Forecasting Political News Engagement on Social Media

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

Understanding how political news consumption changes over time can provide insights into issues such as hyperpartisanship, filter bubbles, and misinformation. To investigate long-term trends of news consumption, we curate a collection of over 60M tweets from politically engaged users over seven years, annotating ∼10% with mentions of news outlets and their political leaning. We then train a neural network to forecast the political lean of news articles Twitter users will engage with, considering both past news engagements as well as tweet content. Using the learned representation of this model, we cluster users to discover salient patterns of long-term news engagement. Our findings include the following: (1) hyperpartisan users are more engaged with news; (2) right-leaning users engage with contra-partisan sources more than left-leaning users; (3) topics such as immigration, COVID-19, Islamaphobia, and gun control are salient indicators of engagement with low quality news sources.

1   Introduction

The transformation of the news ecosystem from print to online media has fundamentally changed how people read about and engage with current events. While this media decentralization has undoubtedly increased one's access to diverse and timely information, it has also led to hyperpartisanship, polarization, and misinformation, all of which are fostered by computer-mediated communication. Recent work has investigated socio-technical issues such as the causes of polarization and "filter bubbles" (Bakshy, Messing, and Adamic 2015; Bail et al. 2018; DellaPosta 2020; Liu et al. 2021; Shivaram et al. 2022), the influence of hyperpartisan media (King, Schneer, and White 2017; Guess et al. 2021), and the factors that contribute to the spread of misinformation (Lobato et al. 2020; Osmundsen et al. 2021a). It has also led to research focusing on identifying interventions that can help foster healthier news engagement habits (Bhargava et al. 2019; Lutzke et al. 2019; Masrour et al. 2020).

Among such scholarship, however, studies have typically been either small controlled trials or observational studies over very short time periods. Those that consider long-term trends are primarily focusing on population-level changes over time. What is lacking is a long-term perspective to detail the evolution of an individual's news consumption and sharing habits. In response, we collect over 60M tweets from politically engaged Twitter users over a seven year period, and we annotate ∼10% with mentions of news outlets and the political lean of those outlets. With this data, we investigate how the types of news sources a user engages with change over long time periods. Our primary contributions are as follows:

- **Novel Dataset:** We curate a new dataset to foster research into long-term news engagement behavior, consisting of ∼6.5M news engagement tweets over seven years. To preserve anonymity, the dataset consists of records of the mentioned news source, partisan lean, day, and anonymized user IDs.[1]

- **Forecasting Model:** We develop LSTM models to forecast news engagement behavior over three-month windows, predicting the number of news mentions from each of seven political stance categories (extreme liberal to extreme conservative). The best performing model has a mean absolute error of 3.7 engagements (out of an average of ∼39 news engagements per window).

- **Discovery of Long-term Patterns:** The intermediate states of the LSTM offer a compact representation of the long-term behavior of each user. We cluster users to discover salient patterns, finding that (1) hyperpartisan users are more engaged with the news; (2) right-leaning users engage with contra-partisan sources more than left-leaning users do; (3) topics such as immigration, COVID-19, Islamaphobia, and gun control are leading indicators of engagement with low quality news sources.

2   Related Work

One area of relevant research considers the factors contributing to engagement with hyperpartisan news, misinformation, and conspiracy theories. In some cases, the content is highlighted, such as its topic and tone (Wischnewski, Bruns, and Keller 2021), as well as how such content aligns with the political views of the user. For example, Osmundsen et al. (2021a) show that hostility towards political opponents drives misinformation sharing, Rathje, Van Bavel, and

---

[1]Code and data are available at: https://github.com/tapilab/icwsm-2024-news-forecasting

van der Linden (2021) find that such out-group language strongly predicts social media engagement, and Brady et al. (2017) and Valenzuela, Piña, and Ramírez (2017) find that moral-emotional language in political messages increases their diffusion. Other related research focuses on the personality traits of the user; e.g., Lobato et al. (2020) show that individuals with personality traits high in traditionalism and low in social dominance were more willing to share misinformation about COVID-19, and a meta-analysis of COVID-19 misinformation by van Mulukom et al. (2020) finds that biases, group identity, and distrust in institutions contribute to misinformation sharing. There are also investigations into how hyperpartisan news and misinformation spreads online (Haber et al. 2021; Introne et al. 2020), as well as possible interventions for minimizing users' engagement with such content (Bak-Coleman et al. 2022; Bhuiyan et al. 2021; Masrour et al. 2020; Pennycook et al. 2021; Aslett et al. 2022; Nyhan 2021). Yet, beyond an examination of how users navigate YouTube to access more extreme content (Ribeiro et al. 2020), there is little research on user-level long-term trends in news engagement.

A second line of research investigates the dynamics surrounding social media-based filter bubbles, polarization, and ideological segregation. Much has been written on these topics, and we highlight a selection of the most relevant findings, beginning with DellaPosta (2020), who shows that polarization arises not simply by hardening opinions on a handful of issues but rather by belief consolidation, in which partisan views on one topic spread to other topics. Robertson et al. (2021) provide evidence that filter bubbles are driven less by the idiosyncrasies of search engines and more by users self-selecting into "echo chambers." This is consistent with the finding of Bakshy, Messing, and Adamic (2015) that polarization on Facebook is mostly driven by homophily of user friendship networks. Relatedly, Bail et al. (2018) show that exposure to opposing views on social media can increase political polarization. Other work has performed linguistic analysis to track how issues are framed differently by news source and over time (Tsur, Calacci, and Lazer 2015; Liu and Huang 2022; Islam, Roy, and Goldwasser 2023).

Much of this prior work focuses on correlations among static user attributes and news sharing behavior, or it conducts randomized controlled trials that investigate a handful of interventions. Given the complex factors influencing news engagement over the short- and medium-terms, we call for new methods to analyze longitudinal, real-world data and explore patterns of news engagement behavior over time. We propose several neural network architectures and learning methods to more accurately predict future news sharing behavior of users as well as to identify latent representations characterizing a user's transitioning news engagement.

