

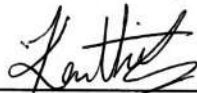
COMPUTATIONAL MODELS OF USER ENGAGEMENT WITH ONLINE NEWS

AN ABSTRACT

SUBMITTED ON 04/10, 2023

TO THE DEPARTMENT OF COMPUTER SCIENCE  
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS  
OF THE SCHOOL OF SCIENCE AND ENGINEERING  
OF TULANE UNIVERSITY  
FOR THE DEGREE OF  
DOCTOR OF PHILOSOPHY

BY



KARTHIK SHIVARAM

APPROVED:



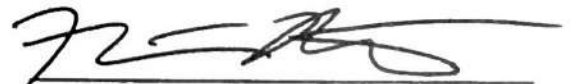
ARON CULOTTA, PH.D.  
CHAIR



MUSTAFA BILGIC, PH.D.



JIHUN HAMM, PH.D.



NICHOLAS MATTEI, PH.D.

# Abstract

The shift from traditional print media to online platforms has revolutionized the way people consume and engage with current events. To enhance user involvement, these platforms typically employ personalization algorithms like recommendation systems, that learn about users' preferences from their past interactions and suggest relevant content. Nevertheless, the use of such algorithms may result in biased engagement patterns caused by data that was influenced by the recommendation system itself, leading to concerns about "filter bubbles" and "echo chambers". Such entities cause users to be over-exposed to information that conforms with their pre-existing beliefs while limiting exposure to opposing viewpoints. As a result, these types of news consumption habits can bias users, leading to negative consequences such as the hyper-partisanship, online polarization, and the spread of misinformation. In this dissertation we aim to better understand factors that affect short-term and long-term news engagement behavior on social media. To achieve this, we conduct simulation studies to understand which aspects of recommendation systems contribute to filter bubble formation. We propose attention-based neural networks to mitigate these effects in content-based recommenders. In addition, long-term news engagement behavior is examined by analyzing observational data collected from Twitter over a decade. Our analysis focuses on a specific type of engagement behavior where users exhibit distrust towards the news media they engage with and examine its impact on engagement diversity. Finally, we propose forecasting methods to predict future news engagement behavior of users which reveal factors that shape long-term news

consumption habits on social media.

COMPUTATIONAL MODELS OF USER ENGAGEMENT WITH ONLINE NEWS

A DISSERTATION

SUBMITTED ON 04/10, 2023

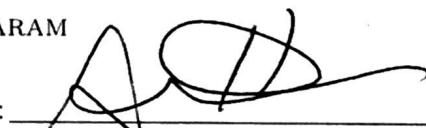
TO THE DEPARTMENT OF COMPUTER SCIENCE  
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS  
OF THE SCHOOL OF SCIENCE AND ENGINEERING  
OF TULANE UNIVERSITY  
FOR THE DEGREE OF  
DOCTOR OF PHILOSOPHY

BY



KARTHIK SHIVARAM

APPROVED:



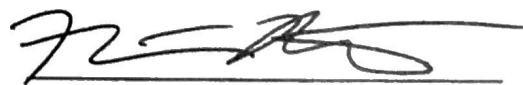
ARON CULOTTA, PH.D.  
CHAIR



MUSTAFA BILGIC, PH.D.



JIHUN HAMM, PH.D.



NICHOLAS MATTEI, PH.D.

© Copyright by Karthik Shivaram, 2023

*All Rights Reserved*

# Acknowledgments

I would like to begin by expressing my deepest gratitude to my advisor Dr. Aron Culotta, for his exceptional guidance and invaluable support which has been an integral part of my academic journey. Since the time I took his courses at IIT to working under his guidance for my research, he has been an important influence in my graduate life over the past five years. He has provided me with countless opportunities to learn and grow, and has challenged me to think critically, while also helping me develop a keen understanding of the research problem at hand. I feel extremely fortunate for the time and effort that he has invested in mentoring and teaching me throughout these years.

I would also like to thank my research collaborators, Dr. Mustafa Bilgic, Dr. Matthew Shapiro, and Ping Liu, for their invaluable help and guidance. Working with each of them has been an enriching experience that has contributed to both my professional and personal growth.

Additionally, I am thankful to my dissertation committee members, Dr. Nicholas Mattei, Dr. Jihun Hamm, and Dr. Mustafa Bilgic, for their support, insightful comments, and constructive feedback throughout my research.

Personally, I am extremely grateful to my parents Madhura and Shivaram for their constant support and encouragement without which this journey would not have been possible. Lastly but not least, I would like to thank my dear friends Skanda, Vignesh, Sanjay (you will be forever missed), Sanjana, Kedar, Akash, Suveen, and Prajaktha for their unwavering support and constant encouragement.

# Contents

<b>Acknowledgments</b>	<b>ii</b>
<b>List of Tables</b>	<b>vii</b>
<b>List of Figures</b>	<b>x</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Dissertation Outline . . . . .	6
1.2 Published Work . . . . .	7
1.3 Other Work . . . . .	8
<b>2 Background and Related Work</b>	<b>10</b>
2.1 Background . . . . .	10
2.2 Related Work . . . . .	13
<b>3 The Interaction between Political Typology and Filter Bubbles in News Recommendation Algorithms</b>	<b>21</b>
3.1 Introduction . . . . .	21
3.2 Data and Annotation . . . . .	23
3.3 Simulation Models . . . . .	27
3.4 Recommender models . . . . .	32
3.5 Problem Formulation . . . . .	34
3.6 Filter Bubble Metrics . . . . .	34

3.7	Experiments and Results . . . . .	35
3.8	Limitations . . . . .	46
3.9	Conclusion . . . . .	47
<b>4</b>	<b>Reducing Cross-Topic Political Homogenization in Content-Based News Recommendation</b>	<b>48</b>
4.1	Introduction . . . . .	48
4.2	Problem Formulation . . . . .	50
4.3	Methods . . . . .	51
4.4	Experiments and Results . . . . .	56
4.5	Conclusion . . . . .	65
<b>5</b>	<b>Characterizing Online Criticism of Partisan News Media using Weakly Supervised Learning</b>	<b>68</b>
5.1	Introduction . . . . .	68
5.2	Data . . . . .	70
5.3	Problem Formulation . . . . .	73
5.4	Methods . . . . .	74
5.5	Experiments and Results . . . . .	83
5.6	Analysis of Media-Targeted Criticism . . . . .	86
5.7	Limitations . . . . .	91
5.8	Conclusion . . . . .	92
<b>6</b>	<b>Forecasting News Engagement Behavior</b>	<b>94</b>
6.1	Introduction . . . . .	94
6.2	Problem Formulation . . . . .	94
6.3	Data . . . . .	95
6.4	Methods . . . . .	97



6.5	Experiments and Results . . . . .	103
6.6	Conclusion . . . . .	117
<b>7</b>	<b>Conclusion</b>	<b>118</b>
7.1	Future Work . . . . .	120
	<b>References</b>	<b>122</b>

# List of Tables

3.1	Statistics of collected news articles. . . . .	25
3.2	Label Distributions of Training Data for Topic Classification . . . . .	25
3.3	Performance of Relevance Classifier . . . . .	26
3.4	The F1 scores of the Topic Classifiers . . . . .	27
3.5	News article topics distribution. . . . .	28
3.6	An example of the utility matrix for a "devout and diverse" user. . . . .	30
4.1	Sample of 50 Polarizing Terms used by STANPP . . . . .	54
4.2	Average model accuracy over 45 topic pairs . . . . .	59
4.3	Average Network Performance across 45 Topic Pairs with Additional Metrics . . . . .	61
4.4	Network hyperparameters considered. . . . .	62
4.5	Top 30 terms with highest attention scores for a topic pair discussing climate change and gun control. . . . .	67
5.1	Tweets collected from 5,470 users and the fraction that reference one of 522 news sources. . . . .	73
5.2	Threshold Parameters for $\phi_{up}$ . . . . .	77
5.3	Labeling function output on unlabeled data . . . . .	79
5.4	Heuristics for $\phi_{tt}$ . . . . .	80
5.5	User Based Features . . . . .	81
5.6	Labeling function accuracy on test data. . . . .	83

5.7	Test set ROC AUC for combinations of model, labeling function, and label denoising methods. . . . .	85
5.8	Hyperparameter Values for Experiments . . . . .	85
5.9	Test set Performance for combinations of model, labeling function, and label denoising methods. . . . .	86
5.10	Progression sequences of first engagement of each type. . . . .	91
6.1	Tweet Distribution by Year . . . . .	96
6.2	Matched Tweet Distribution by Partisan Stance . . . . .	97
6.3	News Engagement Distribution . . . . .	97
6.4	Data Subsets by Time . . . . .	98
6.5	Example Observation Sequence for a given user for $D_1$ . . . . .	99
6.6	Train, validation and test sizes across all datasets . . . . .	104
6.7	Mean Forecast Metrics across all Data-sets (D1 to D4) . . . . .	104
6.8	Statistical Hypothesis Test results for Avg MAE of Models across all datasets using paired T-test . . . . .	104
6.9	Forecast Metrics for all test sets across individual stances (Best scores are highlighted per metric) . . . . .	105
6.10	Model Performance on samples where users go from a Engagement to No-Engagement State . . . . .	112
6.11	Model Performance on samples where users go from a No-Engagement to Engagement State . . . . .	113
6.12	Subset of Top Terms predictive of -3 Engagements . . . . .	115
6.13	Subset of Top Terms predictive of +3 Engagements . . . . .	116

# List of Figures

1.1	Dark Side of Online News Consumption . . . . .	2
1.2	Dissertation Road-map . . . . .	5
3.1	Data Collection and Annotation Pipeline . . . . .	23
3.2	Simulation results by political typology, showing click-through rate vs average document stance for three levels of randomness. . . . .	38
3.3	Click-through rate vs normalized stance entropy for the content-based recommender. . . . .	40
3.4	Hellinger Distance between different Partisan Scores . . . . .	41
3.5	Difference in the number of articles recommended by the content-based and collaborative filtering recommenders as compared to the oracle recommender. Results are the average of 1,000 recommendations for 100 users from three user types: country first conservatives (CFC), devote and diverse (D&D), and opportunity Democrats (OPD). . . . .	44
3.6	Click-through rate vs normalized topic entropy for all recommenders. The content-based recommender exhibits much lower topic diversity than others. . . . .	46
4.1	Network Architectures for STN, STAN and MTAN . . . . .	56
4.2	Experimental Settings Pipeline . . . . .	57
4.3	Distribution of test accuracies across 45 topic pairs for STN, STAN, MTAN and MTANPP . . . . .	60

4.4	Topic similarity vs test accuracy for 20 topic pairs. The topic similarity is measured using Jaccard similarity between sets of overlapping terms for a given topic pair. The trend lines are generated using a lowess regression model. . . . .	63
4.5	Average test accuracy of 45 topic pairs vs loss weights ( $\alpha$ ) used in STANPP and MTAN . . . . .	64
5.1	Example tweet critical of a news source. . . . .	70
5.2	News Sources Partisan Distribution . . . . .	71
5.3	ROC curves of network models trained using different labeling functions. . . . .	84
5.4	Criticism ratio by partisan stance of the user ( <b>left panel</b> ) and in aggregate ( <b>right panel</b> ). . . . .	87
5.5	Criticism shown towards the most mentioned news sources from each partisan stance. . . . .	88
5.6	Comparison of Normalized Stance Entropy before and after removing critical tweets. . . . .	89
5.7	Criticism across time by partisan stance of the user ( <b>top panel</b> ) and news source ( <b>bottom panel</b> ). . . . .	90
6.1	Network Architecture for the Multiple Feature Network . . . . .	103
6.2	MAE Model Performance by Yearly Quarters . . . . .	107
6.3	Truth vs Prediction plots for users with the <b>lowest errors</b> across all test sets for the SFN + C model . . . . .	108
6.4	Truth vs Prediction plots for users with the <b>highest errors</b> across all test sets for the SFN + C model . . . . .	109
6.5	Model Performance at different difficulty levels based on Cosine Distance Ranking . . . . .	110

6.6	Model Performance at different difficulty levels based on Ranking using	
	Baseline Absolute Error . . . . .	111

# Chapter 1

## Introduction

News ecosystems have shifted from traditional print media to online platforms in recent years. This shift has fundamentally changed how people read and engage with current events. Specifically, at least 48% of adults consume news through social media in the United States [179] in 2021. These online platforms mostly employ some form of machine learning based personalization algorithms (e.g., Recommendation Systems) designed to filter and curate news content to increase user engagement [37, 52, 87]. Frequently these types of algorithms undergo training by using users prior interaction behaviors in order to learn their preferences and inclinations, and as a result they can deliver a more personalized experience by recommending relevant content that the users desire. However, this method of interactive retraining relies on confounded data, i.e data generated as a result of the recommender system [28]. As a result, the algorithm biases a user's engagement patterns towards a particular direction that impacts the diversity of content they are exposed to.

This type of news consumption habits have led to concerns revolving around "*filter bubbles*" [130] and "*echo chambers*" [66]. These are scenarios where users are overexposed to information that aligns with their pre-existing beliefs and perspectives while decreasing exposure towards information that contradicts their viewpoints. This can have a "*homogenization*" effect where users become increasingly similar in their behavior and ideologies, limiting exposure to different ideas and perspectives

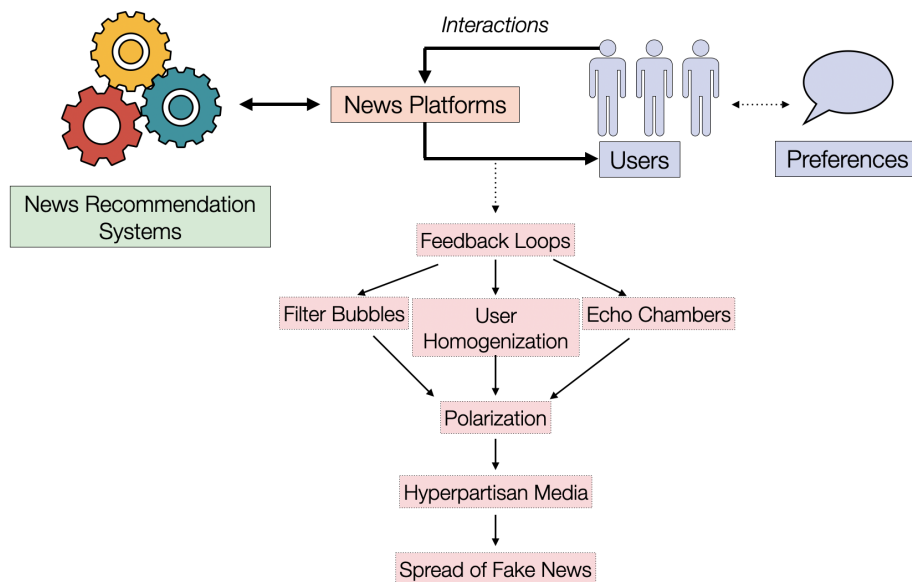


Figure 1.1: Dark Side of Online News Consumption

[125].

These entities are further shown to contribute towards political polarization [6, 7, 35, 44] by decreasing exposure to diverse ideological points of view and forcing extremism in prior opinions and beliefs, which in turn promotes consumption of hyperpartisan news media [71, 99], which is shown to be one of the primary factors that contributes towards the dissemination of misinformation [19, 112, 116] (Figure 1.1). Due to these diverse issues that affect how users consume digital news, it is necessary for us to gain a better understanding of the different factors that impact online news engagement.

The aim of this dissertation has two main components. Firstly, we seek to gain a deeper understanding of the factors that impact news engagement behavior in the short term by conducting simulation-based studies, focusing on filter bubbles and news recommendation systems. Secondly, we investigate long-term news engagement behavior in online social networks by analyzing observational data.

Our initial focus is on examining the short-term effects of news engagement



behavior, particularly in relation to the formation of filter bubbles in news recommendation systems. We seek to gain a deeper understanding of the various biases that contribute to this phenomenon, particularly when users have diverse preferences. To accomplish this, we have compiled a vast dataset of over 900,000 news articles from 41 distinct news sources, which we have classified by topic and partisan inclination. Using simulation-based analyses, we explore the impact of different algorithmic methods on filter bubble formation, taking into account the pre-existing preferences of users based on Pew’s studies of political typologies. Specifically, we find that users with more extreme preferences are presented with less diverse content but demonstrate higher click-through rates than those with less extreme preferences. Furthermore, we discover that content-based and collaborative-filtering recommenders produce significantly different filter bubbles. Lastly, we observe that when users possess contrasting partisan preferences on different topics, these recommenders tend to have a homogenization effect, we denote this as "**cross-topic homogenization**".

Extending this work on filter bubbles and news recommendation systems, we focus our attention on addressing the issue of cross-topic homogenization in content-based news recommendation systems, specifically for users who have diverging political preferences on different topics. For instance, some users may prefer conservative articles on one topic but liberal articles on another. This can result in recommenders suggesting articles with a similar political leaning on both topics, which can have a homogenizing effect, particularly when both topics share politically polarized terms such as "far right" or "radical left". To mitigate this issue, we propose using attention-based neural network models trained in a multi-task based training setting, which can increase attention on words that are specific to the topic while decreasing attention on polarized, topic-general terms. We find that the proposed approach results in more accurate recommendations for simulated users with such diverse preferences.

We next examine long term factors that affect news engagement behavior on social media. Moving away from simulation based studies as discussed above, we use a decade worth's of user data collected from Twitter. We first look at a specific type of news engagement behavior in which users distrust, criticize or ridicule the news source they engage with. This issue of criticism has a impact on (1) *hyperpartisanship, and misinformation* [129, 145], (2) *online polarization* where usually the intent of engagement is ignored [38, 64, 65]. Hence understanding the prevalence and temporal dynamics of media-targeted criticism can help us better measure the health of the information ecosystem. We propose weakly supervised learning methods that leverages multiple noisy labeling functions based on both the tweet's content and the user's historical news sharing behavior and train multiple classifiers. With these classifiers, we then explore how tweets expressing criticism interact with hyperpartisanship and misinformation sharing.

Lastly, we focus on forecasting news engagement behavior to better understand the evolution of user behavior over time. By using deep learning-based sequence models, we aim to predict future news engagement behavior for users and examine the predictive factors that influence user engagement with unreliable or fake news sources. This approach will enable us to gain valuable insights into the long-term factors that impact user engagement behavior. The entire dissertation road-map is shown in Figure 1.2.

The main contributions of this dissertation is as follows :

1. Constructed a novel News Article Dataset of 900K articles labelled across 14 Topics and 5 Partisan Stances.
2. Conducted Simulation Based Analysis to identify primary factors that lead to filter bubble formation in content based and collaborative filtering based news recommendation algorithms when considering users with heterogeneous partisan

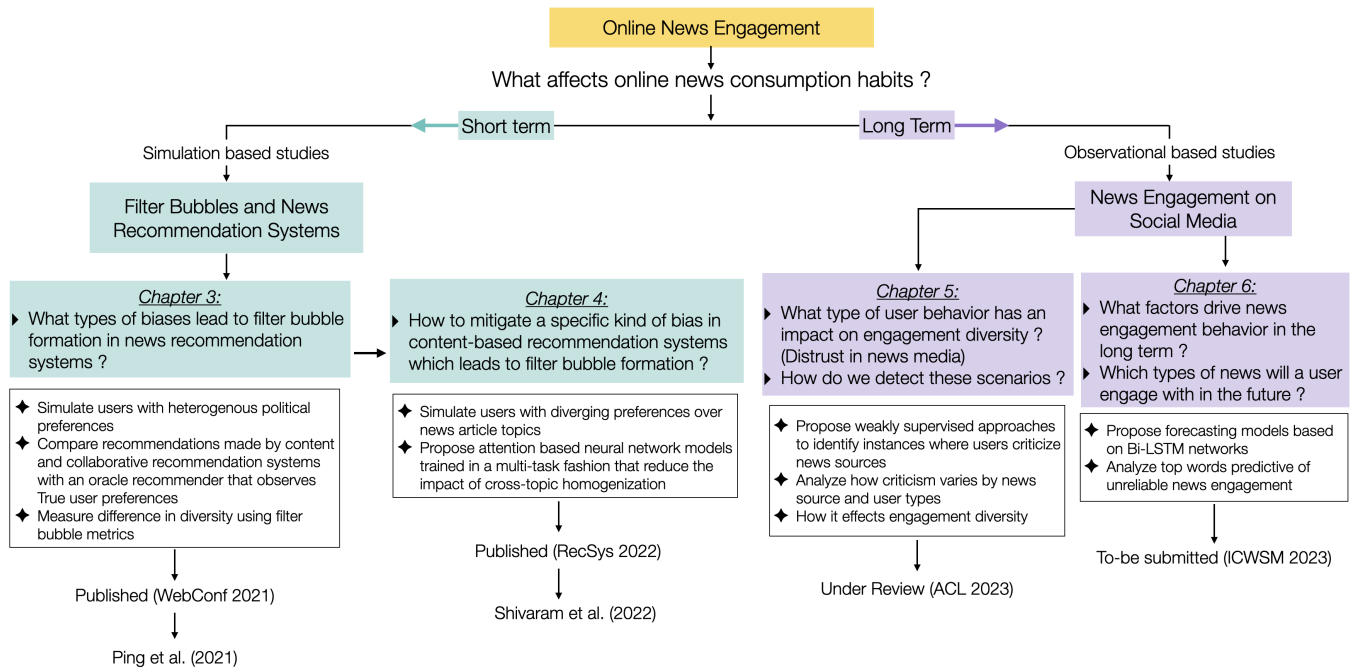


Figure 1.2: Dissertation Road-map

preferences across different political topics.

3. Proposed a attention based neural network models trained in a multi-task approach to mitigate cross-topic homogenization, a specific type of bias in news recommendation system that leads to filter bubble formation [166].
4. Constructed a novel dataset of 6.5 million tweets that engage with one of 522 news sources over a ten year period.
5. Proposed a Weakly Supervised classification approach to identify scenarios of criticism and distrust for online news engagement on twitter.
6. Proposed neural network based forecasting models to predict future news engagement behavior on twitter.

## 1.1 Dissertation Outline

The rest of this dissertation is organized as follows :

- **Chapter 2 - Background and Related Work.** This chapter focuses on the background and related work that is relevant to the subsequent chapters. The discussion centers around Recommendation Systems, Multi-task Learning, and Weakly Supervised Learning, which are employed in this dissertation. The primary focus is to provide a brief overview of these techniques and algorithms. Additionally, relevant related works are discussed, with emphasis on the relationship between filter bubbles and personalization algorithms for news, the various types of popular biases present in news recommendation systems, detecting criticism in online news media, and user modeling for social media.
- **Chapter 3 - The Interaction between Political Typology and Filter Bubbles in News Recommendation Algorithms.** This chapter discusses our work on detecting and characterizing various biases that contribute to the formation of filter bubbles in political news recommendation systems for users with heterogeneous preferences through the use of simulation based analysis.
- **Chapter 4 - Reducing Cross-Topic Political Homogenization in Content-Based News Recommendation.** This chapter discusses our work on applying attention-based neural networks trained using a multi-task approach in order to mitigate the issue of cross-topic homogenization in content based news recommendation systems.
- **Chapter 5 - Characterizing Online Criticism of Partisan News Media using Weakly Supervised Learning.** This chapter discusses our work on applying weakly supervised learning methods in order to detect criticism of partisan news media on twitter.

- **Chapter 6 - Forecasting News Engagement Behavior.** This chapter discusses our work on developing neural network based forecasting models to predict future news engagement behavior on twitter. We also analyze top predictive factors that affect user engagement with unreliable/fake news sources.
- **Chapter 7 - Conclusion.** This chapter summarizes the main findings in this dissertation as well as discusses the future possible research directions of understanding news engagement behavior in online information systems.

## 1.2 Published Work

The work completed and presented in this dissertation (Chapter 3 and 4) has led to the following publications:

- Ping Liu, Karthik Shivaram, Aron Culotta, Matthew A Shapiro and Mustafa Bilgic. "The Interaction between Political Typology and Filter Bubbles in News Recommendation Algorithms." in *The Web Conference (WebConf)*. 2020.  
Code : <https://github.com/IIT-ML/nsf-eager-filter-bubbles>
- Karthik Shivaram, Ping Liu, Matthew A Shapiro, Mustafa Bilgic and Aron Culotta. "Reducing Cross-Topic Political Homogenization in Content-Based News Recommendation". in *ACM Conference in Recommender Systems (RecSys)*. 2022.  
Code : <https://github.com/tapilab/recsys-2022-political>

In addition to these prior publications, the work discussed in Chapter 5 is currently under review.

- Karthik Shivaram, Mustafa Bilgic, Matthew Shapiro and Aron Culotta. "Characterizing Online Criticism of Partisan News Media using Weakly Supervised

Learning". in *Association for Computational Linguistics (ACL)*. 2023.

Code : <https://github.com/karthikshivaram24/news-criticism-detection>

The work discussed in Chapter 6 will be submitted to *The International AAAI Conference on Web and Social Media (ICWSM)*. 2023.

Code : <https://github.com/karthikshivaram24/forecasting-news-engagement>

### 1.3 Other Work

Apart from the research work discussed in this dissertation, additional work has been accomplished during my PhD program, there are as follows :

- **Characterizing variation in toxic language by social context.** How two people speak to one another depends heavily on the nature of their relationship. For example, the same phrase said to a friend in jest may be offensive to a stranger. We apply this simple observation to study toxic comments in online social networks. We curate a collection of 6.7 K tweets containing potentially toxic terms from users with different relationship types, as determined by the nature of their follower-friend connection. And find that such tweets between users with no connection are nearly three times as likely to be toxic as those between users who are mutual friends, and that taking into account this relationship type improves toxicity detection methods by about 5% on average. Furthermore, we provide a descriptive analysis of how toxic language varies by relationship type, finding for example that mildly offensive terms are used to express hostility more commonly between users with no social connection than users who are mutual friends.

This work led to the following publication :

- Bahar Rafdar, Karthik Shivaram and Aron Culotta. "Characterizing vari-

ation in toxic language by social context." in *The International AAAI Conference on Web and Social Media (ICWSM)*. 2020

- **How Does Empowering Users with Greater System Control Affect News Filter Bubbles ?** Algorithmic personalization of news enables online users to find articles of interest; however, studies suggest that algorithmic filtering can lead to “filter bubbles,” where users are presented with content that they are more likely to agree with, leading to ideological segregation and exacerbating societal divisions. In most cases, users are unaware that they are being presented a filtered view and hence are unaware that they are even in a filter bubble. Even when aware, users often lack agency over the filtering process. In this paper, we first design a political news recommendation system augmented with an enhanced interface that exposes the political and topical interests the system inferred from user behavior, and that additionally allows the user to adjust the recommendation system to receive more articles of a particular topic or political stance. We then conduct a user study to understand the impact of the system on the news the user sees. By comparing system behavior with a control group that uses a traditional interface, we find that transparency about the algorithm helps people realize that they are in fact in a filter bubble. We further find that, while a majority of people use the interaction tools to escape from their algorithm-assigned filter bubbles, some actually use these tools to more firmly root themselves in their respective bubbles. Finally, we find that the interaction mechanism increases engagement with the system, leading to a higher click-through-rate even when users are presented with a more diverse set of news content

This work is currently being submitted to the *ACM Conference in Recommender Systems (RecSys)*. 2023.

# Chapter 2

## Background and Related Work

In this chapter, we discuss and provide background and related work concerned with the various problems that are addressed in this dissertation. To begin with, we provide an overview of recommendation systems, multi-task learning, and weakly supervised learning, which are all essential components of this research. Following that, we present several related works that examine filter bubbles in the context of recommendation systems, which we expand upon in Chapter 3. We also explore prior research on identifying biases that impact recommendation systems, algorithms employed in current news recommendation systems, and how they are vulnerable to the problem of cross-topic homogenization as discussed in Chapter 4. Furthermore, we discuss the issue of criticism of online news media (Chapter 5) and how it differs from existing tasks such as sentiment analysis and stance detection. Finally, we examine prior work into modeling user behavior on social media, which we use as a basis for predicting news engagement behavior in Chapter 6.

### 2.1 Background

#### 2.1.1 Recommendation Systems

Recommendation systems are a type of machine learning algorithm that are designed to provide recommendation of objects based on a user's unique preferences. By analyzing a user's past behavior and interactions with the system, it learns their



preferences and makes tailored recommendations accordingly. The use of these systems is extensive across various industries such as e-commerce, social media, and content streaming platforms. They cover a wide range of applications including *news* [79, 95, 141, 143, 187, 188, 189, 202, 205], *movies* [46, 149, 153, 185, 199, 204], *fashion* [32, 33, 80, 196], *friend* [5, 48, 81, 123, 180, 184], and *music recommendation* [22, 74, 170, 178, 183], among others.

From a machine learning perspective these systems can be categorized into 3 main types - content based, collaborative-filtering based and hybrid (content + collaborative) based recommenders.

*Content based* systems [24, 68, 133, 149, 169, 178, 181] recommend items to users based on the attributes of the prior items they have previously interacted with. These types of recommendation systems mainly utilize item attributes in order to recommend relevant content to the user. Multiple types of content based attributes from different modalities are used to train these types of systems ranging from text, images, metadata, audio signals etc.

*Collaborative-filtering based* systems recommend items by measuring similarities between users and items simultaneously. These types of systems learn patterns in engagement behavior by considering users who have similar preferences to recommend items to other users and are mainly trained using data based on user-item interaction in order to recommend relevant content to users. There are 3 main types of collaborative-filtering algorithms which include item-based collaborative filtering, user-based collaborative filtering and matrix factorization. Item-based collaborative filtering [36, 77, 127, 156, 192, 201] methods represent a user profile based on their past interactions with items, and then using the similarities between the target item and the interacted items, determines how relevant the current item is to the user. User-based collaborative filtering methods [14, 15, 160, 164, 182, 198] compares the

similarity between the current user profile and other user profiles based on their past ratings/interactions of items and recommends items that the most similar users like for the current user. Matrix Factorization methods [9, 31, 76, 85, 97, 114] represent both users and items as vectors in a low dimensional representational space and predicts ratings/feedback based on the inner product of these 2 vector representations.

*Hybrid based* [72, 101, 175] systems combine both content-based and collaborative filtering approaches to generate recommendations. This type of system uses content-based approaches to generate an initial set of recommendations, overcoming the "cold-start" problem and then uses collaborative filtering to refine the recommendations based on the user's preferences.

### 2.1.2 Multi-task Learning

Multi-task learning [26] is a type of machine learning approach where a singular model is simultaneously trained to perform multiple related tasks. Training the model in this manner leads it to learn representations that generalize better than just training on a single task as well as leverage correlations between the different task labels. Generally associated with Deep learning [102], it can be divided into 2 main types, hard parameter and soft parameter sharing. *Hard parameter sharing* [110, 111, 144] approaches have a set of shared layers to learn generalized representations which are followed by task specific components before prediction. *Soft parameter sharing* [53, 122, 194] approaches have separate models assigned to each task, and regularization is applied to minimize the difference between the parameters of these models.

