mean) of that class in percentage is computed. Then, we mark the class with an asterisk (⋆) if cv_{mean} is smaller than a threshold. In the following, we decided on the value of 4% as the threshold. Accordingly, the classes for which the cv_{mean} is higher than 4% (due to not having enough samples or having high variances) do not receive the asterisk.

Regarding selecting 4% as the cv_{mean} threshold used for the above statistical tests, we first note that this value is small. For example, for the general population, which has a mean of 25, this threshold is equal to a standard error of 1.0. The use of such a small threshold allows the comparison of all classes to be made in a more reliable way. Furthermore, we have found that with this selection, any two classes with "asterisks" within the same time bucket whose TICR values are at least 0.2 units apart (from each other) always have statistically significantly different means at the 90% confidence level. This finding has been validated for all category pairs and time buckets using t-tests for comparing the means of these classes, and the p-values are always less than 0.1[8].

To come to the above thresholds, we performed pairwise comparisons between all the classes for each bucket using different example thresholds. For each case, this corresponds to calculating a 30×30 table of pairwise tests, in which each cell include the p-value (capturing the statistical significance of the pairwise mean comparisons). Clearly, showing this table – even for a single bucket (and example threshold)– takes a lot of space. For this reason, we instead simply report the determined thresholds (in our case 4% and 0.2 point difference) and mark the classes that satisfied the 4% criteria with an "asterisks". As an example, if we turn our attention to the first bucket, we note that the "most reliable" group's

[8] Here, we use independent samples t-test when the classes are independent and dependent samples t-test when the classes are not independent. Examples when the dependent test is used, include cases when a class is compared to its parent bias or parent reliability class (that it belongs to); e.g., comparison between the "right-unreliable" class and the "right" (over all biases).

TICR mean is higher than that of the "unreliable" class (with more than a 0.2 difference) and that both classes are marked with an "asterisks". Therefore, we can say that these two classes have statistically different means.

More than comparing the interaction levels of classes within the same time bucket, it is also interesting to capture the changes in the interaction level of one class between the time intervals. To cover this aspect, we annotated the cells of the figure with an arrow for any class in a bucket for which the difference between its mean in this bucket and its mean in the previous bucket is statistically significant at the 95% confidence level (i.e., the p-value of the paired t-test is smaller than 0.05). Here, the direction of the arrow indicates whether this variation is increasing (▲) or decreasing (▼). As an example, we note that the "most reliable" class receives these temporal significance indicators between the first two buckets. This class (which was outperforming the other classes in terms of receiving user interactions in the first bucket) hence performs more similar to the other classes in the second bucket. For this group, the decrease pattern between the second and the third bucket is also significant, although the change is not as high. This is mainly due to this class having many samples (23,416 posts) and therefore more easily passing the t-test. The decreasing pattern of this class also continues in the last bucket, but with a sharper slope.

We make several other observations from Fig. 3. As an example, there is a positive correlation between the reliability level of news and the level of interaction they receive in the first bucket. Here, more reliable posts receive interactions at a higher rate during the first hour after posting. In contrast, for the bias parameter, the two extremely biased classes (i.e., "far right" and "far left") receive less interaction rate than the unbiased (balanced) class in this bucket. In the final bucket, the pattern is reversed, suggesting that unbiased postings are more successful in the early stages of their lifespan compared to strongly biased posts. Notably, even in the fourth bucket, unbiased postings receive higher interaction volume because their total interaction (the denominator of the TICR values) is significantly more than those of the other two extremely biased classes (37M vs. 7M and 3M interactions). This is one of the reasons why we chose to provide the temporal dynamics of the TICR values as opposed to the actual interaction values, as the TICR values capture these dynamics more precisely.

Another important observation we want to highlight is that for some classes, identifying the class that a sample trend belongs to is easier to profile when we look at the bias-reliability class not the reliability or bias class independently. As an example, consider the "most unreliable - left" class which receives statistical significance, and "asterisk" in the third bucket. Both of the means of the bias and reliability classes it belongs to is statistically different than this class. This observation suggests that the interaction level is best captured when bias and reliability are evaluated jointly. Another observation worth mentioning is regarding the temporal changes of the most unreliable news. This class has an increasing pattern in terms of the rate of change they experience (for all buckets, it is statistically significant). We have seen the exact opposite trend when it comes to the "most reliable" news sources.

Key observations: The "most reliable" posts and the "most unreliable" posts experience opposite trends in the interaction changes over time. In the first hour following the publication of the post, there is a positive correlation between the reliability of the post and the level of interaction it receives.

3.2 Temporal Dynamics of Different Interactions

We now turn our attention to each individual interaction type, including shares, likes, and comments. In order to capture how much of the total interactions is covered by each of the interactions in each time bucket, we use the same denominator as total interactions for each of these interaction types. As an example, if a post receives a maximum of 600 total interactions during our timeline of study and receives 90 likes during the first time bucket, we say that the TICR of likes for this period is 15%.

Shares: With sharing having perhaps the most direct effect on what news people may be exposed to, we start our analysis with shares. The TICR scores for the number of shares (of posts linking articles) are shown in Fig. 4. First, note that the average TICR value for all posts (the top-right cell in each table) has decreased from 25% to approximately 4% in each bucket (due to normalizing over the total interactions). Comparing the top-right cells in Figs. 4-6, we note that the fraction of shares is almost the same as the number of comments but smaller than the number of likes.

Second, while it should not be expected that the top-right cell of all buckets to have equal values when considering individual interaction classes, we note that they are almost the same (in the range of 4-4.5%). Since we picked the bucket thresholds to have roughly the same number of total interactions over all posts to each bucket (but not necessarily the same volume for each interaction type), this suggests that shares as an aggregate (over all classes of news articles) represents a relatively stable fraction of the total number of interactions.