3   Data

Our dataset is based on tweets where users engage, over a period of several years, with news sources across different partisan and ideological dimensions. For the present study, we focus on "news engagement" events where a user either mentions the official Twitter handle of a news source or

Year	All Tweets	NE Tweets	NE %
2006	3	0	0.00
2007	424	0	0.00
2008	14,649	7	0.05
2009	92,788	189	0.20
2010	187,609	2398	1.28
2011	741,038	16,065	2.17
2012	1,569,018	49,765	3.17
2013	1,853,832	74,069	4.00
2014	2,172,500	120,568	5.55
2015	2,579,583	181,957	7.05
2016	3,502,755	350,700	10.01
2017	4,816,301	585,272	12.15
2018	5,744,817	684,334	11.91
2019	7,889,222	904,271	11.46
2020	16,247,000	1,843,002	11.34
2021	16,021,606	1,654,322	10.33
Total	**63,433,145**	**6,466,919**	**10.20**

Table 1: Tweets collected by year (NE = news engagement).

shares a URL to an article from that source. We first identify a set of English language-based news sources covering the partisan landscape. We then assign political ideology ratings to different news sources based on *allsides.com*, a media bias rating site, where the ideology spectrum $\{-2,-1,0,+1,+2\}$ represents -2 for extreme liberals and +2 for extreme conservatives. We utilize ratings for 419 different news sources; yet, to account for a more diverse range of media quality, we extend the sample to include 103 low-reliability sources collected from Osmundsen et al. (2021b),[2] which were originally identified by Guess et al. (2019) and Grinberg et al. (2019). These low-reliability sources are rated as being either *pro-Republican* or *pro-Democrat*, and we assign the partisan lean of these sources as -3 (pro-Democrat) and +3 (pro-Republican). The final dataset is thus comprised of 522 news sources that have valid Twitter handles and URLs.

We next use the Twitter Streaming API to identify users who engage with each of the above mentioned 522 news sources. We submitted queries to the API in Fall 2021 to identify mentions of each news source, resulting in the identification of 1.67 million users. To account for the diversity of all Twitter users (i.e., users who do not engage with any of the identified news sources), we also collected data for 59K random users using the Twitter API over the same period. We limited the presence of bots by filtering out users that appear to exhibit automated behavior based on frequency of tweets, number of followers, and number of friends (c.f., Appendix). Based on the news sources with which they engage, we then sampled ∼1,200 users by partisan stance, evenly distributed across news sources per stance. We combined this with 1,200 random users sampled from the Twitter API, resulting in an initial set of 9,781 users.

We collected each sampled user's entire Twitter timeline,

| | **News Engagement Tweets** | |
Stance	Number	Percentage
Unreliable Liberal (-3)	112,560	1.74
Extreme Liberal (-2)	1,141,939	17.66
Liberal (-1)	1,977,177	30.57
Non-partisan (0)	1,240,848	19.19
Conservative (1)	569,489	8.81
Extreme Conservative (2)	1,256,521	19.43
Unreliable Conservative (3)	168,385	2.60
Total	**6,466,919**	**100**

Table 2: News engagement tweets by partisan stance.

amounting to a total of 63.5 million tweets. The annual tweet distribution for news engagement tweets is shown in Table 1, where "NE Tweets" refer to those instances when a user engages with a given news source, encompassing 10.2% of the tweets collected. The distribution of tweets by partisan stance can be seen in Table 2, where the most commonly engaged news sources are moderate liberal/Democrat (-1), followed by moderate/non-partisan (0) and strongly conservative/Republican (+2).

Table 1 shows that the majority of tweets in our dataset were posted after 2014, and we thus analyze tweets made after January 1, 2015. To focus on users with a significant history of news engagement, we exclude user accounts that do not engage with news sources at least 50 times since 2015. We also exclude users whose news engagement is not distributed across at least a three-year period. Based on these filtering criteria, 3,806 users were removed from our sample. We also filter out 137 potentially automated accounts in our sample by removing users whose total news engagement volume is three standard deviations greater than the mean news engagement volume (i.e., 3-sigma rule) (Hasan, Orgun, and Schwitter 2018). In sum, our final dataset is comprised of 5,838 users who generate 44.2M total tweets, of which 4.6M are news engagements.

4 Problem Statement

We forecast a user's future news engagement behavior based on their prior activity. To formulate this task, for user u_i in time window t_j, let $y_{t_j}^i \in \mathbb{N}^{1 \times 7}$ be a vector representing the number of news engagements with each of the seven partisan stances $\{-3, -2, -1, 0, 1, 2, 3\}$. As described earlier, a news engagement occurs when a user either mentions the Twitter handle or shares a URL to an article from a news source.

As an input to the model, let $m_{t_j}^i$ represent attributes of all tweets (text, mentions, hashtags, etc.) of user u_i across t_j. Given the observed historic news engagement count vectors $\{y_{t_1}^i, y_{t_2}^i, ..., y_{t_n}^i\}$ from time window t_1 to t_n and tweet attributes $\{m_{t_1}^i, m_{t_2}^i, ..., m_{t_n}^i\}$, our goal is to estimate the probability distribution over the engagement counts of a future time step t_{n+1} for user u_i.

$$P(y_{t_{n+1}}^i | y_{t_1}^i ... y_{t_n}^i, m_{t_1}^i ... m_{t_n}^i) \qquad (1)$$

In other words, given historic tweet attributes and engagement counts, we intend to predict the number of times a

Dataset	Train	Val	Test
D_1	2015-2017	2015-2017	2018
D_2	2016-2018	2016-2018	2019
D_3	2017-2019	2017-2019	2020
D_4	2018-2020	2018-2020	2021

Table 3: Time ranges for each train/test split of the forecasting task.

Figure 1: Train-test splits for D_1 (2015-2018), showing the time ranges used for each instance. Analogous time ranges are used for D_2-D_4.

given user engages with news sources across all stances (-3 to $+3$) in subsequent time steps.

5 Methods

We discuss in this section our data pre-processing techniques, a number of forecasting models employed via multiple neural network architectures, and our baseline method. We also describe the different features that we employ for training our models.

5.1 Pre-processing

We set $n = 8$ (the number of input time steps), where each time step t_j spans three months. Thus, given a user's activity over the previous two years, we predict the next three months of news engagements. To create multiple train/test folds, we split our overall dataset (D) into four folds $\{D_1, D_2, D_3, D_4\}$ shown in Table 3, each of which spans three years. Figure 1 illustrates the input and output for the training and testing instances of D_1. Thus, for each user of D_1, we generate four training instances and four testing instances. This is repeated for each fold D_1 through D_4.

5.2 Features

1. **News Engagement Counts** ($y_{t_j}^i$): For each input time step t_j and user u_i, we include as input features the lagged count vectors of the number of engagements for each partisan stance $p \in \{-3, -2, -1, 0, 1, 2, 3\}$. We denote this $y_{t_j}^i$, where $y_{t_j}^i \in \mathbb{R}^{1 \times 7}$. We also standardize these count values using z-score standardization, where the means and standard deviations are calculated over the input time steps (i.e., t_1 to t_n).

2. **Tweet Text** ($v_{t_j}^i$): We also include features over the tweet text. For each three month time step, we select the 25

most recent tweets for both the news engagement tweets and the non-news engagement tweets. We then pass these two sets of tweets through TwHIN-Bert (Zhang et al. 2022), a transformer based language model fine-tuned on Twitter data, and extract embedding representations for each token of each tweet. We perform two levels of aggregation over these token representations: (1) for each tweet, we concatenate the CLS token and average embedding of the other non-CLS tokens of the tweet; (2) we take an average over these 25 tweet representations. This results in two 1,536-dimension vector representations, one for the engagement tweets (eng_{t_j}) and one for the non-engagement tweets ($neng_{t_j}$). We then concatenate these two embedding vectors to get a single text representation $v_{t_j}^i$, where $v_{t_j}^i \in \mathbb{R}^{1 \times 3072}$.