### 2.1.3 Weakly Supervised Learning

Weakly supervised learning is a sub-field of machine learning that uses noisy sources of supervision to reduce human annotation effort. Common approaches in-

clude (1) *Self-training* [91], a semi-supervised method where a model is first trained on a limited labeled dataset. Next, the model is applied to a larger unlabeled dataset to predict labels, and the most confident predictions are selected and added to the labeled dataset for further model training. (2) *Co-training* [90], a semi-supervised method where two distinct classifiers are trained on different feature subsets, each with a small labeled dataset. In each iteration, the classifiers predict labels for the unlabeled data using their respective feature sets. High-confidence predictions from each classifier are then utilized to augment the labeled dataset for the other classifier. This iterative process continues, with each classifier learning from the other’s augmented dataset. (3) *Crowd-sourcing* [98, 103], Crowd-sourcing is a technique where multiple groups of workers are utilized to label data. Due to the label noise introduced by these workers, these types of methods aim to learn the expertise and reliability of each worker in order to produce more reliable labels to train the model. More recent approaches for weakly supervised learning utilize *multi-task learning* [146, 148] or *generative models* [4, 197] to reduce the effects of label noise.

## 2.2 Related Work

### 2.2.1 Filter Bubbles and Personalization

Filter Bubbles are scenarios where users are over-exposed to information that aligns with their pre-existing beliefs and perspectives, while decreasing exposure towards information that contradicts their viewpoints [130]. These entities are primarily a result of prolonged engagement between individuals and online platforms, which tends to limit the diversity of content that users are exposed to. And diversity here can occur in different forms, for example *Content diversity* in the case of News [117], Movie [142] and Music Recommendations [157] and *Demographic Diversity* in the case of Friend or User recommendations [174, 180]. In this dissertation we mainly

focus on content diversity and propose multiple metrics to measure diversity across topics and partisan stances for recommended news articles (Chapter 3).

These entities occur due to content filtering by recommenders as these systems tend to prioritize content that is similar to what the user has already engaged with, filtering out diverse perspectives that may challenge their views. As a result, the user's exposure to alternative viewpoints is significantly limited decreasing the diversity of content shown. Prior work [20, 73] suggests that this effect can be particularly damaging in the realm of politics and news consumption, where individuals who are exposed only to news that confirms their existing beliefs can become more entrenched in their ideological positions, contributing to online polarization and division [35, 172]. From a online user segregation perspective [78], they occur due to *homophily* [119], the tendency of individuals to associate and seek out others who are similar to themselves in terms of demographic characteristics, beliefs, attitudes, and values. Causing these individuals to be more likely to encounter information and perspectives that align with their views. This can lead to a confirmation bias [126], where people selectively seek out and engage with information that confirms their pre-existing beliefs, while dismissing or ignoring information that contradicts their views. As a result, their exposure to diverse viewpoints and information is significantly limited, further reinforcing their existing beliefs. And this can have drastic affects on online political polarization [7, 38].

Much attention has been given to filter bubbles in the context of social media. For instance, research on filter bubbles has shown that, with regard to Twitter, user segregation is neither uniform across ideological orientations nor across the range of topics available for consumption [12]. On Facebook, Bakshy et al. [7] examined 10 million users to quantify individual exposure to diversified news, finding that liberals are less likely to encounter ideologically cross-cutting news content than conservatives,

a finding consistent with parallel research of Twitter [54]. Yet, online and offline political engagement can increase with exposure to this cross-cutting news, particularly when it originates from individuals not necessarily in one’s own filter bubble, i.e. individuals with whom one has weak connections [121]. Beyond news articles themselves, and highlighting the role of influential elites in filter bubble formation [69], comments about content on Facebook and YouTube can also be predictors of echo-chamber formation [18, 162, 163].

Beyond social media-based experiments, and given that, in the U.S., nearly one-fifth of Democrats and Republicans obtain news in a filter bubble-like dynamic [88], efforts have been made to simulate recommender systems to more closely observe filter bubble dynamics. These simulations are able to control select parameters, altering specific characteristics of the online environment. Epstein et al. [57], for example, evaluated “Search Engine Manipulation Effects” and confirmed that ranking bias shifts the behavior of the voting population, thus increasing the vote share for targeted candidates. This finding has since been confirmed via experiments using representative samples of the American public [167]. Elsewhere, Geschke et al. [67] constructed an agent-based model to test the emergence of the filter bubble effect, while Chaney et al. [28] and Jiang et al. [86] attempted to build a simulation environment defining and measuring the filter bubble effect across a variety of recommender algorithms.

Ultimately, filter bubbles have significant and often confounding effects with regard to how people perceive consensus and mobilize around partisan and policy issues [8, 21, 51, 56, 118, 155]. Without some form of intervention, there are significant implications for how one is able to properly receive and process information, accurate or otherwise. Information distortions may not consistently have lasting effects [154], but filter bubbles can affect voters’ election-related decisions nonetheless [56].

A number of strategies that aim to alleviate filter bubbles have been proposed. Masrour et al. [116] study filter bubbles created by network link prediction algorithms and propose a framework that utilizes adversarial learning to create more heterogeneous links in the network. Bhargava et al. [19] propose providing transparency and content control mechanisms to the users to combat filter bubbles on social media. In the news consumption domain, “bias alerts” sent to users can be considered partially effective in mitigating the voting-related implications described above [57]. Providing accuracy reminders before news is consumed may minimize the likelihood that people will trust and share potentially inaccurate information [50, 135]. Yet, one’s understanding of what is truly inaccurate is confounded by news source. Specifically, Dias et al. [47] find that source identification by users may help identify implausible news content from trusted news sources while simultaneously making it more difficult to identify plausible news content from untrusted news sources. This only reinforces the need to use bias alerts and accuracy reminders before news is consumed and perhaps periodically afterwards, too.

Having identified the need to account for both technological and psychological factors, the work discussed in Chapter 3, examines precisely how machine learning algorithms create a filter bubble effect for individuals with varying political views and vary levels of exposure to the gamut of news content.

### 2.2.2 Bias in News Recommendation Systems

The focus on cross-topic homogenization (as discussed in Chapter 4) is motivated in part by sociological theories suggesting that polarization grows when an individual’s partisan view on one topic “spreads” to another topic [44], for example consider a user with heterogeneous preferences, where they have a liberal preference for news articles about gun-control but a conservative preference for articles about immigration, if a phrase like “extreme right” appears in liberal gun-control articles a

user has liked as well as liberal immigration articles a user has not yet read, the recommender may incorrectly recommend liberal immigration articles due to the phrase "extreme right" as content-based news recommenders learn text features that correlate with user engagement. Furthermore, comprehensive public polling by Pew shows that many Americans do indeed have political stances that vary significantly by topic [136], which is in line with research indicating that the public is less politically monolithic than "elites," and that many citizens do not have fully-formed partisan opinions on many topics [83]. Given these sociological findings, if recommendation systems are systematically biased to show politically homogeneous content across topics, then they may serve as accelerants of partisanship. This is particularly problematic if the user initially does not have fully-formed opinions on a new topic, which is the scenario we aim to simulate in our experiments by having one topic appear less frequently than another. Recent work on content-based recommenders has shown that such cross-topic homogenization can occur in political news recommendation [106]. Our proposed methods are designed to understand how a recommendation system could be trained to reduce the likelihood of this homogenization.

This work also builds on research studying how partisan bias manifests in news media [23, 124, 139]. For example, Budak et al. [23] finds that news sources of different political leanings are distinguished most by "disproportionately criticizing one side." In our data, we observe this in phrases like "far right" and "radical left," topic-independent phrases criticizing the out-group that can lead to cross-topic homogenization. Our proposed methods are specifically designed to reduce the influence of such phrases.

Our work also adds to the growing study of different types of bias in recommendation systems, such as *popularity bias* [1] where popular items are recommended more frequently than less popular items, regardless of their relevance to the user's

interests or preferences, and *exposure bias* [96] where the recommendations made to users are influenced by their previous exposure to certain items, rather than their true preferences. One major factor that leads to these types of biases is the presence of feedback loops [115], which can contribute to the homogenization of users, causing them to consume similar content while sacrificing utility [28, 106] (measure of how useful or valuable a recommended item is to a user). Homogenization can also lead to the creation of “filter-bubbles” [130] and “echo-chambers” [66], which may also influence polarization [27, 35, 41]. This is prominent in the case of news recommendation systems [11, 64, 168], where several prototypes have been developed to give users more control of the recommender system to increase diversity [19, 128]. Most prior work focuses on increasing the overall partisan diversity of content exposure, ignoring cross-topic effects; furthermore, most prior work focuses on collaborative filtering recommendation systems [116]. In contrast, we focus here on mitigating cross-topic homogenization in content-based recommenders (Chapter 4), filling a key gap in the extant literature.

### 2.2.3 Modern News Recommendation Systems

Recently, a variety of deep learning based approaches have been proposed for news recommendation [79, 95, 141, 143, 187, 188, 189, 202, 205]. Most of these methods are content based, using attention-based deep neural networks to learn representations of both the candidate news article and the user’s interest based on click logs. The methods predict future click events based on the similarity between these two representations [187, 188, 205]. Some prior work also uses observed topic information to learn user interests in a hierarchical fashion [141] and also to enrich the news article representation learned [79]. Modern pre-trained language models have also recently been used in order to improve news and user representations [189, 202]. To our knowledge, none of these prior approaches directly address the issue of cross-topic



homogenization. In our experiments in Chapter 4, we compare with a representative baseline from this recent work by Zhang et al. [202], finding that it is also susceptible to cross-topic homogenization.

### 2.2.4 Criticism of Online News Media

Journalism cannot be free from public scrutiny [25], which may at times be highly critical and formalized (i.e., troll-based [138]). Indeed, the negative sentiment expressed by the public toward journalists and the institution of journalism itself may convey outright disgust and shame [165], most especially from those who are least trustful of the news [92]. Extremist political groups have significantly helped foster the view that the media is neither legitimate nor accurate [59]. Identifying media-targeted criticism is important as it can help better analyze factors that influence the spread of online misinformation [62] as well as help better determine the political lean of different online media [173].

Although this task of detecting criticism (as discussed in Chapter 5) is related to stance detection [3, 42, 176] and aspect based sentiment analysis [137, 152, 191] it remains a distinct task as we try to identify scenarios where ridicule, distrust, animosity, or sarcasm is shown towards a news source, rather than identifying the emotional tone as in aspect-based sentiment analysis or the user’s opinion towards the topic of the news source as done in stance detection we propose methods to detect if the user criticizes the news source they engage with.

### 2.2.5 User Behavior Modeling

Prior work in user behavior modeling for social media based applications are typically used for downstream tasks such as user churn rate prediction [2, 93, 193], predicting item consumption and purchasing habits [16, 109], user return prediction [89] and forecasting user engagement [108, 177]. Traditional methods [89, 109] typ-

ically utilize handcrafted features to represent user behavior and apply statistical learning techniques. Recently deep learning based techniques have been applied to model user behavior [108, 177, 193]. The work proposed by Yang et al. [193] utilizes LSTM models to learn effective historical patterns of user activity to predict churn rate for a social media application, Liu et al. [108] utilizes GNN's to model user actions and combines these with LSTM's to capture temporal dynamics of actions in order to forecast future user engagement on Snapchat.

# Chapter 3

## The Interaction between Political Typology and Filter Bubbles in News Recommendation Algorithms

### 3.1 Introduction

Machine learning algorithms provide personalized curation of news, blogs, and social media posts in order to improve user experience. However, there is mounting evidence that this kind of automated filtering leads to '*filter bubbles*', which are scenarios where users are over-exposed to ideas that conform with their preexisting perceptions and beliefs, prompting intellectual isolation [130]. In this chapter, I investigate this phenomenon in the context of political news recommendation algorithms, which can have significant and often confounding effects with regard to how people perceive consensus and mobilize around partisan and policy issues [8, 51, 56, 118, 155].

Prior work typically simplifies this problem space by reducing user preferences to a single partisan score (e.g., strong liberal to strong conservative) [150]. However, this ignores the nuanced and varied preferences across different topics. For example, a user may have conservative views on abortion but liberal views on health care. In this chapter, we are interested in understanding how a user's preferences influence the behavior of recommendation algorithms, and in turn the diversity of news content to

which they are exposed to in the short term.

To investigate this, we first collect over 900K news articles from 41 sources annotated by topic and partisan lean. Then, drawing on recent Pew surveys of political typology [49], we simulate nine classes of users (e.g., solid liberals, disaffected Democrats, country first conservatives, etc.) with differing partisan preferences across 14 news topics. Next we conduct simulation studies to compare the articles recommended by *content* and *collaborative filtering* algorithms with those articles recommended by an “*oracle*” approach that observes the user’s true preferences. This allows us to measure the change in diversity of recommendations introduced by the recommendation system versus what would be expected based solely on the user’s true preferences. Specifically, we compare recommendation diversity and user utility measures to address the following research questions:

- **RQ1 : How do user preferences influence the diversity of recommendations ?**

We find that users with more extreme preferences are shown less diverse content but have higher click-through rates than users with less extreme preferences.

- **RQ2 : How do filter bubbles vary by the type of recommendation system they interact with ?**

We find that the filter bubbles created by content-based recommenders and collaborative filtering are markedly different. Content-based recommendations are susceptible to biases based on how distinctive the partisan language used on a topic is, leading to over-recommendation of the most linguistically polarized topics. Collaborative filtering recommenders, on the other hand, are susceptible to the majority opinion of users, leading to the most popular topics being recommended regardless of user preferences.

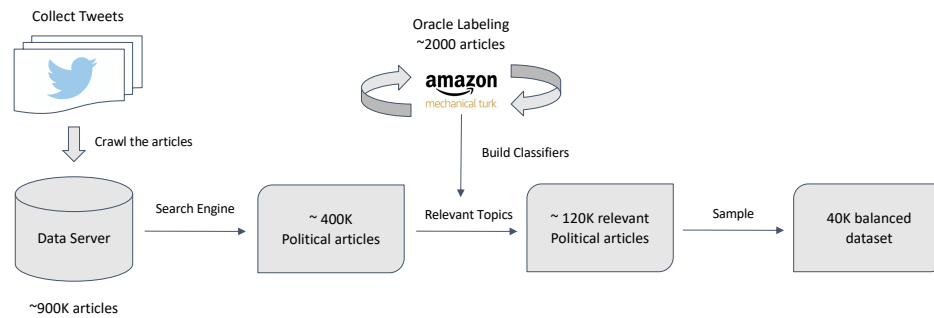


Figure 3.1: Data Collection and Annotation Pipeline

- **RQ3 : How does recommendation diversity vary for users with heterogeneous preferences ?**

We find that when users have divergent views on different topics, recommenders tend to have a homogenization effect. For example, if a user is conservative on most issues, but liberal on health care, they are shown more conservative articles on health care than desired. The reasons again differ based on the type of recommender: for content-based, lexical overlap between topics can mislead the recommender; whereas for collaborative filtering, a small group of users with heterogeneous preferences are "subsumed" by a majority group that has less diverse views.

## 3.2 Data and Annotation

For this study, we require a large set of news articles annotated by both political stance and topic. In this section, we summarize our data collection and annotation process. Our overall approach is to use the news source as a proxy for political stance, and to use text classifiers to assign one or more topics to each article.

### 3.2.1 News article collection

To collect a range of political news articles, we first identified 41 featured news sources from *www.allsides.com*, which annotates each source with a *political stance* in  $\{-2, -1, 0, +1, +2\}$ , ranging from very liberal (-2) to very conservative (+2). The ratings are based in part on user surveys of the perceived slant of the news source.

To collect articles, we next query the Twitter API with the URL of each source to identify tweets that contain links to news articles. We then crawl each URL and collect the title, source, and content of each article. We submitted these queries to Twitter, continuously from September 2019 to August 2020, resulting in over 900K articles. These articles are summarized in Table 3.1. Popular sources from each stance include DailyBeast (-2, 17k articles), New York Times (-1, 47k), Forbes (0, 74k), Fox News (1, 36k), and Brietbart (2, 28k). Each article is annotated with the partisan score of its source.<sup>1</sup>

While this process gives us a broad range of articles from across the political spectrum, it is of course not without some sampling bias. E.g., articles shared on Twitter differ from a uniform random sample of all articles from all news sources. However, given that our focus is on articles likely to be read and shared by users, this sampling methodology seems appropriate for our purposes. To account for the unequal distribution of articles by partisan stance, in our experiments below we sample to have a balanced distribution of articles.

### 3.2.2 Topic classification

From the 900K articles we collected, our next goal is to build a classifier to annotate each article with the topics it discusses. To do so, we trained a two-stage classifier: one to determine if the article is relevant to U.S. politics, and a second to

---

<sup>1</sup>While this may introduce some label noise at the article level [63], we expect this to have limited impact in aggregate.

<b>stance</b>	<b>interpretation</b>	<b># sources</b>	<b># articles</b>	<b>% articles</b>
-2	extreme left	10	93,700	10.1
-1	moderate left	11	282,432	30.3
0	neutral	8	286,639	30.8
+1	moderate right	4	93,279	10.0
+2	extreme right	8	175,998	18.9

Table 3.1: Statistics of collected news articles.

<b>Topic</b>	<b>Negative Labels</b>	<b>Positive Labels</b>
LGBTQIA	1,972	114
abortion	1,909	177
environment	1,963	123
guns	2,014	72
health care	1,947	139
immigration	1,978	108
racism	1,986	100
taxes	1,963	123
technology	2,032	54
trade	2,006	80
trump impeachment	1,803	283
us 2020 election	1,725	361
us military	2,001	85
welfare	2,002	84

Table 3.2: Label Distributions of Training Data for Topic Classification

assign one or more topics to the article.

To collect training data, we first independently annotated a sample of documents with political relevance and topics. Through several discussions and iterative refinement, we arrived at the following list of 14 topics: *abortion*, *environment*, *guns*, *health care*, *immigration*, *LGBTQIA*, *taxes*, *technology*, *trade*, *Trump impeachment*, *US military*, *welfare*, *US 2020 election*, and *racism*.

To increase the training sample, we next sampled additional documents to be annotated using Amazon Mechanical Turk. Using our expert annotations as a guide, we identified 12 high-quality AMT annotators, and had them annotate 3,250 total documents, of which 2,086 were annotated as politically relevant. The label

<b>Accuracy</b>	<b>F1</b>	<b>Recall</b>	<b>Precision</b>
0.7865	0.8307	0.7909	0.8773

Table 3.3: Performance of Relevance Classifier

distribution of this annotated dataset can be seen in Table 3.2.

From these labeled data, we next trained a binary classifier to determine if the article is relevant to U.S. politics or not. For this we used a standard logistic regression model using tf-idf features. Table 3.3 summarizes the accuracy of this classifier.

For topic classification, as it is a multi-label classification task, we trained 14 independent binary classifiers (one per topic). As the label distributions are highly imbalanced, we used SMOTE (Synthetic Minority Oversampling Technique) [30] to over-sample the positive class. Each of these topic classifiers uses logistic regression and tf-idf based features. The settings for the tf-idf vectorizer are as follows: the maximum number of features is 5,000, the maximum document frequency is 0.95, and the minimum document frequency is 30. These classifiers were separately optimized using a 5-fold cross validation loop with grid-search using the F1-score as the optimization metric. Table 3.2.2 shows the final cross-validation results for each topic. While F1 is generally high, we note that the classifier has smaller F1 score for the technology and welfare topics. For technology, this is likely do to ambiguity of whether an article is related to U.S. politics – e.g., an article about Facebook’s earnings is not relevant, but one that discusses new regulations is. For welfare, this topic is much broader than the rest, covering everything from cash assistance programs to homelessness issues. More training data would likely help here.



Table 3.4: The F1 scores of the Topic Classifiers

<b>Topic</b>	<b>F1</b>	<b>Topics</b>	<b>F1</b>
abortion	0.942	environment	0.898
guns	0.906	healthcare	0.785
immigration	0.853	LGBTQIA	0.894
racism	0.776	taxes	0.848
technology	0.538	trade	0.839
impeachment	0.888	US military	0.773
US election 2020	0.847	welfare	0.598

### 3.2.3 Article Sampling

With the two classifiers described above, we then annotated all collected articles with relevance and topic. Table 3.5 shows the predicted topic distribution of those articles determined to be relevant and to have at least one topic assigned. To ensure that the final sample has a uniform distribution of political stance, we randomly sample 8K articles from each stance, resulting in the final topic distribution in the final two columns in the table. (Note that many articles have more than one topic assigned.) Given the high fraction of articles about the 2020 election and Trump’s impeachment, we additionally down-sampled these topics to ensure a broader diversity of articles.

## 3.3 Simulation Models

In order to study the relationship between user preferences and recommendation systems, we would ideally conduct large-scale user studies to observe real-world interactions. However, given the challenges of conducting such studies, we instead build on the growing line of research conducting simulation studies of recommendation systems [28, 82, 86, 158].

To conduct such a simulation, we must make some assumptions about the

topics	Before sampling		After sampling	
	# articles	% articles	# articles	% articles
abortion	3,421	1.7	1,382	2.6
environment	4,329	2.2	1,656	3.2
guns	4,647	2.4	1,787	3.4
healthcare	14,823	7.6	5,444	10.6
immigration	10,736	5.5	4,308	8.3
LGBTQIA	2,848	1.5	1,126	2.1
racism	10,051	5.1	4,069	7.9
taxes	8,187	4.2	3,055	5.9
technology	3,722	1.9	1,379	2.6
trade	6,739	3.4	2,323	4.5
impeachment	45,989	23.4	6,811	13.2
US military	17,205	8.8	9,409	18.3
US election 2020	57,996	29.6	6,501	12.6
welfare	5,413	2.7	2,054	4.0
<b># labels</b>	196,106		51,304	
<b># articles</b>	167,431		40,000	

Table 3.5: News article topics distribution.

interaction model. Our approach largely follows that of prior work [28, 82], though here we use real news articles annotated by stance and topic. We assume that each user has a predefined, fixed set of preferences over articles they would like to read. These preferences are parameterized by the topic and stance of the article; e.g., a user may prefer to read a liberal article about healthcare more than a conservative article about immigration. As we are interested in short-term effects of recommenders, for this study we assume that user preferences do not change over time, though this is of course an important consideration for future studies.

The simulation proceeds by first showing the user an article. We then simulate the user’s response: either “like” or “dislike,” sampled proportional to the user’s preferences. With this feedback, the recommender updates its model to re-sort the remaining articles, then shows the next article to the user.

In the following sections, we describe this process in more detail, including

the user profile model, a user-choice model, and specific recommendation engines we implement.

### 3.3.1 User utility model

We represent each user’s preferences with a two-dimensional matrix of utility values  $U = \{u_{ij}\}$ , where  $u_{ij} \in [0, 1]$  indicates the user’s utility for reading an article on topic  $i$  with political stance  $j$ . (Thus,  $U$  is a  $14 \times 5$  matrix.) Large values indicate greater utility and therefore a larger probability of clicking on an article with topic  $i$  and stance  $j$ .

We wish to investigate how recommender behavior varies with heterogeneous utility matrices. Rather than randomly generate these matrices, in order to make them more reflective of reality, we sampled them based on Pew surveys of U.S. political typologies [49]. This comprehensive survey attempts to identify more nuanced political ideologies than a simple left/right spectrum. The survey contains many questions relevant to our identified topics above. E.g., for abortion, there is a survey question asking whether abortion should be legal in all/most cases. For immigration, there is a question asking whether immigrants strengthen or weaken the country. Pew clustered the responses to identify nine political types: *solid liberals*, *opportunity Democrats*, *disaffected Democrats*, *bystanders*, *devout and diverse*, *new era enterprisers*, *market skeptic Republicans*, *country first conservatives*, and *core conservatives*. These types capture a number of common heterogeneous ideologies – for example, the devout and diverse type leans conservative on issues of abortion and LGBTQIA, but leans liberal on race and health care. Similarly, the market skeptic Republicans lean liberal on issues of trade and taxation.

For each political type, then, we have a list of survey responses indicating the fraction of respondents who agree with the statement (e.g., 92% of solid liberals think that abortion should be legal in all/most cases). In our simulations, to generate a new

topics	-2	-1	0	+1	+2
abortion	0.276	0.411	0.546	<b>0.682</b>	0.546
environment	0.298	0.505	<b>0.711</b>	0.505	0.298
guns	0.332	0.490	<b>0.648</b>	0.490	0.332
healthcare	0.515	<b>0.711</b>	0.515	0.319	0.122
immigration	0.045	0.285	0.525	<b>0.766</b>	0.525
LGBTQIA	0.250	0.423	0.596	<b>0.769</b>	0.596
racism	<b>0.815</b>	0.575	0.335	0.095	0.010
taxes	0.080	0.283	0.486	<b>0.689</b>	0.486
technology	0.228	0.397	0.567	<b>0.737</b>	0.567
trade	0.400	0.511	0.622	<b>0.733</b>	0.622
Trump impeachment	0.313	0.452	<b>0.592</b>	0.452	0.313
US military	0.171	0.362	0.553	<b>0.744</b>	0.553
US election 2020	0.180	0.395	<b>0.610</b>	0.395	0.180
welfare	<b>0.860</b>	0.582	0.304	0.025	0.010

Table 3.6: An example of the utility matrix for a "devout and diverse" user.

user, we first pick a political type, then sample a utility matrix based on these survey responses. We convert these responses into a utility matrix as follows: for each survey question, we separate the responses into quantiles (0-20%, 21-40%, etc.), and assign the response to one of the five political stance categories  $\{-2, -1, 0, +1, +2\}$ . Thus, the fact that 92% of solid liberals think abortion should be legal means that their primary stance is  $-2$  on abortion. To generate the utility value for each topic/stance pair, we first sample a utility value for the primary stance using a Beta distribution centered on their survey response (e.g.,  $Beta(.92, 1)$  for the running example). We then decay this value for the other stances for this topic as a function of standard deviation of responses on this topic (i.e., a measure of how divisive this topic is). We then repeat this process for each topic. Table 3.6 shows an example utility matrix for the devout and diverse profile.

As with any simulation, one can question how reflective the simulated users are of the real world. The key aspect that these utilities do capture, however, is a broad spectrum of ideologies with which we can investigate variation in recommender behavior.

---

**Algorithm 1** The user interaction model
 

---

**Input:**  $u$  – the user vector;  $v$  – the article vector

**Output:**  $B$  – a Boolean variable to indicate whether the user likes this article or not.

```

 $v_{ui} = \text{Beta}^1(\text{dot}(u, \text{normalized}(v)))$ 
 $p_{ui} = v_{ui} \times \text{Beta}^1(0.98)$ 
if  $\text{Random} < p_{ui}$  then
  return Like
else
  return Dislike
end if

```

---

### 3.3.2 User interaction model

Given a user’s utility matrix, we next must simulate their behavior when presented with a recommended article. To do so, we follow the approach of prior work proposed by Chaney et al. [28]. To represent each article, we create a binary matrix of the same shape as the user utility matrix, containing 1 in cell  $(i, j)$  if the article has been assigned topic  $i$  and stance  $j$ . (Recall that the topic is derived from the text classifier, and the stance from the news source.) To sample whether a user will “like” or “dislike” an article, we first flatten both the utility matrix and the item matrix into 1d arrays, then compute the dot product between them. We then sample a value from a *Beta* distribution centered on this dot product value. Finally, a random number is generated and compared to the sampled value to determine the action of the user. Algorithm 1 formalizes this process.

In this algorithm, the function takes the user vector  $u$  and the item vector  $v$ . It calculates the dot product with  $u$  and normalized  $v$  to constrain the output as a probability from 0 to 1. We use a modified *Beta\** distribution (for which the mean and standard deviation are used from prior work [28]) to calculate the probability  $p_{ui}$  the user will click the given article. A random number is generated and used to determine whether the user will click this article, given  $p_{ui}$ .

## 3.4 Recommender models

In order to study the short term effects of recommendation systems , we chose and implemented five recommender systems, which include a random recommender (as a baseline), a content-based recommender, a collaborative filtering recommender, an oracle recommender and a hybrid recommender.

### 3.4.1 Random recommender

A random news recommender randomly samples news articles to recommend from a given candidate pool, this is done without replacement.

### 3.4.2 Content-based recommender

A content-based recommender (CBR) is a user-personalized model that learns the user’s preference, given the user’s previous interactions. We treat this as a binary classification problem – given an article, will the user like or dislike it? As training data, we seed the model with 700 simulated examples per user, sampled uniformly for each topic. We train a standard logistic regression classifier separately for each user, using tf-idf word features from each article. During the simulation, the training data is updated after each user interaction, and the model is retrained. Note that the classifier does not observe the stance and topic assignments for each document – this simulates the situation where neither the structure nor values of the user’s utility matrix are known to the recommender.

### 3.4.3 Collaborative Filtering recommender

A collaborative filtering recommender (CFR) uses the concept of similarities between users and items and recommend similar users the ‘liked’ items from each other’s ‘like’ history. We use nonnegative matrix factorization [58] on the user-item matrix to construct the collaborative filtering recommender.

### 3.4.4 Oracle recommender

We also implement an oracle recommender, which observes the user’s utility matrix and news’ topic and stance matrix. This algorithm samples documents proportional to the user’s probability of liking these documents. This baseline enables us to observe what biases are introduced by the recommender algorithms versus those that are inherent in the user’s pre-existing preferences.

### 3.4.5 Hybrid recommender

A simple way to try to reduce filter bubbles is to inject random recommendations into the user’s article list. We are interested in how the systems behave as the amount of randomness is injected. How quickly does the diversity increase as we introduce randomness? To investigate this, we consider three settings for each recommender above: randomness as 0% (totally personalized), 50% (hybrid), and 100% (totally random).

## 3.5 Problem Formulation

Let  $V$  be a collection of news articles. Each article  $v \in V$  is associated with one or more of 14 topics introduced in Section 3.2.2. Let  $U$  be a group of users. Each user  $u \in U$  belongs to one of the nine political types introduced in Section 3.3.1. In each simulation run, every user  $u$  is recommended  $N$  articles, one at a time. For each recommended article  $i$ , we simulate a binary random variable  $r_i$ , where  $r_i = 1$  mean the user clicks on /likes the article and  $r_i = 0$  means they do not.

## 3.6 Filter Bubble Metrics

In order to study the effect of filter bubbles on different recommendation algorithms and on different political user types, we propose and utilize the following metric based measures.

### 3.6.1 Click-through rate

The click-through rate (CTR) is the fraction of recommended articles that the user clicks on. A high CTR indicates that the algorithm can deliver accurate recommendations to the users, and thus has high utility. The CTR is defined as follows.

$$CTR = \frac{\sum(r_i)}{N}, 1 \leq i \leq N \quad (3.1)$$

### 3.6.2 Average document stance

Average document stance is the average partisan score of the articles that are *shown* to the users. Letting  $s(v_i) \in \{-2, -1, 0, 1, 2\}$  be the partisan score for article  $v_i$ , then the average document stance for a sequence of recommended articles is:



$$\bar{s} = \frac{\sum s(v_i)}{N}, 1 \leq i \leq N \quad (3.2)$$

### 3.6.3 Normalized stance entropy

Let  $p_i$  represent the fraction of articles that are shown to the users that have stance  $i$ . Normalized stance entropy is the entropy of this distribution, normalized by  $\log m$  so that its maximum is 1, where  $m = 5$  in our case, representing the five stances:

$$entropy = \frac{-\sum_{i=1}^m p_i \log p_i}{\log m} \quad (3.3)$$

A high value of normalized stance entropy would indicate a smaller filter bubble effect since the stances of the shown articles are more diverse since a higher entropy value indicates more uniform distribution across stances.

