Characterizing Online Criticism of Partisan News Media Using Weakly Supervised Learning

Karthik Shivaram[1], **Mustafa Bilgic**[2], **Matthew Shapiro**[3], **Aron Culotta**[1]

[1] Department of Computer Science, Tulane University, New Orleans, LA   USA
[2] Department of Computer Science, Illinois Institute of Technology, Chicago, IL   USA
[3] Department of Social Sciences, Illinois Institute of Technology, Chicago, IL   USA
kshvaram@tulane.edu, mbilgic@iit.edu, shapiro@iit.edu, aculotta@tulane.edu

Abstract

We propose novel methods to identify tweets that criticize partisan news sources. Prior work suggests that criticism, ridicule, and distrust of news media all play important roles in hyperpartisanship, misinformation, and filter bubble formation. Thus, understanding the prevalence and temporal dynamics of media-targeted criticism can provide us with updated tools to assess the health of the information ecosystem. There is a scarcity of labeled data for this task, and we develop a weakly supervised learning approach that leverages multiple noisy labeling functions based on both the content of the tweet as well as the historical news sharing behavior of the user. Using this classifier, we explore how tweets expressing criticism vary by user, news source, and time, finding substantial spikes in media criticism during politically polarizing events, such as the investigation into Russian interference in the 2016 U.S. elections and the 2017 "unite the right" rally in Charlottesville. This type of media-targeting criticism is also more likely to occur after users have been exposed to unreliable and hyperpartisan media.

1   Introduction

The media is crucial for a functioning democracy (Street 2010; Gentzkow, Glaeser, and Goldin 2006), but trust in news media has eroded. A recent Pew survey found that only 61% of U.S. adults have some or a lot of trust in the information they get from national news organizations, a drop of 15% from just six years prior (Liedke and Gottfried 2022). Such trends have implications for our political institutions, as distrust and animosity toward media contribute to hyperpartisanship, polarization, and misinformation sharing (Rathje, Van Bavel, and van der Linden 2021; Osmundsen et al. 2021). We currently lack methods to study how perceptions of the media evolve, how they vary according to news source, and how they are influenced by current events. Being able to observe media criticism — historically and in real-time — would allow us to monitor the health of the information ecosystem and identify trends in attitudes towards the media.

To address this methodological gap, this paper offers data and techniques to identify social media messages that express criticism and/or distrust of a news source (e.g.,

Figure 1: Example tweet critical of a news source.

Figure 1). We collect over 3.5M tweets mentioning news sources over the past ten years and then train a neural network to categorize tweets based on their content as well as the user's past engagement behavior. Labeled data is scarce and costly to generate, so we apply a weak supervision approach, using noisy labeling functions based on keywords and user attributes to train the classifier. After validating the classifier on a smaller number of manually-labeled examples, we apply it to all of our historical data, analyzing the prevalence of media-critical tweets by user, news source, and over time. Given these advances in our understanding of news engagement dynamics, the primary contributions of this work can be summarized as follows:

- **Dataset**: We construct a new dataset of 3.5M tweets that engage with one of 522 news sources over a ten year period. Tweet IDs, news sources, and inferred sharing intent are available at the paper repository.[1]

- **Weak supervision:** We find that classification using only text and user-based heuristics can provide accurate labels (89% F_1) but with modest coverage (48%). Training a weakly-trained classifier enables 100% coverage while maintaining high accuracy (84% F_1).

- **Effects on Polarization Estimates:** We find that adjusting for media-criticising tweets provides a different picture of the diversity of user news engagement. Our accounting for news-critical tweets shows that users previously engaging with diverse news sources are actually more hyperpartisan.

- **Criticism by user, news source, and over time:** After applying the classifier to the larger dataset, we find that the most criticised news sources are CNN and MSNBC on the

---

[1]https://github.com/tapilab/icwsm-2024-news-criticism

left and Fox News and OANN on the right; that hyper-partisan users are more likely to post critical tweets; and that the rate of news criticism exhibited several significant spikes during key political events (e.g., during the investigation of Russian involvement in the 2016 U.S. election).

2 Related Work

Political Stance Detection Detecting media criticism may appear related to political stance and bias detection (Demszky et al. 2019; Darwish et al. 2020; He et al. 2021; Li and Goldwasser 2021; Dutta et al. 2022), but it is a novel exercise. Media criticism goes beyond simply identifying the general political lean of a news article or topic, instead representing instances where messages are meant to ridicule, express distrust or sarcasm, or convey animosity *towards a news source*. Thus, we must distinguish between a tweet criticising a politician who appears on CNN and a tweet that criticises CNN directly. To our knowledge, there has been no attempt to generate a classifier for detecting media criticism, which is essential given the implications of negatively framed content in social media (Horner et al. 2021; Ferrara and Yang 2015; Stieglitz and Dang-Xuan 2013).

Media-Targeted Criticism Journalism is scrutinized by the public (Carlson 2016), at times in a highly critical and formalized way. The public expresses negative sentiment toward journalists and the institution of journalism itself with disgust and shame (Shin, Kim, and Joo 2021). This is particularly true for individuals who are least trustful of the news (Karlsson and Clerwall 2019). Criticism of journalism is also motivated by efforts of extremist political groups that foster the view that the media is neither legitimate nor accurate (Figenschou and Ihlebæk 2019). Identifying media-targeted criticism is essential when attempting to analyze factors that influence the spread of online misinformation (Gabriel et al. 2022). It also helps us better understand the political lean of online media (Stefanov et al. 2020), functioning as a more nuanced tool for assessing media-based ideology.

Weakly Supervised Classification Weakly supervised learning, a sub-field of machine learning, uses noisy sources of supervision to minimize human annotation efforts. Prior work applies approaches such as self-training (Karamanolakis et al. 2021), co-training (Karamanolakis, Hsu, and Gravano 2019), and crowd-sourcing (Li, Rubinstein, and Cohn 2019). Other work uses multi-task learning (Ratner et al. 2018, 2019) or generative models (Yu, Ding, and Bach 2022; Bach et al. 2017) to reduce the effects of label noise.

3 Data

Our goal was to collect news engagement tweets that are (a) diverse with respect to the partisan lean of the news source, (b) diverse with respect to the partisan lean of the users, and (c) posted over many years in order to observe long-term trends both at the user level and in aggregate. To do so, we identified a diverse set of Twitter users who engage with political news using the following steps:

Type	Count	News Engagement
All Tweets	36,543,574	3,491,270 (9.5%)
Quotes	1,417,012	171,403 (12.1%)
Retweets	18,823,632	1,965,171 (10.4%)
Replies	8,495,253	670,138 (7.9%)
Status	7,807,677	684,558 (8.8%)

Table 1: Tweets collected from 5,470 users and the fraction that reference one of 522 news sources.

Step 1: Collect news sources. We collected 419 English-based news sources from *allsides.com*, a media rating site used often in prior work (Baly et al. 2020; Liu et al. 2021; Sinno et al. 2022; Lyu and Luo 2022). Each news source is associated with a partisan stance in $\{-2, -1, 0, +1, +2\}$, ranging from extreme liberal (-2) to extreme conservative ($+2$). To cover a more diverse range of media quality, we extend the above collection by including 103 low-reliability sources from Osmundsen et al. (2021), itself originally sourced from Guess et al. (2019) (42 sources) and Grinberg et al. (2019) (125 sources).[2] These are rated as being either *pro-Republican* or *pro-Democrat*. We denote the partisan lean of these sources as -3 (pro-Democrat) and +3 (pro-Republican). We represent fake news with +3 and -3 stances to capture the extreme polarization attributed to such news, as it is considered a more intensified form of partisan bias, appealing strongly to users with corresponding political leanings. This representation aligns with polarization theory, which suggests that fake news is more attractive to polarized users due to its extremity and alignment with their partisan preferences (Osmundsen et al. (2021)). We retrieved the Twitter handle and URL for each of the 522 resulting news sources.

Step 2: Identify users. To identify users who engage with these 522 news sources, we used the Twitter Search API to find tweets that either mention a news source's Twitter account or contain a URL matching the news source's web domain. We submitted queries for each news source in Fall 2021, yielding 1.67M matching tweets.

While these matched users are likely to be actively engaged with news, we also wanted to widen the user set to account for a more diverse set of users. To do so, we used the Twitter Streaming API to sample from all English-language tweets posted during the same time period. We added these 59K tweets to the 1.67M already collected.