3. **Hashtags** ($\#_{t_n}^i$): For the last input time step t_n, we select the top 100 most frequently used hashtags for that specific three-month window. We then pass these hashtags through the language model (TwHIN-Bert) and perform the same aggregation step that we used for our text-based features, obtaining a final vector representation $\#_{t_n}^i$ where $\#_{t_n}^i \in \mathbb{R}^{1 \times 1536}$.

4. **Input Quarter Encoding** ($q_{t_j}^i$): To capture seasonal effects, for each observation sequence we encode the year-quarter of each input time step (Q1 - Q4) as a one-hot encoded vector $q_{t_j}^i$, where $q_{t_j}^i \in \mathbb{R}^{1 \times 4}$.

5. **Forecast Quarter Encoding**: Similarly, for each observation sequence, we encode the year-quarter of the time-step we are forecasting as a one-hot encoding vector.

5.3 Baseline

For comparison, we include a simple approach that sets the prediction to be equal to the values for the final input time step (i.e., $\hat{y}_{t_{n+1}}^i = y_{t_n}^i$).

5.4 Single Feature Network (SFN)

We use a bi-directional LSTM (Bi-LSTM) (Schuster and Paliwal 1997) as our primary forecasting model. We select this model in part due to its application in other types of user activity forecasting – e.g., Yang et al. (2018) fit an LSTM to predict churn rate for a social media application, and Liu et al. (2019) used GNNs to forecast future user engagement on Snapchat. Similar models have performed best on time series benchmark competitions (Oreshkin et al. 2019).[3]

The Single Feature Network (SFN) uses only one type of feature — either text based representation sequences $\{v_{t_1}^i, v_{t_2}^i, ..., v_{t_n}^i\}$ (**SFN+T**) or news engagement count based sequences $\{y_{t_1}^i, y_{t_2}^i, ..., y_{t_n}^i\}$ (**SFN+C**). When using the text based representations we add a linear layer before passing the input sequences into the Bi-LSTM. After passing the input sequences through our Bi-LSTM model, we extract the final hidden states for both the forward and backward layers ($\overrightarrow{h_{t_n}}, \overleftarrow{h_{t_n}}$) and concatenate them to obtain a single

hidden state representation h_{t_n}. This is then passed through a final output layer $<W_{out}, b_{out}>$ to predict the future news engagement count vector $\hat{y}_{t_{n+1}}^i$ for time step t_{n+1}, as shown in Equation 2.

$$\hat{y}_{t_{n+1}}^i = (W_{out} h_{t_n} + b_{out}) \qquad (2)$$

5.5 Multiple Feature Network (MFN)

The Multiple Feature Network (MFN) combines multiple features (tweet texts, news engagement counts, and input quarter encodings) to forecast future news engagement counts. This model's architecture resembles the single feature network (SFN) with a few modifications. Once we extract the final hidden state representation h_{t_n} from our Bi-LSTM layers (discussed above), we concatenate the hashtag representation $\#_{t_n}^i$ of the final input time step and the output quarter encoding $q_{t_{n+1}}^i$ to this hidden state representation h_{t_n}. This is then passed through an intermediate layer $<W_{inter}, b_{inter}>$, the output of which, g_{inter}^i, is then passed through our final output layer $<W_{out}, b_{out}>$ to predict the future news engagement count vector $\hat{y}_{t_{n+1}}^i$ for time step t_{n+1}, as shown in equation 4.

$$g_{inter}^i = (W_{inter}[h_{t_n}, \#_{t_n}^i, q_{t_{n+1}}] + b_{inter}) \qquad (3)$$

$$\hat{y}_{t_{n+1}}^i = (W_{out} g_{inter}^i + b_{out}) \qquad (4)$$

All network models are trained to minimize Mean Absolute Error (MAE) loss.[4] As we predict a vector of engagement counts, the overall MAE loss is a sum across individual MAE losses for each news engagement stance:

$$\text{Total MAE Loss} = \sum_{r=1}^{|p|} \text{MAE}(y_{t_{n+1}}[r], \hat{y}_{t_{n+1}}[r]) \qquad (5)$$

6 Experimental Settings

To perform our forecasting experiments, we first construct our train, test, and validation sets for each fold D_1 through D_4 (Table 3). For validation, we hold out 20% of the users in the training set. Once we have our train, validation, and test sets, we filter out sequences where a user has no news engagement activity over the entire input sequence. The resulting dataset sizes are shown in Table 4.

All experiments use a system with 4 Nvidia A5000 GPUs, 512 GB RAM and an AMD Ryzen Threadripper 3975WX CPU. We report means and standard deviations of scores across five random seeds. Table 9 in the Appendix lists all hyperparameters that were tuned on the validation data.

7 Results and Analysis

We now turn to the results of our experiments and conduct an extended analysis of long-term news engagement behavior of users on Twitter.

---

[3]While transformer-based models are a viable alternative, recent work has questioned the effectiveness of such models for time series forecasting (Zeng et al. 2023).

[4]Other loss functions such as Mean Squared Error (MSE), Mean Absolute Percentage Error (MAPE), and Huber Loss were also considered, but exploratory experiments suggested MAE was less sensitive to outliers.

Dataset	Train	Val	Test
D_1	15,708	3,904	21,881
D_2	17,536	4,345	22,327
D_3	17,862	4,465	22,648
D_4	18,106	4,542	17,378

Table 4: Train, validation, and test set sizes. Each instance is a single user's timeline, with two years of input observations used to forecast the next three months of news engagements.

Model	MAE	MSE
Baseline	3.89	216.33
SFN+C	**3.73**	**207.89**
SFN+T	4.22	258.08
MFN	3.85	220.55

Table 5: Mean absolute error and mean-squared error averaged over all test folds, showing SFN+C as the most accurate model.

7.1 Model Accuracy

Table 5 shows the average error rates of the models across all test datasets, D_1 through D_4, averaged across stance and dataset. We find that SFN+C and MFN perform first- and second-best, respectively, both outperforming the baseline model. Even though the SFN+T model performs the worst, we are surprised to discover that it performs relatively well for a model using text features without news engagement counts. To assess these differences in performance, we conduct paired t-tests over the combined results and find that, with the exception of the Baseline-MFN comparison, all pairwise comparisons yield significantly different average error rates ($p < 0.01$).