Third, across all buckets, the "most unreliable" class is the clear winner. When compared to the other classes of reliability (first row) and even all classes of bias (last column), they receive a greater proportion of the shares during all buckets. Although this pattern could not occur for the "total interactions", it is feasible here since TICR values here are normalized over the total interactions. Referring back to the "total interactions", for which this class saw the lowest ratio of interactions in the first two buckets(when compared with the other reliability classes), we, therefore, expect the number of likes (Fig. 5) and comments (Fig. 6) to be comparatively less (than for the other reliability classes) for these two time buckets. This shows that the "most unreliable" news often is relatively more shared early, despite not seeing as many likes and comments, but that this evens out over time.

Fourth, some classes consistently (throughout the four time buckets) see a larger relative sharing fraction than the other classes. For example, the two extreme bias classes ("far right" and "far left") perform better than the remaining

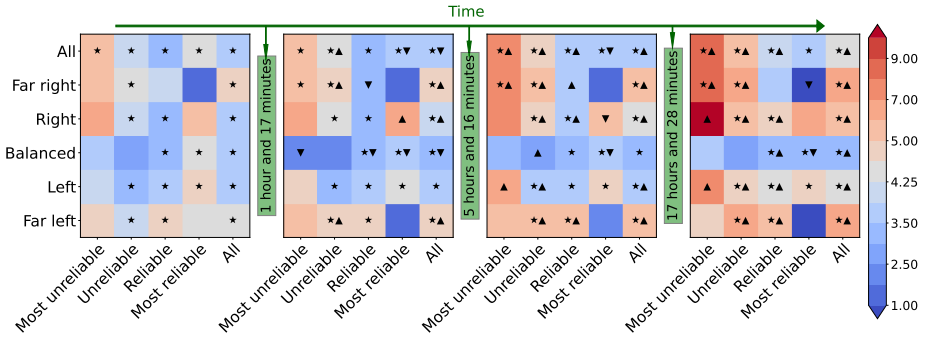


Fig. 4: Temporal dynamics of the total interactions covered for shares (⋆: coefficient of variation of the mean is smaller than 4%, ▲ and ▼: has deviation from the previous time bucket with p-value <0.05).

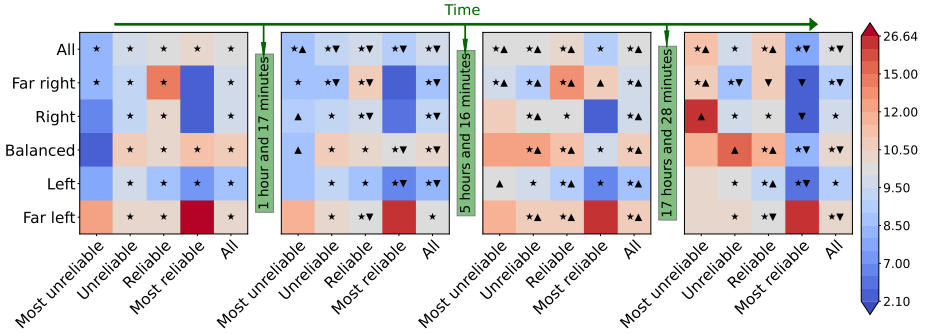


Fig. 5: Temporal dynamics of the total interactions covered for likes (⋆: coefficient of variation of the mean is smaller than 4%, ▲ and ▼: has deviation from the previous time bucket with p-value <0.05).

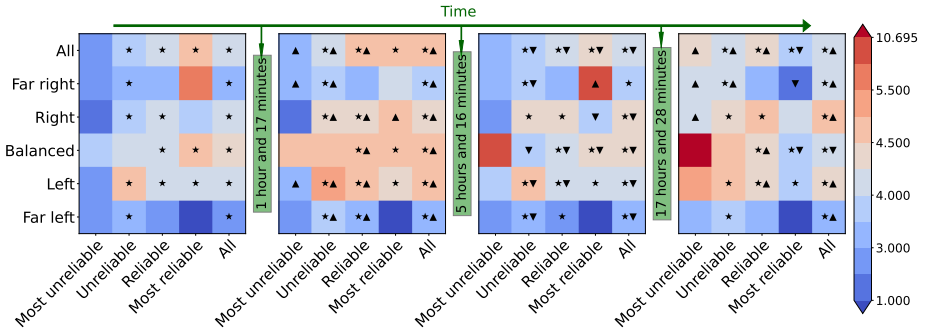


Fig. 6: Temporal dynamics of the total interactions covered for comments (⋆: coefficient of variation of the mean is smaller than 4%, ▲ and ▼: has deviation from the previous time bucket with p-value <0.05).

bias classes in all buckets. Moreover, for both of these extreme bias classes as well as for the "most unreliable" group, the trend of shares rate is increasing as time goes on. As a result of this trend, the last bucket exhibits a negative correlation between the total number of interactions covered for shares and the reliability of news, with the "most unreliable" news seeing relatively more late sharing.

Fifth, we observe several classes with relatively different temporal dynamics than the bias and reliability classes they belong to. As an example, we can clearly observe that for the third bucket, the "unreliable-right" class has a markedly different pattern than both the corresponding bias and reliability classes that it belongs to. This suggests that it is important to consider both these parameters in combination when predicting the sharing of the news in a bucket.

Key observations: Among all the reliability classes, the "most unreliable" posts experience the greatest gains in terms of share rates. During the late stages of the posts' lifetime (17 hours after publishing), there is a negative correlation between reliability levels and share rates. The most reliable postings receive the least normalized number of shares.