### 3.6.4 Normalized topic entropy

Similar to normalized stance entropy, we also measure the diversity of topics. This provides a measure of topical diversity, in addition to stance diversity above. The metric is the same as Equation 3.3, where  $p_i$  is instead the probability of articles having topic  $i$  in a sequence of recommendations, and  $n = 14$  since there are 14 topics. A low value of normalized topic entropy indicates that the recommender is recommending documents in a small set of topics.

$$entropy = \frac{-\sum_{i=1}^n p_i \log p_i}{\log n} \quad (3.4)$$

## 3.7 Experiments and Results

In this section we discuss the different setup and settings that we used to conduct our experiments as well as analyze and discuss the results in relation to the

research questions we discussed previously.

### 3.7.1 Settings

Firstly we generate 100 synthetic users for each political type following the user utility model described in Section 3.3.1. To initialize the recommendation models, we initially bootstrap 50 articles per topic for each user, resulting in 700 articles in total. Then the recommender recommends 1,000 articles, one by one, in a sequence and updates the algorithm after each recommendation. The CBR and CFR have three different randomness settings as we mentioned in the previous section.

Simulation of the oracle recommender is done explicitly as follows. For a given political type, for every article  $v$ , we calculate the probability  $p_v$  that the given political type would click that article if they are shown that article, based on their user profile. To study varying degrees of randomness in the oracle recommender, we compute a sampling weight for each article as  $\exp(w \times p_v)$  where  $w$  is a hyperparameter. We sample  $K$  articles from our dataset, using weighted sampling without replacement. We repeat this process  $M$  times. The probability  $q_v$  that the article will be shown by the oracle is the fraction of samples that contain  $v$ . When  $w = 0$ , each article has  $\exp(0 \times p_v) = 1$  weight, resulting in uniform sampling, and hence results in the random algorithm. As  $w > 0$ , articles that have a higher chance of being clicked gets a higher weight.

Once we have the shown ( $q_v$ ) and click ( $p_v$ ) probabilities, we can calculate the expectations for the CTR and all other metrics for all the political types using the whole dataset. We choose to use  $K$  as 1000, and  $M$  as 5000 in our case. For the hyperparameter  $w$ , we vary the value from 0 (totally random) to 9 (optimal personalized solution). For comparing CBR and CFR to the oracle recommender, we use  $w$  that achieves a similar CTR for that prototype, and analyze where the CBR and CFR differ from the oracle. This analysis allows us to measure the bias introduced by the

recommender beyond that inherent in the user preferences.

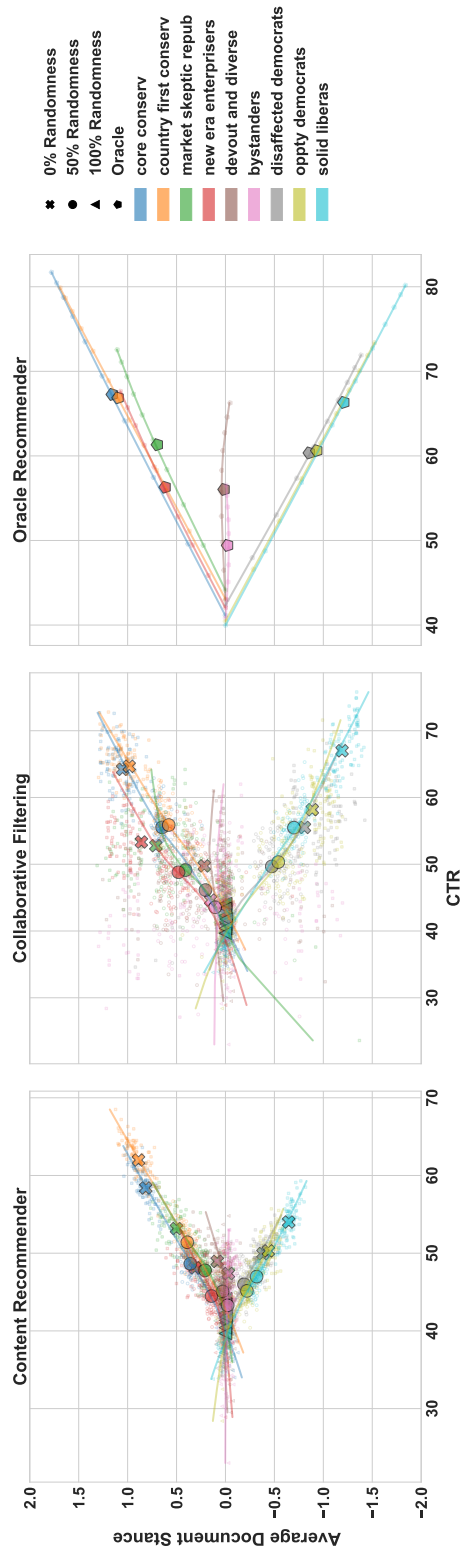


Figure 3.2: Simulation results by political typology, showing click-through rate vs average document stance for three levels of randomness.

### 3.7.2 Results

#### How do user preferences influence the diversity of recommendations?

We first investigate how the user’s political type influences the diversity of the recommended documents. Because there is a strong relationship between diversity and utility (i.e., CTR), we are particularly interested in their trade-off. We consider content-based recommender, collaborative filtering recommender, and the oracle recommender. For each, we have varying levels of randomness through the hybrid recommendation approach. In this way, we can plot how the CTR varies with filter bubble measures such as average document stance, stance entropy, and topic entropy. We would like to determine how this trade-off varies by political type.

Figure 3.2 shows the main results of CTR versus average document stance. Each panel summarizes the results of multiple simulation runs. Each dot represents the result for one user. For content-based recommender and collaborative filter recommender, each political type has three settings, which are 0% randomness, 50% randomness (hybrid recommender), and 100% randomness (random recommender). The larger symbols (e.g., circle, triangle, and cross) represent the centroids of each setting. For the oracle recommender, the randomness is controlled by the  $w$  parameter, where  $w$  ranges from  $w = 0$  (fully random) to  $w = 9$  (user preferences are given high priority). We also fit a LOWESS curve for each political type to visualize the tradeoff between CTR and document stance.

The first observation is that more extreme political types have both higher CTR and higher magnitude document stances. E.g., when no randomness is used, country-first conservatives have over a 60% CTR, and an average partisan score of nearly 1.0 for both content-based and collaborative filtering recommendations. On the other hand, more moderate political types, such as bystanders and devout & diverse, do not attain such high CTRs. These results make clear the intuitive finding that the

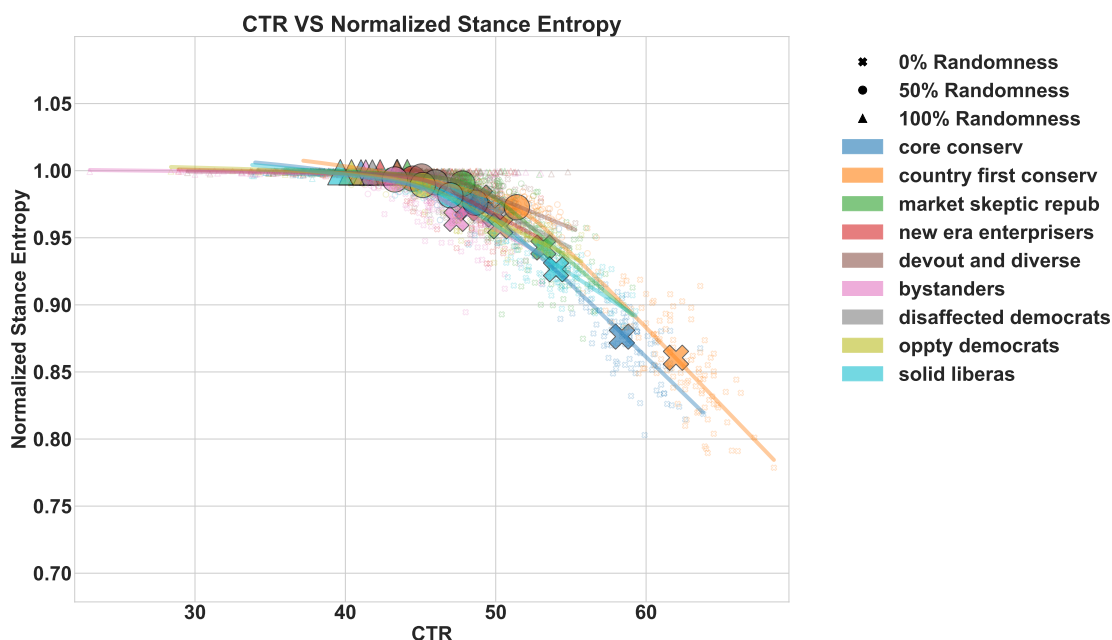


Figure 3.3: Click-through rate vs normalized stance entropy for the content-based recommender.

more extreme a user's preferences are, the more extreme their recommendations will be, and that it is easier to find articles that they are likely to click. We can also see from the third panel that the oracle is able to achieve even higher CTRs, though to do so it must recommend even more extreme and homogeneous documents. Figure 3.3 shows a similar result instead using stance entropy as a measure of diversity. For more extreme users, stance entropy decreases more quickly as CTR increases.

Examining these figures, there is a notable difference in the recommendation behavior for left-leaning versus right-leaning users. In the first panel of Figure 3.2, we see that right-leaning users ultimately exhibit higher CTRs, and more extreme partisan scores, than left-leaning users. Furthermore, we only see this difference in the content recommender, not for collaborative filtering or oracle recommenders. Upon further inspection, we conjecture that this is in part due to the asymmetry in the textual similarities between documents of different partisan scores. In particular, it appears that articles with score 0 are more similar to left-leaning articles (scores

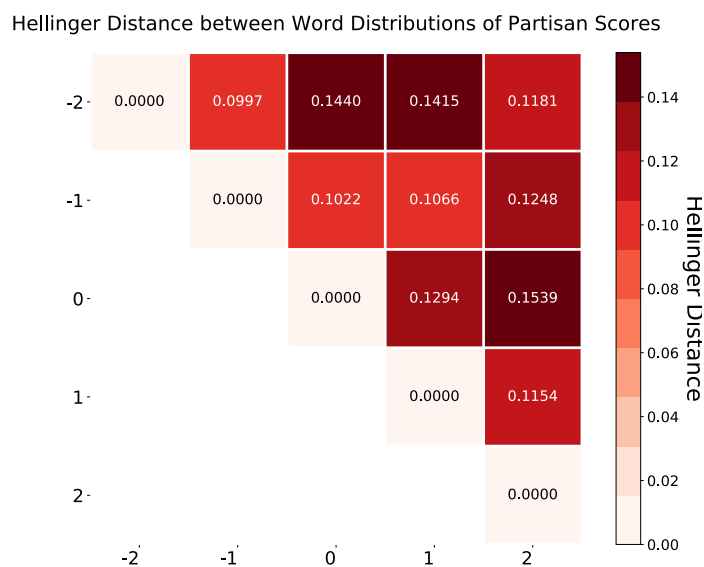


Figure 3.4: Hellinger Distance between different Partisan Scores

-2, -1) than they are to right-leaning articles (scores +1, +2). The result is that the content-based recommender has a more difficult time distinguishing between -2 and 0 articles than it does distinguishing between +2 and 0 articles. To further investigate this, we fit five different multinomial bag-of-words models, one per partisan score, by grouping together all articles with the same partisan score. We then compute the Hellinger distance [17] between each pair of multinomials to determine how similar the word distributions are. We find that the differences between -2 and 0 (.1415) and -1 and 0 (.1022) are substantially smaller than that between +2 and 0 (.1539) and +1 and 0 (.1294), further supporting this interpretation (Figure 3.4).

### How do filter bubbles vary by type of recommendation system?

As we have just seen, different recommendation systems can have different impact on filter bubble formation. In this section, we further compare CBR and CFR to their comparable oracle recommender counterpart to investigate possible biases introduced by CBR and CFR into the recommendation processes. To do so, we first compute the average number of articles recommended from each topic/partisan score

pair for each political type, using the versions of CBR and CFR with the highest overall click-through rate. We then compare these values with the corresponding recommendations provided by the oracle recommender.<sup>2</sup>

Figure 3.5 shows the results for three political types: country-first conservatives (CFC), devout and diverse (D&D), and Opportunity Democrats (OPD). Each cell in the heat map displays the difference between the average number of articles recommended by either CBR/CFR and those recommended by the oracle. For example, in the top left panel, we see that the content-based recommender shows on average 113 more immigration/+2 documents than the oracle does to country-first conservatives. By examining these results, we can identify a few trends that characterize the different sorts of bias introduced by either content-based or collaborative filtering recommenders.

For CBR, a key source of bias is **linguistic polarization**. For some topics, there is a clear distinction between the language used in right-leaning articles versus left-leaning articles. For example, in the immigration topic, terms like “illegal” and “alien” are much more likely to appear in right-leaning articles, while terms like “undocumented” are more common in left-leaning articles. In such cases, it will take few training examples for the recommender to develop an accurate model of user preferences, resulting in an over-recommendation of such topics. Furthermore, this can often result in a feedback loop, wherein immigration/+2 articles are recommended and clicked on, further reinforcing the over-recommendation of such articles.

This behavior is most noticeable in the immigration/+2 cell of the first panel of Figure 3.5. We can further see this behavior in Figure 3.6, which shows that content-based recommenders tend to have lower entropy over topics shown than the

---

<sup>2</sup>We select the randomness hyper-parameter  $w$  to result in an oracle with the same click-through rates as the CBR or CFR method it is being compared with.



other two recommendation models for all of the political types at the extreme ends.

For collaborative filtering, we identify two sources of bias. The first is that the distribution of preferences across all users will influence the popularity of some topics over others. For example, across all political types, abortion and trade have high utilities, so they tend to be over-recommended across all user types. We also observe that minority groups tend to be ‘subsumed’ by larger groups. For example, the devout and diverse group appears to be grouped with more right-leaning groups and hence recommended more right articles across almost all topics, whereas the opportunity Democrats are grouped with left-leaning groups and hence are recommended more left articles across almost all topics, as the bottom row of Figure 3.5 shows.

A final source of bias that affects both recommendation systems is the overall makeup of the pool of articles to be recommended. As Table 3.5 indicates, topics such as US military, US election, and impeachment are the most common. The initial bootstrap for CBR and CFR had equal articles from each topic (50 articles from each topic), hence these topics were underrepresented compared to their representation in the overall pool. Thus, articles from these topics tend to be under-recommended by CBR and CFR systems compared to the oracle recommender, which does not have a bootstrap and hence is unaffected by it.

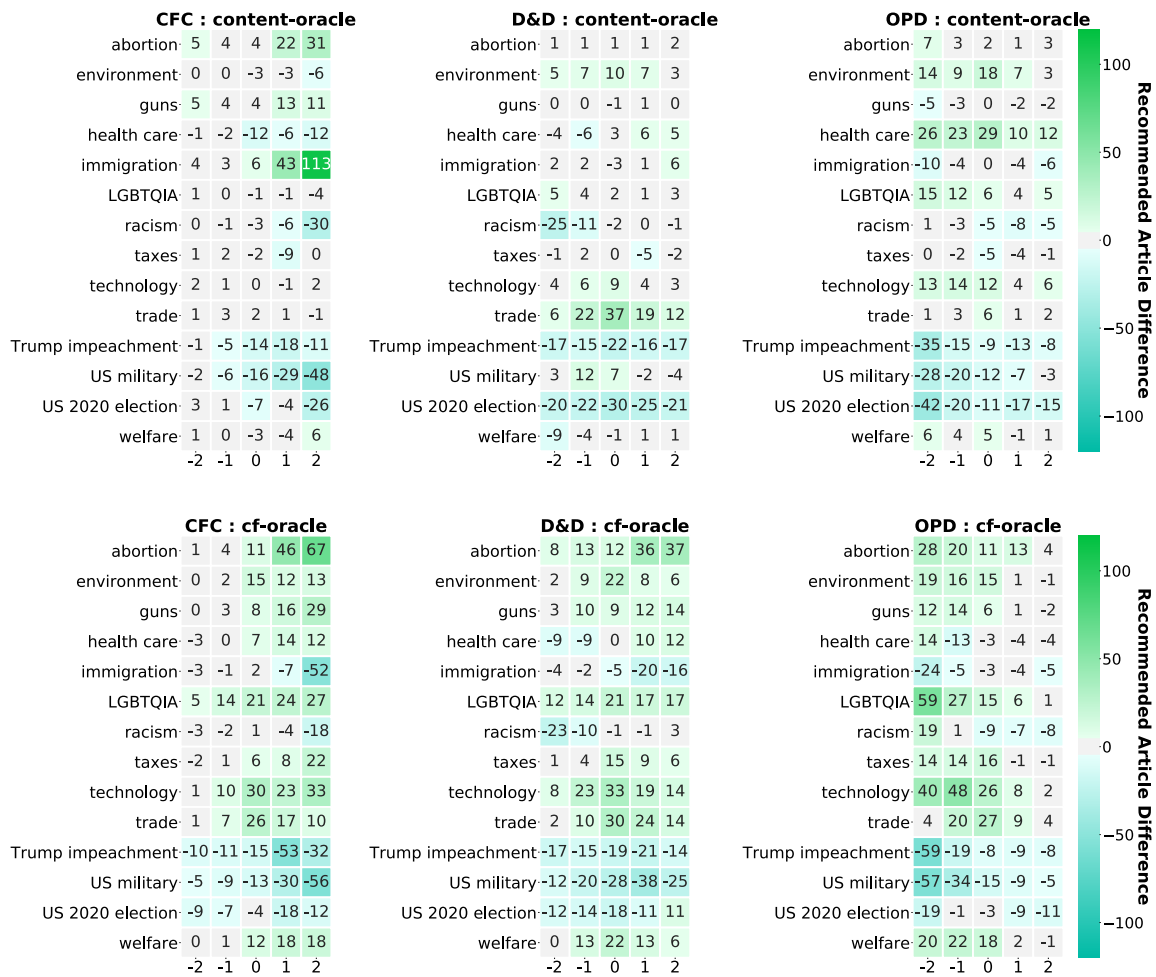


Figure 3.5: Difference in the number of articles recommended by the content-based and collaborative filtering recommenders as compared to the oracle recommender. Results are the average of 1,000 recommendations for 100 users from three user types: country first conservatives (CFC), devote and diverse (D&D), and opportunity Democrats (OPD).

### How does recommendation diversity vary for users with heterogeneous preferences?

The biases described above can also have effects on users with heterogeneous preferences. For example, Devout and Diverse users lean right on most issues, but lean left on issues of race, welfare, and health care. Both content-based and collab-

orative filtering systems under-recommend left leaning articles on these topics, but for different reasons. For collaborative filtering, the devout and diverse users are clustered together with other right-leaning users (e.g., core conservatives). Because those other users have right-leaning preferences for race and welfare, the devout and diverse users are recommended similar articles. Similarly, while the content-based recommender over predicts immigration/+2 for country-first conservatives, the collaborative filtering algorithm instead *under* predicts this category. The CFC type is most distinct because it is more conservative on immigration than "typical" right-leaning users, and so they are grouped together with these more typical users and shown less extreme views on immigration.

The explanation for the content-based recommender is more nuanced. A central issue is that there is keyword overlap across topics that can mislead the recommender. For example, the keyword "baby" correlates with right-leaning articles both for the abortion topic and the health care topic. Because D&D users lean right on abortion issues, after clicking on several right-leaning abortion articles, the recommender may also start to recommend right-leaning health care articles, contrary to their preferences. Similar behavior occurs between the welfare and taxes topic, where the term "socialist" correlates with right-leaning articles for both topics. As D&D users lean right on taxes but left on welfare, left-leaning articles on welfare are under-recommended.

Together, these examples suggest that recommender systems can have a homogenization effect on such users, for example by pushing D&D users to more typical right-leaning articles, and by pushing opportunity democrats to more typical left-leaning articles, even though their true preferences are more mixed. Importantly, we do not see such behavior for the oracle recommender, but rather these are artifacts of the biases of recommendation systems that learn imperfect models of user preferences.

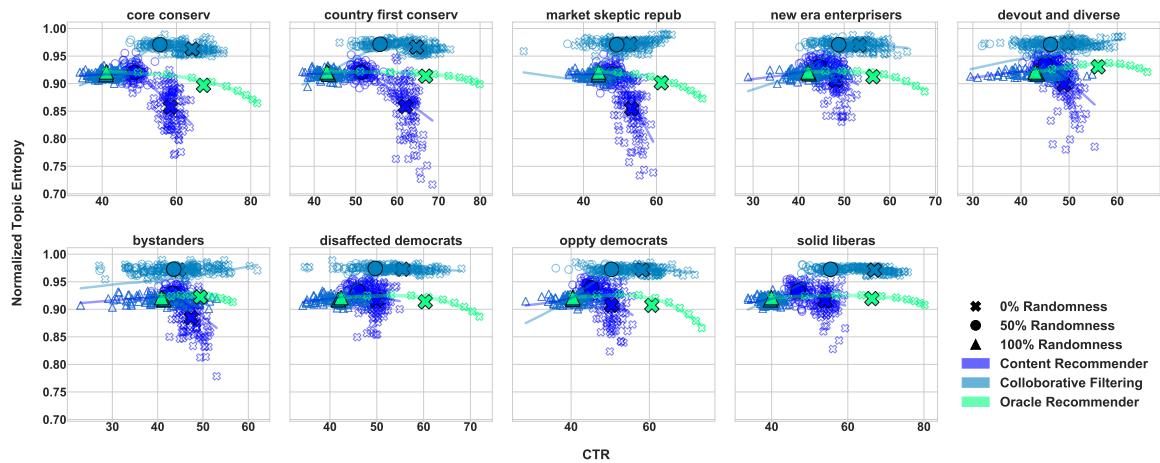


Figure 3.6: Click-through rate vs normalized topic entropy for all recommenders. The content-based recommender exhibits much lower topic diversity than others.

### 3.8 Limitations

We assumed that the news source’s partisan score was reflective of its articles. While this appears to be a reasonable assumption in aggregate, there are undoubtedly some individual errors introduced here. We plan to build partisan score classifiers for each topic to relax this assumption. In the meantime, we need to take into account that the classifiers might introduce their own bias. Further, the user utility model is constant during the recommendation process. Modeling long-term effects requires further assumptions about the causal effect of news consumption on reader beliefs. Typical recommender systems suffer from self-reinforcement because their training data is tainted by skewed recommendations. One might expect that the filter bubbles could cause the user views to become less heterogeneous, further reinforcing and exacerbating filter bubbles. Our work focuses on short-term effects on news consumption, leaving effects on reader beliefs for future work.

### 3.9 Conclusion

In this chapter, we have presented several simulations to understand the relationship between political typology and news recommendation algorithms. We find that users with more extreme views tend to be easier for recommendation systems to model, and thus tend to enjoy higher click-through rates, though this is only possible with less diverse recommendations both in terms of political views and topics. Furthermore, we find that two common classes of recommendation algorithms, content-based and collaborative filtering, can each result in filter bubbles, though of different types and for different reasons. Finally, we find that users with heterogeneous preferences tend to be recommended articles that reflect more homogeneous viewpoints.

# Chapter 4

## Reducing Cross-Topic Political

## Homogenization in Content-Based News

## Recommendation

### 4.1 Introduction

The previous chapter looked at various biases that occur in content and collaborative filtering based news recommendation systems that lead to filter bubble formation. In this chapter we address one of these specific types of biases that affect content-based recommendation systems and propose methods to mitigate its effects. Specifically, we consider users who have diverse political preferences by topic — e.g., users that prefer to read conservative articles on one topic but liberal articles on another. Based on polling by Pew, such users are a sizable portion of the U.S. population [136]. Because content-based recommenders learn text features that correlate with user engagement, we find that they can have a homogenizing effect by recommending articles with the same political lean on both topics. For example, the phrase “extreme right” may appear in liberal articles a user has liked discussing gun control, as well as in liberal articles on immigration that the user has not yet read. If a user in fact prefers conservative articles on immigration, the recommender may thus incorrectly recommend liberal articles due to the presence of the phrase “extreme right.”

This flawed recommendation is further exacerbated when topics are not known *a priori*, which is often the case in political news, where detecting emerging topics is a research challenge of its own [190].

Our goal in this Chapter is to study the phenomenon more closely and propose models to reduce its impact. Our technical approach builds on two threads of machine learning for content based news recommendation – neural attention mechanisms [29, 105, 195] and multitask learning [107, 151, 194]. We adapt these approaches to the cross-topic homogenization problem in two ways: (1) by formulating a penalty term to reduce attention given to topic-independent polarized words; (2) by formulating a secondary prediction task to increase attention given to topic-dependent words.

To do so, we draw upon a collection of 900k news articles as discussed in Chapter 3. We next simulate browsing sessions for users with opposing political preferences for topic pairs, creating a setting in which the system observes more interactions for the first topic than for the second topic. In this way, we are able to measure and focus particularly on the homogenization effect of the first topic on the second. We then propose two attention-based neural network models designed to reduce this homogenization effect. The first model adds a new term to the objective function in order to penalize attention given to topic independent polarized phrases, like “extreme right,” that predict stance across many topics. Conversely, the second approach rewards attention placed on topic dependent terms, like “undocumented” versus “illegal” immigrants, resulting in topic-specific models that are less prone to overgeneralize across topics. We also consider a model that combines the two new learning objectives into a single model. In our experiments using 45 topic pairs, we find that the proposed approach improves accuracy by roughly 5% on the second topic, while still maintaining accuracy comparable to the baseline on the first topic. These results provide evidence that recommendation systems can be designed to mitigate

cross-topic homogenization.

## 4.2 Problem Formulation

We assume a user interaction session consists of a sequence of articles  $\mathbf{a} = \{a_1 \dots a_n\}$  and a corresponding sequence of binary feedback labels  $\mathbf{y} = \{y_1 \dots y_n\}$ , where  $y_i = 1$  means the user liked article  $a_i$ , and  $y_i = 0$  means they did not. We additionally assume that each article  $a_i$  is assigned to exactly one **unobserved** topic  $t_i \in \mathcal{T}$ . To simulate partisan preferences, we assume that a user’s feedback label follows their political preferences for that topic. E.g., if a user prefers conservative articles on topic  $t_i$ , then the feedback label will be  $y = 1$  for conservative articles shown and  $y = 0$  for liberal articles shown.

The phenomenon of interest occurs when a user has opposing political preferences on two topics — e.g., they prefer to read liberal articles on immigration but conservative articles on abortion. This is a challenging case for the recommender — not only are topic assignments unobserved, but topics do not arrive uniformly at random. For example, the system may observe mostly immigration articles and only a few abortion articles. In this setting, the system may incorrectly extrapolate that because the user prefers liberal articles on immigration, they also prefer liberal articles on abortion, leading to poor recommendations. We call this *cross-topic political homogenization*, as the recommender is biased towards showing politically homogeneous articles across the two topics.

To measure system behavior in this setting, we assume we observe a training batch consisting of  $n_1$  article interactions from topic  $t_1$  and  $n_2$  interactions from topic  $t_2$ , where  $n_2 \ll n_1$ . We assume the user has different political preferences for  $t_1$  and  $t_2$  (e.g., they may prefer liberal articles on  $t_1$  and conservative articles on  $t_2$ ). Based on these  $(n_1 + n_2)$  interactions, the system trains a content-based recommender. We



then measure the accuracy of the recommender on a held-out sample of articles from both topics. Accuracy here indicates the fraction of recommended articles that receive positive (simulated) user feedback. We expect overall accuracy to be lower for topic  $t_2$ , both because the system observes fewer user interactions for  $t_2$ , and also because the user’s preferences switch political leanings between topics. This setup can be viewed as a challenging type of cold-start problem; i.e., we have very few training examples from topic  $t_2$ , and those examples conflict politically with the training examples from topic  $t_1$ .

In our experiments, we consider several binary classifiers that predict user interaction label  $y$  given a new article  $a$ . We offer models that attempt to reduce cross-topic homogenization both by reducing attention on topic-independent terms and also by increasing attention on topic-dependent terms.

## 4.3 Methods

We propose multiple network architectures trained in both a single task and multitask fashion to mitigate the effect of cross-topic political homogenization. The network architectures are shown in Figure 4.1. The following subsections discuss these architectures in detail.

### 4.3.1 Baseline 1: Single Task Network (STN)

Our first baseline model performs article classification, where each article  $a_i$  contains  $k$  words  $\{w_{i0} \dots w_{ik}\}$ . We first pass article  $a_i$  through a pre-trained BERT [45] model (uncased, 12-layer, 768-hidden, 12-heads, 110M parameters) to obtain BERT’s word level embeddings  $\{r_{i0} \dots r_{ik}\}$ . We choose  $r_{i0}$  (“CLS” token’s embedding) and pass it through a linear layer  $\langle W_q, b_q \rangle$  with a sigmoid activation to compute the corresponding class probability  $\hat{y}_i$ :

$$\hat{y}_i = \sigma(W_q r_{i0} + b_q) \quad (4.1)$$

where values of  $\hat{y}_i$  close to 1 indicate that the user has high probability of liking article  $a_i$ . This network is trained on the  $(n_1 + n_2)$  labeled articles from prior user interactions, using binary cross-entropy ( $bce(y_i, \hat{y}_i)$ ) as the loss function.

### 4.3.2 Baseline 2: Single Task Attention Network (STAN)

Our second baseline augments the prior model with an attention layer. This model is inspired by the approach proposed by Yang et al. [195], but without the hierarchical aspect. In this network an extra linear layer  $\langle W_a \rangle$  is used to calculate word attention weights  $u_{it}$  given word embedding  $r_{it}$  as the input. We next normalize these word attention weights to get  $\hat{u}_{it}$  by applying a softmax transformation:

$$u_{it} = W_a r_{it} \quad (4.2)$$

$$\hat{u}_{it} = \frac{\exp(u_{it})}{\sum_{t=1}^k \exp(u_{it})} \quad (4.3)$$

Next the attention context vector  $u_i$  is obtained by taking the weighted average between the word attention weights and the article word embeddings:

$$u_i = \sum_{t=1}^k \hat{u}_{it} r_{it} \quad (4.4)$$

This resulting vector  $u_i$  encapsulates all information of the words and their corresponding context in the article. Finally, this vector is passed through an output layer  $\langle W_l, b_l \rangle$  with a sigmoid activation to obtain  $\hat{y}_i = \sigma(W_l u_i + b_l)$ . This network also uses binary cross-entropy loss.

### 4.3.3 Proposed Method 1: Single Task Attention Network with Polarization Penalty (STANPP)

Our first proposed approach modifies the STAN model to reduce attention on topic-independent polarized terms. This is accomplished in a two-step process: first, we identify a candidate set of such polarized terms, then we augment the objective function to penalize attention on them and related terms.

In order to identify topic-independent polarized terms, we assume we have access to a large collection of articles labeled by stance but not by topic (e.g., the partisan lean of a news source provides a strong source of such supervision). Terms that predict stance reliably across this collection are likely to be topic-independent. While any number of feature selection approaches could be used here, in the experiments below we simply select the top 200 terms according to a Chi-Squared test, used to measure the dependence between terms and political stance (see Table 4.1). Terms such as “socialist,” “right-wing,” and “conservative” exemplify the topic-independent, polarized language we wish to reduce attention towards. Additionally, polarizing figures such as Alexandria Ocasio-Cortez and Rudy Giuliani also appear across many topics while strongly correlating with the political stance of the article. (I.e., conservative articles tend to be critical of Alexandria Ocasio-Cortez, while liberal articles tend to be critical of Rudy Giuliani.)