Table 6 breaks down error rates by fold and partisan stance. We can see that SFN+C has the lowest error for most cases. Exceptions include -3/+3, where the LSTM models generally do not outperform the baseline. We attribute this in part to class imbalance — the overall infrequency of -3/+3 stances (c.f., Table 2) lead to under prediction of engagement.

Figure 2 plots test error by quarter. We observe that error magnitude generally increases over time due largely to the overall increase in engagement volume. The highest increase in errors is measured for 2020, possibly resulting from a number of events that increased news engagement, namely, COVID-19 and Trump's impeachment. Over time, the SFN+C model performs best, with considerable improvements over other models from 2020-Q4 to 2021-Q2.

7.2 User-level Error Analysis

To gauge model performance for individual users, we plot the true and predicted counts for a sample of users in Figure 3. These users were selected because they exhibited the highest error rates according to the SFN+C model. Each subplot in Figure 3 represents a different user-stance combination. For most stances, we observe that our models perform well when accounting for the overall trend of engagement counts. The main challenge for these models are sudden
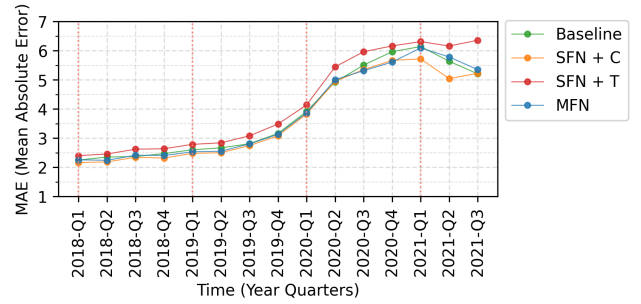

Figure 2: Mean absolute error by quarter. The rising error is in part due to increased user activity since 2020.

spikes in engagement. Yet, the SFN+C model seems to handle engagement spikes better than the SFN+T model, likely due to the fact that text based features do not capture the intensities of the news engagement for these users. We also observe that neither the SFN+C nor SFN+T models perform very well for the -3 category, which could result from the low volume of these types of engagements in our sample.

7.3 Predicting Sudden Shifts in News Engagement

Sudden shifts in engagement may be difficult to predict, but they are of practical interest given that they represent a user's news engagement transition point. To examine this phenomenon more closely, we compare model performance across a range of transition points. To identify transitions where there is a shift in engagement between the input time steps (i.e., t_1 to t_n) and the forecast window (t_{n+1}), we rank samples by measuring the cosine distance between the engagement count vector of the last time-step of the input sequence ($y_{t_n}^i$) and the count vector to be predicted ($\hat{y}_{t_{n+1}}^i$). We next compute error metrics at all ranks of these transition points, ordered from most to least transition. Figure 4 plots the results for test set D_1 across the baseline and the SFN+C model, showing that our SFN+C model performs better than the baseline at all rankings. Further, the biggest improvements over the baseline occur at instances of greatest transition, a finding that we also discover for other test sets and models. Compared to the baseline, the proposed models are more helpful when forecasting users' sudden engagement shifts.

A subset of transition points are sequences where a user shifts from some news engagement to no news engagement at all or vice versa. These shifts likely suggest that the user is becoming engaged or disengaged with news, which we expect to occur during notable events (e.g., elections and political controversies). In these instances, some users become swept up in the issues, while others avoid the flood of information. We measure the performance of these methods by first selecting all samples where there is some news engagement activity in a particular input step but no news engagement in the output step, or where there is no news engagement in the input step but news engagement in the output step. The results in Table 7 show improvements over the

Dataset	Model	-3	-2	-1	0	1	2	3
D_1	Baseline	**.20**	3.14	5.09	3.21	1.32	3.08	**.49**
	SFN+C	.21±.005	**3.00±.004**	**4.80±.012**	**2.98±.008**	**1.29±.007**	**2.93±.003**	.54±.012
	SFN+T	.27±.004	3.29±.031	5.40±.045	3.36±.043	1.40±.014	3.30±.053	.57±.021
	MFN	.27±.001	3.08±.029	4.97±.116	3.09±.076	1.34±.025	3.05±.046	.57±.028
D_2	Baseline	**.36**	3.70	6.24	3.80	2.01	3.10	.45
	SFN+C	.39±.042	**3.46±.112**	**5.88±.335**	**3.55±.132**	**1.90±.155**	**2.99±.082**	**.43±.008**
	SFN+T	.40±.001	3.91±.015	6.78±.025	3.98±.007	2.18±.011	3.50±.049	.52±.016
	MFN	.41±.0	3.55±.018	6.10±.024	3.64±.010	1.99±.007	3.07±.034	.44±.005
D_3	Baseline	**.62**	6.39	10.84	6.98	3.57	**6.51**	**.68**
	SFN+C	.64±.007	**6.01±.029**	**10.46±074**	6.74±.053	**3.49±.014**	**6.51±.065**	.70±.002
	SFN+T	.69±.001	6.55±.019	11.40±.035	7.27±.023	3.82±.010	7.34±.114	.77±.003
	MFN	.69±.001	6.06±.020	10.53±.111	**6.73±032**	3.49±021	6.58±.058	.73±.004
D_4	Baseline	**1.39**	6.07	9.97	7.63	4.92	8.40	**1.23**
	SFN+C	1.44±.017	**5.73±.099**	**9.77±.317**	**7.32±.043**	4.75±.040	**8.31±.137**	1.35±.026
	SFN+T	1.50±.005	6.67±.073	11.69±.473	8.33±.147	5.04±.037	9.02±.157	1.40±.003
	MFN	1.52±.002	6.04±.097	10.51±.185	7.89±.259	**4.73±.041**	8.57±.152	1.41±.011

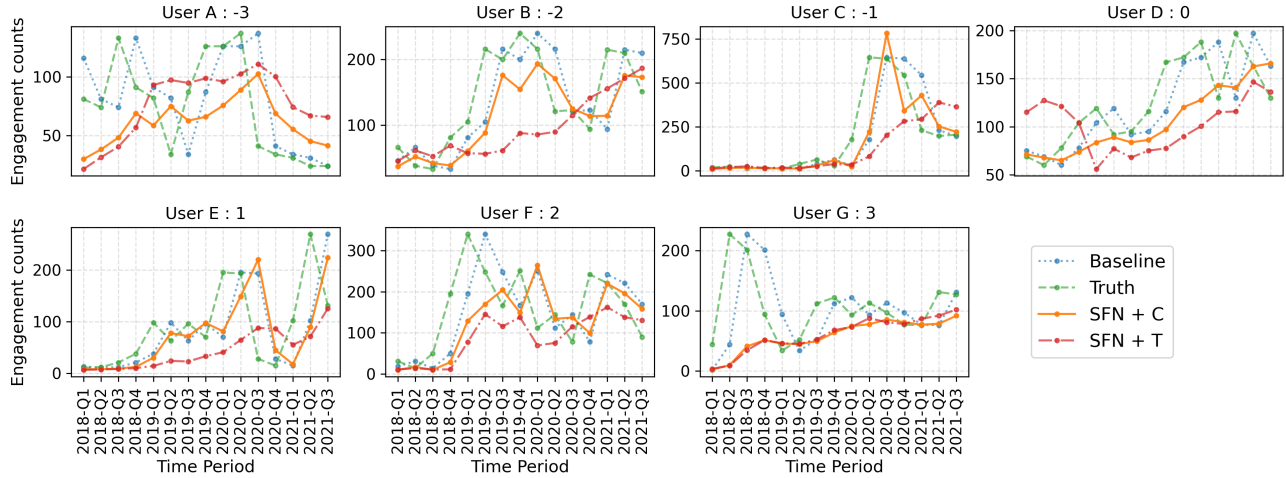Table 6: Mean absolute error by partisan stance (smallest errors in bold).