Likes: A more passive way to (indirectly) impact how visible posts on Facebook is to like various posts. One reason for this is that posts with many likes are more likely to occur higher up in the timelines of friends. A like also represents a user's (in most cases positive) interaction with the news. Fig. 5 shows the temporal dynamics of the total interactions covered for the number of likes. First, again it is evident that considering bias and reliability simultaneously will yield more reliable results. As an example, in the second bucket, the "unreliable-balanced" class deviates from the bias and reliability classes it belongs to.

Second, in the first bucket, a positive correlation is observed between the reliability level and the rate of likes in the early stages of the posts' lifetime (initial hour). In other words, during the initial time period, people more frequently like reliable news. This is in contrast to the share rates (Fig. 4), which happens more for unreliable posts during the very first stages of the posts' lifetime.

These observations may suggest that the sharing patterns and like patterns are substantially different and depend on the reliability and bias of the news. Yet, some similarities between their patterns can also be observed. For example, if we consider "all" posts, both metrics observe an increase in the third bucket.

Key observations: During the first hour following the publication of a post, there is a positive correlation between the post's reliability and the number of likes it receives. Throughout this period the "most reliable" posts experience the most like rates.

Comments: Similar to likes, comments provide an indirect way of exposing friends to various posts. However, in contrast to likes, a single user can give several comments on the same posts. Here, we treat all comments the same but

note that the somewhat larger fraction of comments in part may come from users making several comments on the same post.

Fig. 6 displays the temporal dynamics of TICR values for the number of comments. There are multiple observations to be made from this figure. First, the "most reliable" group, which was not successful in terms of shares, performs the best in the first three buckets of comment results. As a result, for a typical post in this category, we expect to see a higher normalized rate of comments during the first 17 hours after publishing it. Second, we observe (from the last columns) that the two extreme bias classes perform poorly, except in the final bucket of the "far-right" class. This is the opposite of the pattern we have observed for the shares of these classes. Third, we again observe deviations for several of the bias-reliability classes from the bias or reliability classes that they belong to. An example of this is the "unreliable- right" class in the last bucket.

Other interaction types: While Facebook also allows other interaction types, these typically see smaller interaction volumes and, therefore may have a less clear impact on the dissemination patterns of news. We include results for some of the other used interactions in Appendix A.2.

3.3 Outlet-specific Results

In addition to examining the temporal patterns of interactions across different types of news, we have also studied how bias and reliability of the news published by specific media outlets impacted people's interactions with the posts sharing that news over time.

For this analysis, we selected six outlets with the goal to achieve a fair comparison of popular outlets with different political biases. First, we selected the top-10 outlets with the most articles in our dataset to ensure that each selected outlet had sufficient samples for statistical significance. Second, we omitted yahoo.com which is among the top-10 outlets as it is more known as a news aggregator (rather than a news source). Third, from the remaining outlets, we selected to provide the analysis results for (1) two right-biased outlets (Fox News and New York Post), (2) two left-biased outlets (The New York Times and CNN), and two outlets that could represent (mostly) unbiased news sources (NPR and Reuters). For the classification of the outlets, we used Ad Fontes Media outlet-based ratings.[9] Table 4 lists these sources and high-level statistics extracted from our dataset, including their bias class (from Ad Fontes ratings), the number of articles in our data, the number of posts sharing these articles, the number of interactions related to these posts, the number of posts per article, the number of interactions per post, and their popularity in terms of their monthly visits.

We next present temporal analysis results for the total interactions of articles published by The New York Times (left-biased), Fox News (right-biased), and NPR (unbiased). Results for the other three outlets are found in Appendix A.3. Furthermore, using the code we publish, interested researchers can conduct

[9] Ad Fontes Media provides evaluations of both publishers and individual articles.

Table 4: Statistics of the outlets (†: M stands for million, ‡: Website monthly visits reported by similarweb.com (Oct. 2022)).

Outlet	N.Y.Times	Fox News	CNN	N.Y.Post	NPR	Reuters
Bias Class	Left	Right	Left	Right	Unbiased	Unbiased
Articles No	300	209	206	135	176	125
Posts No	6354	2797	2501	3085	2582	777
Interaction No	4.74 M†	6.18 M	3.45 M	1.68 M	3.88 M	0.48 M
Post per Article	21	13	12	23	15	6
Interaction per Post	746	2209	1378	543	1499	615
Monthly Visits‡	618.60 M	280.30 M	569.10 M	144.20 M	115.50 M	89.30 M

similar analyses for the remaining media outlets we studied, although not all of the results will be statistically significant.

The New York Times: Fig. 7 shows the results for The New York Times. We note that the white boxes represent categories of news for which we did not have data. As perhaps expected, for The New York Times, we did not have data for any of the right-biased categories (irrespective of reliability).

First, note that we are reporting the TICR statistics for the total interactions. As discussed previously, given the selection of bucket thresholds, we anticipate around 25% of TICR of the total interactions for all buckets when considering the overall population (reported in the top-right cell of each bucket). However, when considering individual publishers, this is not necessarily the case. For example, as seen in Fig. 7, the user engagement with posts linking news articles by The New York Times that are older than 17 hours is lower than average. Instead, posts linking their news appear most successful during the third bucket (5-17 hours after posts first appear). Second, a definite association between interactions with The New York Times news related posts receive and their reliability can also be seen when we focus on the early stages of postings (first two buckets) and late stages (17 hours onward) and exclude the most unreliable news (which has only six articles in our dataset). Here, the first stages' correlation is positive, whereas for the late stage this correlation is highly negative. While aiming to receive early engagements, it is clear that the reliability of the news plays a significant factor in the actual interaction levels achieved.