Step 3: Filter users. For studying long term trends, we retained only those users identified in the previous step with accounts at least five years old. Additionally, to limit the impact of automated accounts, we removed users that appear to exhibit automated behavior based on frequency of tweets, number of followers, and number of friends (c.f., Appendix A). After filtering, we then sampled users to diversify by partisan stance and news source, ensuring that the dataset is not dominated by a handful of news sources. To do so, we sampled ∼600 users from each partisan stance (based on the

---

[2]We took 167 from Osmundsen et al. (2021) and dropped those that either (a) were already in AllSides or (b) did not have a Twitter account or website. This reduced the list to 103 sources.

news source they mention); within each partisan stance, we sampled an equal number of users for each news source. We added to this set a random sample of 1,200 users identified from the Streaming API in the previous step. This resulted in a final set of 5,470 users representing a diverse set of political interests and engagement.

Step 4: Collect and annotate timelines. After the users were sampled, we collected each user's entire timeline to identify a larger set of news engagement tweets. From the 5,470 users, we collected nearly 37M tweets spanning ten years. For each tweet, we searched for a mention or URL that refers to any of the 522 news sources identified in Step 1 and labeled each matching tweet with its corresponding partisan score (i.e., the score of the referenced news source). Of the 37M tweets, we observe in Table 1 that 3.5M engage with at least one of the 522 news sources, suggesting that these users are, by design, quite engaged with political news and thus should not be considered representative of all Twitter users or the U.S. population as a whole. (Please see §9 for more discussion of such limitations.)

4 Problem Formulation

With the above data, we express our problem as follows: For each tweet that mentions a news source, we must determine whether or not that tweet is critical of the news source. A *critical* tweet is one that expresses disapproval of the media source. Note that this is distinct from tweets that criticize entities described in a news article, as well as from tweets that criticize the event the article describes. We use the term *critical* to encompass a variety of connotations, such as ridicule, distrust, animosity, and sarcasm. As usual, these expressions range from the direct ("@FoxNews is garbage.") to the subtle ("Nice to see @CNN continuing with their objective, unbiased, journalism. {cough, cough}").

We formulate this as a binary classification task. For each tweet t_i mentioning a news source, we assign a class label $y_i \in \{0, 1\}$, where $y_i = 1$ indicates that the author is criticising the news source, and $y_i = 0$ indicates the absence of criticism.

Based on our initial exploration of the data, we make the following simplifying assumptions to formulate a more tractable task: (1) we remove direct retweets of news sources, as these do not add any additional context to assess intent (e.g., "RT @CNN: breaking news ...") ; (2) we remove tweets that are part of threaded replies, as it is challenging to determine who the target of criticism is (e.g., "@JoeSmith @CNN You are garbage"); and (3) we restrict analysis to tweets that either reply directly to a news source (e.g., "@FoxNews #FakeNews") or mention the news source in the body of the tweet (e.g., "When will @CNN stop lying?").

Finally, in line with prior work showing that engagement occurs most frequently with ideologically extreme content (Eady et al. 2021), we discovered early on that the rate of criticism is significantly higher for more partisan news sources. To focus on the most salient subset of data, we thus target the 216 news sources with partisan stance in $\{-3, -2, +2, +3\}$. These sources include popular outlets such as CNN, MSNBC, and Slate on the left, and Fox,

OANN, and Breitbart on the right. With these assumptions, our final sample consists of 1.2M tweets that qualify for classification.

5 Methods

Given the lack of labeled data, and given the presence of several strong classification signals based on user and keyword features, weakly supervised learning provides an efficient methodology to train a classification model for this task (Kim and Ghahramani 2012; Ratner et al. 2018; Li, Rubinstein, and Cohn 2019). The overall approach is to first define a set of *labeling functions* that can provide noisy labels for a large subset of data. For example, the presence of the term #FakeNews may serve as a strong labeling function. As well, when a user who mostly engages with strongly conservative media mentions a strongly liberal news source, it is probable that the intent is to criticise the liberal news source.

We train a classifier instead of using these labeling functions directly because a classifier can generalize beyond the specific rules encoded in the labeling functions, allowing it to make predictions on unseen data with greater performance, also a classifier can integrate the signals from multiple labeling functions, potentially resolving conflicts and inconsistencies among them to provide more robust and reliable predictions.

Once these labeling functions are defined, they are used to create training data for a classifier. To account for label noise, we compare several weakly supervised learning methods designed for such scenarios. We next describe the labeling functions and classification methods in turn.

5.1 Labeling Functions

We provide here an overview of the labeling functions used to train the classifier. (For more details, please see Appendix B.) We define a labeling function ϕ as a collection of heuristics that maps a given tweet to a corresponding label $y \in \{0, 1, -1\}$. For our purposes, 0 represents the absence of criticism, 1 represents the presence of criticism, and -1 represents the inability of the labeling function to assign labels (abstention) due to either missing information or certain cut-off thresholds not being met, as described below. We implement labeling functions based on the three following types of information:

User features (ϕ_{up}): This labeling function relies on a user's historic news engagements and the political accounts they follow. First, we estimate the partisan stance of the user (conservative or liberal) based on whom they follow and the partisan lean of the news sources with which they engage. For example, if more than 90% of the political accounts they follow are liberal, and if more than 90% of their news engagement tweets are liberal, the user is labeled as liberal.[3] This labeling function annotates each news engagement tweet that is aligned with the user's partisan lean as 0. Tweets of the opposite partisan lean are annotated as 1. This function abstains for tweets from users for whom we could not infer partisan lean based on the thresholds above.

---

[3] The 90% threshold is a tuning parameter to tradeoff precision and coverage.

Text Features (ϕ_{tt}): This labeling function relies on the text of the tweet in which the user mentions a news source. We consider keywords both indicative of criticism (e.g., "propaganda," "fake news") as well as those indicative of support (e.g., "must watch", "worth reading"). In addition to individual words/phrases, we also consider word collocations — e.g., when "false" and "story" appear in any order, the text is labeled as critical.

Union of the above (ϕ_{un}): This labeling function takes the union of the prior two functions ϕ_{up} and ϕ_{tt}, ignoring conflicting assignments. That is, if the two functions agree, or if one of them abstains, the predicted label is returned; otherwise, it abstains. By removing conflicting labels, we expand the coverage of the single labeling functions and reduce label noise.

To assess the coverage of each function, we apply them to the unlabeled data filtered as described in §4 and report the estimated label distribution in Table 2. We observe that ϕ_{tt} and ϕ_{up} exhibit similar levels of coverage, labeling 30% and 26% of the data, respectively, and each labeling 3% of the data as critical. As a fraction of the labeled instances, excluding abstentions, each method labels about 11% as critical. The union function ϕ_{un} improves coverage to 48%, while also assigning about 11% of the labeled instances to the critical class. This limited overlap between the instances labeled by the user and text labeling functions suggests that they are complementary. We discuss the accuracy of these labeling functions on manually annotated data in §6.

5.2 Classification Models

In this section, we discuss the different models trained based on the labeling functions from the previous section. We consider separate neural networks based only on user features or only on text features, as well as a network that combines the two. Unlike the labeling functions, these models are binary classifiers: critical vs. not critical.

User Network This network model takes as its input hand-crafted features based on the user's Twitter profile as well as their historic news engagements. These include features such as the distribution of partisan stances among a user's mentions or follows, the partisan stance of the tweet being classified, and how the news source is referenced (e.g., direct reply or mention in tweet body).[4]

We use a simple fully-connected network for this model. For each tweet, we extract the above features, f_i, and pass them through one hidden layer with relu activation, followed by a classification layer with sigmoid activation:

$$z_u = \text{relu}(W_u f_i + b_u) \qquad \hat{y}_i = \sigma(W_o z_u + b_o)$$

Text Network This network uses the actual text of the tweet (and any referenced tweet) to perform classification. To improve generalizability, we first pre-process all tweets by replacing Twitter handles with a placeholder token. We then pass each tweet through a version of the RoBERTa language model pre-trained on English tweets (Barbieri et al. 2020) to obtain word-level representations $\{a_0, a_1, ...a_p\}$, after which we perform a pooling aggregation to obtain a single vector representation r_i. This is passed through one hidden layer and one classification layer with sigmoid activations:

$$z_t = \sigma(W_t r_i + b_t) \qquad \hat{y}_i = \sigma(W_g z_t + b_g)$$

Combined Network This network combines user features and text based representations together in order to identify if a given tweet contains the presence of criticism towards an engaged news source. For each tweet we obtain the intermediate representations z_u (from the User Network) and z_t (from the Text Network) and pass them through linear layers to obtain z_{cu} and z_{ct}. These are then concatenated to obtain $z_c = [z_{cu}, z_{ct}]$ and passed through to the final output layer to compute the corresponding class label $\hat{y}_i$:

$$z_{cu} = \sigma(W_{cu} z_u + b_{cu}) \quad z_{ct} = \sigma(W_{ct} z_t + b_{ct})$$
$$\hat{y}_i = \sigma(W_h z_c + b_h)$$

We use binary cross-entropy as the objective function to train all networks.