Given this set of  $R$  polarized terms, we next augment the STAN model to reduce the magnitude of attention they and related terms are given. We first embed each of the polarized terms using BERT to obtain word vectors  $\{r_1 \dots r_R\}$ . Then, for each document  $a_i$ , we measure the similarity between the attention context vector  $u_i$  from the STAN model with each of the polarized word vectors  $r_j$  by taking the sigmoid of their dot product  $\sigma(u_i \cdot r_j)$ . The loss for a single document is then a linear combination of the *bce* loss and the average similarity between the attention vector

Table 4.1: Sample of 50 Polarizing Terms used by STANPP

---

abortion accused adam administration admitted alexandria allegations alleged amy andrew biden bush campaigns chuck conservative conspiracy controversial dca democrat democrats donald emails facts failed fbi foundation fox giuliani gop hunter illegal impeach interference joe kamala liberal nancy ocasiocortez pelosi probe radical republican republicans rightwing rudy scandal schiff socialist terrorist vermont

---

and the polarized words.

$$L_{\text{STANPP}} = (1 - \alpha)bce(y_i, \hat{y}_i) + \alpha \left( \frac{1}{R} \sum_{j=1}^R \sigma(u_i \cdot r_j) \right) \quad (4.5)$$

Here  $\alpha$  is a hyperparameter tuned on validation data, as described in the experiments below. Thus, the loss function aims to jointly minimize classification error while making the document representation dissimilar to the polarized terms.

#### 4.3.4 Proposed Method 2: Multitask Attention Network (MTAN)

Rather than penalize topic-independent terms, our second proposed approach instead rewards topic-dependent terms. Since we do not observe topic labels, we cannot use them directly to do so. Instead, we create a multitask model that predicts both the article label as well as a masked word from the article headline. The intuition is that this will encourage the model to pay attention to words that are specific to this article, and that such terms are likely to be topic-dependent.

For the headline word prediction task, we use the “binary negative sampling” approach from word2vec [120]. For each article  $a_i$ , we sample a word  $h_i$  from the headline of the article and mask it. For a pair  $(a_i, h_i)$ , we create a binary classification task to determine whether word  $h_i$  came from the headline of article  $a_i$ .

We then create two samples for each article, one positive  $(a_i, h_i)$  and one neg-

ative  $(a_i, h'_i)$ . The negative headline words are sampled from a vocabulary consisting of all headline terms in our dataset, excluding those present in the headline of  $a_i$ . The candidate headline words  $h_i$  are passed through a pre-trained BERT embedding model to get the corresponding word embedding  $r_{h_i}$ . Next we take the dot product  $g_i$  between the candidate headline word embedding  $r_{h_i}$  and the attention context vector  $u_i$  from our STAN subnetwork<sup>1</sup> to measure how similar these two vectors are:

$$g_i = u_i \cdot r_{h_i}.$$

Finally, this dot product  $g_i$  is passed through a linear layer  $\langle W_c, b_c \rangle$  with sigmoid activation to obtain  $\hat{y}_{h_i}$ , the predicted probability that the candidate headline word  $h_i$  belongs to the headline of the news article  $a_i$ :  $\hat{y}_{h_i} = \sigma(W_c g_i + b_c)$ .

We compute a linear combination of the losses of each of the subnetworks in this architecture as the total loss to optimize:

$$L_{\text{MTAN}} = (1 - \alpha) \cdot bce(y_i, \hat{y}_i) + \alpha \cdot bce(y_{h_i}, \hat{y}_{h_i}) \quad (4.6)$$

where  $y_{h_i}$  is the true binary label for the candidate headline word  $h_i$ , and  $\alpha$  is a hyperparameter tuned on validation data, as described in 4.4.

### 4.3.5 Proposed Method 3: Multitask Attention Network with Polarization Penalty (MTANPP)

This network combines STANPP and MTAN. It has the same architecture as MTAN but with the extra penalty term from the STANPP network:

---

<sup>1</sup>We remove the masked word prior to embedding.

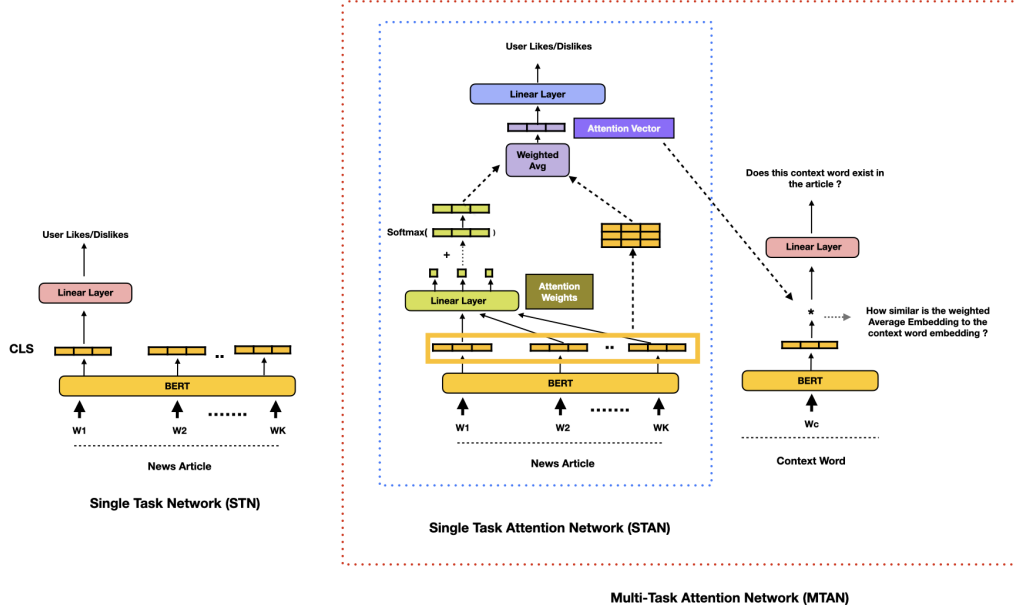


Figure 4.1: Network Architectures for STN, STAN and MTAN

$$L_{\text{MTANPP}} = (1 - (\alpha_1 + \alpha_2)) \cdot bce(y_i, \hat{y}_i) + \alpha_1 \cdot bce(y_{h_i}, \hat{y}_{h_i}) + \alpha_2 \left( \frac{1}{R} \sum_{j=1}^R \sigma(u_i \cdot r_j) \right) \quad (4.7)$$

## 4.4 Experiments and Results

### 4.4.1 Data

We use the news article dataset as defined in Chapter 3. This dataset contains 900k news articles collected from 41 different news sources with corresponding political stance scores ranging over a 5-point scale (-2,-1,0,1,2) where -2 denotes extremely liberal and +2 denotes extremely conservative. To focus on heterogeneous preferences, we drop neutral articles (0) and collapse +2,+1 articles into a “conservative” class, and -2,-1 articles into the “liberal” class. We uniformly sample 100K of these news articles for this study.

To simulate users with heterogeneous political stances across topics, we first need to assign topics to each document. We adopt a simple, transparent approach by using  $k$ -means to cluster the 100k articles into 100 clusters.<sup>2</sup> We first represent each article by concatenating the headline with the first 10 sentences, perform standard tokenization to remove punctuation, then create tf-idf vectors using scikit-learn’s [134] tf-idf vectorizer, with  $min\_df$  of 30 and  $max\_df$  of 0.9. We then run  $k$ -means clustering with  $k = 100$ . To ensure sufficient cluster sizes and sufficient samples from liberal/conservative stances, we filter these clusters to those with at least 400 articles, and sample uniformly so that each cluster has an equal number of liberal and conservative articles. From the clusters that remain, we sample 45 pairs of clusters at random for the basis of our experiments. A manual inspection of these clusters indicates many coherent topics on issues such as immigration, the 2020 election, gun rights, abortion, and healthcare.

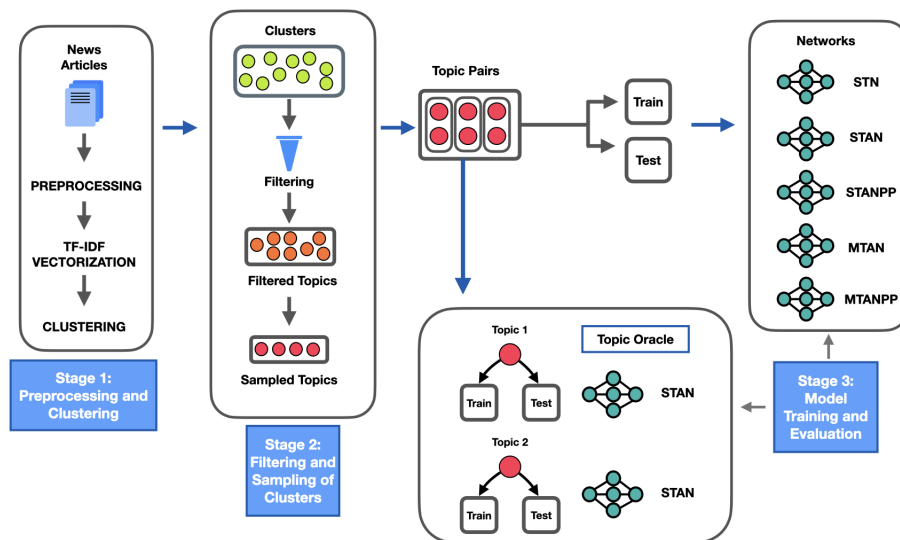


Figure 4.2: Experimental Settings Pipeline

<sup>2</sup>More complicated clustering methods could be used, but our approach is independent of how the topics are determined.

### 4.4.2 Settings

We measure the performance of the above networks using a setting where 90% of the articles in the training and validation data are from a randomly sampled topic 1 and 10% are from a randomly sampled topic 2. The small number of training examples from topic 2 makes this a challenging problem, similar to a cold-start setting. The test set is comprised of an equal distribution of articles from topic 1 and topic 2. We simulate user preferences such that their political preferences for topic 1 articles are the opposite of their preferences for topic 2. We repeat experiments for 45 pairs of topics described in the previous section. Thus, each run consists of a different (topic1, topic2) pair, chosen from our list of discovered topics. Throughout, we refer to topic 1 as the majority topic in the training data and topic 2 as the minority topic, though we run experiments for 45 distinct topic pairs.

To tune each network, we hold out 10% of the training data as a validation set. We perform hyperparameter tuning using grid search over each topic pair using values shown in Table 4.4 and select the best set of parameters based on the accuracy scores on the validation dataset.

After predicting on the test set, we compare the overall accuracy of each approach, as well as investigate how the accuracy varies by topic. Our goal is to improve accuracy on topic 2 without harming accuracy on topic 1. To better assess the ceiling of improvement that is possible, we also fit a model we call the **Topic Oracle**, which, unlike the other methods, is able to observe the topic assignment of each article. To fit this model, we train STAN models separately for topic 1 and topic 2 using the same training data as above. At testing time, we apply the model appropriate for the topic of each test article. The predictions of the Topic Oracle on topic 1 are therefore not influenced by topic 2, and vice versa. This provides a rough upper bound on how well we can expect a model to perform at reducing the impact



Table 4.2: Average model accuracy over 45 topic pairs

<b>Network</b>	<b>Topic 1 Accuracy</b>	<b>Topic 2 Accuracy</b>	<b>Total Accuracy</b>
UNBERT	0.647	0.447	0.547
STN	0.613	0.489	0.551
STAN	0.682	0.498	0.590
STANPP (ours)	0.664	0.525	0.594
MTAN (ours)	<b>0.693</b>	0.531	0.612
MTANPP (ours)	0.687	<b>0.552</b>	<b>0.619</b>
Topic Oracle	0.701	0.596	

of cross-topic homogenization. We additionally compare with **UNBERT** [202], a representative example of recent work using BERT for news recommendation. This approach learns a BERT-based representations of a user based on the articles they’ve liked, then pairs this with an article representation to predict whether they will like a new article. As with the other models, its hyperparameters are tuned on validation data.<sup>3</sup> We use the pytorch [131] and huggingface [186] libraries to implement our networks. All our models are trained using a Nvidia RTX 3090 GPU over a period of 5 days. An overview of our experimental setup is given in Figure 4.2.

### 4.4.3 Results

Table 4.2 reports the accuracy of each approach averaged across 45 different topic pairs. Figure 4.3 shows boxplots of the same results to visualize the variance across topic pairs and Table 4.3 shows additional measures including precision, recall, and F1.

By comparing STAN and Topic Oracle, we can see the considerable impact cross-topic homogenization can have. For Topic 2, which has fewer training samples, accuracy drops from .596 to .498 when training data from topic 1 is included, indicat-

---

<sup>3</sup>Following the implementation [202], this model is trained using only article headlines, due to its high computational complexity by sequence length.

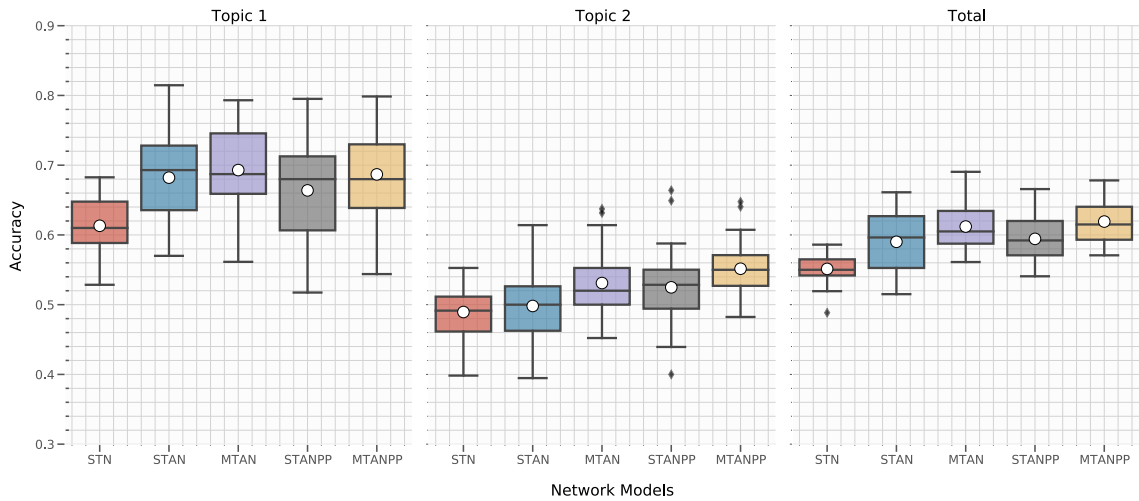


Figure 4.3: Distribution of test accuracies across 45 topic pairs for STN, STAN, MTAN and MTANPP

ing that the content from topic 1 is prohibiting an accurate model for topic 2 articles. This shows how the recommendations for an emerging topic can be quite poor, as the system defaults to recommendations in line with preferences on prior topics.

We observe that on average the proposed STANPP, MTAN, and MTANPP networks tend to have higher accuracy (3%-6%) for recommending topic 2 articles compared to the baseline STN and STAN networks. We also observe an increase in accuracy across topic 1 recommendations for the MTAN, STAN and STANPP networks compared to the STN network (1%-8%). Furthermore, combining STANPP and MTAN into MTANPP appears to do as well or better than each in isolation.

We computed pairwise  $t$ -tests for each pair of models. For topic 2 accuracy, all results are significant at the 5% level except for the differences between STN and STAN and between MTAN and STANPP. For total accuracy, all results are significant except for STAN and STANPP. For topic 1 accuracy, three differences are insignificant: STAN vs MTAN, STAN vs MTANPP, MTAN vs MTANPP. We also see that compared to UNBERT, our proposed approaches perform better across all metrics of comparison.

Table 4.3: Average Network Performance across 45 Topic Pairs with Additional Metrics

Network	Score Type	F1	Precision	Recall	AUC
UNBERT	<b>Topic 1</b>	0.629	0.635	0.634	0.646
	<b>Topic 2</b>	0.410	0.429	0.404	0.447
	<b>Total</b>	0.522	0.535	0.519	0.547
STN	<b>Topic 1</b>	0.602	0.620	0.610	0.613
	<b>Topic 2</b>	0.461	0.482	0.483	0.489
	<b>Total</b>	0.535	0.555	0.546	0.551
STAN	<b>Topic 1</b>	0.670	0.693	0.660	0.682
	<b>Topic 2</b>	0.426	0.476	0.433	0.498
	<b>Total</b>	0.563	0.608	0.547	0.590
MTAN (ours)	<b>Topic 1</b>	0.676	<b>0.716</b>	0.653	<b>0.693</b>
	<b>Topic 2</b>	0.473	0.522	0.466	0.531
	<b>Total</b>	0.583	<b>0.637</b>	0.559	0.612
STANPP (ours)	<b>Topic 1</b>	0.661	0.673	0.663	0.664
	<b>Topic 2</b>	0.466	0.521	0.466	0.525
	<b>Total</b>	0.573	0.612	0.564	0.594
MTANPP (ours)	<b>Topic 1</b>	<b>0.681</b>	0.696	<b>0.676</b>	0.687
	<b>Topic 2</b>	<b>0.522</b>	<b>0.563</b>	<b>0.509</b>	<b>0.552</b>
	<b>Total</b>	<b>0.605</b>	0.630	<b>0.592</b>	<b>0.619</b>

By comparing with the Topic Oracle, we see that the best of the proposed models approaches the accuracy of the topic aware oracle (topic 1: .693 vs .701; topic 2: .552 vs .596). These results also highlight the difficulty of this problem setting, which we attribute to two factors: First, the training data have few examples from topic 2 (often less than 100). Second, the article collection contains a wide variety documents, most of which are not opinion pieces. Thus, the difference between -1 and +1 articles can be difficult to discern based on linguistic evidence, requiring instead a nuanced understanding of the political and policy landscape.

### Shift in Attention

To further understand model behavior, we examine how attention varies by model to confirm whether the loss functions are having the intended effects. To do

Table 4.4: Network hyperparameters considered.

Hyperparameter Values	
learning rate	1e-2 to 1e-5
epochs	3, 5, 10, 20, 30, 50
batch size	8, 16, 32
dropout	0.0, 0.1, 0.3, 0.5
l2 penalty	0.01, 0.05, 0.1
loss weight ( $\alpha$ )	0.01, 0.03, 0.05, 0.07, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9

so, we analyze the change in the attention rank of terms, where the ranking is done based on cumulative attention scores. For a term  $w_t$ , let  $\hat{u}_{it}$  be the corresponding normalized word attention weight for the term  $w_t$  contained in the news article  $d_i$ . Assume there are  $V$  unique terms in the vocabulary. Then, the cumulative attention  $C_t$  for term  $w_t$  across  $n$  documents is calculated as:

$$C_t = \frac{\sum_{i=1}^n \hat{u}_{it}}{\sum_{j=1}^V \sum_{i=1}^n \hat{u}_{ij}} \quad (4.8)$$

For illustrative purposes, we analyze the shift in ranks based on these cumulative attention scores for a topic pair where topic 1 discusses **gun control** and topic 2 discusses **climate change**. Table 4.5 shows the top 30 terms with the highest cumulative attention scores using our attention network models. For the STAN network, most of the top terms are either very specific to topic 1 (e.g., gun, shooting, firearm) or are terms that are polarized and occur across documents (e.g., trump, democrats, left). For the STANPP network we see that terms that are ranked highly are more topic specific (e.g., gun, violence, rifle) and have more focus on topic 2 (e.g., fossil, protection, environmental, climate, fuel, energy, emissions). We see similar trends for the MTAN and MTANPP networks. This indicates that both the single task attention network with the updated loss and the multitask attention network seem to

shift attention away from more polarized terms that occur across topics and towards terms that are more topic specific.

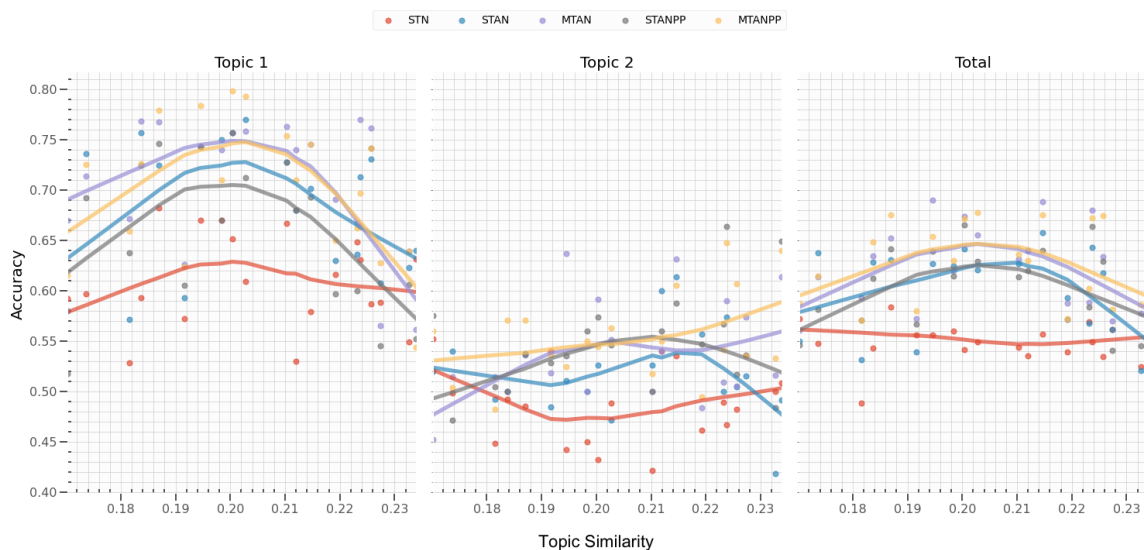


Figure 4.4: Topic similarity vs test accuracy for 20 topic pairs. The topic similarity is measured using Jaccard similarity between sets of overlapping terms for a given topic pair. The trend lines are generated using a lowess regression model.

### Effect of topic similarity

We next investigated how the models perform based on the similarity between topic 1 and topic 2. Intuitively, we expect that if the topics are very different, and share few terms, then there is little opportunity for homogenization, and thus we do not expect our models to provide much improvement. On the other hand, if the topics are too similar, then disentangling them will prove challenging. To measure this, we use a simple method to quantify the overlap in predictive terms across two topics. We fit two logistic regression classifiers, one per topic, to predict the political stance of each article. We then select the top terms from each classifier by picking those whose coefficient has magnitude greater than 0.01.<sup>4</sup> Given these two sets of terms,

---

<sup>4</sup>This is a somewhat arbitrary threshold; similar trends were found with different thresholds.

we compute their Jaccard similarity to measure the overlap of each cluster pair. Thus, topics are similar if they share terms predictive of political stance. Figure 4.4 shows the results for 20 cluster pairs, fit with lowess regression to visualize trends. While there is noticeable variance across cluster pairs, the trends generally match our expectations. The biggest gains occur in the middle of the  $x$ -axis, where the topics are neither too similar nor too dissimilar. In future work, it may be helpful to develop diagnostics to determine the divergence between the training and testing set to guide model tuning.

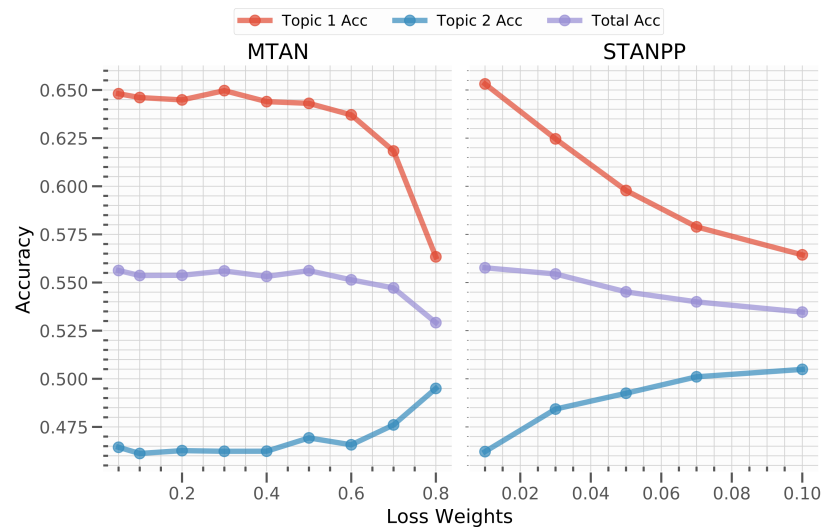


Figure 4.5: Average test accuracy of 45 topic pairs vs loss weights ( $\alpha$ ) used in STANPP and MTAN

### Effect of loss weights

Both the STANPP and MTAN models use a linear interpolation of loss terms (Equations 4.5 and 4.6). While these  $\alpha$  weights are tuned on the validation set, in order to understand their impact on accuracy, we additionally plot results as we vary the  $\alpha$  terms in each equation. We first fix all other hyperparameters in Table 4.4 found by optimizing on the validation set. Then, we enumerate  $\alpha$  values and plot accuracy on the test set in Figure 4.5. We observe that STANPP performs best with small values of  $\alpha$ . When  $\alpha$  is too large, the accuracy on topic 1 begins to drop.

While topic 2 accuracy continues to increase, the cost to topic 1 accuracy begins to overwhelm the tradeoff. In contrast, MTAN appears relatively stable over a range of  $\alpha$  values, until a drop-off once  $\alpha$  is greater than 0.7. This suggests that MTAN may be more suitable in settings where it is difficult to carefully tune  $\alpha$ .

## 4.5 Conclusion

In this chapter, we have identified a specific mechanism that can lead to political homogenization in news recommendation systems, and we have proposed attention-based neural networks to reduce this behavior. The proposed approach exhibits reduction in the impact of political homogenization for simulated users with opposing political leanings across topics. While promising, a considerable amount of work is needed to better understand this phenomenon. First of all, user studies are required to both confirm the propensity of such homogenization as well as to better measure the impact of the proposed approaches, the user study can resemble a randomized control trial where we would have a control and treatment group of users and the treatment effect would be to expose these users to our adapted models (STANPP, MTAN and MTANPP). Second, there is a need to focus on the existing debate about the role of attention in explaining model decisions [84, 161], although these issues appear to be more important in tasks of greater complexity than text classification. Finally, news sources are not monolithic in the viewpoints they publish, which can introduce some bias in the article labels [63], although in aggregate we expect this to have a limited effect.

A natural extension of the work discussed in both Chapter 3 and the current chapter is to perform user-studies in order to confirm our findings. We perform a user based study as discussed in Chapter 1 under section 1.3 where we try to reduce the impact of filter bubbles using interaction and transparency mechanisms. From

this study we find that providing users with these mechanisms improves the user's awareness about filter bubbles.



Table 4.5: Top 30 terms with highest attention scores for a topic pair discussing climate change and gun control.

STAN			STANPP			MTAN			
Terms	Avg Attention	Terms	Avg Attention	Terms	Avg Attention	Terms	Avg Attention	Terms	Avg Attention
gun	0.144	democrat	0.010	fossil	0.256	second	0.012	gun	0.564
rouke	0.044	people	0.010	protection	0.137	fuel	0.011	rouke	0.059
percent	0.024	firearm	0.010	environmental	0.087	white	0.011	guns	0.030
trump	0.018	united	0.010	gun	0.084	rifle	0.010	firearm	0.026
background	0.018	fossil	0.009	rights	0.052	trump	0.009	sanders	0.025
guns	0.017	carbon	0.009	control	0.048	energy	0.008	said	0.020
said	0.015	mass	0.008	violence	0.046	donald	0.006	rep	0.013
sanders	0.014	rights	0.008	political	0.030	emissions	0.005	firearms	0.011
private	0.013	left	0.008	democratic	0.026	national	0.005	activists	0.009
industry	0.012	government	0.008	amendment	0.022	carbon	0.004	trump	0.009
shooting	0.012	democratic	0.008	world	0.019	right	0.004	weapons	0.008
firearms	0.011	joe	0.008	change	0.019	fuels	0.004	just	0.008
democrats	0.011	thursday	0.008	climate	0.019	assault	0.004	background	0.007
rep	0.010	change	0.007	public	0.016	elizabeth	0.003	released	0.006
gov	0.010	dead	0.007	nearly	0.015	years	0.002	don	0.006
								rifle	0.006
								joe	0.005
								work	0.005
								energy	0.005
								bernie	0.005
								left	0.005
								wednesday	0.005
								senator	0.004
								thursday	0.004
								deal	0.004
								activist	0.004
								doesn	0.004
								rights	0.003
								republican	0.003
								rifles	0.003

# Chapter 5

## Characterizing Online Criticism of Partisan News Media using Weakly Supervised Learning

### 5.1 Introduction

The earlier Chapters (3,4) focused on the impact of filter bubbles and recommendation systems on news engagement behavior in the short term. However, this Chapter takes an alternate approach to examine a particular type of news engagement behavior on social media. The study is based on a decade’s worth of observational data collected from Twitter, and it analyzes situations where users express criticism towards the news media they interact with. The Chapter delves into methods that can be utilized to identify such scenarios in online social networks, moving away from the simulation-based studies of the previous Chapters.

Public distrust and animosity towards news media contribute to hyperpartisanship, polarization, and misinformation [129, 145]. These effects have been exacerbated: a 2022 Pew survey reports 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 [104]. This distrust also has a partisan divide — 77% of Democrats versus 42% of Republicans trust national news organizations. How-

ever, we lack a compelling understanding of the connections between media-targeted criticism/distrust and the health of the information ecosystem, particularly across temporal and partisan dimensions.

We highlight a widespread but overlooked nuance of online communication. Efforts to characterize the level of political polarization in online social media platforms (see, for example, [38, 64, 65]) have largely ignored the *intent of engagement* (shares, mentions, etc.) – e.g., the distinction between ridiculing or agreeing with a news source, an example of which is highlighted in Figure 5.1. This Chapter thus discusses data and methods to detect engagement intent, specifically whether a social media user is expressing criticism/distrust of a news source. Based on our collection of over 3.5M Twitter-based news source mentions over the past ten years, we train a neural network to classify the news sharing intent of each mention based on its linguistic context as well as the user’s past engagement behavior. Because labeled data is scarce, we apply weak supervision approaches, relying on noisy labeling functions based on keywords and user-based features to train the classifier. After validating the classifier on a smaller number of manually-labeled examples, we then apply it to all of our historical data, allowing us to ultimately analyze the prevalence of media-critical tweets by user and 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 novel dataset of 3.5M tweets that engage with one of 522 news sources over a ten year period. We will share tweet IDs, news sources, and inferred sharing intent to foster future research in this area.
- **Weak supervision:** We find that weak supervision using both text and user-based heuristics can provide accurate labels (89% F1) with modest coverage (48%). Fitting a weakly-trained classifier improves to full coverage while maintaining high accuracy



Figure 5.1: Example tweet critical of a news source.

(84% F1).

- **Effects on Polarization Estimates:** We find that adjusting for these critical tweets provides a different picture of the diversity of user news engagement. Users who previously appeared to engage with diverse news sources are seen as more hyperpartisan when accounting for tweets that are critical of opposing news sources.
- **Criticism by user, news source, and time:** Applying the classifier to the larger dataset, we find that the most criticized news sources are CNN and MSNBC on the left and Fox News and OANN on the right; that hyperpartisan 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 and during Brett Kavanaugh’s Supreme Court nomination).