Figure 3: True vs. predicted values for users with the highest errors, according to the SFN+C model.

baseline for nearly all datasets. Notably, the SFN+T model outperforms both the SFN+C model and the baseline for D_3 and D_4 when predicting a no engagement-to-engagement shift. We surmise that text features provide a key signal for users' increasing engagement with news topics.

7.4 Text Analysis

To examine text features in more detail, we identify terms that are leading indicators of engagement with unreliable news sources (-3 or +3). Identifying salient terms from LSTMs is challenging, and we perform a number of steps to identify and analyze terms that are associated with un-realiable news. First, we rank all instances by the model's forecasted engagement counts for news sources with -3 and +3 news content. We then select the top 300 and bottom 300 instances, i.e., those times when users are forecast to most-engage and least-engage with unreliable content. To identify the terms that distinguish these two sets of users, for each user, we concatenate the terms from all of the tweets in the input window and perform a chi-square analysis to select the most distinguishing terms for each user group. The most representative terms used in the context of unreliable liberal (-3) and unreliable conservative (+3) news sources are presented in Tables 8a and 8b, respectively.

We observe that there are differences in terms of focus and referencing method. For example, content that precedes engagement with unreliable liberal sources highlights COVID-19, specific politicians, specific policies (e.g., health policy, tax policy, gun policy), Trump's impeachment, the distinction between truth and lies, the private sector, racism, Russia and Putin, and Muslims. For conservative sources, the content focuses extensively on specific groups with a potential role to play within American political institutions, including globalists, Marxists, Antifa, and communists. Beyond those groups, there is also reference to groups that could be based within or outside the United States, including illegal immigrants, Hezbollah, Islamists, traffickers (drug and migrant), the drug cartel, and individuals affiliated with Hezbollah,
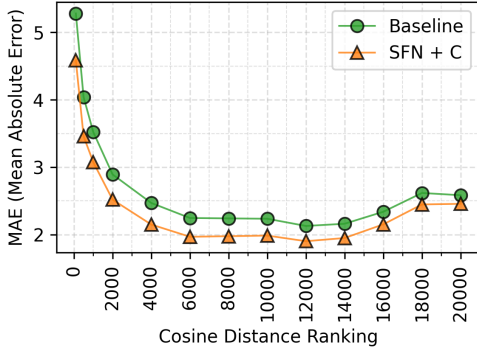
Figure 4: Error by transition points based on cosine distance ranking (lower ranking means more abrupt transitions).

Dataset	Models	$E \rightarrow N$ MAE	$N \rightarrow E$ MAE
D_1	Baseline	0.653	**0.818**
	SFN+C	0.558	0.819
	SFN+T	0.660	0.828
	MFN	**0.549**	0.842
D_2	Baseline	0.651	0.930
	SFN+C	**0.503**	**0.921**
	SFN+T	0.674	0.958
	MFN	0.542	0.938
D_3	Baseline	1.327	2.144
	SFN+C	**1.016**	2.121
	SFN+T	1.267	**2.120**
	MFN	1.029	2.139
D_4	Baseline	1.307	2.535
	SFN+C	**1.268**	2.496
	SFN+T	1.566	2.485
	MFN	1.561	**2.474**

Table 7: Mean absolute error for instances where users shift from Engagement to No Engagement ($E \rightarrow N$) and from No Engagement to Engagement ($N \rightarrow E$).

Venezuela, Iranians, Libya, and jihad.

7.5 Temporal Cluster Analysis

Turning now to the identification of salient patterns of news engagement over time, rather than simply analyzing partisan scores over the 2018-2021 period, we integrate tweet content to discover more nuanced engagement patterns. We build upon prior work that uses the latent representation learned by a neural network as the input to a clustering algorithm (Huang et al. 2014; Xu et al. 2015). To do so, we use the MFN model to represent each user. Specifically, for each prediction window for user i, we compute the final network layer g^i_{inter} (Equation 4), which is the model's representation of all of the content and engagement features from the prior two years. Regarding the user's representation over time, we concatenate the g^i_{inter} vector for each of the 16 prediction windows from 2018-2021 (four per year), resulting in a single vector for each user.

policy, senators, jail, director, test, muslim, lie, 2016, leaders, attack, technologies, services, dead, administration, coronavirus, won, senator, centre, sector, role, democracy, biden, donald, russia, billion, justice, impeached, obama, fake, general, tax, limited, army, journey, yesterday, fucking, various, truth, rights, sir, companies, bank, votes, democratic, biggest, officials, address, protect, intelligence, hospital, asking, deaths, congratulations, startups, impeachment, political, cristiano, nra, military, supporters, senate, voters, massive, healthcare, january, total, guy, trump, chinese, growth, russian, cases, digital, racist, technology, court, muslims, putin, america, sen, hell, #iot, successful, election, fox, investigation, lying, press, graham, rep, ronaldo, texas, republican, leader, ahead, gop, congress

(a)

globalist, lefts, marxist, illegals, islam, hamas, marxists, leftists, soros, msm, censoring, veritas, sharia, globalists, #antifa, ccp, farleft, aborted, cartels, noncitizens, hezbollah, venezuela, islamist, declassified, amnesty, traffickers, fisa, iranians, cartel, deletes, levin, spied, rino, indoctrination, libya, marxism, jihad, islamic, smuggling, passports, declassify, bombshell, leftwing, lid, russiagate, durham, spying, #walkaway, strzok, censors, dominion, leftist, erupts, biological, harvesting, unborn, communists, communism, bribes, behar, stabbed, omars, satanic, nadler, sanctuary, rashida, flashback, pcr, alqaeda, brotherhood, censorship, totalitarian, #foxnews, wuhan, assange, clapper, antitrump, accusers, informant, rinos, censored, maxine, bureaucrats, deception, explodes, lefties, destroys, cabal, irans, parenthood, accuser, infanticide, kerry, dossier, faucis, jabs, lawabiding, sharpton

(b)

Table 8: Terms that are leading indicators of engagement with unreliable (a) liberal and (b) conservative news sources.