5.3 Label Denoising

Given the label noise inherent in the labeling functions above, we additionally experiment with several label denoising approaches. The general pipeline consists of fitting a probabilistic model that combines the labels generated by different labeling functions and denoises them to return soft (weighted) labels (Zhang et al. 2021). These soft labels are then used to train our classification models. We consider four different approaches that are appropriate for our task: Dawid Skene (DS) (Dawid and Skene 1979), IBCC (Kim and Ghahramani 2012), EBCC (Li, Rubinstein, and Cohn 2019), and Data Programming (DP) (Ratner et al. 2018).[5] These methods leverage agreements and conflicts between the different labeling functions in order to reduce label noise. For example, if two heuristics predict that a tweet is critical but a third does not, these denoising methods would learn how to resolve conflicts based on the frequency of disagreements for each heuristic. We use both ϕ_{up} and ϕ_{tt} generated labels to train these models, and we use the implementations provided by WRENCH (Zhang et al. 2021) (with default hyperparameters), a weak supervision benchmark library for classification tasks.

6 Classification Experiments

In this section we describe the experiments to validate the classification approach. To construct a smaller dataset for tuning and validation, we first randomly sample 300 tweets from the full dataset and manually annotate them. Because there is high class imbalance (only about 10% critical), this resulted in only a small number of critical examples. Thus, we augment these data by using our labeling functions to identify a sample of critical and non-critical examples, which we then manually annotate. After this annotation process we sub-sample from this resulting set to have balanced label distributions, splitting them into validation and

---

[4]A complete list of features is available in Appendix C.

[5]While other methods exist, they require a larger number of labeling functions to be effective.

function	critical	not critical	abstain
ϕ_{tt}	5,087 (3%)	39,356 (27%)	103,594 (70%)
ϕ_{up}	4,600 (3%)	34,601 (23%)	108,836 (74%)
ϕ_{un}	7,872 (5%)	63,181 (43%)	76,984 (52%)

Table 2: Labeling function output on unlabeled data

Function	Coverage	F1	Prec	Rec	Acc
ϕ_{up}	0.388	0.845	0.862	0.853	0.860
ϕ_{tt}	0.391	0.869	0.879	0.869	0.869
ϕ_{un}	0.600	0.888	0.888	0.888	0.888

Table 3: Labeling function accuracy on test data.

test datasets, ensuring that there are no overlapping users across our training, testing, and validation sets. The final dataset sizes for test and validation are, respectively, 312 and 233. The inter-annotator agreement of two annotators (one an author and another an unaffiliated student) has a Cohen's kappa of 0.71, indicating substantial agreement.

To train each weakly supervised model, we apply our labeling functions (ϕ_{up}, ϕ_{tt}, ϕ_{un}) to the unlabeled tweets filtered according to §4. Then, we sample a balanced distribution of labels across critical and non-critical classes for each of the three labeling functions. The final number of weakly supervised training examples used for each labeling function is 9,200 for ϕ_{up}, 10,174 for ϕ_{tt}, and 17,962 for ϕ_{un}. We train all possible combinations of our different network models and labeling functions, using the Adam (Kingma and Ba 2014) optimizer. We tune each network across different hyperparameter values (see Appendix, Table 9) and select the best parameters based on validation accuracy to identify our best models. These networks are implemented in Pytorch (Paszke et al. 2019).

Table 3 reports the performance of our labeling functions alone on the test set. We observe that the labeling functions are rather reliable (.845-.888 F_1 for the subset that they are able to classify), with the text heuristics slightly more accurate than the user heuristics; both have modest coverage (~40% of instances are able to be annotated by each labeling functions). The union function offers an apparent improvement over both, increasing coverage by ~20% and accuracy by 1-2%.[6]

Turning to the classification models, Figure 2 shows the ROC curves for each model/labeling function combination. We see that the text and combined networks perform better than the user network at different threshold values and across all three labeling functions. The worst performing labeling function across all three networks is ϕ_{tt}, which may be due to the poor generalization of the keyword-based heuristics.

Table 4 shows each classification model's ROC AUC score along with the Label Denoising enhancements. We observe that the text and combined networks are comparable in performance across different training settings. We also see that using soft-labels generated by the different label mod-
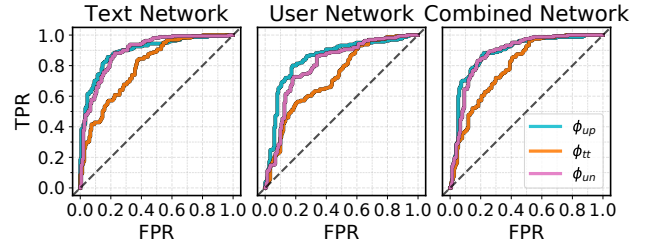
Figure 2: ROC curves of network models trained using different labeling functions.

Network	ROC AUC
Text Network + ϕ_{up}	0.800 ± 0.028
Text Network + ϕ_{tt}	0.719 ± 0.008
Text Network + ϕ_{un}	0.810 ± 0.017
Text Network + DS	**0.840 ± 0.007**
Text Network + IBCC	0.822 ± 0.014
Text Network + EBCC	0.836 ± 0.006
Text Network + DP	0.812 ± 0.013
User Network + ϕ_{up}	0.749 ± 0.049
User Network + ϕ_{tt}	0.657 ± 0.022
User Network + ϕ_{un}	0.744 ± 0.025
Combined Network + ϕ_{up}	0.810 ± 0.015
Combined Network + ϕ_{tt}	0.723 ± 0.006
Combined Network + ϕ_{un}	0.796 ± 0.008
Combined Network + DS	0.816 ± 0.015
Combined Network + IBCC	0.784 ± 0.038
Combined Network + EBCC	0.810 ± 0.023
Combined Network + DP	0.826 ± 0.015

Table 4: Test set ROC AUC for combinations of model, labeling function, and label denoising methods.

els (*Dawid Skene, IBCC, EBCC,* and *Data Programming*) improves performance compared to just training the models with hard labels generated by a single labeling function. The best performing model is the text network that uses soft-labels predicted by the *Dawid Skene* label model, achieving an ROC AUC score of 0.840, averaged across five random seed settings. This is a 3% improvement over using the text network on the ϕ_{un} labeling function alone.

7 Analysis of Media-Targeted Criticism

As our main interest for this work is to characterize criticism shown towards partisan news media, we next use our best performing model to classify all data filtered according to §4. The resulting dataset contains ~1.2 million tweets, of which ~1.16 million were labelled as non-criticism and ~45K were labelled as criticism. (Appendix A.1 reports additional validation measures for these classifications.)