When considering bias, it is clear that less biased news receives higher interactions in the early time slots (although The New York Times belongs to the left party). Again, the trend of deviation of a class from the reliability and bias class that it belongs to can be seen in different buckets. As an example during hours 1 to 5 (after publishing a post), a typical post belonging to the "most reliable-left" class does not follow the pattern of the "most reliable" nor the "left group".

Fox News: Fig. 8 shows the temporal results for Fox News. For the first and the last bucket, similar to The New York Times, we see a correlation between reliability engagement, when again discarding the non-significant results of the

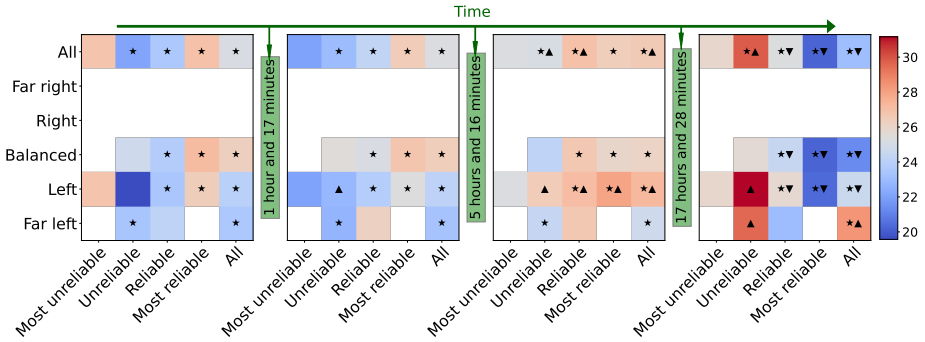


Fig. 7: Temporal dynamics of total interactions for The New York Times (white boxes: no data available, ⋆: coefficient of variation of the mean is smaller than 4%, ▲ and ▼: has a deviation from the previous time bucket with p-value <0.05).

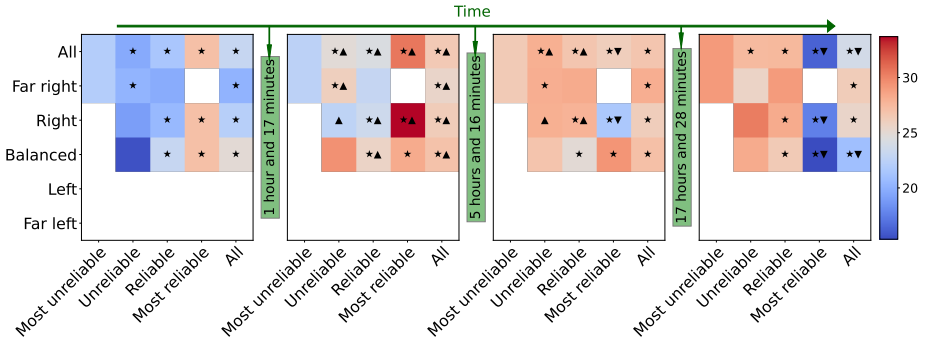


Fig. 8: Temporal dynamics of the total interactions covered for Fox News (white boxes: no data available, ⋆: coefficient of variation of the mean is smaller than 4%, ▲ and ▼: has a deviation from the previous time bucket with p-value <0.05).

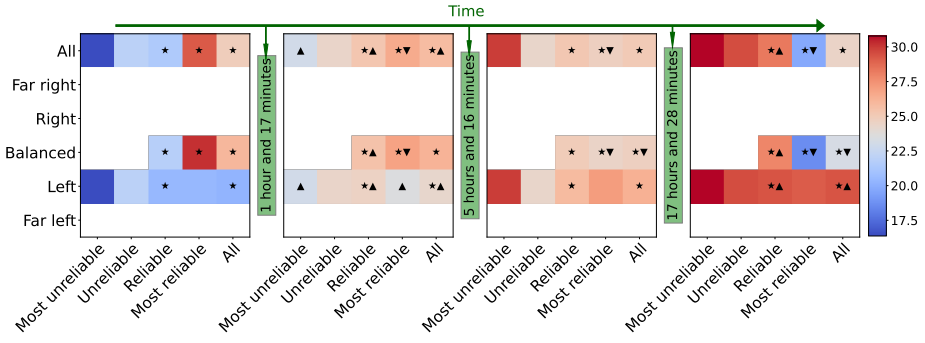


Fig. 9: Temporal dynamics of the total interactions covered for NPR (white boxes: no data available, ⋆: coefficient of variation of the mean is smaller than 4%, ▲ and ▼: has a deviation from the previous time bucket with p-value <0.05).

"most unreliable" class. After around five hours, the "most reliable" class loses its first-place ranking to the "unreliable" group. The large increase in unreliable news after 5 hours is statistically supported. Again, we can see a big, normalized decline in the most reliable news 17 hours after posting. Similarly to what we have observed for The New York Times, we may observe that in the earliest phases of a post's lifetime, balanced news is more engaging than biased ones, although Fox News itself is a right-biased biased news outlet. Here, statistical evidence supports the divergence we see for the bias class from the average population, until 5 hours after posting.

Key observation: In spite of Fox News and the New York Times being biased publishers, for both, related unbiased posts receive a higher interaction rate than biased ones in the first hour following posting.

NPR: Finally, we used NPR as an example of an outlet with very limited bias. As seen in Fig. 9, again balanced news receives higher interaction rates than the unbiased ones in the very first bucket, and the trend changes in the last bucket. The biased news published by this outlet tends to receive the most interaction during the late stages of the posts' lifetime. Moreover, the statistically significant decreasing pattern of the interaction rate with the "most reliable" news is worth noting.