Salient user clusters are determined by applying k-means clustering (K=20) to these user vectors.[5]

Figure 5 visualizes both the news engagement and terms used in each cluster. Each heat map displays the number of news engagements per partisan stance in each time window, averaged over the users in the cluster. These counts are discretized into five bins for visualization. The content above each cluster contains the cluster number, the percentage of users belonging to the cluster, the average stance of news engagements in that cluster, and the top-three terms most indicative of the cluster for each year. The terms are ranked by calculating chi-squared statistics for terms appearing in one cluster versus any other cluster. The clusters are presented in the order of their average stance, from most liberal to most conservative.

As an example, we observe that cluster #1 contains 4.3% of the users. The average partisan lean of news engagements in this cluster is -.87, which is the most liberal cluster identified. The heatmap indicates that users in this cluster primarily engage with partisan stances -2, -1, and 0. We can also see that users were somewhat engaged early

---

[5]We experimented with other values of K, which produced similar results. We settled on 20 given that it offered a good tradeoff between cluster cohesion and the number of clusters.
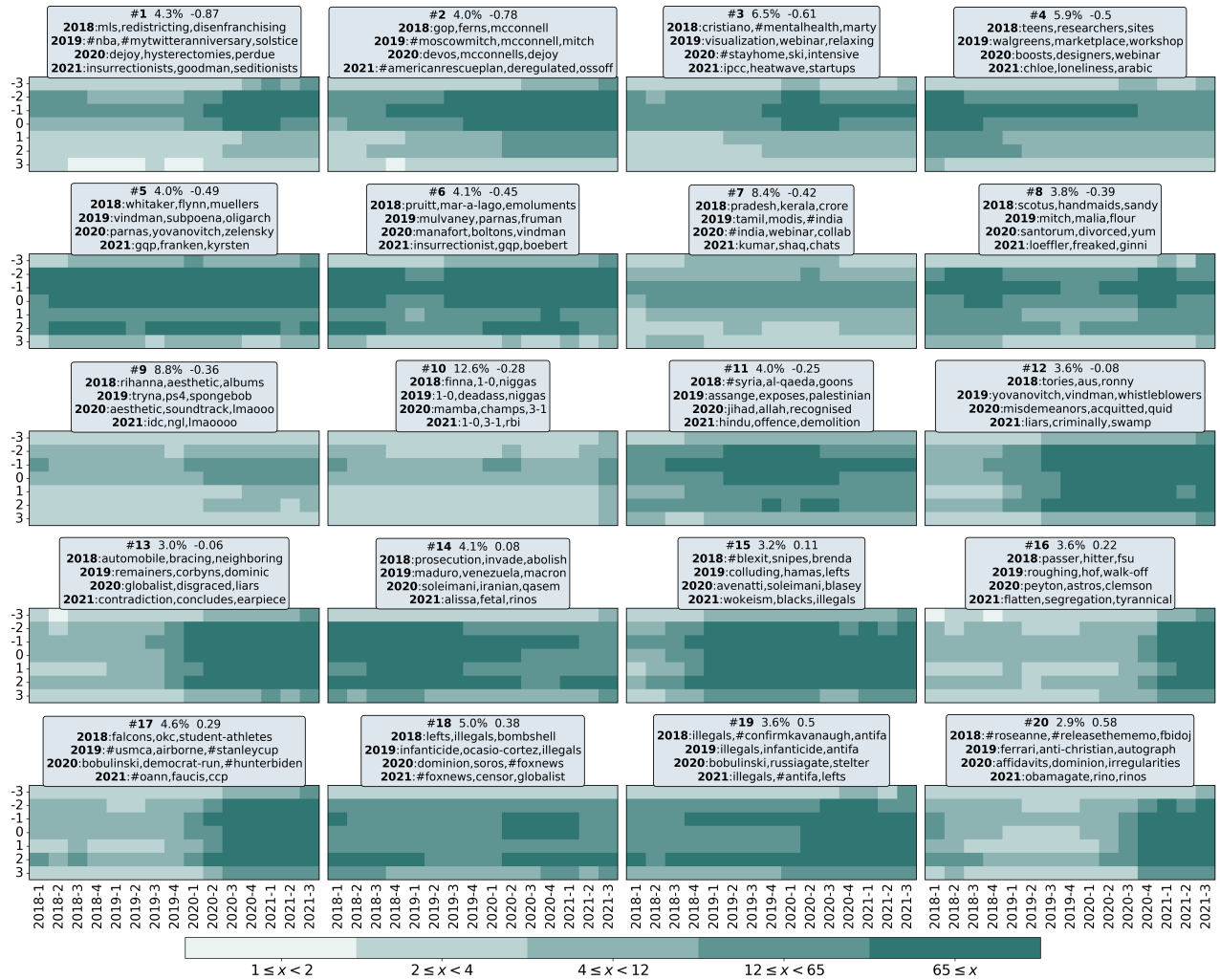
Figure 5: Visualization of 20 discovered clusters of users based on the learned representation of the forecasting model. Over time and partisan stance, each heatmap/cluster includes: percentage of all users within the cluster, the average partisan lean of users in the cluster, and the most distinctive terms for each year.

in 2018, but then had less engagement until a spike beginning in late 2020. Looking at the most distinctive terms for each year, we note that much of 2018-2019 was focused on sports ("mls"=Major League Soccer, #nba=National Basketball Association), with few political term ("disenfranchising," "redistricting"). From 2020, however, the language became more focused on politics: "hysterectomies" refers to a news report of detained immigrants being pressured into gynecological procedures; "DeJoy" was the U.S. Postmaster General, who was involved in a controversy about changes in mail delivery that impacted mail-in voting, "Perdue" refers to a close Senate election in Georgia. In 2021, all three top words refer to the January 6 attack on the U.S. Capitol.

At the other end of the political spectrum, cluster #17 also contains users who were not at first very politically engaged.

Keywords focused initially on sports ("Falcons" (football) and #stanleycup (hockey)); yet, in 2020, users became engaged with topics such as Hunter Biden, son of future President Joe Biden, along with his business partner Tony Bobulinski, both of whom were accused of corruption by the Trump campaign. By 2021, these users were focused on the government's response to the pandemic (Fauci), the conservative news network OANN, and tensions between the U.S. and China ("ccp"=Chinese Communist Party).

By using the LSTM representation, we are also able to distinguish between clusters that have similar partisan engagement patterns but discuss different topics. For example, clusters #18 and #19 are both strongly conservative with some variation regarding when users first become politically engaged. While both clusters focus on immigration ("illegals") and abortion ("infanticide"), #18 is more focused on

George Soros, #19 is more focused on far-left groups (Antifa) and investigations into Biden ("bobulinski," "russiagate").