7.1 Criticism by User Partisan Stance

To analyze how criticism varies across users with different partisan preferences, we bin all users into five bins based on their average partisan stance, which is estimated by the stance of news sources a user engages with through direct
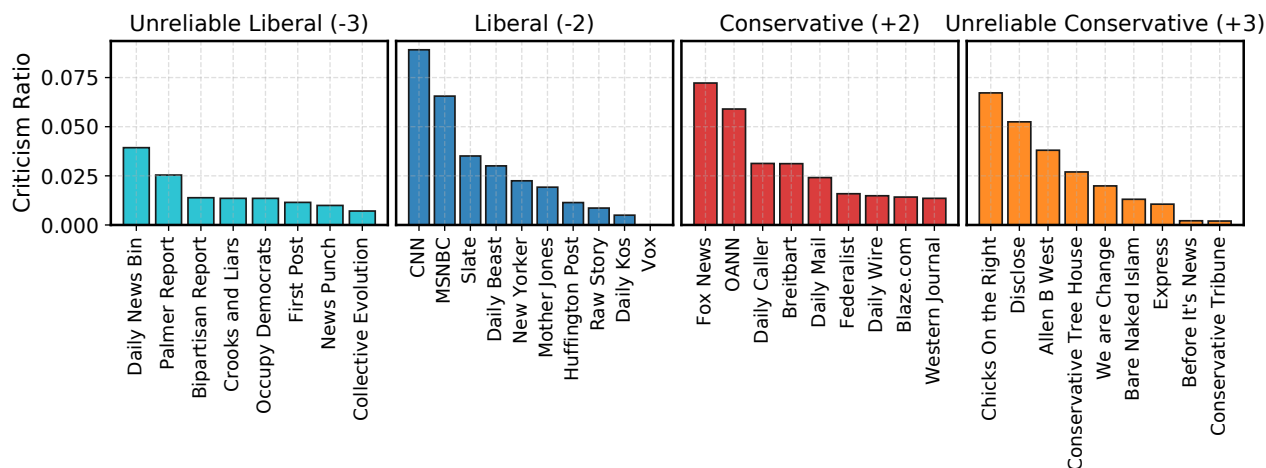
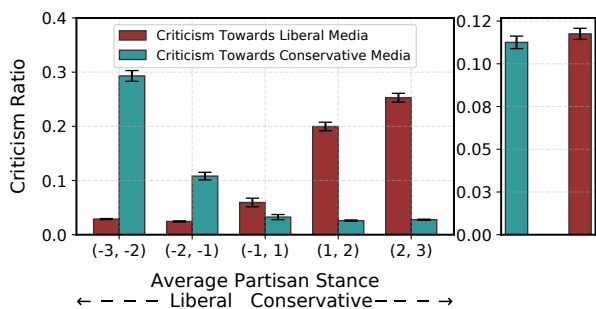Figure 3: Criticism shown towards the most mentioned news sources from each partisan stance.

Figure 4: Criticism ratio by partisan stance of the user (left panel) and in aggregate (right panel).

retweets.[7] For each bin, we compute the criticism ratio, defined as the proportion of news engagement tweets that criticise a news source. The results in Figure 4 show that users with extreme partisan preferences (Bins 1 and 5) are much more likely to express criticism than users with more moderate preferences. Another interesting observation is that users that are moderately liberal (bin 2) exhibit less criticism compared to moderate conservatives (bin 4). The rightmost panel of Figure 4 shows the overall criticism ratios, ignoring user bins, indicating a slightly higher level of criticism towards liberal media than towards conservative media, though the differences do not appear to be significant.

Note that Figure 4 also indicates a small rate of "self-criticism" — e.g., the leftmost bar indicates that a small ($< .04$) fraction of tweets from strongly liberal users about liberal media is critical; the same is true regarding conservative users and conservative news. While some of this may be due to unavoidable classification errors, a manual inspection confirmed that many of these are accurate. For example, many far right conservatives criticised Fox News for not sup-

---

[7]We assume direct retweets is a reliable indicator of user support for a particular news source.

porting the discredited theory that the election was rigged: "@FoxNews Please tell Chris Wallace there is serious fraud and the election is not even close to being decided! You folks are just hacks now, especially CW, not fair and balanced and unafraid that's for sure."

7.2 Criticism by News Source

Figure 3 plots the criticism ratio for the top ten most mentioned news sources from each partisan stance. Among liberal news media, CNN receives the highest rate of criticism ($\sim$9% of all engagements), followed by MSNBC ($\sim$6%). For conservative media, Fox ($\sim$7%) is targeted the most, followed by OANN ($\sim$6%). We also note that criticism shown towards unreliable conservative sources is greater than that shown towards unreliable liberal sources, although overall engagement with +3 news sources relative to -3 news sources is also greater.

The results suggest that more popular news sources receive higher rates of criticism, consistent with a recent Pew survey of media distrust (Jurkowitz et al. 2020), which estimates rates of cross-partisan distrust in each news source. The rankings of news sources in that survey align with the criticism ratios we find here: Fox (61% of liberals distrust), Breitbart (36%), DailyCaller (9%); CNN (58% of conservatives distrust), MSNBC (47%), Huffington Post (34%), Vox (11%). Thus, even though some sources may be seen as more extreme (e.g., DailyCaller vs. FoxNews; MSNBC vs. CNN), their lower popularity results in lower rates of criticism.

7.3 Effect of Criticism on Diversity Measures

Identifying tweets containing media-targeted criticism may affect estimates of user polarization (Conover et al. 2011; Garimella and Weber 2017; Garimella et al. 2021). For example, a user who posts supportive tweets for liberal news sources and critical tweets for conservative news sources may appear to have diverse news consumption if the intent of the tweet is ignored. We thus examine how measures of user diversity change after removing critical tweets.
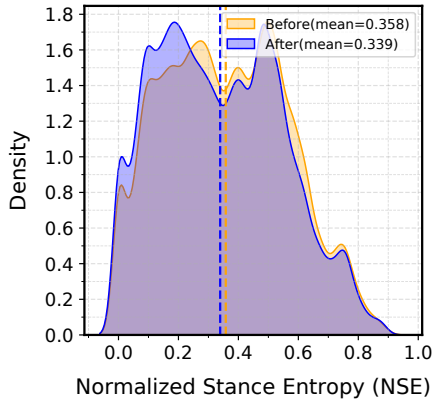
Figure 5: Comparison of Normalized Stance Entropy before and after removing critical tweets.

Figure 6: Criticism across time by partisan stance of the user (top panel) and news source (bottom panel).

To measure the diversity of a user's news engagements, we employ the normalized stance entropy measure (**NSE**) used in prior work on filter bubbles (Liu et al. 2021): NSE $= \frac{-\sum_{i=1}^{m} p_i \log p_i}{\log m}$, where p_i is the fraction of a user's engagements that belong to a particular stance $i \in \{-3, -2, 2, 3\}$, and $m = 4$ is the total number of partisan stances. NSE has a maximum value of 1, and higher values indicate more diverse news engagement.

Figure 5 shows the distribution of NSE scores before and after removing critical tweets. We can see that, after removing critical tweets, NSE drops from .358 to .339 on average (statistically significant at $p < .0001$ according to a t-test). The most noticeable change is the reduction in users in the middle range (.25 to .5), and a corresponding 13% increase in the number of users with low diversity (0 to .25). These results indicate that measures of user diversity do decrease once critical tweets are accounted for.

7.4   Criticism Over Time

To investigate how criticism towards news media changes over time, we calculate the fraction of critical tweets per month. Figure 6 plots both the source of criticism (top panel) and the target of criticism (bottom panel). We observe that criticism towards media has increased over time, with liberal media receiving higher criticism ratios than conservative media for most time periods.

We also observe substantial spikes in these time series (mid-2017, late-2018 and mid-2019). To identify possible events corresponding to these spikes, we extract the most common terms used during these time windows. For the mid-2017 spike, prominent terms like {*Trump, President, Obama, Comey, Russia, police, Charlottesville, Syria, Muslim*} reference events such as the investigations into Russian involvement in the 2016 U.S. elections, the "unite the right" rally in Charlottesville, and the Syrian War. For the 2018 spike, terms like {*Kavanaugh, vote, women, Mueller, investigation*} refer to the sexual misconduct allegations against Supreme Court judge Brett Kavanaugh and Robert Meuller's investigation into Russian interference in the 2016 U.S. elec-
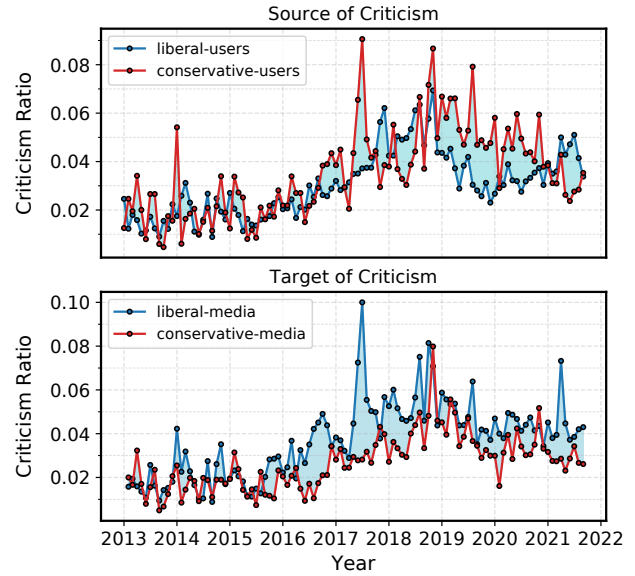
tion. For the mid-2019 spike, terms like {*Mueller, racist, 2020, Biden, children, border, Epstein*} refer to the detention of child migrants during the border crises and the arrest of Jeffrey Epstein for sex trafficking crimes. We thus find that certain high-profile, hyper-partisan events can lead to spikes in media criticism.