## 5.2 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 sought to identify a diverse set of Twitter users who engage with political news, using the following steps:

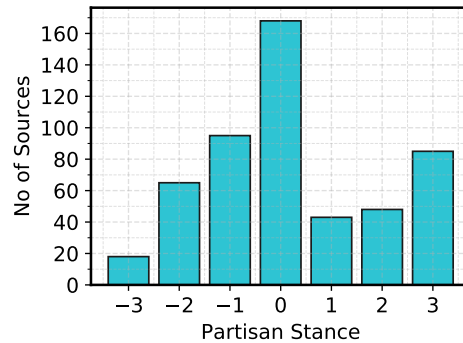


Figure 5.2: News Sources Partisan Distribution

**Step 1: Collect news sources.** We collected 419 English-based news sources from *allsides.com*, a media rating site used often in prior work [10, 106, 113, 171]. 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 capture a wider range of media quality, we added to this set 103 sources identified by Guess et al. [70] as news sources of low reliability (e.g., those that publish stories determined to have little factual basis). We denote the partisan stance of these news sources as  $-3$  (unreliable liberal sources) and  $+3$  (unreliable conservative sources). Figure 5.2 shows the distribution of partisan stances over the final set of 522 news sources. Before proceeding to the next step, we retrieved the Twitter handle and url for each news source.

**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 diversify the user set to identify a broader 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 tweets collected above.

**Step 3: Filter users.** To allow us to study long term trends, we retained only those users identified in the previous step with accounts at least five years old. Additionally 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 [40] for this filtering step, 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 ( $\leq 1000$ )
2. Following Size ( $\leq 1000$ )
3. Daily Tweet Activity ( $\leq 10$ )
4. Total Tweets authored during the life of the account ( $\geq 1000$  and  $\leq 30000$ )

After filtering, we finally sampled users to diversify by partisan stance and news source, ensuring that one or two news sources do not dominate the dataset. To do so, we sampled  $\sim 600$  users from each partisan stance (based on the 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 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 next collected each user’s entire timeline to identify a larger set of news engagement tweets. From these 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 from 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, 3.5M engage with one of

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

the 522 news sources (Table 5.1), 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 §5.7 for more discussion of such limitations.)

### 5.3 Problem Formulation

With the above data, we now 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. 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 criticizing 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 [55], we find that the rate of criticism is higher for more partisan news sources. We thus restrict our attention to the 216 news sources with partisan stance in  $\{-3, -2, +2, +3\}$ , to focus on the most salient subset of data. 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 universe consists of 1.2M tweets that are candidates for classification.

## 5.4 Methods

Given the lack of labeled data, and 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 dataset [98, 103, 147].

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. Similarly, 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. 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. Below we describe the labeling functions and classification methods in turn.



### 5.4.1 Labeling Functions

We define a labeling function  $\phi$  as a collection of heuristics that maps a given tweet to a corresponding label  $y \in \{0, 1, -1\}$ . Here 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 ( $\phi_{up}$ ):

This labeling function relies on a user’s historic news engagements and the political accounts they follow.<sup>1</sup> 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 they engage with. 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.<sup>2</sup> This labeling function annotates each news engagement tweet that is aligned with the user’s partisan lean as 0, and tweets of the opposite partisan lean as 1. This function abstains for tweets from users for whom we could not infer partisan lean based on the thresholds above. Formally we denote this labeling function as  $\phi_{up}$ , 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.

Here  $e_{up}^a$  represents the heuristic that measures the partisan distribution of all the historic tweets of a given user  $u_j$  where they engage with a news source and compares the difference in the normalized distribution between the conservative  $c_{ac}^{u_j}$

---

<sup>1</sup>We use a dataset of politician Twitter accounts with party affiliation – <https://www.propublica.org/datastore/dataset/politicians-tracked-by-politwoops>

<sup>2</sup>The 90% threshold is a tuning parameter to tradeoff precision and coverage, described at length in the Appendix.

and liberal engagements  $c_{al}^{u_j}$  against a chosen threshold value  $\delta_a$ . It assigns a value of 1 if the current 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 otherwise. For example, if the user is strongly conservative ( $c_{ac}^{u_j} \gg c_{al}^{u_j}$ ) then the minority stance would be liberal and vice versa. We also check to see if the total number a user's tweets are greater than a threshold  $\rho_a$  to make sure enough user information is present for the heuristic to work effectively.

$$e_{up}^a(t_i^u) = \begin{cases} 1, & \text{if } |c_{ac}^{u_j} - c_{al}^{u_j}| \leq \delta_a \wedge s_i^{u_j} = ps_{min}^a(u_j) \\ -1, & \text{if } c_{ac}^{u_j} + c_{al}^{u_j} < \rho_a \\ 0, & \text{otherwise} \end{cases} \quad (5.1)$$

$e_{up}^{drt}$  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). This heuristic is shown in equation 5.2.

$$e_{up}^{drt}(t_i^{u_j}) = \begin{cases} 1, & \text{if } |c_{drtc}^{u_j} - c_{drtl}^{u_j}| \leq \delta_{drt} \wedge s_i^{u_j} = ps_{min}^{drt}(u_j) \\ -1, & \text{if } c_{drtc}^{u_j} + c_{drtl}^{u_j} < \rho_{drt} \\ 0, & \text{otherwise} \end{cases} \quad (5.2)$$

$e_{up}^{pf}$  represents a heuristic that measures the partisan distribution of a set of politician Twitter accounts a user follows  $u_j^{pf}$  and compares the difference in the normalized distribution between the conservative politician accounts  $c_{pc}^{u_j}$  and liberal politician accounts  $c_{pl}^{u_j}$  against a chosen threshold value  $\delta_{pf}$ . It assigns a value of 1 if the current 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  $ps_{min}^{pf}(u_j)$  based on the user's followed politician

Table 5.2: Threshold Parameters for  $\phi_{up}$ 

Threshold Parameter	Values
$\delta$	0.05, 0.1, 0.15, 0.2, 0.25, 0.3, 0.35
$\rho$	5, 10, 15, 20, 25, 30

accounts and 0 otherwise. We also check to see if the total number of politician Twitter accounts the user follows are greater than a threshold  $\rho_{pf}$ .

$$e_{up}^{pf}(t_i^{u_j}) = \begin{cases} 1, & \text{if } |c_{pc}^{u_j} - c_{pl}^{u_j}| \leq \delta_{pf} \wedge s_i^{u_j} = ps_{min}^{pf}(u_j) \\ -1, & \text{if } ||u_j^{pf}|| < \rho_{pf} \\ 0, & \text{otherwise} \end{cases} \quad (5.3)$$

Using the results of each of these heuristics, the labeling function  $\phi_{up}$  uses a unanimous voting scheme to assign the final labels for a user’s tweet  $t_i^u$ .

$$\phi_{up}(t_i^{u_j}) = \begin{cases} 1, & \text{if } e_{up}^a(t_i^{u_j}) == e_{up}^{drt}(t_i^{u_j}) == e_{up}^{pf}(t_i^{u_j}) == 1 \\ 0, & \text{if } e_{up}^a(t_i^{u_j}) == e_{up}^{drt}(t_i^{u_j}) == e_{up}^{pf}(t_i^{u_j}) == 0 \\ -1, & \text{otherwise} \end{cases} \quad (5.4)$$

All the threshold parameters ( $\delta, \rho$ ) are tuned over a manually annotated dataset across different parameter ranges (Table 5.2).

### Text Features ( $\phi_{tt}$ ):

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. We denote

this labeling function as  $\phi_{tt}$ , 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 that utilizes a collection of keywords and phrases to label both the positive ( $e_{tt}^{d+}$ ) and negative 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 heuristics where a single item in this set contains a rule of the form  $w_r : \{w_0^o, w_1^o, \dots, w_m^o\}$ , here  $w_r$  represents a root word and  $w_i^o$  represents an optional word. These heuristics first check if the tweet text contains  $w_r$  and then checks to see if any of the  $w_i^o$ 's are also present in the tweet text. Similar to  $e_{tt}^d$ , this heuristic set contains different sets of rules for both the positive ( $e_{tt}^{c+}$ ) and negative ( $e_{tt}^{c-}$ ) class.

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

Using the results from these heuristics, the labeling function  $\phi_{tt}$  uses a "logical or" scheme to assign the final labels for a user's tweet  $t_i^{u_j}$ .

$$\phi_{tt}(t_i^{u_j}) = \begin{cases} 1, & \text{if } e_{tt}^{d+}(t_i^{u_j}) == 1 \vee e_{tt}^{c+}(t_i^{u_j}) == 1 \vee e_{tt}^{p+}(t_i^{u_j}) == 1 \\ 0, & \text{if } e_{tt}^{d-}(t_i^{u_j}) == 0 \vee e_{tt}^{c-}(t_i^{u_j}) == 0 \vee e_{tt}^{p-}(t_i^{u_j}) == 0 \\ -1, & \text{otherwise} \end{cases} \quad (5.5)$$

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 positive and negative classes both fire, we assign a label of -1. For the complete list of heuristics refer Table 5.4.

Table 5.3: Labeling function output on unlabeled data

<b>function</b>	<b>pos</b>	<b>neg</b>	<b>abstain</b>
$\phi_{tt}$	5,087 (3%)	39,356 (27%)	103,594 (70%)
$\phi_{up}$	4,600 (3%)	34,601 (23%)	108,836 (74%)
$\phi_{un}$	7,872 (5%)	63,181 (43%)	76,984 (52%)

### Union of the above ( $\phi_{un}$ ):

This labeling function takes the union of the prior two functions  $\phi_{up}$  and  $\phi_{tt}$ , 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 §5.3 and report the estimated label distribution in Table 5.3. We observe that  $\phi_{tt}$  and  $\phi_{up}$  exhibit similar levels of coverage, labeling 30% and 26% of the data, respectively, and each labeling 3% of the data as positive. As a fraction of the labeled instances, excluding abstentions, each method labels about 11% as positive. The union function  $\phi_{un}$  improves coverage to 48%, while also assigning about 11% of the labeled instances to the positive class. We will discuss the accuracy of these labeling functions on manually annotated data in §5.5.

### 5.4.2 Classification models

In this section we discuss the different models we train 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.

Table 5.4: Heuristics for  $\phi_{tt}$ 

Heuristic Type	Heuristics	Class 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 : news, stories, reporting, narrative, media fake : news, reports, stories, story, report, media, network hoax : news, reports, stories, story, report, media conspiracy : theories, theory fictitious : news, report, story, narrative, media misrepresent : news, facts, truth, story, report, narrative, media misinform : public,people,america exaggerate : news, report, story, narrative mislead : public,people,america made up : lies, crap, shit make up : lies, crap, shit brainwash : people, public, america spread : lies, propaganda, conspiracies, shit, fear deceive : people, public, america biased : news, report, narrative, network, media, shit one sided : news, report, narrative, network, media bullshit : news, report, narrative, network, media crap : news, report, narrative, network, media shit : news, report, narrative, network, media garbage : news, report, narrative, network, media	1
	breaking : news, exclusive, report, story watch : now, live good : news, report, story, journalism, narrative, article, piece, video great : news, report, story, journalism, narrative, article, piece, video best : news, report, video inspiring : news, report, story, journalism, narrative, article, piece, video incredible : news, report, story, journalism, narrative, article, piece, video real : news, report, story, journalism, narrative, article, piece thanks : news, report, story, journalism, narrative, article, piece thx : news, report, story, journalism, narrative, article, piece latest : news, report, story, narrative, article, piece, scoop fantastic : news, report, story, narrative, article, piece, scoop	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 5.5: User Based Features

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

**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). A complete list of features is available in Table 5.5.

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:

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

**Text Network :** This network uses the actual text of the tweet (and the 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 [13] to obtain word level representations  $\{a_0, a_1, \dots, a_p\}$ , 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:

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

**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.

### 5.4.3 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 [200]. 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) [43], IBCC [98], EBCC [103],



Table 5.6: Labeling function accuracy on test data.

<b>Function</b>	<b>Coverage</b>	<b>F1</b>	<b>Prec</b>	<b>Rec</b>	<b>Acc</b>
$\phi_{up}$	0.388	0.845	0.862	0.853	0.860
$\phi_{tt}$	0.391	0.869	0.879	0.869	0.869
$\phi_{un}$	0.600	0.888	0.888	0.888	0.888

and Data Programming (DP) [147].<sup>3</sup> These methods leverage agreements and conflicts between the different labeling functions in order to reduce label noise. We use both  $\phi_{up}$  and  $\phi_{tt}$  generated labels to train these models, and we use the implementations provided by WRENCH [200](with default hyperparameters), a weak supervision benchmark library for classification tasks.

## 5.5 Experiments and Results

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, this resulted in only a small number of positive examples. Thus, we augment these data by using our labeling functions to identify a sample of positive and negative 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 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.

To train each weakly supervised model, we apply our labeling functions ( $\phi_{up}, \phi_{tt}, \phi_{un}$ ) to the unlabeled tweets filtered according to §5.3. Then, we sample a balanced distribution of labels across positive and negative classes for each of the three

---

<sup>3</sup>While other methods exist, they require require a larger number of labeling functions to be effective

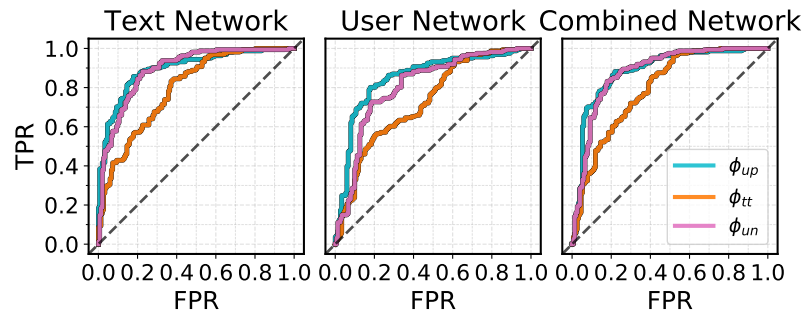


Figure 5.3: ROC curves of network models trained using different labeling functions.

labeling functions. The final number of weakly supervised training examples used for each labeling function is 9,200 for  $\phi_{up}$ , 10,174 for  $\phi_{tt}$ , and 17,962 for  $\phi_{un}$ . We train all possible combinations of our different network models and labeling functions, using the Adam [100] optimizer. We tune each network across different hyperparameter values as shown in Table 5.8. And select the best parameters based on validation accuracy to identify our best models. These networks are implemented in Pytorch [132].

Table 5.6 first reports the performance of our labeling functions alone on the test set. We observe that the labeling functions are rather reliable (.845-.888 F1), with the text heuristics slightly more accurate than the user heuristics; both have modest coverage ( $\sim 40\%$ ). The union function offers an apparent improvement over both, increasing coverage by  $\sim 20\%$  and accuracy by 1-2%.<sup>4</sup>

Turning to the classification models, Figure 5.3 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  $\phi_{tt}$ , which may be due to the poor generalization of the keyword-based heuristics.

---

<sup>4</sup>The coverage is higher here than in Table 5.3 since the manually labeled data includes both uniformly sampled tweets as well as those annotated by the labeling functions.

Network	ROC AUC
Text Network + $\phi_{up}$	0.800 $\pm$ 0.028
Text Network + $\phi_{tt}$	0.719 $\pm$ 0.008
Text Network + $\phi_{un}$	0.810 $\pm$ 0.017
Text Network + DS	<b>0.840 <math>\pm</math> 0.007</b>
Text Network + IBCC	0.822 $\pm$ 0.014
Text Network + EBCC	0.836 $\pm$ 0.006
Text Network + DP	0.812 $\pm$ 0.013
User Network + $\phi_{up}$	0.749 $\pm$ 0.049
User Network + $\phi_{tt}$	0.657 $\pm$ 0.022
User Network + $\phi_{un}$	0.744 $\pm$ 0.025
Combined Network + $\phi_{up}$	0.810 $\pm$ 0.015
Combined Network + $\phi_{tt}$	0.723 $\pm$ 0.006
Combined Network + $\phi_{un}$	0.796 $\pm$ 0.008
Combined Network + DS	0.816 $\pm$ 0.015
Combined Network + IBCC	0.784 $\pm$ 0.038
Combined Network + EBCC	0.810 $\pm$ 0.023
Combined Network + DP	0.826 $\pm$ 0.015

Table 5.7: Test set ROC AUC for combinations of model, labeling function, and label denoising methods.

Table 5.7 shows each classification model’s ROC AUC score along with the Label Denoising enhancements.<sup>5</sup> 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 models (*Dawid Skene*, *IBCC*, *EBCC*, and

---

<sup>5</sup>Other metrics are available in the Appendix, Table 5.9.

Table 5.8: Hyperparameter Values for Experiments

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

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

Table 5.9: Test set Performance for combinations of model, labeling function, and label denoising methods.

*Data Programming*) helps improve 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 average ROC AUC score of 0.840 across different random seed settings.

## 5.6 Analysis of Media-Targeted Criticism

As our main interest for this work is to help answer important questions regarding issues of criticism shown towards partisan news media, we use our best performing model to infer which tweets indicate criticism, from all data filtered according to §5.3. The resulting dataset contains  $\sim 1.2$  million tweets, of which  $\sim 1.16$  million were labelled as non-criticism, and  $\sim 45K$  were labelled as criticism. With these data, we next consider how news criticism varies by user, news source, and time.

### 5.6.1 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

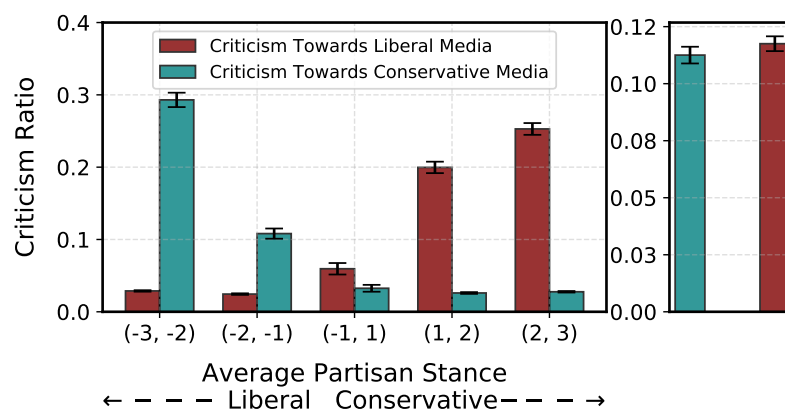


Figure 5.4: Criticism ratio by partisan stance of the user (**left panel**) and in aggregate (**right panel**).

by the stance of news sources a user engages with through direct retweets.<sup>6</sup> For each bin, we compute the criticism ratio, defined as the proportion of news engagement tweets that criticize a news source. The results in Figure 5.4 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 right panel of Figure 5.4 shows the overall criticism ratios, ignoring user bins. This indicates a slightly higher level of criticism towards liberal media than towards conservative media, though the differences do not appear to be significant.

### 5.6.2 Criticism by News Source

Figure 5.5 plots the criticism ratio for the top ten most mentioned news sources from each partisan stance. Among reliable liberal news media, **CNN** receives the most 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

---

<sup>6</sup>We assume direct retweets is a reliable indicator of user support for a particular news source.

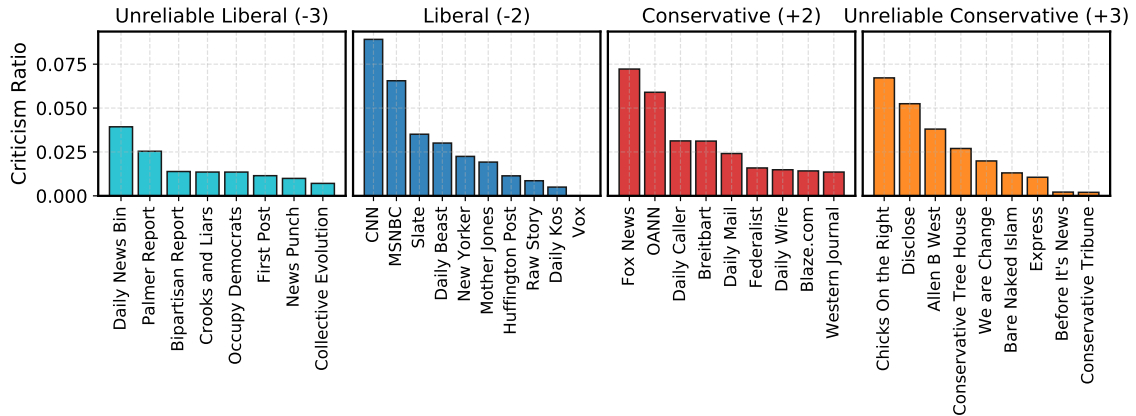


Figure 5.5: Criticism shown towards the most mentioned news sources from each partisan stance.

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.

### 5.6.3 Effect of criticism on diversity measures

Identifying tweets containing media-targeted criticism may affect estimates of news engagement diversity, which in turn provides fresh insights about the nature of filter bubbles [60]. We thus examine how filter bubble measures change after removing critical tweets.

To measure the diversity of a user’s news engagements, we use the normalized stance entropy measure (**NSE**) used in prior work on filter bubbles by Liu et al. [106]:  $NSE = \frac{-\sum_{i=1}^m p_i \log p_i}{\log m}$ , 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 5.6 shows the distribution of NSE scores before and after removing critical tweets. We can see that, after removing critical tweets, NSE is reduced with means of .358 to .339, the standard error for the means are 0.00373 (Before) and

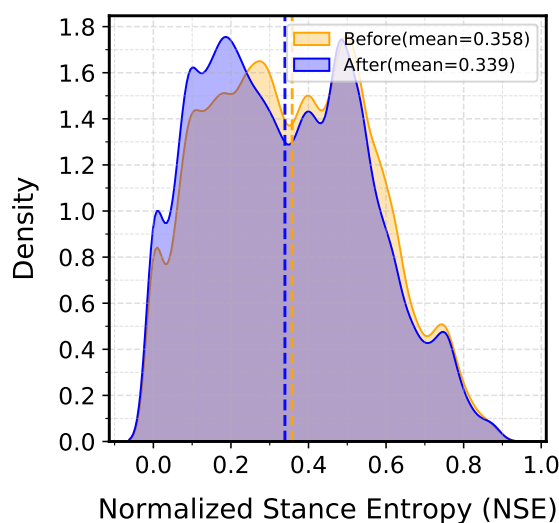


Figure 5.6: Comparison of Normalized Stance Entropy before and after removing critical tweets.

0.00375 (After) and both are found to be significant with  $p < 0.0001$  using 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), signaling that users do have less diverse news engagements when criticism is considered.

#### 5.6.4 Criticism Over Time

To investigate how criticism towards news media changes over time, we calculate the fraction of critical tweets per month. Figure 5.7 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 during these time windows. For the mid-2017 spike, prominent terms like  $\{Trump, President, Obama, Comey, Russia, police, Charlottesville,$

*Syria, Muslim*} indicate 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 Mueller’s investigation into Russian interference in the 2016 U.S. election. 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.

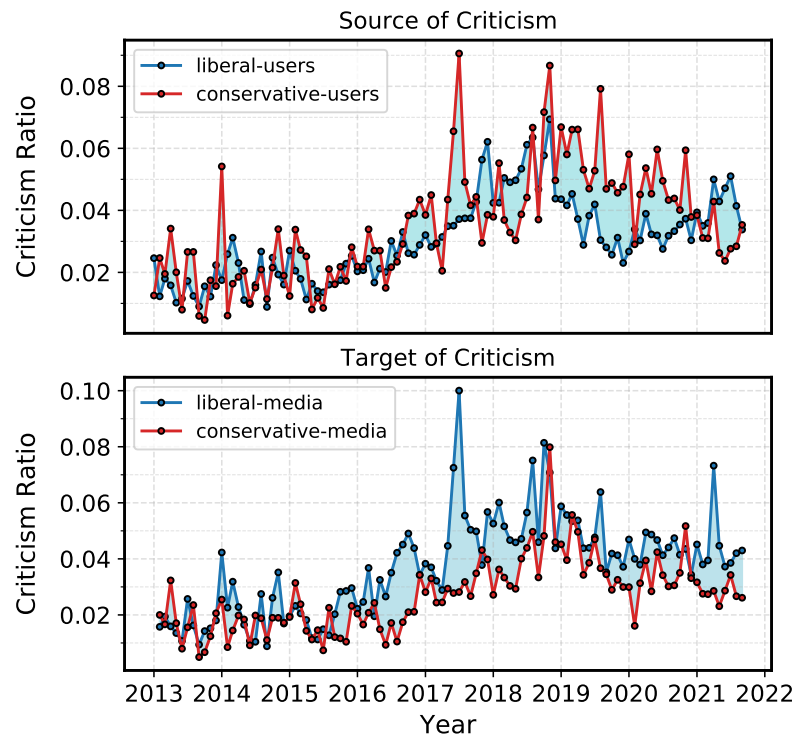


Figure 5.7: Criticism across time by partisan stance of the user (**top panel**) and news source (**bottom panel**).

### 5.6.5 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 sep-



User Group	Sequence	% of Users
<b>Liberal</b>	$-2 \rightarrow -3 \rightarrow C_C$	<b>49.07</b>
	$-2 \rightarrow C_C \rightarrow -3$	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
<b>Conservative</b>	$2 \rightarrow 3 \rightarrow C_L$	<b>46.33</b>
	$2 \rightarrow C_L \rightarrow 3$	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 5.10: Progression sequences of first engagement of each type.

arate users into liberal and conservative groups based on the stance of their direct retweets (as in §5.6.1). For liberal users, we consider all who have engaged with -3 and -2 sources as well as 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 5.10 indicate that, for both sides, the most common sequence is partisan  $\rightarrow$  unreliable  $\rightarrow$  critical. This is followed by partisan  $\rightarrow$  critical  $\rightarrow$  unreliable. Overall, these results suggest that most users begin engaging with unreliable news prior to expressing criticism of cross-partisan media.

## 5.7 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 did not attempt 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 §5.3, this was done in part to focus our efforts on media most often mentioned in critical tweets. Future work should consider additional news sources.
- **Classifier noise:** As our experiments indicate, the classifier is imperfect, and these errors can propagate to the analysis in §5.6. Future work could apply adjustment methods to calibrate estimates of critical tweets [61, 94].

Regarding social impact, 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 [34, 39], a feature that is unlikely to be eliminated in online news consumption patterns anytime soon.

## 5.8 Conclusion

In this Chapter, we have proposed a methodology for identifying tweets that criticize partisan news media, and we have conducted a descriptive analysis to understand how such tweets vary by user, news source, and time. Our classification experiments indicate that weak supervision can effectively train such a classifier with limited and manually-annotated data.

Some of the substantive results are intuitive – e.g., hyperpartisan users are more likely to criticize media from the other side (Figure 5.4), with CNN and Fox receiving the largest shares of critical tweets (Figure 5.5). Other results are more nuanced – e.g., unreliable news sources (-3, +3) do not necessarily receive more criti-

cism 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.

# Chapter 6

## Forecasting News Engagement Behavior

### 6.1 Introduction

The previous chapter focused on a particular news engagement behavior where users express criticism towards the news media they engage with. This chapter aims to deepen our understanding of how users interact with news content online and to uncover the factors that drive their behavior. In pursuit of this objective, we focus our efforts towards exploring methods that can forecast future news engagement behavior on Twitter. By examining patterns of user behavior over time, these methods can provide valuable insights into the key factors that influence online news engagement behavior in the long term.

### 6.2 Problem Formulation

We formulate this problem of predicting future news engagement behavior as a forecasting task where we define news engagement as a scenario in which a given user either mentions a news source or shares an article published by it. Let  $y_{t_j}^i$  represent a count vector of all partisan stances  $p, p \in \{-3, -2, -1, 0, 1, 2, 3\}$  of engaged news sources for a given user  $u_i$  across time step  $t_j$ . And 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  $\{y_{t_1}^i, y_{t_2}^i, \dots, y_{t_n}^i\}$  and tweet at-

tributes  $\{m_{t_1}^i, m_{t_2}^i, \dots, m_{t_n}^i\}$ , our goal is to estimate the probability distribution over the engagement counts of a future time step  $t_{n+1}$  for user  $u_i$ .

$$P(y_{t_{n+1}}^i | y_{t_1}^i \dots y_{t_n}^i, m_{t_1}^i \dots m_{t_n}^i) \quad (6.1)$$

For example, given the recent historic tweet attributes and engagement counts we want to predict how many times a given user would engage with news sources across all stances (-3 to +3) for the next consecutive time step.

### 6.3 Data

We use an extended version of the dataset discussed in Chapter 5. In addition to the already collected historic Twitter profiles of 5470 users, we additionally collect 4311 users in a similar fashion as described in Chapter 5, so in total we have 9781 unique users in our data collection ( $D$ ). The tweet distribution by year is shown in Table 6.1 where the total number of tweets collected thus far is 63.5 million. In this Table matched tweets refer to the tweets where a user engages with a given news source and encompasses 10.2 % of our entire data collection. Here we define news engagement as a scenario where a user either mentions the official Twitter handle of the news source or shares an url of a news article published by the news source in a given tweet. The distribution of engagement types can be seen in Table 6.3. Each news source is associated with a partisan stance score ranging from -3 to +3 (liberal to conservative with varying degree of extremeness) and we map every matched tweet to a corresponding stance depending on which news source the user engaged with in the current tweet. The distribution of tweets by partisan stance can be seen in Table 6.2, where majority of the engaged tweets belong to the -1 partisan stance followed by 0 and +2.

Given that the majority of the matched tweets in our dataset occurred after

Year	All Tweets	Matched Tweets	Matched %
2006	3	0	0.000
2007	424	0	0.000
2008	14,649	7	0.048
2009	92,788	189	0.204
2010	187,609	2398	1.278
2011	741,038	16,065	2.168
2012	1,569,018	49,765	3.172
2013	1,853,832	74,069	3.995
2014	2,172,500	120,568	5.550
2015	2,579,583	181,957	7.054
2016	3,502,755	350,700	10.012
2017	4,816,301	585,272	12.152
2018	5,744,817	684,334	11.912
2019	7,889,222	904,271	11.462
2020	16,247,000	1,843,002	11.344
2021	16,021,606	1,654,322	10.326
<b>Total</b>	<b>63,433,145</b>	<b>6,466,919</b>	<b>10.195</b>

Table 6.1: Tweet Distribution by Year

2015, as indicated in Table 6.1, we have decided to focus on tweets that were created on or after January 1, 2015, for the purpose of this study. In order to ensure that we only include users who have a significant history of engaging with news sources, we apply two filtering criteria: (1) we exclude user accounts that do not engage with news sources at least 50 times, and (2) we exclude accounts where the news engagements are not spread across at least 3 years in time, this results in 3806 users being filtered out and 5975 users remain.

To reduce the impact of automated accounts we filter out user accounts which we consider as outliers, we accomplish this by removing all user accounts whose total news engagement volume is 3 standard deviations greater than the mean news engagement volume (i.e 3-sigma rule) [75, 140], we end up removing 137 users that were identified as outliers by the above criteria. The final number of users that remain in our data collection is 5838.

Stance	Matched Tweets	Matched %
-3	112,560	1.741
-2	1,141,939	17.658
-1	1,977,177	30.574
0	1,240,848	19.188
1	569,489	8.806
2	1,256,521	19.430
3	168,385	2.604
<b>Total</b>	<b>6,466,919</b>	<b>100</b>

Table 6.2: Matched Tweet Distribution by Partisan Stance

News Engagement Type	Tweets #	Tweets %
Url	3,550,470	54.902
Mention	1,750,292	27.065
Both (Url + Mention)	1,166,157	18.033
<b>Total</b>	<b>6,466,919</b>	<b>100</b>

Table 6.3: News Engagement Distribution

## 6.4 Methods

In this section, we discuss how we preprocess our data, the different forecasting models we utilize, which are based on multiple neural network architectures and our baseline method. We also discuss the different features we use for model training.