There are two intuitive findings from this visualization. First, users who engage with less hyperpartisan news also have lower engagement overall. For example, clusters #7, #9, #10 have the lowest engagement and also tend to engage mostly with -1 and 0 partisan stances. This is in line with prior work suggesting that the emotional language of hyperpartisan news encourages greater engagement online (Hasell 2021; Eady et al. 2021; Weismueller et al. 2022). Second, engagement with contra-partisan news varies across the political spectrum. For example, comparing liberal clusters #1-#6 with conservative clusters #14-#20, we observe that liberal clusters have substantially less engagement with conservative stances than conservative clusters have with liberal stances. Cluster #19, for example, engages heavily with both -1 and +3 sources, while cluster #3 engages little with conservative sources. This finding has additional implications given that Bail et al. (2018) found that exposure to opposing views on social media can increase political polarization. This is also indicative of animosity towards cross-partisan news and politicians, which has been implicated in misinformation sharing and filter bubbles (Osmundsen et al. 2021a; Rathje, Van Bavel, and van der Linden 2021).

While follow-up studies are required to more rigorously investigate these and other hypotheses, this analysis provides insights into the patterns and prevalence of long-term news engagement behaviors.

8 Discussion and Limitations

We have offered a methodology for studying how users engage with political news over long time periods. The results suggest that the future news engagement behaviors of users can be predicted reasonably well based on prior user behavior. However, sudden shifts in behavior are still difficult to predict, and these can often occur due to external political events (e.g., elections, protests, etc.). Clustering and visualizing users based on the latent representation learned by the forecasting model can serve as an exploratory data analysis method to motivate future work in this area.

8.1 Limitations

There are several important limitations to this work to consider. First, by design, our dataset focuses on users with high news engagement (although we also included a random sample of Twitter users). Of course, such users are quite different from the wider population, and so one should be cautious when generalizing our results beyond the sample. Second, we have not attempted to understand the *intent* behind each news engagement. As discussed in the previous section, many users engage with cross-partisan news sources to ridicule rather than support them. Distinguishing among these cases will be essential for future work on this topic, e.g., by using techniques from Shivaram et al. (2024). Third, it is quite possible that the text-based models exhibit varying error rates across demographic groups, though we have not empirically investigated this. As Table 6 shows, error rates

vary by partisan stance; hence, forecast quality could vary by a user's demographic characteristics just as it varies by a user's political preferences.

8.2 Ethics and Broader Impacts

- **Negative societal impacts:** While our goal is to understand extant user behavior, forecasting a user's future news engagements can be viewed as a type of user profiling. This may present societal risks if such a technology is used to, for example, censor users by political ideology or conduct targeted advertising to widen partisan divisions. Similarly, it could be used to target users susceptible to misinformation sharing.

- **Cost of misclassification:** There is currently limited cost to the errors of the proposed approach, but if such a model were to be incorporated into a process to restrict the spread of misinformation, then forecasting -3/+3 engagements erroneously could lead to over-aggressive content filtering.

- **Privacy and consent**: We have released an anonymized dataset where each record contains the news source, partisan lean, day, and a unique, anonymous user ID. To comply with terms of service, no raw tweets will be shared. All data is derived from publicly available sources and no interventions are performed. As such, the study was determined to be exempt by the institution's IRB committee. While we have done our best to anonymize the data, it is conceivable that one could guess some of the user names, by cross-referencing the sequence of news sources they engaged with over time.

- **FAIR:** Our data is released as a simple CSV file with full documentation. It is thus findable (via link from the paper on the AAAI/ICWSM website), accessible (CSV is an open format), interoperable (CSV can be opened by any system), re-usable (future research on news engagement is possible with the data).

References

Aslett, K.; Guess, A. M.; Bonneau, R.; Nagler, J.; and Tucker, J. A. 2022. News credibility labels have limited average effects on news diet quality and fail to reduce misperceptions. *Science Advances*, 8(18): eabl3844.

Bail, C. A.; Argyle, L. P.; Brown, T. W.; Bumpus, J. P.; Chen, H.; Hunzaker, M. F.; Lee, J.; Mann, M.; Merhout, F.; and Volfovsky, A. 2018. Exposure to opposing views on social media can increase political polarization. *Proceedings of the National Academy of Sciences*, 115(37): 9216–9221.

Bak-Coleman, J. B.; Kennedy, I.; Wack, M.; Beers, A.; Schafer, J. S.; Spiro, E. S.; Starbird, K.; and West, J. D. 2022. Combining interventions to reduce the spread of viral misinformation. *Nature Human Behaviour*, 1–9.

Bakshy, E.; Messing, S.; and Adamic, L. A. 2015. Exposure to ideologically diverse news and opinion on Facebook. *Science*, 348(6239): 1130–1132.

Bhargava, R.; Chung, A.; Gaikwad, N. S.; Hope, A.; Jen, D.; Rubinovitz, J.; Saldías-Fuentes, B.; and Zuckerman, E. 2019. Gobo: A System for Exploring User Control of Invisible Algorithms in Social Media. In *Conference Companion Publication of the 2019 on Computer Supported Cooperative Work and Social Computing*, 151–155.

Bhuiyan, M. M.; Horning, M.; Lee, S. W.; and Mitra, T. 2021. NudgeCred: Supporting News Credibility Assessment on Social Media Through Nudges. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2): 1–30.

Brady, W. J.; Wills, J. A.; Jost, J. T.; Tucker, J. A.; and Bavel, J. J. V. 2017. Emotion shapes the diffusion of moralized content in social networks. *Proceedings of the National Academy of Sciences*, 114(28): 7313–7318.

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.

DellaPosta, D. 2020. Pluralistic collapse: The "oil spill" model of mass opinion polarization. *American Sociological Review*, 85(3): 507–536.

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.

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*.

Guess, A. M.; Barberá, P.; Munzert, S.; and Yang, J. 2021. The consequences of online partisan media. *Proceedings of the National Academy of Sciences*, 118(14).

Haber, J.; Singh, L.; Budak, C.; Pasek, J.; Balan, M.; Callahan, R.; Churchill, R.; Herren, B.; and Kawintiranon, K. 2021. Research note: Lies and presidential debates: How political misinformation spread across media streams during the 2020 election. *Harvard Kennedy School Misinformation Review*.

Hasan, M.; Orgun, M. A.; and Schwitter, R. 2018. A survey on real-time event detection from the Twitter data stream. *Journal of Information Science*, 44(4): 443–463.

Hasell, A. 2021. Shared emotion: The social amplification of partisan news on Twitter. *Digital Journalism*, 9(8): 1085–1102.

Huang, P.; Huang, Y.; Wang, W.; and Wang, L. 2014. Deep embedding network for clustering. In *2014 22nd International conference on pattern recognition*, 1532–1537. IEEE.