4 Prediction of the Maximum Interaction's Volume

Another interesting aspect when comparing bias-reliability classes is the extent to which a post's maximum interaction value (denominator of TICR) is predictable from the post's interaction at each moment. To quantify the proportion of the variation in the denominator of TICR that can be explained by the current interaction a post has received, we next present a correlation-based analysis.

First, we divide the time axis into exponentially increasing time buckets. For the first bucket, we use a size of 15 minutes, and then we use a factor of 1.2 to increase the bucket sizes. Then, in each bucket and for each group, we compute the coefficient of determination (r^2) as the squared value of the Pearson correlation coefficient between the current interaction values of the posts of the class and the total interaction they receive in the future. Finally, we recorded the moments in which the (r^2) reached .6 and .8, respectively. Table 5 summarize the results. While more advanced prediction models might be used in practice, not limiting the discussion to a particular predictive model provides quantifiable insights into the extent to which we can rely on predictive models to estimate the total number of interactions from the current value of the interaction a post received (even with simple models). We next share some of our key observations.

First, note that in most classes a Pearson correlation coefficient of 0.8 (r^2 of 0.6) is achieved within one hour of posting, suggesting that the total number of interactions is relatively well predicted very early. Second, if all posts are

Table 5: Minimum time required for reaching high correlations between the current and ultimate interactions (m: minutes, h: hours, and d: days).

		Reliability									
		Most unreliable		Unreliable		Reliable		Most reliable		All	
		r^2>.6	r^2>.8	r^2>.6	r^2>.8	r^2>.6	r^2>.8	r^2>.6	r^2>.8	r^2>.6	r^2>.8
Bias	Far left	15m	21m	25m	1h, 51m	15m	15m	31m	31m	15m	31m
	Left	15m	15m	25m	2h, 13m	31m	2h, 13m	15m	37m	31m	1h, 17m
	Balanced	9h, 35m	1d, 10h	15m	13h, 48m	6h, 39m	16h, 33m	21m	1h, 51m	1h, 17m	11h, 30m
	Right	15m	15m	21m	1h, 51m	25m	2h, 40m	18m	18m	21m	2h, 13m
	Far right	15m	15m	15m	15m	15m	18m	15m	37m	15m	15m
	All	15m	31m	15m	53m	37m	11h, 30m	21m	1h, 4m	31m	6h, 39m

taken into consideration, this can be accomplished within 30 minutes of posting. Third, considering all reliability classes (last row), we can see that we can achieve this level of predictability within around 40 minutes of posting. Fourth, as we examine all bias classes in the last column, we can see that the biased classes are able to reach this level earlier than the unbiased classes.

Fifth, note that reaching the high value r^2 level of 0.8 for the general population (last row and last column of the table) is feasible within 7 hours after the posting. With regard to our definition of 4-bucket thresholds, we can say that for all the classes except for 3 we can reach the 0.6 level of r^2 in the first bucket. In the second bucket, it is also feasible to achieve an r^2 level of 0.8, except for the six classes. Finally, we note that for all classes except one, we can reach the r^2 level of 0.8 before the fourth bucket, allowing us to apply patterns observed in this bucket more broadly.

5 Related work

This paper relates to the works modeling and understanding the behavior of users, their interactions with various kinds of news and contents, and the factors that play roles in this context. For example, Aldous et al. [1] focus on the topic and emotional factors and analyze their effects on posting on five social media platforms (Facebook, Instagram, Twitter, YouTube, and Reddit) to demonstrate that user engagement is strongly influenced by the content's topic, with certain topics being more engaging on a particular platform. Their work shows that the engagement level is impacted differently on various platforms and by different topics. They also demonstrate that post emotion is indeed a significant factor. Karami et al. [10] demonstrate how social engagement may be used as a distin-

guishing characteristic between false and true news spreaders. However, they do not consider the temporal patterns of different user interactions in their study.

The most comparable work to ours is the recent work by Edelson et al. [4]. Their large-scale study explores how consumers engage with news inside the Facebook news ecosystem, as well as with specific pieces of news from unreliable suppliers and also between the suppliers and their audiences. However, their methodology is distinct from ours in that they base their study on publisher ratings rather than independent bits of news, they use binary classes for reliability, and they do not account for the temporal dynamics of the user interactions. Galen et al. [21] carried out a similar investigation as Edelson et al. on Reddit rather than Facebook. They also employ publisher-based rankings and demonstrate that low-factual content receives 20% fewer upvotes and 30% fewer cross-posting exposures than neutral or more factual information.

In another line of research, Allcott et al. [2] examine how users engage with fake news information and websites. Their findings indicate that through the end of 2016, user interactions with fraudulent information increased consistently on both Facebook and Twitter. Since then, engagements on Facebook have decreased significantly while continuing to increase on Twitter. Another group of studies related to our work are the ones which examine the temporal dynamics of user interactions but in different contexts. For example, Vassio et al. [19] examine how influencer-generated material draws interactions over time. Their findings indicate that while the growth rate of interactions naturally decays with time, the decay rate differs substantially between posts and social media platforms. As another related work and with a different methodology from the above works, in [12] the authors use NLP techniques to analyze over 2,5 million social media comments. The results show that Social media misinformation is largely disregarded by users.

6 Limitations

Our study has four main limitations that the researchers should consider when generalizing the findings. First, we dropped the posts with less than 10 total interactions from our study. While these types of postings constitute a significant portion of the total number of posts on Facebook, they make up a very small fraction of the total interactions (less than 5% in our dataset) and typically are of little interest to both Facebook content moderators (wanting to ban large interactions with misinformation) and also content publishers.

Second, similar to some other works (e.g, [4]), we limited the study to news postings and interactions on Facebook public forums (the most popular social media platform [17]). Therefore, interactions with news articles on other social media platforms and on the publisher's website were not considered. We consider a combined analysis that also takes into account these aspects as an interesting future work. It should also be noted that our study is based on the CrowdTangle dataset and does not consider every public page on Facebook. Yet, CrowdTangle

covers many pages from the whole public pages distribution. As as example, they index more than 99% of the pages with more than 25K followers [5].