### 6.4.1 Data Pre-processing

For this study we set  $n=8$  (the number of input timesteps), and each time step  $t_j$  encompasses a 3 month-period. As we are using data collected from 2015 - 2021, there is a issue of data drift due to the large shift in engagement volume throughout the years (as shown in Table 6.1), to control for this phenomena, we split our overall dataset ( $D$ ) into 4 subsets  $\{D_1, D_2, D_3, D_4\}$  where each of them covers observations across a 3 year time period (train - 2 year, test - 1 year) across all users as shown in Table 6.4 and train separate models over each of these smaller datasets. Here

each observation consists of features across 8 time steps, (8, 3-month windows that spans across 2 years) which is used as the input and the next consecutive 3 month window is used as our forecasting horizon. An example of this for a given user in  $D_1$  is shown in Table 6.5 in terms of time based windows (this process is repeated across  $D_2, D_3, D_4$ ). From this Table we can see that each user has 4 observation sequences used for training and 4 for testing purposes.

<b>Dataset</b>	<b>Train</b>	<b>Val</b>	<b>Test</b>
$D_1$	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 6.4: Data Subsets by Time



Input time steps (Year - 3 Month Quarter)										Forecast Horizon
	$t_1$	$t_2$	$t_3$	$t_4$	$t_5$	$t_6$	$t_7$	$t_8$	$t_9$	
<b>Train</b>	2015 - 1	2015 - 2	2015 - 3	2015 - 4	2016 - 1	2016 - 2	2016 - 3	2016 - 4	2017 - 1	2017 - 1
	2015 - 2	2015 - 3	2015 - 4	2016 - 1	2016 - 2	2016 - 3	2016 - 4	2017 - 1	2017 - 2	2017 - 2
	2015 - 3	2015 - 4	2016 - 1	2016 - 2	2016 - 3	2016 - 4	2017 - 1	2017 - 2	2017 - 3	2017 - 3
	2015 - 4	2016 - 1	2016 - 2	2016 - 3	2016 - 4	2017 - 1	2017 - 2	2017 - 3	2017 - 4	2017 - 4
<b>Test</b>	2016 - 1	2016 - 2	2016 - 3	2016 - 4	2017 - 1	2017 - 2	2017 - 3	2017 - 4	2018 - 1	2018 - 1
	2016 - 2	2016 - 3	2016 - 4	2017 - 1	2017 - 2	2017 - 3	2017 - 4	2018 - 1	2018 - 2	2018 - 2
	2016 - 3	2016 - 4	2017 - 1	2017 - 2	2017 - 3	2017 - 4	2018 - 1	2018 - 2	2018 - 3	2018 - 3
	2016 - 4	2017 - 1	2017 - 2	2017 - 3	2017 - 4	2018 - 1	2018 - 2	2018 - 3	2018 - 4	2018 - 4

Table 6.5: Example Observation Sequence for a given user for  $D_1$

## 6.4.2 Features

1. **News Engagement Counts**( $y_{t_j}^i$ ) : For each input time step  $t_j$  for user  $u_i$  for a given observation sequence, we use a count vector of all partisan stances  $p, p \in \{-3, -2, -1, 0, 1, 2, 3\}$  of engaged news sources denoted as  $y_{t_j}^i$  where  $y_{t_j}^i \in R^{1 \times 7}$ . 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 Texts**( $v_{t_j}^i$ ) : For each input time step  $t_j$  for user  $u_i$  for a given observation sequence, we select the top 25 most recent tweets for both the news engagement tweets and non-news engagement tweets for that specific 3-month window. We next pass these 2 sets of tweets through a transformer based language model called TwHIN-Bert [203] and extract embedding representations for each token of each tweet. We perform 2 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 next take an average over these 25 tweet representations. This results in two 1536 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 2 embedding vectors to get a single text representation  $v_{t_j}^i$ , where  $v_{t_j}^i \in R^{1 \times 3072}$ .
3. **Hashtags**( $\#_{t_n}^i$ ): For the last input time step  $t_n$  for user  $u_i$  for a given observation sequence, we select the top 100 most frequently used hashtags for that specific 3-month window. We then pass these hashtags through our language model (TwHIN-Bert) and perform an identical aggregation step we used for our text based features, to obtain a final vector representation  $\#_{t_n}^i$  where  $\#_{t_n}^i \in R^{1 \times 1536}$ .

4. **Input Quarter Encoding**( $q_{t_j}^i$ ): For each observation sequence we encode the year quarter of each input time step as a one-hot encoding vector  $q_{t_j}^i$ , where  $q_{t_j}^i \in R^{1 \times 4}$ .
5. **Forecast Quarter Encoding** : Similarly as above, for each observation sequence we encode the year quarter of the time-step we are forecasting as a one-hot encoding vector.

### 6.4.3 Baseline

For our baseline, we use a traditional approach which uses the news engagement count vector of the last time step  $y_{t_n}^i$  of the input sequence as the predicted label (i.e  $\hat{y}_{t_{n+1}}^i = y_{t_n}^i$ ).

### 6.4.4 Single Feature Network (SFN)

This network model uses a single feature (either tweet texts or news engagement counts) to forecast future news engagement counts. We use a bi-directional lstm (Bi-LSTM) [159] as our forecasting model. The inputs to these models are either text based representation sequences  $\{v_{t_1}^i, v_{t_2}^i, \dots, v_{t_n}^i\}$  (**SFN + T**) or news engagement count based sequences  $\{y_{t_1}^i, y_{t_2}^i, \dots, y_{t_n}^i\}$ (**SFN + C**). When using the text based representations we add a linear layer before passing the input sequences into the bidirectional LSTM model. 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 both to obtain a single hidden state representation  $h_{t_n}$ . This is then passed through a final output layer  $\langle W_{out}, b_{out} \rangle$  to predict the future news engagement count vector  $\hat{y}_{t_{n+1}}^i$  for time step  $t_{n+1}$  as shown in equation 6.2.

$$\hat{y}_{t_{n+1}}^i = (W_{out}h_{t_n} + b_{out}) \quad (6.2)$$

### 6.4.5 Multiple Feature Network (MFN)

This network model uses multiple features (tweet texts ( $v_{t_j}^i$ ), news engagement counts ( $y_{t_j}^i$ ) and input quarter encodings ( $q_{t_j}^i$ )) to forecast future news engagement counts. This has a similar architecture as 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 (as 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  $h_{t_n}$  and then pass this through our final output layer  $\langle W_{out}, b_{out} \rangle$  to predict the future news engagement count vector  $\hat{y}_{t_{n+1}}^i$  for time step  $t_{n+1}$  as shown in equation 6.3.

$$\hat{y}_{t_{n+1}}^i = (W_{out}[h_{t_n}, \#_{t_n}^i, q_{t_{n+1}}^i] + b_{out}) \quad (6.3)$$

The architecture for this network is shown in Figure 6.1. Both of the network models (SFN & MFN) are trained using a Mean Absolute Error (MAE) loss. Other loss functions such as Mean Squared Error (MSE), Mean Absolute Percentage Error (MAPE) and Huber Loss were also experimented with, but we found that MAE loss worked the best for training these models.

As we predict a vector of engagement counts, the overall MAE loss is a sum across individual MAE losses for each news engagement stance. As shown in equation 6.4.

$$\text{Total MAE Loss} = \sum_{r=1}^{|p|} \text{MAE}(y_{t_{n+1}}[r], \hat{y}_{t_{n+1}}[r]) \quad (6.4)$$

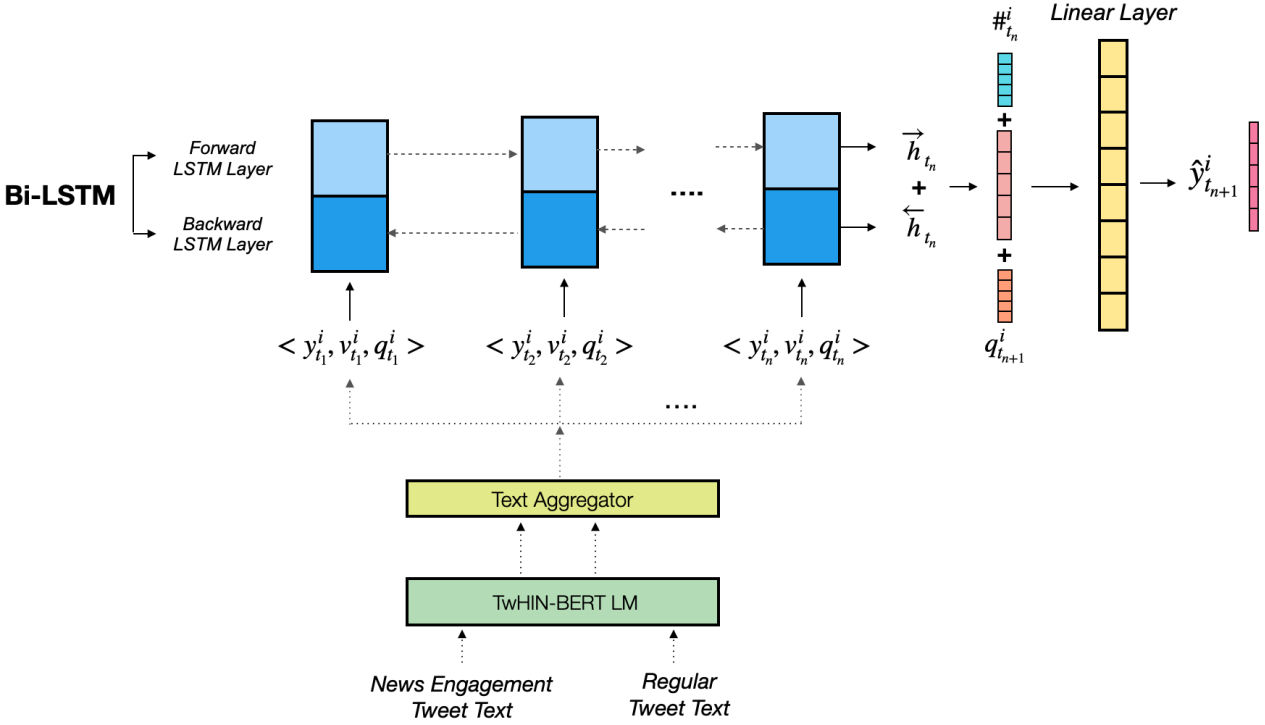


Figure 6.1: Network Architecture for the Multiple Feature Network

## 6.5 Experiments and Results

### 6.5.1 Settings

To perform our forecasting experiments, we first construct our train and validation sets for each sub-dataset (D1,D2,D3,D4). We first select all samples that occur within the given train/validation time period (as shown in Table 6.4) and split them by user, selecting 80% of the users to create our train set and 20% for our validation set. For our test dataset we make sure for each user (across both train and validation) we have forecast windows over the 4 quarters of the year we are forecasting as shown in Table 6.5. Once we have our train, validation and test for each sub dataset, we next filter out sequences where the user has no news engagement activity over the input. The resulting dataset sizes are shown in Table 6.6.

We then train our models over each dataset and tune each model’s hyperpa-

Dataset	Train	Val	Test
D1	15,708	3904	21,881
D2	17,536	4345	22,327
D3	17,862	4465	22,648
D4	18,106	4542	17,378

Table 6.6: Train, validation and test sizes across all datasets

rameters using grid search. The best hyperparameters are chosen based on validation MAE. We then measure forecasting performance for each model across multiple metrics such as mean absolute error (MAE), mean squared error (MSE) and mean absolute percentage error (MAPE). We use *Adam* as the optimizer to train our networks.

Model	Avg MAE	Avg MAPE	Avg MSE
Baseline	3.978	0.509	225.865
SFN + C	<b>3.835</b>	<b>0.444</b>	<b>218.564</b>
SFN + T	4.308	0.539	264.023
MFN	3.943	0.485	226.907

Table 6.7: Mean Forecast Metrics across all Data-sets (D1 to D4)

Model Comparisons	P-value <0.05	P-value <0.01
Baseline vs SFN + C	TRUE	TRUE
Baseline vs SFN + T	TRUE	TRUE
Baseline vs MFN	FALSE	FALSE
SFN + C vs SFN + T	TRUE	TRUE
SFN + C vs MFN	TRUE	FALSE
SFN + T vs MFN	TRUE	TRUE

Table 6.8: Statistical Hypothesis Test results for Avg MAE of Models across all datasets using paired T-test

## 6.5.2 Results

The forecasting results for each test dataset are shown in Table 6.9. Since we have 7 different stances to measure news engagement, we mainly compare the aggregate metrics across all stances. For the test set of D1 we observe that the SFN + C model performs the best across all different metrics (Avg MAE : 2.239, Avg MSE

Test Dataset	Metric	Model	-3	-2	-1	0	1	2	3	Avg
D1	MAE	Baseline	<b>0.197</b>	3.143	5.089	3.212	1.324	3.083	<b>0.485</b>	2.362
		SFN + C	0.210	<b>2.960</b>	<b>4.813</b>	<b>3.032</b>	<b>1.242</b>	<b>2.889</b>	0.527	<b>2.239</b>
		SFN + T	0.264	3.265	5.440	3.295	1.418	3.326	0.602	2.515
		MFN	0.210	3.001	4.907	3.133	1.275	3.011	0.527	2.295
	MAPE	Baseline	0.069	0.579	0.720	0.622	0.334	0.453	0.138	0.416
		SFN + C	<b>0.048</b>	<b>0.461</b>	<b>0.635</b>	<b>0.539</b>	<b>0.233</b>	<b>0.338</b>	<b>0.090</b>	<b>0.335</b>
		SFN + T	0.104	0.517	0.748	0.591	0.285	0.426	0.177	0.407
		MFN	0.054	0.469	<b>0.635</b>	0.575	0.254	0.391	0.101	0.354
	MSE	Baseline	<b>2.791</b>	101.655	194.288	80.426	29.651	161.562	13.353	83.389
		SFN + C	4.628	<b>92.066</b>	<b>173.509</b>	<b>73.989</b>	29.686	<b>147.507</b>	23.220	<b>77.801</b>
		SFN + T	4.938	107.794	208.864	82.296	39.126	192.568	22.760	94.050
		MFN	4.311	93.277	178.151	79.201	<b>29.484</b>	151.619	<b>22.352</b>	79.771
D2	MAE	Baseline	<b>0.363</b>	3.701	6.240	3.802	2.016	3.095	0.449	2.810
		SFN + C	0.409	<b>3.509</b>	<b>6.036</b>	<b>3.612</b>	1.948	<b>2.957</b>	0.436	<b>2.701</b>
		SFN + T	0.409	3.908	6.697	3.954	2.137	3.440	0.494	3.006
		MFN	0.392	3.585	6.118	3.648	<b>1.945</b>	2.994	<b>0.417</b>	2.728
	MAPE	Baseline	0.096	0.596	0.717	0.636	0.390	0.458	0.135	0.432
		SFN + C	<b>0.060</b>	0.493	<b>0.636</b>	<b>0.544</b>	<b>0.285</b>	<b>0.389</b>	<b>0.090</b>	<b>0.357</b>
		SFN + T	0.095	0.591	0.734	0.627	0.353	0.536	0.143	0.440
		MFN	0.064	<b>0.490</b>	0.682	0.582	0.304	0.415	0.097	0.376
	MSE	Baseline	<b>8.394</b>	<b>135.683</b>	<b>229.573</b>	95.525	56.045	153.819	<b>10.009</b>	<b>98.435</b>
		SFN + C	14.231	142.533	231.820	91.489	58.266	<b>150.741</b>	14.041	100.446
		SFN + T	12.166	167.877	300.168	108.147	69.531	190.636	15.081	123.372
		MFN	12.463	143.127	234.574	<b>91.072</b>	<b>55.358</b>	153.633	11.799	100.289
D3	MAE	Baseline	<b>0.620</b>	6.389	10.836	6.983	3.567	6.506	<b>0.675</b>	5.082
		SFN + C	0.674	<b>5.988</b>	<b>10.381</b>	<b>6.720</b>	<b>3.484</b>	<b>6.501</b>	0.727	<b>4.925</b>
		SFN + T	0.689	6.519	11.296	7.257	3.830	7.476	0.771	5.405
		MFN	0.674	6.075	10.490	6.731	3.485	6.592	0.722	4.967
	MAPE	Baseline	0.160	0.742	0.843	0.754	0.534	0.576	0.175	0.541
		SFN + C	<b>0.100</b>	<b>0.578</b>	<b>0.768</b>	<b>0.656</b>	<b>0.448</b>	<b>0.450</b>	<b>0.114</b>	<b>0.445</b>
		SFN + T	0.148	0.669	0.890	0.748	0.484	0.492	0.182	0.516
		MFN	0.148	0.632	0.828	0.697	0.462	0.471	0.127	0.481
	MSE	Baseline	<b>17.862</b>	391.679	714.744	310.403	139.905	<b>604.606</b>	<b>25.189</b>	314.913
		SFN + C	29.541	<b>344.048</b>	<b>670.690</b>	302.278	141.910	624.751	45.324	<b>308.363</b>
		SFN + T	26.298	413.575	840.848	356.255	178.143	775.066	44.065	376.321
		MFN	25.037	351.587	691.502	<b>298.807</b>	<b>139.058</b>	646.398	42.457	313.549
D4	MAE	Baseline	<b>1.390</b>	6.074	9.972	7.631	4.923	8.400	<b>1.226</b>	5.659
		SFN + C	1.484	<b>5.658</b>	<b>9.712</b>	<b>7.292</b>	<b>4.605</b>	<b>8.213</b>	1.348	<b>5.473</b>
		SFN + T	1.514	6.621	11.809	8.557	5.116	9.107	1.407	6.304
		MFN	1.511	6.169	10.557	7.600	4.712	8.555	1.364	5.781
	MAPE	Baseline	0.188	0.819	<b>0.960</b>	<b>0.909</b>	0.661	<b>0.749</b>	0.242	0.647
		SFN + C	<b>0.179</b>	<b>0.786</b>	1.027	0.951	<b>0.612</b>	0.757	<b>0.171</b>	<b>0.640</b>
		SFN + T	0.206	0.950	1.327	1.231	0.733	0.903	0.211	0.794
		MFN	0.204	0.918	1.210	1.053	0.652	0.847	0.216	0.729
	MSE	Baseline	<b>97.218</b>	393.098	591.616	397.215	278.096	1030.315	<b>59.511</b>	406.724
		SFN + C	107.879	<b>357.631</b>	<b>542.017</b>	<b>361.471</b>	<b>247.393</b>	<b>1015.021</b>	82.120	<b>387.647</b>
		SFN + T	108.030	431.993	745.833	445.685	287.034	1133.112	84.761	462.350
		MFN	107.845	405.158	621.267	385.696	255.259	1043.199	79.704	414.018

Table 6.9: Forecast Metrics for all test sets across individual stances (Best scores are highlighted per metric)

: 77.801, Avg MAPE : 0.335) followed by the MFN model. For the test set of D2 we find that for both MAE and MAPE the SFN+C model performs the best (Avg MAE : 2.701, Avg MSE : 98.435, Avg MAPE : 0.357) followed by MFN except for the MSE metric where the baseline model has the lowest error. For the test set of D3 the SFN + C model performs the best (Avg MAE : 4.925, Avg MSE : 308.363, Avg MAPE : 0.445) across all metrics followed by the MFN model. Similar observations are seen for the test set of D4, but for this test set the MFN model performs worst than the baseline across all metrics. At a stance level we observe that across most of the test datasets the baseline tends to perform better for the -3 and +3 engagement counts and this could be due to lower engagement volume for these stances across our data collection (so fewer user samples who engage with this category of news sources). We also observe that errors increase through time (errors for D1 < errors for D2 < errors for D3 < errors for D4) due to the considerable shift in engagement volume.

We next measure the average performance of the models across all test datasets and is shown in Table 6.7 (here we have 2 levels of aggregation, first average is taken across the stances for each test dataset and then another average is performed over the scores of each test dataset). We find that SFN+C performs the best followed by the baseline model. The MFN model performs worst than the baseline due to it's low performance on the test set in D4. Eventhough the SFN+T model performs the worst, it's still surprising to find that it performs relatively well for a model that only utilizes text based representations. In order to confirm the difference in performance between each of our models, we conduct paired statistical hypothesis tests over the combined results of each model and the results are shown in Table 6.8. We find the baseline and SFN + C model MAE's are statistically different at both 95% and 99% confidence levels. Similar results for other model pairs are observed, except for the comparison between the Baseline and MFN model (due to the bad performance of



MFN for the test set in D4).

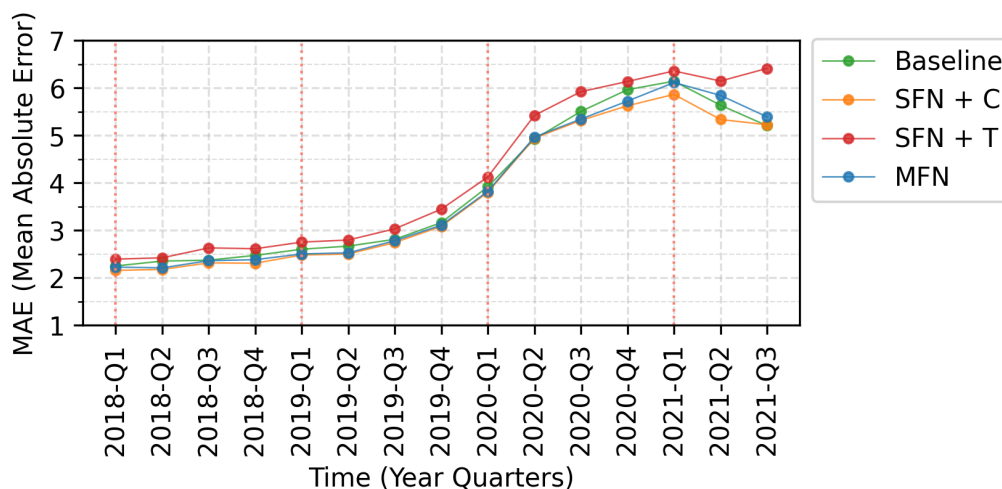


Figure 6.2: MAE Model Performance by Yearly Quarters

We also measure performance by yearly quarters and is shown in Figure 6.2, from this Figure we can observe that errors increase as we progress through time, this is due to the large increase in engagement volume as we progress through the years. The highest increase in errors is measured for the quarters of 2020 which could be due to various events causing sudden social outbreaks (such as COVID-19, impeachment of Trump, etc.) . The SFN+C model performs the best across most of the quarters with a larger difference in performance in comparison to the other models measured during 2020-Q4 to 2021-Q2.

To get a better understanding of how our models perform at a user level, we analyze model performance using truth vs prediction plots across all test sets for a subset of users. For each stance we sort users by the their MAE values (mean across absolute errors measured across all test instances) using the predictions from the SFN+C model. We then select users who have the lowest and highest MAE values for each news engagement stance. To ensure we select users who have substantial news engagement volume for each forecasting horizon, we filter out users who do not engage at-least 10 times with a news source for each horizon.

Figure 6.3 represents the truth vs prediction plots for users with the lowest MAE across all test sets for the SFN+C model, while Figure 6.4 represents users with the highest errors. In these Figures each subplot represents a different user and stance combination. From these Figures we observe that our models perform quite well when capturing the overall trend of engagement counts across most stances (eg: Fig 6.3 - User C :-1, User D: 0, User E: 1. Fig 6.4 - User B: -2, User C: -1, User E: 1, User F:2). The main challenge for these models seem to be a sudden peak in engagement behavior during certain time periods, the SFN + C model seems to handle this issue better than the SFN + T model which may be due to the text based features not being able to capture the intensities of the news engagement for these users. We also observe that both models (SFN + C , SFN + T) don't perform too well for the -3 category which may be due to the low volume of these types of engagements in our data collection.

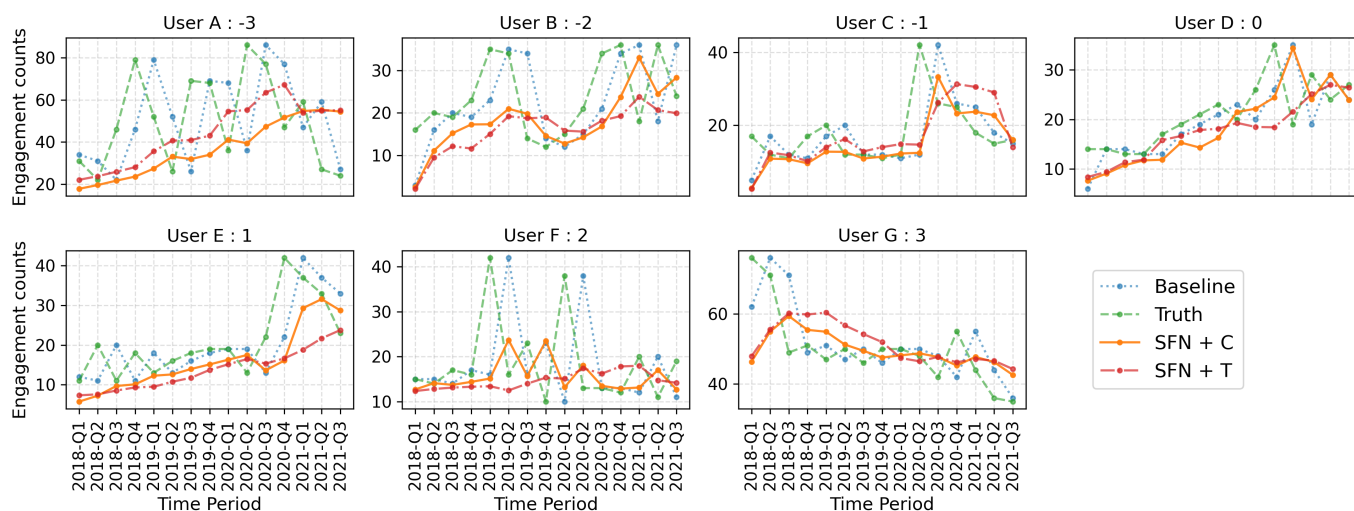


Figure 6.3: Truth vs Prediction plots for users with the **lowest errors** across all test sets for the SFN + C model

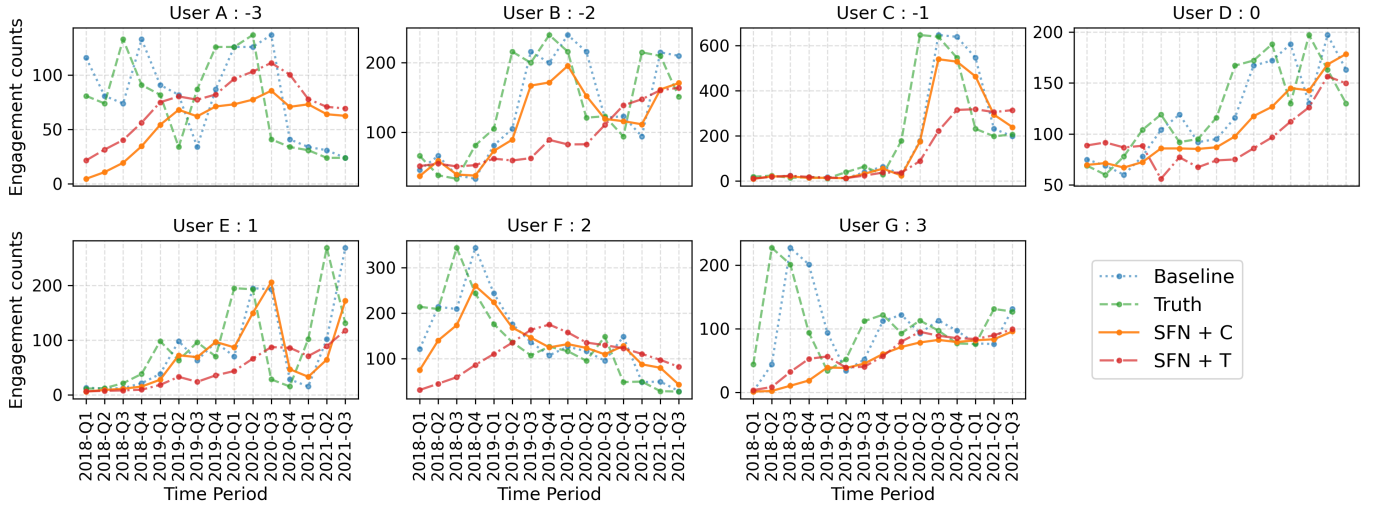


Figure 6.4: Truth vs Prediction plots for users with the **highest errors** across all test sets for the SFN + C model

### 6.5.3 Difficulty Analysis

In order to gain a deeper understanding of scenarios where our proposed models perform better than the baseline we conduct additional analysis by comparing forecasting performance at different degrees of difficulty.

#### Cosine Distance Ranking

One way we measure model performance at different difficulties is by ranking individual instances in our test sets based on a hardness metric. To identify hard samples where there is a considerable 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 forecasted engagement count vector ( $\hat{y}_{t_{n+1}}^i$ ). We next measure the avg MAE scores across test samples at different rankings. The results across all the different test sets are shown in Figure 6.5. From these plots we observe that our proposed models perform better than the baseline up to a ranking of 14K for D1, D2 and D3, largest difference between the baseline and our models are observed

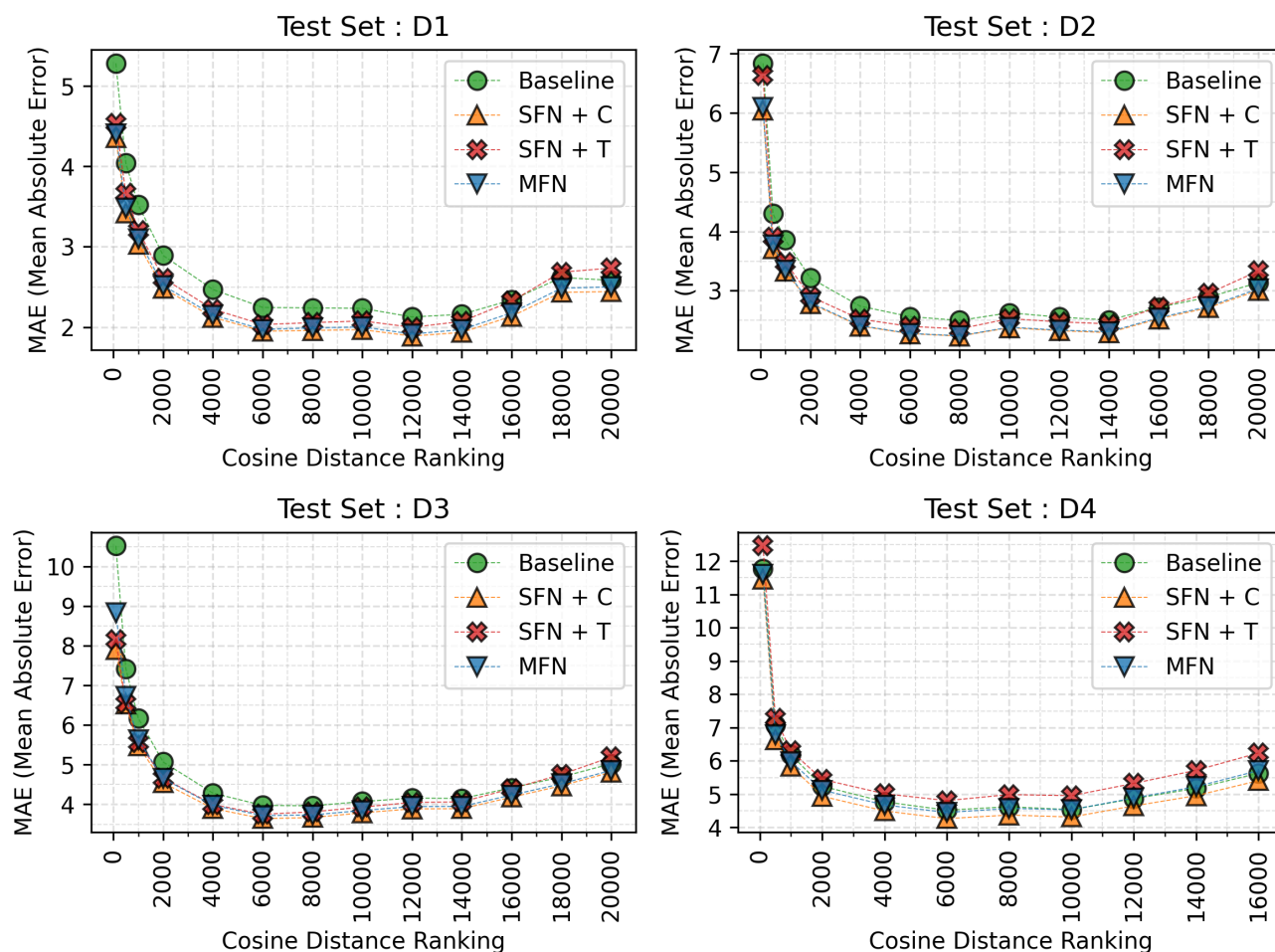


Figure 6.5: Model Performance at different difficulty levels based on Cosine Distance Ranking

for the test sets in D1 and D2. For the test set of D4, the SFN + C and MFN models perform slightly better than the baseline up to a ranking of 10K.