7.5   Progression Towards Criticism

Finally, we study how individual users engage with news over time, specifically the context surrounding the first occurrence of a critical tweet. To assess whether this takes place before or after users engage with partisan/unreliable media, we separate users into liberal and conservative groups based on the stance of their direct retweets (as in §7.1). For liberal users, we consider all who have engaged with -3 and -2 sources as well as those who have posted at least one tweet critical of conservative media (denoted C_C). Analogously, for conservative users, we consider those who have engaged with +3 and +2 sources while posting at least one tweet critical of liberal media (denoted C_L). We then find the first occurrence of each engagement type for each user and count the frequency of each sequence. The results in Table 5 indicate that, for both sides, the most common sequence is partisan → unreliable → critical. The next most-common sequence is partisan → critical → unreliable. These results suggest that engagement with low reliable, hyper-partisan media often precedes media criticism, though a more rigorous study is required to isolate causal effects.

8   Discussion and Conclusions

In this paper, we have proposed a methodology for identifying tweets that criticise partisan news media, and we have analyzed how such tweets vary by user, news source, and

User Group	Sequence	% of Users
Liberal	$-2 \rightarrow -3 \rightarrow C_C$	**49.07**
	$-2 \rightarrow C_C \rightarrow -3$	45.91
	$C_C \rightarrow -2 \rightarrow -3$	2.77
	$-3 \rightarrow -2 \rightarrow C_C$	2.11
	$C_C \rightarrow -3 \rightarrow -2$	0.13
	$-3 \rightarrow C_C \rightarrow -2$	0.00
Conservative	$2 \rightarrow 3 \rightarrow C_L$	**46.33**
	$2 \rightarrow C_L \rightarrow 3$	33.79
	$3 \rightarrow 2 \rightarrow C_L$	13.03
	$C_L \rightarrow 2 \rightarrow 3$	6.07
	$3 \rightarrow C_L \rightarrow 2$	0.49
	$C_L \rightarrow 3 \rightarrow 2$	0.29

Table 5: Progression sequences of first engagement of each type.

time. Our classification experiments indicate that weak supervision can effectively train such a classifier with limited manually-annotated data.

Some of the substantive results are intuitive – e.g., hyperpartisan users are more likely to criticise media from the other side (Figure 4), with CNN and Fox receiving the largest shares of critical tweets (Figure 3). Other results are more nuanced – e.g., unreliable news sources (-3, +3) do not necessarily receive more criticism than reliable news sources (-2, +2). Furthermore, we found substantial changes in critical tweets over time, including the tripling of the criticism ratio toward liberal media in mid-2017 and the doubling of the criticism ratio toward conservative media in late-2018. Finally, our accounting for media-oriented critical tweets reveals that user news engagement is not as politically diverse as one might otherwise expect.

9 Limitations

The limitations of this work include the following:

- User sample: As we aimed to identify users who engage with political news, the results should not be interpreted as representative of all of Twitter or the U.S. While most users appear to be based in the U.S., we made no effort to exclude users from other countries.
- Media sample: While we considered a wide range of news sources, for our primary results we focus on engagement with partisan and unreliable sources, omitting media rated as -1, 0, or 1 by AllSides. As described in §4, this was done in part to focus our efforts on media most often mentioned in critical tweets. Future work should consider including additional news sources.
- Article stance: While we attribute the overall stance of a news source to each of its articles, this assumption may not hold for all cases. Individual news articles can vary in their political leanings. Future work should consider analyzing the stances of news articles at a more granular level to account for this variability.
- News source stance: This study assumes that the partisan stances of news sources are static over time. However, in reality, these stances may shift due to various factors such as changes in policies or broader political dynamics. Fu-

ture work could assume dynamic partisan stances to better capture the evolving nature of news sources and their potential impact on user engagement.
- Classifier noise: As our experiments indicate, the classifier is imperfect, and these errors can propagate to the analysis in §7. Future work could apply adjustment methods to calibrate estimates of critical tweets (Forman 2005; Keith and O'Connor 2018).

We acknowledge that we do not consider the issue of whether criticism of the media fosters greater levels of democracy but rather whether a key feature of democracy – criticism in media – might play a role in promoting the consumption of more polarizing news. That said, criticism of the media itself is accepted by the public as a key feature to improve journalism (Craft, Vos, and Wolfgang 2016; Cheruiyot 2019), a feature that is unlikely to be eliminated in online news consumption patterns anytime soon.

10 Ethics Statement

Understanding how hyperpartisanship and misinformation evolve over time has important implications for civic discourse and political engagement. This work analyzes public, historical data from Twitter users but does not intervene in any way. As such, the study was determined to be exempt by the institution's IRB committee. Nevertheless, inferences made from online data are error prone and, with improper use, could lead to disparate treatment or targeting of users. We have released tweet IDs for this dataset for research purposes, but not the raw content, which is in line with Twitter's terms of service and FAIR principles.

References

Bach, S. H.; He, B.; Ratner, A.; and Ré, C. 2017. Learning the structure of generative models without labeled data. In *International Conference on Machine Learning*, 273–282. PMLR.

Baly, R.; Da San Martino, G.; Glass, J.; and Nakov, P. 2020. We Can Detect Your Bias: Predicting the Political Ideology of News Articles. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 4982–4991. Online: Association for Computational Linguistics.

Barbieri, F.; Camacho-Collados, J.; Neves, L.; and Espinosa-Anke, L. 2020. Tweeteval: Unified benchmark and comparative evaluation for tweet classification. *arXiv preprint arXiv:2010.12421*.

Carlson, M. 2016. Metajournalistic Discourse and the Meanings of Journalism: Definitional Control, Boundary Work, and Legitimation. *Communication Theory*, 26(4): 349–368.

Cheruiyot, D. 2019. *Criticising journalism: popular media criticism in the digital age*. Ph.D. thesis, Karlstads Universitet.

Conover, M.; Ratkiewicz, J.; Francisco, M.; Gonçalves, B.; Menczer, F.; and Flammini, A. 2011. Political polarization on twitter. In *Proceedings of the international aaai conference on web and social media*, volume 5, 89–96.

Craft, S.; Vos, T. P.; and Wolfgang, J. D. 2016. Reader comments as press criticism: Implications for the journalistic field. *Journalism*, 17(6): 677–693.

Cresci, S.; Di Pietro, R.; Petrocchi, M.; Spognardi, A.; and Tesconi, M. 2015. Fame for sale: Efficient detection of fake Twitter followers. *Decision Support Systems*, 80: 56–71.

Darwish, K.; Stefanov, P.; Aupetit, M.; and Nakov, P. 2020. Unsupervised user stance detection on twitter. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, 141–152.

Dawid, A. P.; and Skene, A. M. 1979. Maximum likelihood estimation of observer error-rates using the EM algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 28(1): 20–28.

Demszky, D.; Garg, N.; Voigt, R.; Zou, J.; Shapiro, J.; Gentzkow, M.; and Jurafsky, D. 2019. Analyzing Polarization in Social Media: Method and Application to Tweets on 21 Mass Shootings. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2970–3005. Minneapolis, Minnesota: Association for Computational Linguistics.

Dutta, S.; Caur, S.; Chakrabarti, S.; and Chakraborty, T. 2022. Semi-supervised stance detection of tweets via distant network supervision. In *Proceedings of the fifteenth ACM international conference on web search and data mining*, 241–251.

Eady, G.; Bonneau, R.; Tucker, J.; and Nagler, J. 2021. News Sharing on Social Media: Mapping the Ideology of News Media Content, Citizens, and Politicians. OSF Preprints.

Ferrara, E.; and Yang, Z. 2015. Measuring emotional contagion in social media. *PloS one*, 10(11): e0142390.