Third, despite the t-test results indicating that the results are significant for several classes, the significance of the results may differ between different classes. To help interpret the significance of individual results the interested reader can consider also the number of articles in our dataset for each specific class. To help the interested reader to reproduce the results and more easily consider such additional dimensions, we will share our code. Here it should also be noted that we utilized the TICR distributions of the posts, not the aggregated results across the articles. One reason for this is that the number of posts for the flagged classes was sufficient for the findings to frequently have p-values less than 0.05.

Finally, The study focuses on the impact of bias and reliability on user engagement but does not account for other potential factors such as the relevance, timeliness, or credibility of the news source, as well as the user's individual preferences and views. Further research can consider these factors and their impact on user engagement, as well as investigate the effects of alternative labeling methods or different time frames compared to those used in the current study.

7 Ethical Considerations

All data was collected via public APIs while adhering to the rate limits of the companies hosting the data. The study is done at the aggregate level and no specific individuals are revealed. The likelihood of a substantial portion of the analyzed posts having been removed from Facebook is low due to the 28-day temporal separation between the publication date of the article and the date of data collection.

8 Conclusions

This paper presented a large-scale investigation of the temporal dynamics of various user interactions with Facebook posts belonging to different classes of bias and reliability. Using a carefully designed methodology, our investigation has answered and provided statistically supported insights into the research questions outlined in the introduction. For example, we demonstrated that user engagement with news for various classes of bias and reliability varies over time and highlighted these differences (RQ1). We have also done a study for different interaction types and observed that various interactions for the same class have different temporal interaction patterns (RQ3). Various statistically significant patterns were identified in the answers to the above questions which examine the four dimensions of this study: bias, reliability, time, and interaction type.

First, the results illustrate the importance of incorporating time into future research. As an example, we saw that the "most reliable" posts and the "most unreliable" posts exhibit opposite trends in terms of total interaction dynamics. The results also show that the temporal patterns of user interaction varied among the various user interactions, highlighting that users tend to interact

differently with news of different levels of bias and reliability. A key benefit of this identification is that it allows users to profile temporal engagement patterns with varying types of news, including Facebook content moderators, by identifying different temporal patterns for different classes of interactions (e.g., shares, likes) to different posts. Moreover, this study highlights the importance of incorporating bias and reliability concurrently in future studies by showing that bias-reliability classes have statistically significant differences from bias and reliability classes with which they are associated.

As all of the temporal patterns addressed in our research are dependent on the total interactions covered metric, which requires direct access to the value of the total interactions a post receives, we have quantified the predictability of total interactions from intermediate interaction values. Except for a few specific classes, there are strong correlations between the current and total engagement that a post receives within a few hours after its posting. Additionally, we have quantified this effect for various classes (RQ2).

To conclude, as the first study to address all four dimensions of bias, reliability, time, and interaction type in a single investigation, this work quantified the effect of these factors on the interaction level dynamics a post receives.

Acknowledgements The authors express their gratitude to CrowdTangle for providing the Facebook data. They also extend their thanks to the four anonymous reviewers for their insightful comments that helped improve the paper.

References

1. Aldous, K.K., An, J., Jansen, B.J.: What really matters?: characterising and predicting user engagement of news postings using multiple platforms, sentiments and topics. Behaviour & Information Technology pp. 1–24 (2022)
2. Allcott, H., Gentzkow, M., Yu, C.: Trends in the diffusion of misinformation on social media. Research and Politics **6**(2) (4 2019). https://doi.org/10.1177/2053168019848554
3. Barfar, A.: Cognitive and affective responses to political disinformation in Facebook. Computers in Human Behavior **101**, 173–179 (12 2019). https://doi.org/10.1016/j.chb.2019.07.026
4. Edelson, L., Nguyen, M.K., Goldstein, I., Goga, O., McCoy, D., Lauinger, T.: Understanding engagement with U.S. (mis)information news sources on Facebook. Proceedings of the ACM SIGCOMM Internet Measurement Conference, IMC pp. 444–463 (2021). https://doi.org/10.1145/3487552.3487859
5. Elena: What data is CrowdTangle tracking? https://help.crowdtangle.com/en/articles/1140930-what-data-is-crowdtangle-tracking (8 2021)
6. Elena: Crowdtangle-about us. https://help.crowdtangle.com/en/articles/4201940-about-us (2022)
7. Ferrara, E., Interdonato, R., Tagarelli, A.: Online popularity and topical interests through the lens of instagram. In: Proceedings of the 25th ACM conference on Hypertext and social media. pp. 24–34 (2014)