### Baseline Absolute Error Ranking

Another ranking measure we utilize is based on identifying samples that are hard for the baseline model. For this we utilize the absolute errors of the baseline model as the ranking metric. The results across all the different test sets are shown in Figure 6.6. From this Figure we observe that our proposed models perform better than the baseline significantly up to a ranking of 2000, after this point the performance

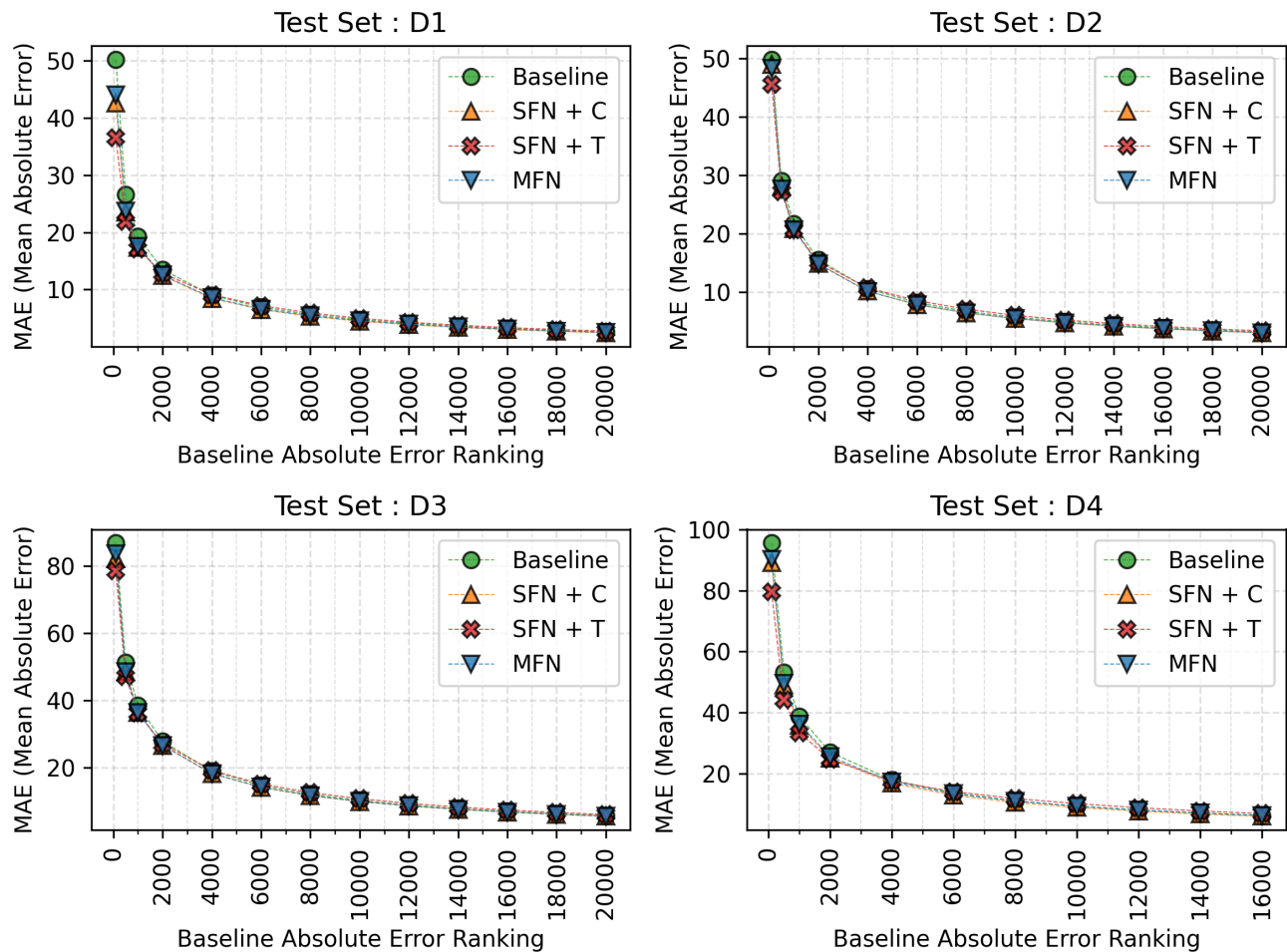


Figure 6.6: Model Performance at different difficulty levels based on Ranking using Baseline Absolute Error

differences between the models are small.

### Engagement to No-Engagement Transitions

One category of difficult observations to forecast are sequences where a user goes from engaging with news to no engagements in the consecutive time period. To measure model performances for these types of engagement patterns, we first sample our test datasets to select instances where users have some news engagement activity in the last input time step of their input sequences (i.e.  $y_{t_n}^i$ ) but zero news engagement for their forecasting window (i.e.  $y_{t_{n+1}}^i$ ). We then calculate the forecasting metrics over

these instances for each test dataset. The results are shown in Table 6.10. From these results we observe that our proposed models perform better than the baseline when considering Avg MAE and Avg MSE (across stances) for the test sets in D1, D2 and D3. For the test set in D4 all the models perform worst than the baseline for Avg MAE but the SFN + C beats it when considering Avg MSE. To summarize, for news engagement scenarios where users go from having some engagement activity to none, our proposed models perform better than the baseline approach across all of the test sets except for D4.

Test Dataset	Sample Size	Model	Avg MAE	Avg MSE
D1	1638	Baseline	0.653	35.115
		SFN + C	<b>0.489</b>	21.998
		SFN + T	0.650	26.794
		MFN	0.522	<b>21.845</b>
D2	1332	Baseline	0.651	9.166
		SFN + C	<b>0.508</b>	<b>7.047</b>
		SFN + T	0.637	9.805
		MFN	0.540	7.093
D3	816	Baseline	1.327	167.761
		SFN + C	<b>0.986</b>	<b>37.362</b>
		SFN + T	1.284	66.252
		MFN	1.146	65.717
D4	557	Baseline	<b>1.307</b>	41.798
		SFN + C	1.409	<b>36.915</b>
		SFN + T	1.622	46.540
		MFN	1.662	45.620

Table 6.10: Model Performance on samples where users go from a Engagement to No-Engagement State

### No-Engagement to Engagement Transitions

Similar to the above another category of engagement patterns that are difficult to forecast would be when users go from not engaging with news to suddenly engaging with news in the consecutive time period. To measure model performances for these types of engagement patterns, we first sample our test datasets to select instances where users have no news engagement activity in the last input time step of their input

sequences (i.e  $y_{t_n}^i$ ) but some news engagement for their forecasting window (i.e  $y_{t_{n+1}}^i$ ). We then calculate the forecasting metrics over these instances for each test dataset. The results are shown in Table 6.11. For the test set in D1 , the best performing model is SFN + C when considering Avg MAE, when considering AVG MSE both the SFN + C and MFN models beat the baseline and have similar performance. For the test set in D2 , SFN + C model performs better than the baseline across both metrics. For the test set in D3 and D4, we observe similar patterns. Overall we observe that compared to the engagement to no-engagement scenario (as discussed above), the difference between the performance of the models are smaller indicating that all approaches suffer when trying to forecast this specific type of engagement scenario.

Test Dataset	Sample Size	Model	Avg MAE	Avg MSE
D1	1541	Baseline	0.818	19.506
		SFN + C	<b>0.813</b>	19.263
		SFN + T	0.845	19.773
		MFN	0.822	<b>19.220</b>
D2	1464	Baseline	0.930	51.301
		SFN + C	<b>0.921</b>	50.966
		SFN + T	0.945	<b>50.917</b>
		MFN	0.930	51.095
D3	1131	Baseline	2.144	237.208
		SFN + C	2.125	236.446
		SFN + T	<b>2.121</b>	<b>233.230</b>
		MFN	2.139	235.672
D4	634	Baseline	2.535	368.216
		SFN + C	2.462	363.940
		SFN + T	2.501	<b>362.628</b>
		MFN	<b>2.457</b>	362.790

Table 6.11: Model Performance on samples where users go from a No-Engagement to Engagement State

#### 6.5.4 Top Predictive Terms Analysis

To get a better understanding of what factors affect future news engagement with unreliable / fake news (+3,-3), we analyze the top terms that are predictive of

engagement with these types of news sources. We first combine all our test instances into a single test set and then rank each of the instances by predicted engagements for both the +3 stance and -3 stance. We select the top 200 test instances which are representative of high engagements and the bottom 200 test instances which are representative of low engagements for both the +3 and -3 stance. For these test sequences we next select the corresponding tweets that were used to train our SFN + T models (all tweets used to generate the text features for the input sequence of each instance). We next perform a chi-square test and select the top 500 terms (according to their chi-square value) that are predictive of user engagements with +3 (unreliable conservative) and -3 (unreliable liberal) news sources. A subset of these top terms predictive of +3 engagements is shown in Table 6.13 and for -3 engagements is shown in Table 6.12, these terms are manually split into 3 categories based on whether they are words, hashtags or mentions.

For terms that are predictive of +3 engagements (Table 6.13), we observe that they revolve around (1) **Highly polarizing political issues** such as (i) *gun-control* (gun, anti-gun, anti-gunners) , (ii) *taxes* (tax, billionaires, #taxreform, #goptaxscam) and (iii) immigration (illegals, migrants, #openbordersinc), (iv) race (2) **Polarizing news events** such as (i) *russia's interference in the 2020 US elections* (mueller, russia, russian, voting, #trumprussia), (ii) *black lives matter* (blm,#backtheblue, #bluelivesmatter), (iii) *sexual misconduct allegations against Supreme Court judge Brett Kavanaugh* (kavanaugh), (iv) *Arrest of Jeffrey Epstein for sex trafficking crimes*, (v) *Impeachment of Donald Trump* (treason, trump, impeached, unfit, complicit, trumpanzee, #impeachtrump, #dumptrump, #traitortrump, #liarinchief, #racistinchief), (vi) *2020 Elections* (elected, voting, #voteblue2020, #wewillrememberinnovember, #flushtheturdnovemberthird, #bidenharris2020), (vi) *Covid-19* (covid-19, vaccines) , (3) **Distrust towards news media** (#fakenews, #corruptmedia, #journalismis-



dead, lies, scam, corrupt, revealed, busted, liar, exposes) (4) **Islamophobia** (arab, isis, jihad, islamic). (5) **Christianity** (christians, vatican, jesus, bible, god, church, psalm, evangelical, #jesus, #biblephrophecy) (6) **Covid-19** (covid-19, vaccines), (7) **Engagement with hyper-partisan news sources** (cnn, foxnews, #foxnews, #tucker, @msnbc, @briertbartnews, @cnnbrkm, @newyorker, @foxandfirends, @foxnewsnight, @washingtonpost).

For terms that are predictive of -3 engagements (Table 6.12), we observe a large overlap with the +3 engagement terms that depict political events and interaction with hyperpartisan news media, but with additional focus on sports (eagles, #flyeaglesfly, #packers, @espn).

Token Type	Top Tokens
<b>Words</b>	congress, baseball, trumps, eagles, communal, anti, mueller, terrorism, cannabis, white, lockdown, islam, taliban, terrorists, homebuyers, minority, pandemic, black, nfl, blm, trump, leveraging, kobe, journalists, ice, carson, ringer, jesus, employment, ai, economy, democrats, donald, ballots, republicans, fbi, aliens, lgbt, shame, border, christ, immigrant, terror, fox, army, religion, democrat, socialist, homeless, impeachment, racist, women, bush, giuliani, housing, assange, god, eagles, illegal, impeached, gay, minorities, mcconnell, worship, healthcare, investigative, russia, russians, season
<b>Hashtags</b>	#christian, #flyeaglesfly, #gop, #corporateaccountability, #gobells, #packers, #covid19, #evangelical, #eagles, #trumprussia, #oann, #journalism, #noagenda, #usa, #maga, #gopackgo
<b>Mentions</b>	@realdonaldtrump, @donaldjtrumpjr, @realjameswoods, @aaronwilsonnfl, @thehill, @michaelvaughan, @msnbc, @economictimes, @breitbartnews, @seanhannity, @marklevinshow, @espn, @packers, @foxnews, @abcnews, @potus, @bipartisan

Table 6.12: Subset of Top Terms predictive of -3 Engagements

Token Type	Top Tokens
<b>Words</b>	<p>treason, trumps, liberal, scandal, jerusalem, iran, mueller, cannabis, vatican, conservative, cambridge, fishing, voting, black, blm, isis, lies, biden, scam, christians, police, democracy, chris, ivanka, ukrainian, putin, arab, corrupt, fairness, kavanaugh, disingenuous, pelosi, retirement, hateful, brussels, gop, mcconnell, revealed, lockdown, jihad, deal, senators, marijuana, kushner, liberty, fact, debt, epstein, nfl, apartheid, busted, senate, irony, liar, jesus, illegals, bible, freedom, crimes, exposed, racist, alt-leftist, disgraceful, assange, god, union, coronavirus, cia, impeached, tax, elected, unfit, complicit, healthcare, gun, qanon, exposes, anti-gun, covid-19, vaccines, election, criminal, islam, excuse, potus, uncovers, vaccine, islamic, trump, crooked, hillary, billionaires, shocking, investigated, donald, ocasio-cortez, godcast, facts, fox, church, impeach, horrible, cnn's, weed, covid, migrants, russia, zionist, sleazy, russian, corbyn, white, dictators, wuhan, genocide, fake, anti-gunners, coward, meddling, democrats, psalm, republicans, fbi, trumpanzee, woke, brexit, benghazi, pastor, chrislam, betrayal, house, evil, traitors, evangelical, republican, israel</p>
<b>Hashtags</b>	<p>#goptaxscam, #ncpol, #god, #wearebetterthanthat, #taxreform, #gop, #trumpcare, #goptraitors, #wordsdanasaidthatkilledkane, #backtheblue, #foxnews, #impeachtrump, #trumprussia, #jesus, #dumptrump, #f1, #trending, #bible, #trunews, #tcot, #cubs, #trump, #liarinchief, #racistinchief, #corruptdems, #liberalhypocrisy, #bibleprophecy, #theresistance, #nowtheendbegins, #tucker, #smartnews, #cannabiscommunity, #goblue, #journalismisdead, #wewillrememberinnovember, #notgoingaway, #russiagate, #cannabis, #fakenews, #hailstatedog, #endtimes, #worldcup, #traitortrump, #cannabisculture, #voteblue2020, #bluelivesmatter, #corruptmedia, #democrats, #putinsgop, #openbordersinc, #ccot, #liberalinsanity, #bidenharris2020, #freetheherb, #ridinwithbiden, #statepension, #resist, #flushtheturdnovemberthird</p>
<b>Mentions</b>	<p>@thomhartmann, @washingtonpost, @abc, @youtube, @gop, @msnbc, @bethlynch2020, @breitbartnews, @sencapito, @chrishayes, @newsmax, @aynrandpaulryan, @govmikehuckabee, @peculiarbaptist, @aoc, @cnnbrk, @ianbremmer, @davidfrum, @bluelivesmtr, @devinnunes, @alyssamilano, @hillaryclinton, @joenbc, @gopleader, @joebiden, @charliekirk11, @cnn, @cbsnews, @newyorker, @pamelageller, @christianpost, @espn, @barackobama, @davidcorndc, @donaldjtrumpjr, @cnnpolitics, @foxandfriends, @thehill, @franklingraham, @huffpostpol, @100perfedup, @thebushcenter, @sensanders, @realglenmacnow, @rushlimbaugh, @doj, @nfl, @potus, @realdonaldtrump, @nbcnews, @cubs, @guardian, @billoreilly, @senwarren, @jaketapper, @gopchairwoman, @jebbush, @nytimes, @govwhitmer, @davidboreanaz, @foxnews, @speakerryan, @foxnewsnight, @ronbrownstein</p>

Table 6.13: Subset of Top Terms predictive of +3 Engagements

## 6.6 Conclusion

In this chapter, we have proposed methods for forecasting future news engagement activity for users on Twitter. We have also conducted quantitative analysis to determine where our proposed methods beat the baseline when considering difficult to forecast scenarios. Our forecasting experiments indicate that utilizing deep forecasting models such as Bi-LSTM can effectively be used to forecast future news engagement activity. We find that just utilizing prior news engagement counts as features performs substantially better than incorporating other forms of information such as text and hashtags.

We find that for scenarios where users have considerable shift in engagement, our proposed models tend to perform well compared to the baseline. We also find that for engagement scenarios where users go from having some engagement activity to no-engagement activity in the consecutive time period, our proposed models perform better than the baseline approach, while for scenarios where users go from having some news engagement activity to no engagement activity in the consecutive time period all approaches tend to perform in range with the baseline approach. Finally we also find (i) news that discuss polarizing political issues and news events (ii) Distrust towards news media (iii) Religion and (iv) Interactions with hyper-partisan news sources contribute towards users engaging with unreliable news.

# Chapter 7

## Conclusion

In this dissertation, we have proposed and developed computational methods to understand online news engagement in social media. We have analyzed the short term effects of news engagement through the lens of filter bubbles and news recommendation systems using simulation based studies. Additionally, we delved into analyzing long-term news engagement behaviors through the use of observational data, focusing on identifying a particular engagement behavior where the user exhibits a lack of trust towards the news source they engage with, ultimately impacting the diversity of their engagement. Finally, we proposed forecasting models that help provide better insight into the factors that influence user engagement behavior.

From a filter bubble and news recommendation systems perspective we specifically have presented several simulations to understand the relationship between political typology and news recommendation algorithms. We find that users who hold more extreme views are more easily modeled by recommendation systems, leading to higher click-through rates. However, this only occurs with less diverse recommendations in terms of political views and topics. Furthermore, we find that both content-based and collaborative filtering recommendation systems can each result in filter bubbles, though of different types and for different reasons. Finally, we find that users with heterogeneous preferences tend to be recommended articles that reflect more homogeneous viewpoints. We also identified a specific mechanism that

can lead to political homogenization in news recommendation systems, and proposed attention-based neural networks to reduce this behavior. The proposed approach exhibits reduction in the impact of political homogenization for simulated users with opposing political leanings across topics.

From a user behavior modeling perspective we have proposed a methodology for identifying tweets that criticize partisan news media, and we have conducted a descriptive analysis to understand how such tweets vary by user, news source, and time. Some of the substantive results are intuitive – e.g., hyperpartisan users are more likely to criticize media from the other side. 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. We also proposed methods for forecasting future news engagement activity for users on Twitter, and conducted a quantitative analysis to determine where the proposed methods beat the baseline approach in difficult-to-forecast scenarios. The results indicate that deep forecasting models such as Bi-LSTM can effectively forecast future news engagement activity. Moreover, using prior news engagement counts as features alone performs better than incorporating other forms of information such as text and hashtags.

## 7.1 Future Work

### 7.1.1 Filter Bubbles and News Recommendation Systems

The next steps to extend the work discussed in this dissertation would be to study a more *diverse range of modern day recommendation systems* that utilize a more complex modeling procedure such as hybrid (content + collaborative), reinforcement learning and deep learning based recommendation systems. From a simulation perspective, since we mainly assume that a user's news preferences are static across time for this work, in contrast the next obvious step would be to conduct these studies while considering *dynamic user preferences*. Another avenue for extending this work could be through *algorithmic auditing*, which can help quantify the extent to which each of the various bias that were identified actually occur in deployed real world recommender systems.

### 7.1.2 Modeling News Engagement Behavior

The predictive modeling for future news engagement behavior has the potential for expansion in various areas. One such area is the identification of *progression stages*, where models can aid in the detection of clear stages of news engagement for users. These stages can be based on factors like engagement levels (high or low), interaction with news sources with a particular stance (such as +2 to +3), or intent behind engagement (such as trust, support, distrust, or criticize). This information can provide a deeper understanding of the drivers behind user engagement with hyperpartisan and fake news media. Another avenue for future work in this domain can be through estimating *cause and effects* of news engagement on social media, our forecasting models can be used to approximate the counter-factual outcomes, which in turn can help us to measure the effect of various treatments. For example, one objective could be to better understand the role hyper-partisan media plays in

misinformation sharing, to test this hypothesis we could assume engagement with hyper-partisan media (+2 and -2 sources) as the treatment variable.

## References

- [1] Himan Abdollahpouri. Popularity bias in ranking and recommendation. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, AIES '19*, page 529–530, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450363242. doi: 10.1145/3306618.3314309. URL <https://doi.org/10.1145/3306618.3314309>.
- [2] Wai-Ho Au, Keith CC Chan, and Xin Yao. A novel evolutionary data mining algorithm with applications to churn prediction. *IEEE transactions on evolutionary computation*, 7(6):532–545, 2003.
- [3] Isabelle Augenstein, Tim Rocktäschel, Andreas Vlachos, and Kalina Bontcheva. Stance detection with bidirectional conditional encoding. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 876–885, 2016.
- [4] Stephen H Bach, Bryan He, Alexander Ratner, and Christopher Ré. Learning the structure of generative models without labeled data. In *International Conference on Machine Learning*, pages 273–282. PMLR, 2017.
- [5] Hakan Bagci and Pinar Karagoz. Context-aware friend recommendation for location based social networks using random walk. In *Proceedings of the 25th international conference companion on world wide web*, pages 531–536, 2016.
- [6] Christopher A Bail, Lisa P Argyle, Taylor W Brown, John P Bumpus, Haohan



- Chen, MB Fallin Hunzaker, Jaemin Lee, Marcus Mann, Friedolin Merhout, and Alexander Volfovsky. Exposure to opposing views on social media can increase political polarization. *Proceedings of the National Academy of Sciences*, 115(37):9216–9221, 2018.
- [7] Eytan Bakshy, Solomon Messing, and Lada A Adamic. Exposure to ideologically diverse news and opinion on facebook. *Science*, 348(6239):1130–1132, 2015.
- [8] Delia Baldassarri and Andrew Gelman. Partisans without constraint: Political polarization and trends in american public opinion. *American Journal of Sociology*, 114(2):408–446, 2008.
- [9] Linas Baltrunas, Bernd Ludwig, and Francesco Ricci. Matrix factorization techniques for context aware recommendation. In *Proceedings of the fifth ACM conference on Recommender systems*, pages 301–304, 2011.
- [10] Ramy Baly, Giovanni Da San Martino, James Glass, and Preslav Nakov. 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)*, pages 4982–4991, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.404. URL <https://aclanthology.org/2020.emnlp-main.404>.
- [11] Jack Bandy and Nicholas Diakopoulos. Auditing news curation systems: A case study examining algorithmic and editorial logic in apple news. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 36–47, 2020.
- [12] Pablo Barberá, John T Jost, Jonathan Nagler, Joshua A Tucker, and Richard

- Bonneau. Tweeting from left to right: Is online political communication more than an echo chamber? *Psychological science*, 26(10):1531–1542, 2015.
- [13] Francesco Barbieri, Jose Camacho-Collados, Leonardo Neves, and Luis Espinosa-Anke. Tweeteval: Unified benchmark and comparative evaluation for tweet classification. *arXiv preprint arXiv:2010.12421*, 2020.
- [14] Alejandro Bellogín and Javier Parapar. Using graph partitioning techniques for neighbour selection in user-based collaborative filtering. In *Proceedings of the sixth ACM conference on Recommender systems*, pages 213–216, 2012.
- [15] Alejandro Bellogín, Pablo Castells, and Iván Cantador. Neighbor selection and weighting in user-based collaborative filtering: a performance prediction approach. *ACM Transactions on the Web (TWEB)*, 8(2):1–30, 2014.
- [16] Austin R Benson, Ravi Kumar, and Andrew Tomkins. Modeling user consumption sequences. In *Proceedings of the 25th International Conference on World Wide Web*, pages 519–529, 2016.
- [17] Rudolf Beran. Minimum hellinger distance estimates for parametric models. *The annals of Statistics*, pages 445–463, 1977.
- [18] Alessandro Bessi, Fabiana Zollo, Michela Del Vicario, Michelangelo Puliga, Antonio Scala, Guido Caldarelli, Brian Uzzi, and Walter Quattrociocchi. Users polarization on facebook and youtube. *PloS one*, 11(8):e0159641, 2016.
- [19] Rahul Bhargava, Anna Chung, Neil S Gaikwad, Alexis Hope, Dennis Jen, Jamin Rubinovitz, Belén Saldías-Fuentes, and Ethan Zuckerman. 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*, pages 151–155, 2019.

- [20] Engin Bozdag and Jeroen Van Den Hoven. Breaking the filter bubble: democracy and design. *Ethics and information technology*, 17:249–265, 2015.
- [21] Pablo Briñol, Derek D. Rucker, Zakary L. Tormala, and Richard E. Petty. *Individual differences in resistance to persuasion: The role of beliefs and meta-beliefs*, pages 83–104. Routledge Taylor & Francis Group, December 2003. ISBN 9781410609816.
- [22] Jiajun Bu, Shulong Tan, Chun Chen, Can Wang, Hao Wu, Lijun Zhang, and Xiaofei He. Music recommendation by unified hypergraph: combining social media information and music content. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 391–400, 2010.
- [23] Ceren Budak, Sharad Goel, and Justin M Rao. Fair and balanced? quantifying media bias through crowdsourced content analysis. *Public Opinion Quarterly*, 80(S1):250–271, 2016.
- [24] Iván Cantador, Alejandro Bellogín, and David Vallet. Content-based recommendation in social tagging systems. In *Proceedings of the fourth ACM conference on Recommender systems*, pages 237–240, 2010.
- [25] Matt Carlson. Metajournalistic Discourse and the Meanings of Journalism: Definitional Control, Boundary Work, and Legitimation. *Communication Theory*, 26(4):349–368, 11 2016. ISSN 1050-3293. doi: 10.1111/comt.12088. URL <https://doi.org/10.1111/comt.12088>.
- [26] Rich Caruana. Multitask learning. *Machine learning*, 28(1):41–75, 1997.
- [27] L. Elisa Celis, Sayash Kapoor, Farnood Salehi, and Nisheeth Vishnoi. Controlling polarization in personalization: An algorithmic framework. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*,

- FAT\* '19, page 160–169, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450361255. doi: 10.1145/3287560.3287601. URL <https://doi.org/10.1145/3287560.3287601>.
- [28] Allison JB Chaney, Brandon M Stewart, and Barbara E Engelhardt. How algorithmic confounding in recommendation systems increases homogeneity and decreases utility. In *Proceedings of the 12th ACM Conference on Recommender Systems*, pages 224–232, 2018.
- [29] Sneha Chaudhari, Varun Mithal, Gungor Polatkan, and Rohan Ramanath. An attentive survey of attention models. *arXiv preprint arXiv:1904.02874*, 2019.
- [30] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.
- [31] Chong Chen, Min Zhang, Yongfeng Zhang, Yiqun Liu, and Shaoping Ma. Efficient neural matrix factorization without sampling for recommendation. *ACM Transactions on Information Systems (TOIS)*, 38(2):1–28, 2020.
- [32] Wen Chen, Pipei Huang, Jiaming Xu, Xin Guo, Cheng Guo, Fei Sun, Chao Li, Andreas Pfadler, Huan Zhao, and Binqiang Zhao. Pog: personalized outfit generation for fashion recommendation at alibaba ifashion. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 2662–2670, 2019.
- [33] Xu Chen, Hanxiong Chen, Hongteng Xu, Yongfeng Zhang, Yixin Cao, Zheng Qin, and Hongyuan Zha. Personalized fashion recommendation with visual explanations based on multimodal attention network: Towards visually explainable recommendation. In *Proceedings of the 42nd International ACM SIGIR*

- Conference on Research and Development in Information Retrieval*, pages 765–774, 2019.
- [34] David Cheruiyot. *Criticising journalism: popular media criticism in the digital age*. PhD thesis, Karlstads Universitet, 2019.
- [35] Uthsav Chitra and Christopher Musco. Analyzing the impact of filter bubbles on social network polarization. In *Proceedings of the 13th International Conference on Web Search and Data Mining*, pages 115–123, 2020.
- [36] Evangelia Christakopoulou and George Karypis. Local item-item models for top-n recommendation. In *Proceedings of the 10th ACM conference on recommender systems*, pages 67–74, 2016.
- [37] Wei Chu and Seung-Taek Park. Personalized recommendation on dynamic content using predictive bilinear models. In *Proceedings of the 18th international conference on World wide web*, pages 691–700, 2009.
- [38] Michael Conover, Jacob Ratkiewicz, Matthew Francisco, Bruno Gonçalves, Filippo Menczer, and Alessandro Flammini. Political polarization on twitter. In *Proceedings of the international aaai conference on web and social media*, volume 5, pages 89–96, 2011.
- [39] Stephanie Craft, Tim P Vos, and J David Wolfgang. Reader comments as press criticism: Implications for the journalistic field. *Journalism*, 17(6):677–693, 2016. doi: 10.1177/1464884915579332. URL <https://doi.org/10.1177/1464884915579332>.
- [40] Stefano Cresci, Roberto Di Pietro, Marinella Petrocchi, Angelo Spognardi, and Maurizio Tesconi. Fame for sale: Efficient detection of fake twitter followers. *Decision Support Systems*, 80:56–71, 2015.