Figenschou, T. U.; and Ihlebæk, K. A. 2019. Challenging Journalistic Authority. *Journalism Studies*, 20(9): 1221–1237.

Forman, G. 2005. Counting positives accurately despite inaccurate classification. In *European conference on machine learning*, 564–575. Springer.

Gabriel, S.; Hallinan, S.; Sap, M.; Nguyen, P.; Roesner, F.; Choi, E.; and Choi, Y. 2022. Misinfo Reaction Frames: Reasoning about Readers' Reactions to News Headlines. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 3108–3127.

Garimella, K.; Smith, T.; Weiss, R.; and West, R. 2021. Political polarization in online news consumption. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 15, 152–162.

Garimella, V. R. K.; and Weber, I. 2017. A long-term analysis of polarization on Twitter. In *Eleventh international AAAI conference on web and social media*.

Gentzkow, M.; Glaeser, E. L.; and Goldin, C. 2006. The rise of the fourth estate. How newspapers became informative and why it mattered. In *Corruption and reform: Lessons from America's economic history*, 187–230. University of Chicago Press.

Grinberg, N.; Joseph, K.; Friedland, L.; Swire-Thompson, B.; and Lazer, D. 2019. Fake news on Twitter during the 2016 US presidential election. *Science*, 363(6425): 374–378.

Guess, A.; Lyons, B.; Montgomery, J. M.; Nyhan, B.; and Reifler, J. 2019. Fake news, Facebook ads, and misperceptions. *Democracy Fund*.

He, Z.; Mokhberian, N.; Câmara, A.; Abeliuk, A.; and Lerman, K. 2021. Detecting Polarized Topics Using Partisanship-aware Contextualized Topic Embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, 2102–2118.

Horner, C. G.; Galletta, D.; Crawford, J.; and Shirsat, A. 2021. Emotions: The unexplored fuel of fake news on social media. *Journal of Management Information Systems*, 38(4): 1039–1066.

Jurkowitz, M.; Mitchell, A.; Shearer, E.; and Walker, M. 2020. U.S. Media Polarization and the 2020 Election: A Nation Divided. https://www.pewresearch.org/journalism/2020/01/24/u-s-media-polarization-and-the-2020-election-a-nation-divided/. Accessed: 2024-01-02.

Karamanolakis, G.; Hsu, D.; and Gravano, L. 2019. Leveraging Just a Few Keywords for Fine-Grained Aspect Detection Through Weakly Supervised Co-Training. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 4611–4621. Hong Kong, China: Association for Computational Linguistics.

Karamanolakis, G.; Mukherjee, S.; Zheng, G.; and Awadallah, A. H. 2021. Self-training with weak supervision. *arXiv preprint arXiv:2104.05514*.

Karlsson, M.; and Clerwall, C. 2019. Cornerstones in Journalism. *Journalism Studies*, 20(8): 1184–1199.

Keith, K.; and O'Connor, B. 2018. Uncertainty-aware generative models for inferring document class prevalence. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 4575–4585.

Kim, H.-C.; and Ghahramani, Z. 2012. Bayesian classifier combination. In *Artificial Intelligence and Statistics*, 619–627. PMLR.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Li, C.; and Goldwasser, D. 2021. MEAN: Multi-head Entity Aware Attention Networkfor Political Perspective Detection in News Media. In *Proceedings of the Fourth Workshop on NLP for Internet Freedom: Censorship, Disinformation, and*

Propaganda, 66–75. Online: Association for Computational Linguistics.

Li, Y.; Rubinstein, B.; and Cohn, T. 2019. Exploiting worker correlation for label aggregation in crowdsourcing. In *International conference on machine learning*, 3886–3895. PMLR.

Liedke, J.; and Gottfried, J. 2022. U.S. adults under 30 now trust information from social media almost as much as from national news outlets. https://www.pewresearch.org/fact-tank/2022/10/27/u-s-adults-under-30-now-trust-information-from-social-media-almost-as-much-as-from-national-news-outlets/. Accessed: 2022-10-27.

Liu, P.; Shivaram, K.; Culotta, A.; Shapiro, M. A.; and Bilgic, M. 2021. The Interaction between Political Typology and Filter Bubbles in News Recommendation Algorithms. In *Proceedings of the Web Conference 2021*, 3791–3801.

Lyu, H.; and Luo, J. 2022. Understanding Political Polarization via Jointly Modeling Users, Connections and Multimodal Contents on Heterogeneous Graphs. In *Proceedings of the 30th ACM International Conference on Multimedia*, MM '22, 4072–4082. New York, NY, USA: Association for Computing Machinery. ISBN 9781450392037.

Osmundsen, M.; Bor, A.; Vahlstrup, P. B.; Bechmann, A.; and Petersen, M. B. 2021. Partisan polarization is the primary psychological motivation behind political fake news sharing on Twitter. *American Political Science Review*, 115(3): 999–1015.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Rathje, S.; Van Bavel, J. J.; and van der Linden, S. 2021. Out-group animosity drives engagement on social media. *Proceedings of the National Academy of Sciences*, 118(26).

Ratner, A.; Hancock, B.; Dunnmon, J.; Goldman, R.; and Ré, C. 2018. Snorkel metal: Weak supervision for multi-task learning. In *Proceedings of the Second Workshop on Data Management for End-To-End Machine Learning*, 1–4.

Ratner, A.; Hancock, B.; Dunnmon, J.; Sala, F.; Pandey, S.; and Ré, C. 2019. Training complex models with multi-task weak supervision. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 4763–4771.

Shin, W.; Kim, C.; and Joo, J. 2021. Hating journalism: Anti-press discourse and negative emotions toward journalism in Korea. *Journalism*, 22(5): 1239–1255.

Sinno, B.; Oviedo, B.; Atwell, K.; Alikhani, M.; and Li, J. J. 2022. Political Ideology and Polarization: A Multi-dimensional Approach. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 231–243. Seattle, United States: Association for Computational Linguistics.

Stefanov, P.; Darwish, K.; Atanasov, A.; and Nakov, P. 2020. Predicting the topical stance and political leaning of media using tweets. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 527–537.

Stieglitz, S.; and Dang-Xuan, L. 2013. Emotions and information diffusion in social media—sentiment of microblogs and sharing behavior. *Journal of management information systems*, 29(4): 217–248.

Street, J. 2010. *Mass media, politics and democracy*. Bloomsbury Publishing.

Yu, P.; Ding, T.; and Bach, S. H. 2022. Learning from multiple noisy partial labelers. In *International Conference on Artificial Intelligence and Statistics*, 11072–11095. PMLR.

Zhang, J.; Yu, Y.; Li, Y.; Wang, Y.; Yang, Y.; Yang, M.; and Ratner, A. 2021. Wrench: A comprehensive benchmark for weak supervision. *arXiv preprint arXiv:2109.11377*.

Ethics Checklist

1. For most authors...

(a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes, see Ethics section.

(b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes

(c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes, see Introduction.

(d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? Yes, see Limitations.

(e) Did you describe the limitations of your work? Yes, see Limitations.

(f) Did you discuss any potential negative societal impacts of your work? Yes, see Ethics.

(g) Did you discuss any potential misuse of your work? Yes, see Ethics.

(h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? Yes, see Ethics.

(i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes

2. Additionally, if your study involves hypotheses testing...

(a) Did you clearly state the assumptions underlying all theoretical results? Yes, see Sections 6 and 7.

(b) Have you provided justifications for all theoretical results? Yes, see Sections 6 and 7.

(c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? Yes, see Limitations.

(d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? Yes, see Limitations.

(e) Did you address potential biases or limitations in your theoretical framework? Yes, see Limitations.

(f) Have you related your theoretical results to the existing literature in social science? Yes, see Related Work.

(g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? Yes, see Section 8.

3. Additionally, if you are including theoretical proofs...

(a) Did you state the full set of assumptions of all theoretical results? NA

(b) Did you include complete proofs of all theoretical results? NA

4. Additionally, if you ran machine learning experiments...

(a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? No, but have released code and data.

(b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? Yes, see Appendix.

(c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? Yes, see Section 6.

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? Yes, see Appendix.

(e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? Yes, see Section 6.

(f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? Yes, see Section 10.

5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets...

(a) If your work uses existing assets, did you cite the creators? Yes, see Data.