8. Ferreira, C.H., Murai, F., Silva, A.P., Almeida, J.M., Trevisan, M., Vassio, L., Mellia, M., Drago, I.: On the dynamics of political discussions on Instagram: A network perspective. Online Social Networks and Media **25**(December 2020), 100155 (2021). https://doi.org/10.1016/j.osnem.2021.100155, https://doi.org/10.1016/j.osnem.2021.100155
9. Gallup and Knight Foundation: Americans Views 2020: Trust, Media and Democracy. A Deepening Divide. Tech. rep., Knight Foundation (2020), https://knightfoundation.org/reports/american-views-2020-trust-media-and-democracy/
10. Karami, M., Nazer, T.H., Liu, H.: Profiling fake news spreaders on social media through psychological and motivational factors. In: Proceedings of the 32nd ACM Conference on Hypertext and Social Media. pp. 225–230 (2021)
11. Kubin, E., von Sikorski, C.: The role of (social) media in political polarization: a systematic review. Annals of the International Communication Association **45**(3), 188–206 (2021)
12. Metzger, M.J., Flanagin, A.J., Mena, P., Jiang, S., Wilson, C.: From dark to light: The many shades of sharing misinformation online. Media and Communication **9**(1), 134–143 (2021)
13. Mitra, T., Gilbert, E.: CREDBANK: A large-scale social media corpus with associated credibility annotations. Proceedings of the 9th International Conference on Web and Social Media, ICWSM 2015 pp. 258–267 (2015)
14. Otero, V.: Ad fontes media's multi-analyst content analysis white paper (2021), https://adfontesmedia.com/white-paper-2021
15. Otero, V.: Ad Fontes Media's Multi-Analyst Content Analysis White Paper (2021), https://adfontesmedia.com/white-paper-2021
16. Shearer, E.: 86% of Americans get news online from smartphone, computer or tablet | Pew Research Center (1 2021), https://www.pewresearch.org/fact-tank/2021/01/12/more-than-eight-in-ten-americans-get-news-from-digital-devices/
17. statista: Most popular social networks worldwide. https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users (2022)
18. Trevisan, M., Vassio, L., Giordano, D.: Debate on online social networks at the time of COVID-19: An Italian case study. Online Social Networks and Media **23**(April), 100136 (2021). https://doi.org/10.1016/j.osnem.2021.100136, https://doi.org/10.1016/j.osnem.2021.100136
19. Vassio, L., Garetto, M., Chiasserini, C., Leonardi, E.: Temporal dynamics of posts and user engagement of influencers on facebook and instagram. In: Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. pp. 129–133 (2021)
20. Vosoughi, S., Roy, D., Aral, S.: The spread of true and false news online. Science **359**(6380), 1146–1151 (2018). https://doi.org/10.1126/science.aap9559
21. Weld, G., Glenski, M., Althoff, T.: Political bias and factualness in news sharing across more than 100,000 online communities. In: Proceedings of the International AAAI Conference on Web and Social Media. vol. 15, pp. 796–807 (2021)
22. Wischnewski, M., Bruns, A., Keller, T.: Shareworthiness and Motivated Reasoning in Hyper-Partisan News Sharing Behavior on Twitter. Digital Journalism **9**(5), 549–570 (2021). https://doi.org/10.1080/21670811.2021.1903960

A Appendix

A.1 Procedure of Computing the Canonical Form of an Article Url

The following procedure is taken to transform URLs to canonical form. We begin by converting all text to lowercase. We then delete the protocol schema (e.g. 'http://') and remove any prefix instances of the strings 'www.' that may be present. Next, we remove any # signs from the URL except for the domains that it could not be removed from the canonical form (e.g., some of *edsource* or *npr* domains URLs). Then, we remove all URL query parameters except for the domains for which this was part of their canonical form (e.g., for some of *abcnews.go.com* domain URLs). As an example, the canonical form of the URL "https://www.nytimes.com/2020/10/24/technology/epoch-times-influence-falun-gong.html?referringSource=articleShare" that we used to collect posts was "nytimes.com/2020/10/24/technology/epoch-times-influence-falun-gong.html".

A.2 Temporal Dynamics of the other Forms of Interactions

In section 3.2 we studied the temporal dynamics of "likes", "shares", and "comments" as the most common interactions users make with Facebook posts. With the same conventions discussed in section 3.2, we here present the results for 2 other common interactions which are "angry", and "haha" in Figs. 10-11. Researchers interested in extracting the statistical analysis results for other types of interactions may use our code. Several observations can be drawn from these results. First, among all the time buckets for the "angry" results, biased posts outperform balanced posts when we simply consider the aggregated bias classes (the right-most column). In other words, biased posts make users angrier than unbiased posts. Second, the general trend for "angry" interactions for the whole population (the top right cell) is that it increases during middle buckets and then decreases after around 17 hours after posting. In other words, angry interactions with posts are more likely to happen during the second and third time buckets. Third, when we consider the left group, the most reliable class gets the most "angry" interactions. A fourth observation is that, when focusing on the "haha" interaction dynamics, there is a general tendency toward decreasing interaction rates for the whole population (right topmost cell). In other words, compared to the other buckets, a greater number of "haha"s is received during the first hour following posting. It should be noted that most of the classes that received arrows and therefore have significant trend changes follow this decreasing pattern. Finally, when we consider just the aggregated bias classes (the right columns), it is apparent that the "right" class received higher rates of "haha"s during all buckets and compared to the other bias classes which have significant means.

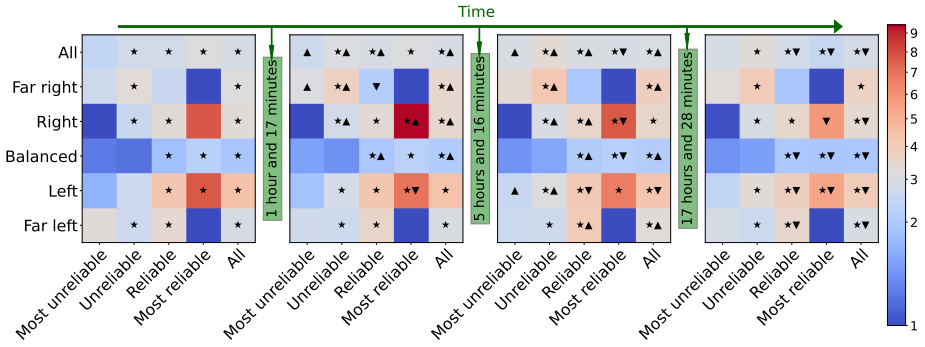


Fig. 10: Temporal dynamics of the total interactions covered for angry counts (⋆: coefficient of variation of the mean is smaller than 4%, ▲ and ▼: has deviation from the previous time bucket with p-value <0.05).