- [41] P. Dandekar, A. Goel, and D. T. Lee. Biased assimilation, homophily, and the dynamics of polarization. *Proceedings of the National Academy of Sciences*, 110(15):5791–5796, Mar 2013. ISSN 1091-6490. doi: 10.1073/pnas.1217220110. URL <http://dx.doi.org/10.1073/pnas.1217220110>.
- [42] Kareem Darwish, Peter Stefanov, Michaël Aupetit, and Preslav Nakov. Unsupervised user stance detection on twitter. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 141–152, 2020.
- [43] Alexander Philip Dawid and Allan M Skene. 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, 1979.
- [44] Daniel DellaPosta. Pluralistic collapse: The “oil spill” model of mass opinion polarization. *American Sociological Review*, 85(3):507–536, 2020.
- [45] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.
- [46] Qiming Diao, Minghui Qiu, Chao-Yuan Wu, Alexander J Smola, Jing Jiang, and Chong Wang. Jointly modeling aspects, ratings and sentiments for movie recommendation (jmars). In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 193–202, 2014.
- [47] Nicholas Dias, Gordon Pennycook, and David G Rand. Emphasizing publishers does not effectively reduce susceptibility to misinformation on social media. *Harvard Kennedy School Misinformation Review*, 1(1), 2020.
- [48] Daizong Ding, Mi Zhang, Shao-Yuan Li, Jie Tang, Xiaotie Chen, and Zhi-Hua Zhou. Baydnn: Friend recommendation with bayesian personalized ranking

- deep neural network. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 1479–1488, 2017.
- [49] C Doherty, J Kiley, and B Johnson. Political typology reveals deep fissures on the right and left: Conservative republican groups divided on immigration, ‘openness’. *Pew Research Center*, 2017.
- [50] James N. Druckman. Communicating policy-relevant science. *PS: Political Science & Politics*, 48(S1):58–69, 2015.
- [51] James N Druckman and Arthur Lupia. Using frames to make scientific communication more effective. *The Oxford handbook of the science of science communication*, pages 243–252, 2017.
- [52] Susan Dumais, Thorsten Joachims, Krishna Bharat, and Andreas Weigend. Sigir 2003 workshop report: implicit measures of user interests and preferences. In *ACM SIGIR Forum*, volume 37, pages 50–54. ACM New York, NY, USA, 2003.
- [53] Long Duong, Trevor Cohn, Steven Bird, and Paul Cook. Low resource dependency parsing: Cross-lingual parameter sharing in a neural network parser. In *Proceedings of the 53rd annual meeting of the Association for Computational Linguistics and the 7th international joint conference on natural language processing (volume 2: short papers)*, pages 845–850, 2015.
- [54] Gregory Eady, Jonathan Nagler, Andy Guess, Jan Zilinsky, and Joshua A. Tucker. How many people live in political bubbles on social media? evidence from linked survey and twitter data. *SAGE Open*, 9(1):2158244019832705, 2019.
- [55] Gregory Eady, Richard Bonneau, Joshua Tucker, and Jonathan Nagler. News

- sharing on social media: Mapping the ideology of news media content, citizens, and politicians. OSF Preprints, 2021.
- [56] Robert Epstein and Ronald E Robertson. The search engine manipulation effect (seme) and its possible impact on the outcomes of elections. *Proceedings of the National Academy of Sciences*, 112(33):E4512–E4521, 2015.
- [57] Robert Epstein, Ronald E Robertson, David Lazer, and Christo Wilson. Suppressing the search engine manipulation effect (seme). *Proceedings of the ACM on Human-Computer Interaction*, 1(CSCW):1–22, 2017.
- [58] Cédric Févotte and Jérôme Idier. Algorithms for nonnegative matrix factorization with the  $\beta$ -divergence. *Neural computation*, 23(9):2421–2456, 2011.
- [59] Tine Ustad Figenschou and Karoline Andrea Ihlebæk. Challenging journalistic authority. *Journalism Studies*, 20(9):1221–1237, 2019. doi: 10.1080/1461670X.2018.1500868. URL <https://doi.org/10.1080/1461670X.2018.1500868>.
- [60] Seth Flaxman, Sharad Goel, and Justin M Rao. Filter bubbles, echo chambers, and online news consumption. *Public opinion quarterly*, 80(S1):298–320, 2016.
- [61] George Forman. Counting positives accurately despite inaccurate classification. In *European conference on machine learning*, pages 564–575. Springer, 2005.
- [62] Saadia Gabriel, Skyler Hallinan, Maarten Sap, Pemi Nguyen, Franziska Roesner, Eunsol Choi, and Yejin Choi. 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)*, pages 3108–3127, 2022.
- [63] Soumen Ganguly, Juhi Kulshrestha, Jisun An, and Haewoon Kwak. Empirical evaluation of three common assumptions in building political media bias



- datasets. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 939–943, 2020.
- [64] Kiran Garimella, Tim Smith, Rebecca Weiss, and Robert West. Political polarization in online news consumption. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 15, pages 152–162, 2021.
- [65] Venkata Rama Kiran Garimella and Ingmar Weber. A long-term analysis of polarization on twitter. In *Eleventh international AAAI conference on web and social media*, 2017.
- [66] R Kelly Garrett. Echo chambers online?: Politically motivated selective exposure among internet news users. *Journal of computer-mediated communication*, 14(2):265–285, 2009.
- [67] Daniel Geschke, Jan Lorenz, and Peter Holtz. The triple-filter bubble: Using agent-based modelling to test a meta-theoretical framework for the emergence of filter bubbles and echo chambers. *British Journal of Social Psychology*, 58(1):129–149, 2019.
- [68] Prem K Gopalan, Laurent Charlin, and David Blei. Content-based recommendations with poisson factorization. *Advances in neural information processing systems*, 27, 2014.
- [69] A Guess. (almost) everything in moderation: New evidence on americans’ online media diets., 2018.
- [70] Andrew Guess, Benjamin Lyons, Jacob M Montgomery, Brendan Nyhan, and Jason Reifler. Fake news, facebook ads, and misperceptions. *Democracy Fund*, 2019.

- [71] Andrew M Guess, Pablo Barberá, Simon Munzert, and JungHwan Yang. The consequences of online partisan media. *Proceedings of the National Academy of Sciences*, 118(14), 2021.
- [72] Asela Gunawardana and Christopher Meek. A unified approach to building hybrid recommender systems. In *Proceedings of the third ACM conference on Recommender systems*, pages 117–124, 2009.
- [73] Mario Haim, Andreas Graefe, and Hans-Bernd Brosius. Burst of the filter bubble? effects of personalization on the diversity of google news. *Digital journalism*, 6(3):330–343, 2018.
- [74] Casper Hansen, Christian Hansen, Lucas Maystre, Rishabh Mehrotra, Brian Brost, Federico Tomasi, and Mounia Lalmas. Contextual and sequential user embeddings for large-scale music recommendation. In *Proceedings of the 14th ACM Conference on Recommender Systems*, pages 53–62, 2020.
- [75] Mahmud Hasan, Mehmet A Orgun, and Rolf Schwitter. A survey on real-time event detection from the twitter data stream. *Journal of Information Science*, 44(4):443–463, 2018.
- [76] Xiangnan He, Hanwang Zhang, Min-Yen Kan, and Tat-Seng Chua. Fast matrix factorization for online recommendation with implicit feedback. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 549–558, 2016.
- [77] Xiangnan He, Zhankui He, Jingkuan Song, Zhenguang Liu, Yu-Gang Jiang, and Tat-Seng Chua. Nais: Neural attentive item similarity model for recommendation. *IEEE Transactions on Knowledge and Data Engineering*, 30(12): 2354–2366, 2018.

- [78] Bas Hofstra, Rense Corten, Frank Van Tubergen, and Nicole B Ellison. Sources of segregation in social networks: A novel approach using facebook. *American Sociological Review*, 82(3):625–656, 2017.
- [79] Linmei Hu, Chen Li, Chuan Shi, Cheng Yang, and Chao Shao. Graph neural news recommendation with long-term and short-term interest modeling. *Information Processing & Management*, 57(2):102142, 2020.
- [80] Yang Hu, Xi Yi, and Larry S Davis. Collaborative fashion recommendation: A functional tensor factorization approach. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 129–138, 2015.
- [81] Mingqing Huang, Qingshan Jiang, Qiang Qu, Lifei Chen, and Hui Chen. Information fusion oriented heterogeneous social network for friend recommendation via community detection. *Applied Soft Computing*, 114:108103, 2022.
- [82] Eugene Ie, Chih wei Hsu, Martin Mladenov, Vihan Jain, Sanmit Narvekar, Jing Wang, Rui Wu, and Craig Boutilier. Recsim: A configurable simulation platform for recommender systems. *arXiv*, 2019.
- [83] William G Jacoby. Neither liberal nor conservative: Ideological innocence in the american public. *Political Science Quarterly*, 133(4):758–761, 2018.
- [84] Sarthak Jain and Byron C Wallace. Attention is not explanation. *arXiv preprint arXiv:1902.10186*, 2019.
- [85] Mohsen Jamali and Martin Ester. A matrix factorization technique with trust propagation for recommendation in social networks. In *Proceedings of the fourth ACM conference on Recommender systems*, pages 135–142, 2010.
- [86] Ray Jiang, Silvia Chiappa, Tor Lattimore, András György, and Pushmeet Kohli.

- Degenerate feedback loops in recommender systems. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 383–390, 2019.
- [87] Christopher C Johnson. Logistic matrix factorization for implicit feedback data. *Advances in Neural Information Processing Systems*, 27:78, 2014.
- [88] M Jurkowitz and A Mitchell. About one-fifth of democrats and republicans get political news in a kind of media bubble. *Pew Research Center*, 2020.
- [89] Komal Kapoor, Mingxuan Sun, Jaideep Srivastava, and Tao Ye. A hazard based approach to user return time prediction. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1719–1728, 2014.
- [90] Giannis Karamanolakis, Daniel Hsu, and Luis Gravano. 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)*, pages 4611–4621, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1468. URL <https://aclanthology.org/D19-1468>.
- [91] Giannis Karamanolakis, Subhabrata Mukherjee, Guoqing Zheng, and Ahmed Hassan Awadallah. Self-training with weak supervision. *arXiv preprint arXiv:2104.05514*, 2021.
- [92] Michael Karlsson and Christer Clerwall. Cornerstones in journalism. *Journalism Studies*, 20(8):1184–1199, 2019. doi: 10.1080/1461670X.2018.1499436. URL <https://doi.org/10.1080/1461670X.2018.1499436>.

- [93] Jaya Kawale, Aditya Pal, and Jaideep Srivastava. Churn prediction in mmorpgs: A social influence based approach. In *2009 international conference on computational science and engineering*, volume 4, pages 423–428. IEEE, 2009.
- [94] Katherine Keith and Brendan O’Connor. Uncertainty-aware generative models for inferring document class prevalence. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4575–4585, 2018.
- [95] Dhruv Khattar, Vaibhav Kumar, Vasudeva Varma, and Manish Gupta. Weave&rec: A word embedding based 3-d convolutional network for news recommendation. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pages 1855–1858, 2018.
- [96] Sami Khenissi and Olfa Nasraoui. Modeling and counteracting exposure bias in recommender systems, 2020.
- [97] Donghyun Kim, Chanyoung Park, Jinoh Oh, Sungyoung Lee, and Hwanjo Yu. Convolutional matrix factorization for document context-aware recommendation. In *Proceedings of the 10th ACM conference on recommender systems*, pages 233–240, 2016.
- [98] Hyun-Chul Kim and Zoubin Ghahramani. Bayesian classifier combination. In *Artificial Intelligence and Statistics*, pages 619–627. PMLR, 2012.
- [99] Gary King, Benjamin Schneer, and Ariel White. How the news media activate public expression and influence national agendas. *Science*, 358(6364):776–780, 2017.
- [100] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

- [101] R Kiran, Pradeep Kumar, and Bharat Bhasker. Dnnrec: A novel deep learning based hybrid recommender system. *Expert Systems with Applications*, 144: 113054, 2020.
- [102] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521 (7553):436–444, 2015.
- [103] Yuan Li, Benjamin Rubinstein, and Trevor Cohn. Exploiting worker correlation for label aggregation in crowdsourcing. In *International conference on machine learning*, pages 3886–3895. PMLR, 2019.
- [104] Jacob Liedke and Jeffrey Gottfried. 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/>, 2022.
- [105] Zhouhan Lin, Minwei Feng, Cicero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. A structured self-attentive sentence embedding. *arXiv preprint arXiv:1703.03130*, 2017.
- [106] Ping Liu, Karthik Shivaram, Aron Culotta, Matthew A Shapiro, and Mustafa Bilgic. The interaction between political typology and filter bubbles in news recommendation algorithms. In *Proceedings of the Web Conference 2021*, pages 3791–3801, 2021.
- [107] Shikun Liu, Edward Johns, and Andrew J. Davison. End-to-end multi-task learning with attention, 2019.
- [108] Yozen Liu, Xiaolin Shi, Lucas Pierce, and Xiang Ren. 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*, pages 2023–2031, 2019.

- [109] Caroline Lo, Dan Frankowski, and Jure Leskovec. Understanding behaviors that lead to purchasing: A case study of pinterest. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 531–540, 2016.
- [110] Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Philip S Yu. Learning multiple tasks with multilinear relationship networks. *Advances in neural information processing systems*, 30, 2017.
- [111] Yongxi Lu, Abhishek Kumar, Shuangfei Zhai, Yu Cheng, Tara Javidi, and Rogério Feris. Fully-adaptive feature sharing in multi-task networks with applications in person attribute classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5334–5343, 2017.
- [112] Lauren Lutzke, Caitlin Drummond, Paul Slovic, and Joseph Árvai. Priming critical thinking: Simple interventions limit the influence of fake news about climate change on facebook. *Global Environmental Change*, 58:101964, 2019.
- [113] Hanjia Lyu and Jiebo Luo. 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, page 4072–4082, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450392037. doi: 10.1145/3503161.3547898. URL <https://doi.org/10.1145/3503161.3547898>.
- [114] Hao Ma, Haixuan Yang, Michael R Lyu, and Irwin King. Sorec: social recommendation using probabilistic matrix factorization. In *Proceedings of the 17th*

- ACM conference on Information and knowledge management*, pages 931–940, 2008.
- [115] Masoud Mansoury, Himan Abdollahpouri, Mykola Pechenizkiy, Bamshad Mobasher, and Robin Burke. *Feedback Loop and Bias Amplification in Recommender Systems*, page 2145–2148. Association for Computing Machinery, New York, NY, USA, 2020. ISBN 9781450368599. URL <https://doi.org/10.1145/3340531.3412152>.
- [116] Farzan Masrour, Tyler Wilson, Heng Yan, Pang-Ning Tan, and Abdol Esfahanian. Bursting the filter bubble: Fairness-aware network link prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 841–848, 2020.
- [117] Nicolas Mattis, Philipp Masur, Judith Möller, and Wouter van Atteveldt. Nudging towards news diversity: A theoretical framework for facilitating diverse news consumption through recommender design. *new media & society*, page 14614448221104413, 2022.
- [118] Aaron M McCright and Riley E Dunlap. The politicization of climate change and polarization in the american public’s views of global warming, 2001–2010. *The Sociological Quarterly*, 52(2):155–194, 2011.
- [119] Miller McPherson, Lynn Smith-Lovin, and James M Cook. Birds of a feather: Homophily in social networks. *Annual review of sociology*, 27(1):415–444, 2001.
- [120] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.



- [121] Seong Jae Min and Donghee Yvette Wohn. All the news that you don't like: Cross-cutting exposure and political participation in the age of social media. *Computers in Human Behavior*, 83:24 – 31, 2018. ISSN 0747-5632.
- [122] Ishan Misra, Abhinav Shrivastava, Abhinav Gupta, and Martial Hebert. Cross-stitch networks for multi-task learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3994–4003, 2016.
- [123] Michael Moricz, Yerbolat Dosbayev, and Mikhail Berlyant. Pymk: friend recommendation at myspace. In *Proceedings of the 2010 ACM SIGMOD International Conference on Management of data*, pages 999–1002, 2010.
- [124] Jonathan Mummolo. News from the other side: How topic relevance limits the prevalence of partisan selective exposure. *The Journal of Politics*, 78(3): 763–773, 2016.
- [125] Efrat Nechushtai, Rodrigo Zamith, and Seth C Lewis. More of the same? homogenization in news recommendations when users search on google, youtube, facebook, and twitter. *Mass Communication and Society*, pages 1–27, 2023.
- [126] Raymond S Nickerson. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of general psychology*, 2(2):175–220, 1998.
- [127] Xia Ning and George Karypis. Slim: Sparse linear methods for top-n recommender systems. In *2011 IEEE 11th international conference on data mining*, pages 497–506. IEEE, 2011.
- [128] Ruchi Ookalkar, Kolli Vishal Reddy, and Eric Gilbert. Pop: Bursting news filter bubbles on twitter through diverse exposure. In *Conference Companion Publication of the 2019 on Computer Supported Cooperative Work and Social Computing*, pages 18–22, 2019.

- [129] Mathias Osmundsen, Alexander Bor, Peter Bjerregaard Vahlstrup, Anja Bechmann, and Michael Bang Petersen. Partisan polarization is the primary psychological motivation behind political fake news sharing on twitter. *American Political Science Review*, 115(3):999–1015, 2021. doi: 10.1017/S0003055421000290.
- [130] Eli Pariser. *The filter bubble: What the Internet is hiding from you*. Penguin UK, 2011.
- [131] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035, 2019. URL <http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf>.
- [132] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.
- [133] Michael J Pazzani and Daniel Billsus. Content-based recommendation systems. *The adaptive web: methods and strategies of web personalization*, pages 325–341, 2007.
- [134] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron

- Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830, 2011.
- [135] Gordon Pennycook, Jonathon McPhetres, Yunhao Zhang, Jackson G Lu, and David G Rand. Fighting covid-19 misinformation on social media: Experimental evidence for a scalable accuracy-nudge intervention. *Psychological science*, 31(7):770–780, 2020.
- [136] Pew Research Center. Beyond red vs. blue: The political typology. <https://www.pewresearch.org/politics/2021/11/09/beyond-red-vs-blue-the-political-typology-2/>, 2021. Accessed: 2021-12-01.
- [137] Minh Hieu Phan and Philip O Ogunbona. Modelling context and syntactical features for aspect-based sentiment analysis. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3211–3220, 2020.
- [138] Whitney Phillips. *This is why we can't have nice things: Mapping the relationship between online trolling and mainstream culture*. Mit Press, 2015.
- [139] Markus Prior. Media and political polarization. *Annual Review of Political Science*, 16:101–127, 2013.
- [140] Friedrich Pukelsheim. The three sigma rule. *The American Statistician*, 48(2): 88–91, 1994.
- [141] Tao Qi, Fangzhao Wu, Chuhan Wu, Peiru Yang, Yang Yu, Xing Xie, and Yongfeng Huang. Hierec: Hierarchical user interest modeling for personalized news recommendation. *arXiv preprint arXiv:2106.04408*, 2021.

- [142] Lijing Qin and Xiaoyan Zhu. Promoting diversity in recommendation by entropy regularizer. In *Twenty-Third International Joint Conference on Artificial Intelligence*. Citeseer, 2013.
- [143] Zhaopeng Qiu, Yunfan Hu, and Xian Wu. Graph neural news recommendation with user existing and potential interest modeling. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 16(5):1–17, 2022.
- [144] Linhao Qu, Shaolei Liu, Manning Wang, and Zhijian Song. Transmef: A transformer-based multi-exposure image fusion framework using self-supervised multi-task learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2126–2134, 2022.
- [145] Steve Rathje, Jay J Van Bavel, and Sander van der Linden. Out-group animosity drives engagement on social media. *Proceedings of the National Academy of Sciences*, 118(26), 2021.
- [146] Alex Ratner, Braden Hancock, Jared Dunnmon, Roger Goldman, and Christopher Ré. Snorkel metal: Weak supervision for multi-task learning. In *Proceedings of the Second Workshop on Data Management for End-To-End Machine Learning*, pages 1–4, 2018.
- [147] Alexander Ratner, Stephen H Bach, Henry Ehrenberg, Jason Fries, Sen Wu, and Christopher Ré. Snorkel: Rapid training data creation with weak supervision. In *Proceedings of the VLDB Endowment. International Conference on Very Large Data Bases*, volume 11, page 269. NIH Public Access, 2017.
- [148] Alexander Ratner, Braden Hancock, Jared Dunnmon, Frederic Sala, Shreyash Pandey, and Christopher Ré. Training complex models with multi-task weak

- supervision. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 4763–4771, 2019.
- [149] SRS Reddy, Sravani Nalluri, Subramanyam Kuniseti, S Ashok, and B Venkatesh. Content-based movie recommendation system using genre correlation. In *Smart Intelligent Computing and Applications: Proceedings of the Second International Conference on SCI 2018, Volume 2*, pages 391–397. Springer, 2019.
- [150] Cristian G Rodriguez, Jake P Moskowicz, Rammy M Salem, and Peter H Ditto. Partisan selective exposure: The role of party, ideology and ideological extremity over time. *Translational Issues in Psychological Science*, 3(3):254, 2017.
- [151] Sebastian Ruder. An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*, 2017.
- [152] Sebastian Ruder, Parsa Ghaffari, and John G Breslin. A hierarchical model of reviews for aspect-based sentiment analysis. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 999–1005, 2016.
- [153] Alan Said, Shlomo Berkovsky, and Ernesto W De Luca. Putting things in context: Challenge on context-aware movie recommendation. In *Proceedings of the workshop on context-aware movie recommendation*, pages 2–6, 2010.
- [154] Matthew J. Salganik and Duncan J. Watts. Leading the herd astray: An experimental study of self-fulfilling prophecies in an artificial cultural market. *Social Psychology Quarterly*, 71(4):338–355, 2008. ISSN 01902725.
- [155] Glenn S Sanders and Brian Mullen. Accuracy in perceptions of consensus:

- Differential tendencies of people with majority and minority positions. *European journal of social psychology*, 13(1):57–70, 1983.
- [156] Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl. Item-based collaborative filtering recommendation algorithms. In *Proceedings of the 10th international conference on World Wide Web*, pages 285–295, 2001.
- [157] Markus Schedl and David Hauger. Tailoring music recommendations to users by considering diversity, mainstreaminess, and novelty. In *Proceedings of the 38th international acm sigir conference on research and development in information retrieval*, pages 947–950, 2015.
- [158] Sven Schmit and Carlos Riquelme. Human interaction with recommendation systems. In *International Conference on Artificial Intelligence and Statistics*, pages 862–870, 2018.
- [159] Mike Schuster and Kuldip K Paliwal. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, 45(11):2673–2681, 1997.
- [160] Suvash Sedhain, Aditya Krishna Menon, Scott Sanner, and Lexing Xie. Autotrec: Autoencoders meet collaborative filtering. In *Proceedings of the 24th international conference on World Wide Web*, pages 111–112, 2015.
- [161] Sofia Serrano and Noah A Smith. Is attention interpretable? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2931–2951, 2019.
- [162] Matthew A. Shapiro and Han Woo Park. More than entertainment: Youtube and public responses to the science of global warming and climate change. *Social Science Information*, 54(1):115–145, 2015.

- [163] Matthew A. Shapiro and Han Woo Park. Climate change and youtube: Deliberation potential in post-video discussions. *Environmental Communication*, 12(1):115–131, 2018.
- [164] Yue Shi, Martha Larson, and Alan Hanjalic. Exploiting user similarity based on rated-item pools for improved user-based collaborative filtering. In *Proceedings of the third ACM conference on Recommender systems*, pages 125–132, 2009.
- [165] Wooyeol Shin, Changwook Kim, and Jaewon Joo. Hating journalism: Anti-press discourse and negative emotions toward journalism in korea. *Journalism*, 22(5):1239–1255, 2021. doi: 10.1177/1464884920985729. URL <https://doi.org/10.1177/1464884920985729>.
- [166] Karthik Shivaram, Ping Liu, Matthew Shapiro, Mustafa Bilgic, and Aron Culotta. Reducing cross-topic political homogenization in content-based news recommendation. In *Proceedings of the 16th ACM Conference on Recommender Systems*, pages 220–228, 2022.
- [167] Yotam Shmargad and Samara Klar. Sorting the news: How ranking by popularity polarizes our politics. *Political Communication*, 37(3):423–446, 2020.
- [168] Yotam Shmargad and Samara Klar. Sorting the news: How ranking by popularity polarizes our politics. *Political Communication*, 37(3):423–446, 2020.
- [169] Jiangbo Shu, Xiaoxuan Shen, Hai Liu, Baolin Yi, and Zhaoli Zhang. A content-based recommendation algorithm for learning resources. *Multimedia Systems*, 24(2):163–173, 2018.
- [170] Jagendra Singh, Mohammad Sajid, Chandra Shekhar Yadav, Shashank Sheshar Singh, and Manthan Saini. A novel deep neural-based music recommendation

- method considering user and song data. In *2022 6th International Conference on Trends in Electronics and Informatics (ICOEI)*, pages 1–7. IEEE, 2022.
- [171] Barea Sinno, Bernardo Oviedo, Katherine Atwell, Malihe Alikhani, and Junyi Jessy Li. 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*, pages 231–243, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.17. URL <https://aclanthology.org/2022.naacl-main.17>.
- [172] Dominic Spohr. Fake news and ideological polarization: Filter bubbles and selective exposure on social media. *Business information review*, 34(3):150–160, 2017.
- [173] Peter Stefanov, Kareem Darwish, Atanas Atanasov, and Preslav Nakov. Predicting the topical stance and political leaning of media using tweets. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 527–537, 2020.
- [174] Ana-Andreea Stoica, Christopher Riederer, and Augustin Chaintreau. Algorithmic glass ceiling in social networks: The effects of social recommendations on network diversity. In *Proceedings of the 2018 World Wide Web Conference*, pages 923–932, 2018.
- [175] Florian Strub, Romaric Gaudel, and Jérémie Mary. Hybrid recommender system based on autoencoders. In *Proceedings of the 1st workshop on deep learning for recommender systems*, pages 11–16, 2016.
- [176] Qingying Sun, Zhongqing Wang, Qiaoming Zhu, and Guodong Zhou. Stance



- detection with hierarchical attention network. In *Proceedings of the 27th international conference on computational linguistics*, pages 2399–2409, 2018.
- [177] Xianfeng Tang, Yozen Liu, Neil Shah, Xiaolin Shi, Prasenjit Mitra, and Suhang Wang. Knowing your fate: Friendship, action and temporal explanations for user engagement prediction on social apps. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 2269–2279, 2020.
- [178] Aaron Van den Oord, Sander Dieleman, and Benjamin Schrauwen. Deep content-based music recommendation. *Advances in neural information processing systems*, 26, 2013.
- [179] Mason Walker and Katerina Eva Malsa. News consumption across social media in 2021. *Pew Research Center*, 2021.
- [180] Shengxian Wan, Yanyan Lan, Jiafeng Guo, Chaosheng Fan, and Xueqi Cheng. Informational friend recommendation in social media. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, pages 1045–1048, 2013.
- [181] Donghui Wang, Yanchun Liang, Dong Xu, Xiaoyue Feng, and Renchu Guan. A content-based recommender system for computer science publications. *Knowledge-Based Systems*, 157:1–9, 2018.
- [182] Hao Wang, Xingjian Shi, and Dit-Yan Yeung. Collaborative recurrent autoencoder: Recommend while learning to fill in the blanks. *Advances in Neural Information Processing Systems*, 29, 2016.
- [183] Xinxi Wang and Ye Wang. Improving content-based and hybrid music recom-

- mendation using deep learning. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 627–636, 2014.
- [184] Zhibo Wang, Jilong Liao, Qing Cao, Hairong Qi, and Zhi Wang. Friendbook: a semantic-based friend recommendation system for social networks. *IEEE transactions on mobile computing*, 14(3):538–551, 2014.
- [185] Shouxian Wei, Xiaolin Zheng, Deren Chen, and Chaochao Chen. A hybrid approach for movie recommendation via tags and ratings. *Electronic Commerce Research and Applications*, 18:83–94, 2016.
- [186] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Huggingface’s transformers: State-of-the-art natural language processing, 2020.
- [187] Chuhan Wu, Fangzhao Wu, Mingxiao An, Jianqiang Huang, Yongfeng Huang, and Xing Xie. Npa: neural news recommendation with personalized attention. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 2576–2584, 2019.
- [188] Chuhan Wu, Fangzhao Wu, Suyu Ge, Tao Qi, Yongfeng Huang, and Xing Xie. Neural news recommendation with multi-head self-attention. 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)*, pages 6389–6394, 2019.
- [189] Chuhan Wu, Fangzhao Wu, Tao Qi, and Yongfeng Huang. Empowering news

- recommendation with pre-trained language models. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1652–1656, 2021.
- [190] Guixian Xu, Yueting Meng, Zhan Chen, Xiaoyu Qiu, Changzhi Wang, and Haishen Yao. Research on topic detection and tracking for online news texts. *IEEE access*, 7:58407–58418, 2019.
- [191] Hu Xu, Bing Liu, Lei Shu, and S Yu Philip. Bert post-training for review reading comprehension and aspect-based sentiment analysis. 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)*, pages 2324–2335, 2019.
- [192] Feng Xue, Xiangnan He, Xiang Wang, Jiandong Xu, Kai Liu, and Richang Hong. Deep item-based collaborative filtering for top-n recommendation. *ACM Transactions on Information Systems (TOIS)*, 37(3):1–25, 2019.
- [193] Carl Yang, Xiaolin Shi, Luo Jie, and Jiawei Han. 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*, pages 914–922, 2018.
- [194] Yongxin Yang and Timothy M Hospedales. Trace norm regularised deep multi-task learning. *arXiv preprint arXiv:1606.04038*, 2016.
- [195] Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. Hierarchical attention networks for document classification. In *Proceedings of the 2016 conference of the North American chapter of the association*

- for computational linguistics: human language technologies*, pages 1480–1489, 2016.
- [196] Ruiping Yin, Kan Li, Jie Lu, and Guangquan Zhang. Enhancing fashion recommendation with visual compatibility relationship. In *The world wide web conference*, pages 3434–3440, 2019.
- [197] Peilin Yu, Tiffany Ding, and Stephen H Bach. Learning from multiple noisy partial labelers. In *International Conference on Artificial Intelligence and Statistics*, pages 11072–11095. PMLR, 2022.
- [198] Jia Zhang, Yaojin Lin, Menglei Lin, and Jinghua Liu. An effective collaborative filtering algorithm based on user preference clustering. *Applied Intelligence*, 45: 230–240, 2016.
- [199] Jiang Zhang, Yufeng Wang, Zhiyuan Yuan, and Qun Jin. Personalized real-time movie recommendation system: Practical prototype and evaluation. *Tsinghua Science and Technology*, 25(2):180–191, 2019.
- [200] Jieyu Zhang, Yue Yu, Yinghao Li, Yujing Wang, Yaming Yang, Mao Yang, and Alexander Ratner. Wrench: A comprehensive benchmark for weak supervision. *arXiv preprint arXiv:2109.11377*, 2021.
- [201] Jinbo Zhang, Zhiqing Lin, Bo Xiao, and Chuang Zhang. An optimized item-based collaborative filtering recommendation algorithm. In *2009 IEEE International Conference on Network Infrastructure and Digital Content*, pages 414–418. IEEE, 2009.
- [202] Qi Zhang, Jingjie Li, Qinglin Jia, Chuyuan Wang, Jieming Zhu, Zhaowei Wang, and Xiuqiang He. Unbert: User-news matching bert for news recommendation.

In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI)*, pages 3356–3362, 2021.

- [203] Xinyang Zhang, Yury Malkov, Omar Florez, Serim Park, Brian McWilliams, Jiawei Han, and Ahmed El-Kishky. Twihin-bert: A socially-enriched pre-trained language model for multilingual tweet representations. *arXiv preprint arXiv:2209.07562*, 2022.
- [204] Lili Zhao, Zhongqi Lu, Sinno Jialin Pan, Qiang Yang, and Wei Xu. Matrix factorization+ for movie recommendation. In *IJCAI*, pages 3945–3951, 2016.
- [205] Qiannan Zhu, Xiaofei Zhou, Zeliang Song, Jianlong Tan, and Li Guo. Dan: Deep attention neural network for news recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 5973–5980, 2019.