(b) Did you mention the license of the assets? NA

(c) Did you include any new assets in the supplemental material or as a URL? Yes, we have released data and code at https://github.com/tapilab/icwsm-2024-news-criticism.

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? Yes, see Section 10.

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? Yes, see Section 10.

(f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR? Yes, see Section 10.

(g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset? See https://github.com/tapilab/icwsm-2024-news-criticism.

6. Additionally, if you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots? NA

(b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? NA

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? NA

(d) Did you discuss how data is stored, shared, and deidentified? NA

A   Bot Heuristics

After our initial seed user collection, we filter these to remove suspected bot accounts as well as those likely to be celebrities or organizations. We use a set of heuristics from the literature for this filtering step (Cresci et al. 2015), where we look at different characteristics of each account and compare them against different cut-off values. The characteristics and their corresponding cut-off values are as follows

1. Follower Size (≤ 1000)

2. Following Size (≤ 1000)

3. Daily Tweet Activity (≤ 10)

4. Total Tweets authored during the life of the account (≥ 1000 and ≤ 30000)

A.1   Additional Validation

To further validate the quality of the predictions the set of 1.2M tweets, we performed two additional checks. First, we manually annotated a random sample of 150 tweets predicted as criticism and 150 predicted as non-criticism, finding an accuracy score of 85%, in line with the the results above.

Next, we analyze the top terms in the tweets predicted to be critical and those predicted to be not critical. To do so, we computed chi-squared values for all terms to identify those that are significantly more likely in one class versus the other. For the critical class, the top terms are {fake, propaganda, #fakenews, lies, hey, tagged, news, FALSE, conspiracy, dear}; for the non-critical class, the top terms are {via, breaking, new, biden, says, bill, #oann, senate, texas, federal}. While many of these top terms are expectedly included in the heuristic functions (Table 7), it is notable that many additional terms are discovered to predict critical tweets, including {hey, tagged, dear, sponsors, misleading, coverup}. Terms like "hey" and "dear" arise from tweets like "Hey @cnn, fire your reporters." Combined with the accuracy results in Table 4, these additional checks give us more confidence in the quality of the data annotated by the classifier.

B   Labeling Functions

In the following subsections we define and describe each of the labeling functions in detail.

Threshold	Values
δ	0.05, 0.1, 0.15, 0.2, 0.25, 0.3, 0.35
ρ	5, 10, 15, 20, 25, 30

Table 6: Thresholds for ϕ_{up} tuned on validation data.

B.1 Based on User Profile (ϕ_{up})

This labeling function first uses heuristics based on a user's historic news engagements and political accounts they follow to estimate their partisan lean. Then, it labels tweets that engage with news sources of the opposite partisan lean as critical, and those that engage with the same partisan lean as not critical. We denote this labeling function as ϕ_{up}, where $\{e_{up}^a, e_{up}^{drt}, e_{up}^{pf}\}$ is the collection of heuristics it uses to assign a label to a given tweet, based on three different methods to estimate a user's partisan lean.

First, e_{up}^a measures the partisan distribution of *all* historic tweets of user u_j that mention a news source. It computes the fraction of tweets that engage with conservative ($c_{ac}^{u_j}$) and liberal ($c_{al}^{u_j}$) news sources. If the minimum of these fractions is less than or equal to a threshold δ_a, then the heuristic labels all tweets that mentions a news source. It assigns a value of 1 (critical) if the tweet $t_i^{u_j}$ of the user engages with a news source whose stance $s_i^{u_j}$ is equal to the minority partisan stance $ps_{min}^a(u_j)$ of the user's news engagements and 0 (not critical) otherwise. For example, if the user is strongly conservative ($c_{ac}^{u_j} >> c_{al}^{u_j}$), then the minority stance would be liberal and vice versa. We also check to see if the total number of a user's news engagement tweets are greater than a threshold ρ_a to make sure enough user information is present for the heuristic to work effectively.

$$e_{up}^a(t_i^u) = \begin{cases} -1 & \text{if number of news tweets} < \rho_a \\ & \text{or } \mathbf{min}(c_{ac}^{u_j}, c_{al}^{u_j}) > \delta_a \\ 1 & \text{if } s_i^{u_j} = ps_{min}^a(u_j) \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

Next, e_{up}^{drt} represents a similar heuristic as e_{up}^a but instead of considering all the tweets where the user engages with a news source, we only consider *direct retweets* (these are retweets whose original author is an official Twitter account of a news source). Focusing on direct retweets reduces noise introduced by quote tweets or news source mentions, which may be critical. Analogous thresholds δ_{drt} and ρ_{drt} are used to ensure that the minority partisan class is small enough, as well as to ensure that the overall volume of direct retweets is sufficient to estimate a user's partisan lean.

The third heuristic e_{up}^{pf} is analogous to the above, but instead estimates a user's partisan lean based on the set of political Twitter accounts a user follows. Using a dataset of politician Twitter accounts with party affiliation[8], it follows the same procedure as above to compute the fraction of liberal and conservative accounts a user follows, again assigning tweets as critical if they are cross-partisan according to

Name	Heuristics	Label
e_{tt}^d	cover up, covering up, shameful reporting, fakenews, fraud news, racist news, fraud network, racist network, not reporting, untrusted news, shit news, half truths, tell the truth, cant handle the truth, bunch of crap, stop lying, brainwashed, misinformation, disinformation, exaggerations, scaremongering, propaganda, fearmongering, hypocrisy, boycott	1
	watch this, must watch, live update, listen to, please read, read this, must read, worth reading, please share, study finds, top stories, top story, shocking news	0
e_{tt}^c	(false, fake, hoax, fictitious, misrepresent, one sided, bullshit, crap, shit, garbage, exaggerate) $\wedge$ (news, stories, reporting, narrative, media, network, reports) (conspiracy) $\wedge$ (theories, theory) (misinform, mislead) $\wedge$ (public, people, america) (made up, make up) $\wedge$ (lies, crap, shit) (brainwash, deceive) $\wedge$ (people, public, america) (spread) $\wedge$ (lies, propaganda, conspiracies, shit, fear) (biased) $\wedge$ (news, report, narrative, network, media, shit)	1
	(breaking) $\wedge$ (news, exclusive, report, story) (watch) $\wedge$ (now, live) (good, great, inspiring, incredible, real, thanks, thx, latest, fantastic) $\wedge$ (news, report, story, journalism, narrative, article, piece, video) (best) $\wedge$ (news, report, video)	0
e_{tt}^p	expose @NS, exposing @NS, exposes @NS, @NS exposed, @NS sucks, @NS is a joke, @NS fuck you, fuck you @NS, screw you @NS, @NS screw you, fuck @NS, @NS crap, @NS is crap, crap from @NS, @NS should fire, cant trust @NS, can not trust @NS, dont trust @NS, do not trust @NS	1
	via @NS	0

Table 7: Text heuristics (ϕ_{tt})

the estimated user stance, and 0 otherwise. Thresholds δ_{pf} and ρ_{pf} are used for this heuristic.

The final labeling function ϕ_{up} uses a unanimous vote to assign the final labels for a user's tweet t_i^u.

$$\phi_{up}(t_i^{u_j}) = \begin{cases} 1, & \text{if } e_{up}^a(t_i^{u_j}) == e_{up}^{drt}(t_i^{u_j}) \\ & == e_{up}^{pf}(t_i^{u_j}) == 1 \\ 0, & \text{if } e_{up}^a(t_i^{u_j}) == e_{up}^{drt}(t_i^{u_j}) \\ & == e_{up}^{pf}(t_i^{u_j}) == 0 \\ -1, & otherwise \end{cases} \quad (2)$$

The threshold parameters (δ, ρ) are tuned across different ranges (Table 6) to maximize accuracy on the manually annotated validation set. The final settings are: $\delta_a = .1$, $\delta_{drt} = .15$, $\delta_{pf} = .05$ and $\rho_a = 10$, $\rho_{drt} = 25$, $\rho_{pf} = 5$.

B.2 Based on Tweet Text (ϕ_{tt})

This labeling function uses multiple text based heuristics to annotate tweets (Table 7). We denote this labeling function as ϕ_{tt}, where $\{e_{tt}^d, e_{tt}^c, e_{tt}^p\}$ represents the types of heuristics it uses to assign a label to a given tweet.