A.3 Temporal Dynamics of CNN and The New York Post and Reuters

The temporal dynamic results for CNN (as our second left-based example) and New York Post (as our second right-based outlet) and Reuters (as the second least biased publisher) are presented in Figs. 12-14. In contrast to the other biased example outlets, we observed both right-biased and left-biased articles published by New York Post.

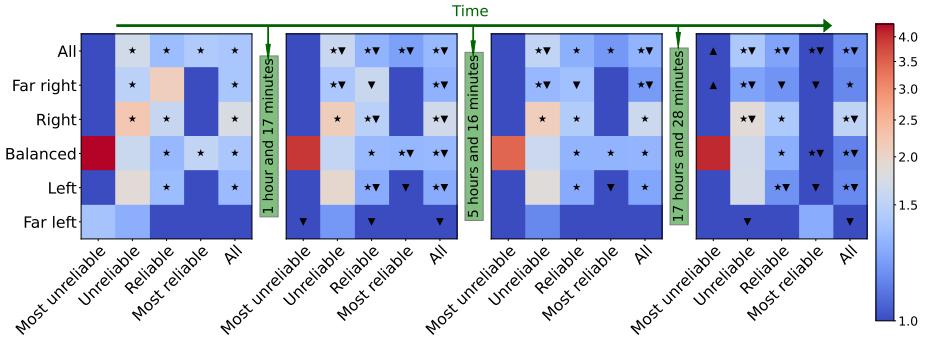


Fig. 11: Temporal dynamics of the total interactions covered for haha counts (⋆: coefficient of variation of the mean is smaller than 4%, ▲ and ▼: has deviation from the previous time bucket with p-value <0.05).

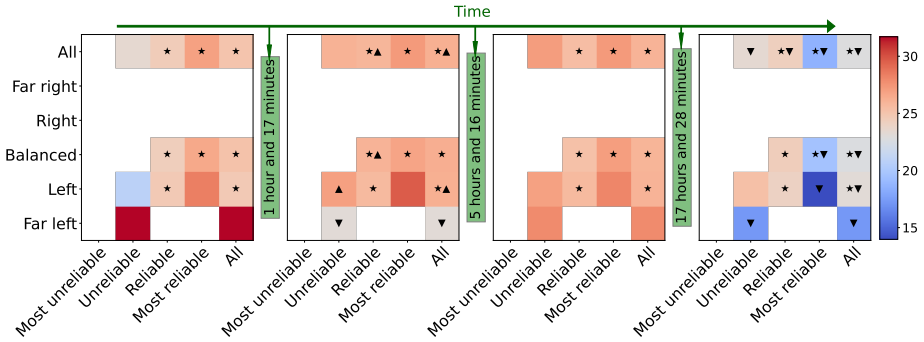


Fig. 12: Temporal dynamics of the total interactions covered for CNN ($\star$: coefficient of variation of the mean is smaller than 4%, ▲ and ▼: has deviation from the previous time bucket with p-value <0.05).

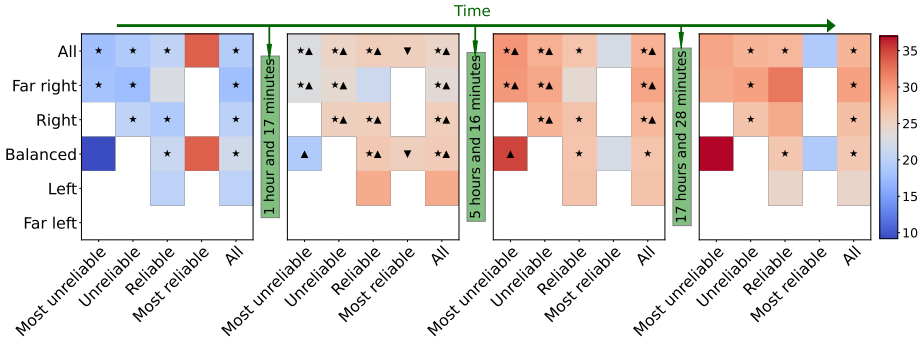


Fig. 13: Temporal dynamics of the total interactions covered for The New York Post ($\star$: coefficient of variation of the mean is smaller than 4%, ▲ and ▼: has deviation from the previous time bucket with p-value <0.05).

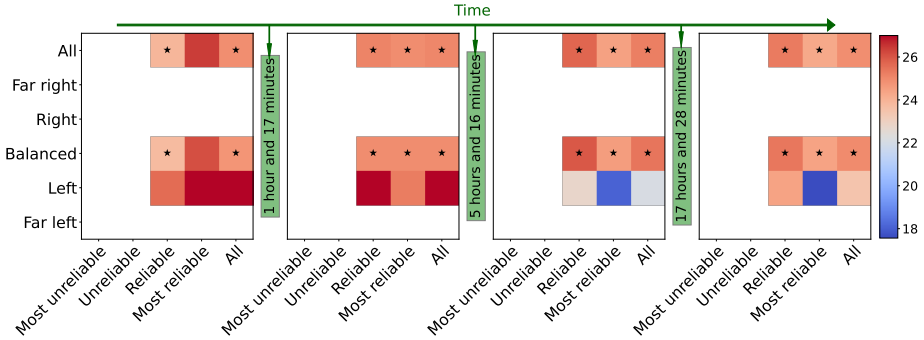


Fig. 14: Temporal dynamics of the total interactions covered for Reuters ($\star$: coefficient of variation of the mean is smaller than 4%, ▲ and ▼: has deviation from the previous time bucket with p-value <0.05).