Introne, J.; Korsunska, A.; Krsova, L.; and Zhang, Z. 2020. Mapping the narrative ecosystem of conspiracy theories in online anti-vaccination discussions. In *International Conference on Social Media and Society*, 184–192.

Islam, T.; Roy, S.; and Goldwasser, D. 2023. Weakly Supervised Learning for Analyzing Political Campaigns on Facebook. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 17, 411–422.

King, G.; Schneer, B.; and White, A. 2017. How the news media activate public expression and influence national agendas. *Science*, 358(6364): 776–780.

Liu, M.; and Huang, J. 2022. "Climate change" vs. "global warming": A corpus-assisted discourse analysis of two popular terms in The New York Times. *Journal of World Languages*, 8(1): 34–55.

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.

Liu, Y.; Shi, X.; Pierce, L.; and Ren, X. 2019. Characterizing and forecasting user engagement with in-app action graph: A case study of snapchat. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2023–2031.

Lobato, E. J.; Powell, M.; Padilla, L. M.; and Holbrook, C. 2020. Factors predicting willingness to share COVID-19 misinformation. *Frontiers in psychology*, 11: 2413.

Lutzke, L.; Drummond, C.; Slovic, P.; and Árvai, J. 2019. Priming critical thinking: Simple interventions limit the influence of fake news about climate change on Facebook. *Global Environmental Change*, 58: 101964.

Masrour, F.; Wilson, T.; Yan, H.; Tan, P.-N.; and Esfahanian, A. 2020. Bursting the Filter Bubble: Fairness-Aware Network Link Prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 841–848.

Nyhan, B. 2021. Why the backfire effect does not explain the durability of political misperceptions. *Proceedings of the National Academy of Sciences*, 118(15).

Oreshkin, B. N.; Carpov, D.; Chapados, N.; and Bengio, Y. 2019. N-BEATS: Neural basis expansion analysis for interpretable time series forecasting. In *International Conference on Learning Representations*.

Osmundsen, M.; Bor, A.; Vahlstrup, P. B.; Bechmann, A.; and Petersen, M. B. 2021a. Partisan Polarization Is the Primary Psychological Motivation behind Political Fake News Sharing on Twitter. *American Political Science Review*, 115(3): 999–1015.

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

Pennycook, G.; Epstein, Z.; Mosleh, M.; Arechar, A. A.; Eckles, D.; and Rand, D. G. 2021. Shifting attention to accuracy can reduce misinformation online. *Nature*, 592(7855): 590–595.

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).

Ribeiro, M. H.; Ottoni, R.; West, R.; Almeida, V. A.; and Meira Jr, W. 2020. Auditing radicalization pathways on YouTube. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 131–141.

Robertson, R. E.; Green, J.; Ruck, D.; Ognyanova, K.; Wilson, C.; and Lazer, D. 2021. Engagement Outweighs Exposure to Partisan and Unreliable News within Google Search. *arXiv preprint arXiv:2201.00074*.

Schuster, M.; and Paliwal, K. K. 1997. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, 45(11): 2673–2681.

Shivaram, K.; Bilgic, M.; Shapiro, M. A.; and Culotta, A. 2024. Characterizing Online Criticism of Partisan News Media Using Weakly Supervised Learning. In *Proceedings of the International Conference on Web and Social Media*.

Shivaram, K.; Liu, P.; Shapiro, M.; Bilgic, M.; and Culotta, A. 2022. Reducing Cross-Topic Political Homogenization in Content-Based News Recommendation. In *Proceedings of the 16th ACM Conference on Recommender Systems*, 220–228.

Tsur, O.; Calacci, D.; and Lazer, D. 2015. A frame of mind: Using statistical models for detection of framing and agenda setting campaigns. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 1629–1638.

Valenzuela, S.; Piña, M.; and Ramírez, J. 2017. Behavioral Effects of Framing on Social Media Users: How Conflict, Economic, Human Interest, and Morality Frames Drive News Sharing. *Journal of Communication*, 67(5): 803–826.

van Mulukom, V.; Pummerer, L.; Alper, S.; Cavojova, V.; Farias, J. E. M.; Kay, C. S.; Lazarevic, L.; Lobato, E. J. C.; Marinthe, G.; Banai, I. P.; Šrol, J.; and Žeželj, I. 2020. Antecedents and consequences of COVID-19 conspiracy theories: a rapid review of the evidence. *PsyArXiv*.

Weismueller, J.; Harrigan, P.; Coussement, K.; and Tessitore, T. 2022. What makes people share political content on social media? The role of emotion, authority and ideology. *Computers in Human Behavior*, 129: 107150.

Wischnewski, M.; Bruns, A.; and Keller, T. 2021. Shareworthiness and Motivated Reasoning in Hyper-Partisan News Sharing Behavior on Twitter. *Digital Journalism*, 1–23.

Xu, J.; Wang, P.; Tian, G.; Xu, B.; Zhao, J.; Wang, F.; and Hao, H. 2015. Short text clustering via convolutional neural networks. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, 62–69.

Yang, C.; Shi, X.; Jie, L.; and Han, J. 2018. I know you'll be back: Interpretable new user clustering and churn prediction on a mobile social application. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 914–922.

Zeng, A.; Chen, M.; Zhang, L.; and Xu, Q. 2023. Are transformers effective for time series forecasting? In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, 11121–11128.

Zhang, X.; Malkov, Y.; Florez, O.; Park, S.; McWilliams, B.; Han, J.; and El-Kishky, A. 2022. TwHIN-BERT: A Socially-Enriched Pre-trained Language Model for Multilingual Tweet Representations. *arXiv preprint arXiv:2209.07562*.

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 Discussion and Limitations.

   (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 Discussion and Limitations.

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

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

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

   (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 Discussion and Limitations.

   (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 Model Accuracy.

   (b) Have you provided justifications for all theoretical results? Yes, see Model Accuracy

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

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

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

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)? Yes.

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

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

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

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

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.

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

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

   (f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR? Yes, see Discussion and Limitations

   (g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset? Yes.

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

Technical Appendix

Bot Heuristics

We filter out suspected bot accounts from our initial collection of users, as well as those likely to be celebrities or organizations, by using a set of heuristics from the literature for this filtering step (Cresci et al. 2015). Specifically, we compare the characteristics of each account with several cut-off values 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)

Hyperparameters

We search over the hyperparameters shown in Table 9 to train our models, picking the best settings based on mean absolute error on the validation set.

Parameter	Values
Hidden Units - LSTM	32, 64, 128, 256, 512
Hidden Units - Linear	128, 512
Bidirectional	True, False
LSTM Layers	1, 2
Activation	Sigmoid, Relu
Learning Rate	1e-3, 1e-4, 1e-5
Batch Size	32, 64, 128
Early Stopping Patience	3, 5

Table 9: Hyperparameter values tuned on validation data.