Here e_{tt}^d represents a set of heuristics that perform a direct string match to label both the critical (e_{tt}^{d+}) and non-critical classes (e_{tt}^{d-}). It labels a user's tweet $t_i^{u_j}$ if any of these keywords or phrases are present in the text of the tweet.

e_{tt}^c represents a set of word collocation heuristics, each of which is a conjunction of two disjunctive clauses: $\{w_1^a \vee \ldots \vee w_k^a\} \wedge \{w_1^b \vee \ldots \vee w_m^b\}$, e.g., $\{\text{fake} \vee \text{false}\} \wedge \{\text{news} \vee \text{media} \vee \text{stories}\}$. Similar to e_{tt}^d, this heuristic set contains different sets of rules for both the critical (e_{tt}^{c+}) and non-critical (e_{tt}^{c-}) class.

Lastly, e_{tt}^p represents a set of phrase heuristics which check to see if a news source is mentioned with certain neighboring words in a specific order.

The final labeling function ϕ_{tt} uses a "logical or" to assign the final labels for a user's tweet $t_i^{u_j}$.

$$\phi_{tt}(t_i^{u_j}) = \begin{cases} 1, & \text{if } e_{tt}^{d+}(t_i^{u_j}) == 1 \\ & \vee\, e_{tt}^{c+}(t_i^{u_j}) == 1 \\ & \vee\, e_{tt}^{p+}(t_i^{u_j}) == 1 \\ 0, & \text{if } e_{tt}^{d-}(t_i^{u_j}) == 0 \\ & \vee\, e_{tt}^{c-}(t_i^{u_j}) == 0 \\ & \vee\, e_{tt}^{p-}(t_i^{u_j}) == 0 \\ -1, & otherwise \end{cases} \quad (3)$$

If any of these heuristics fail to assign a label due to them not being present in the text of tweet or if the heuristics for both the critical and non-critical classes both fire, we assign a label of -1. For the complete list of heuristics see Table 7.

B.3 Union of Labeling Functions (ϕ_{un})

This is a labeling function that performs a union over the labels generated by ϕ_{up} and ϕ_{tt}. It assigns a label of 1 if either of the labeling functions assigns a label of 1 and they don't conflict (similarly for class 0). It assigns a label of -1 when both labeling functions assign a label of -1 or they have conflicting labels (i.e $\phi_{up}(t_i^{u_j}) \neq \phi_{tt}(t_i^{u_j})$). We denote this labeling function as ϕ_{un}.

C User Features

The features for the User Network are described in Table 8.

D Experimental Settings

Table 9 shows all hyperparameters, optimized on the held-out validation set. The best setting is then used to compute accuracy on the test set (Table 4). All experiments use a system with 4 Nvidia A5000 GPUs, 512 GB RAM and an AMD Ryzen Threadripper 3975WX CPU. We report mean and standard deviation of scores across five random seeds.

Name	Description
News Source Engagement Partisan Distribution	The distribution of partisan stances of all the news sources the user engages with across all his tweets
Followed Politician Accounts Partisan Distribution	The distribution of partisan stances of all politicians the user follows
Followed News Source Accounts Partisan Distribution	The distribution of partisan stances of all news sources the user follows
Engaged News Source Partisan Stance	The partisan stance of the news source the user currently engages with in the current tweet
Tweet Type	The type of tweet (i.e retweet, replied_to, status, quote)
News Source Engagement Type	How the user engages with the news source in the current tweet (i.e mention, URL)
Is Direct Reply	If the current tweet is a direct reply to a news source twitter account
Multiple News Source Engagement	If multiple news sources are mentioned in the current tweet
Public Metrics	The public metrics of the current tweet
Engaged News Source Fraction	The fraction of engagements of the current news source engaged in the tweet

Table 8: User Based Features

Hyperparameter	Values
Learning Rate	1e-2 to 1e-6
Epochs	50
Early Stopping Patience	3,5,7
Hidden Units	64,128,256,512,1024
Pre-trained Freezing	True, False
Hidden Activations	Relu, Sigmoid
Batch Size	8,16,64,128,256
Dropout	0.05,0.1,0.2,0.3

Table 9: Hyperparameter Values for Experiments

E Additional Performance Metrics of the Classification Models

Table 10 shows other measures of classification performance on the test set for all considered models.

F Change in Criticism of Popular News Sources

Based on the results from Figure 3, we also analyzed how the criticism shown towards these news sources changed through time. The resulting graph can be seen in Figure 7.
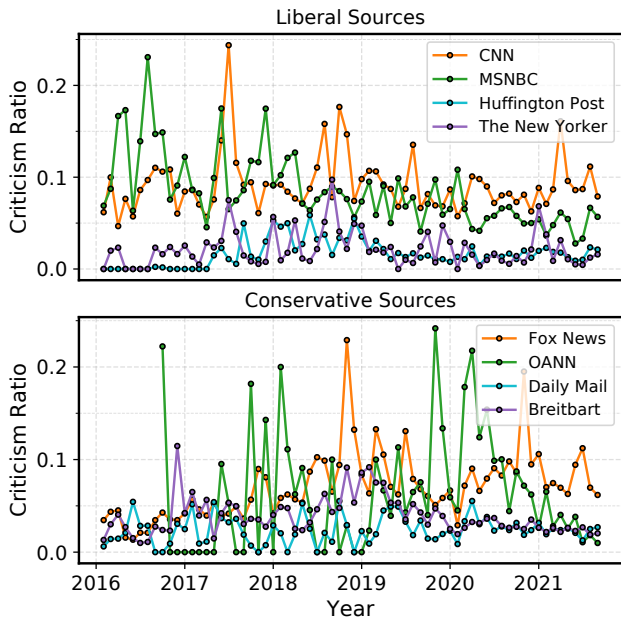

Figure 7: Criticism of Popular News Sources through Time

Network	Accuracy	F1	Precision	Recall
Text Network + ϕ_{up}	0.800 ± 0.049	0.800 ± 0.028	0.805 ± 0.023	0.801 ± 0.026
Text Network + ϕ_{tt}	0.723 ± 0.008	0.717 ± 0.007	0.738 ± 0.016	0.723 ± 0.009
Text Network + ϕ_{un}	0.813 ± 0.015	0.810 ± 0.017	0.825 ± 0.010	0.813 ± 0.015
Text Network + DS	$\mathbf{0.840 \pm 0.007}$	$\mathbf{0.840 \pm 0.007}$	$\mathbf{0.840 \pm 0.007}$	$\mathbf{0.840 \pm 0.007}$
Text Network + IBCC	0.824 ± 0.015	0.824 ± 0.015	0.828 ± 0.017	0.824 ± 0.015
Text Network + EBCC	0.837 ± 0.006	0.837 ± 0.006	0.839 ± 0.006	0.837 ± 0.006
Text Network + DP	0.812 ± 0.013	0.812 ± 0.013	0.812 ± 0.012	0.812 ± 0.013
User Network + ϕ_{up}	0.747 ± 0.049	0.747 ± 0.050	0.751 ± 0.047	0.747 ± 0.050
User Network + ϕ_{tt}	0.662 ± 0.018	0.643 ± 0.033	0.699 ± 0.022	0.662 ± 0.018
User Network + ϕ_{un}	0.746 ± 0.025	0.745 ± 0.025	0.747 ± 0.026	0.746 ± 0.025
Combined Network + ϕ_{up}	0.812 ± 0.014	0.811 ± 0.015	0.816 ± 0.013	0.812 ± 0.014
Combined Network + ϕ_{tt}	0.728 ± 0.006	0.719 ± 0.009	0.755 ± 0.019	0.728 ± 0.006
Combined Network + ϕ_{un}	0.799 ± 0.007	0.797 ± 0.007	0.808 ± 0.008	0.799 ± 0.007
Combined Network + DS	0.816 ± 0.014	0.818 ± 0.015	0.816 ± 0.014	0.816 ± 0.014
Combined Network + IBCC	0.785 ± 0.035	0.783 ± 0.037	0.792 ± 0.032	0.785 ± 0.036
Combined Network + EBCC	0.810 ± 0.023	0.810 ± 0.023	0.810 ± 0.023	0.810 ± 0.022
Combined Network + DP	0.826 ± 0.014	0.826 ± 0.014	0.826 ± 0.015	0.826 ± 0.014

Table 10: Test set Performance for combinations of model, labeling function, and label denoising methods.