TU WIEN Informatics

Exploring Group Fairness in News Media Recommendations: Algorithms, Metrics, and Grouping

DIPLOMARBEIT

zur Erlangung des akademischen Grades

Diplom-Ingenieur

im Rahmen des Studiums

Data Science

eingereicht von

Blake Huebner, MSc, BSc
Matrikelnummer 12139066

an der Fakultät für Informatik

der Technischen Universität Wien

Betreuung: Univ.Ass.in Mag.a rer.nat. Dr.in techn. Julia Neidhardt
Mitwirkung: Projektass. Dipl.-Ing. Thomas Elmar Kolb, BSc

Wien, 16. März 2023

Blake Huebner	Julia Neidhardt

TU WIEN Informatics

Exploring Group Fairness in News Media Recommendations: Algorithms, Metrics, and Grouping

DIPLOMA THESIS

submitted in partial fulfillment of the requirements for the degree of

Diplom-Ingenieur

in

Data Science

by

Blake Huebner, MSc, BSc
Registration Number 12139066

to the Faculty of Informatics

at the TU Wien

Advisor: Univ.Ass.in Mag.a rer.nat. Dr.in techn. Julia Neidhardt
Assistance: Projektass. Dipl.-Ing. Thomas Elmar Kolb, BSc

Vienna, 16th March, 2023

Blake Huebner	Julia Neidhardt

Erklärung zur Verfassung der Arbeit

Blake Huebner, MSc, BSc

Hiermit erkläre ich, dass ich diese Arbeit selbständig verfasst habe, dass ich die verwendeten Quellen und Hilfsmittel vollständig angegeben habe und dass ich die Stellen der Arbeit – einschließlich Tabellen, Karten und Abbildungen –, die anderen Werken oder dem Internet im Wortlaut oder dem Sinn nach entnommen sind, auf jeden Fall unter Angabe der Quelle als Entlehnung kenntlich gemacht habe.

Wien, 16. März 2023

Blake Huebner

Danksagung

Ich möchte mich bei meiner Betreuerin Julia Neidhardt für ihre ständige Unterstützung und ihr Feedback herzlich bedanken. Ein weiterer Dank, geht an Thomas E. Kolb für seine kontinuierliche Unterstützung während der gesamten Projektdauer. Es war für mich eine Freude, meine Masterarbeit im Christian Doppler Lab for Recommender Systems, einschließlich der Laborpartnern, schreiben zu können. Die Zusammenarbeit mit dem Team, war für mich ein großes Vergnügen.

Abschließend möchte ich Nora Pauelsen meinen besonderen Dank aussprechen, mit welcher ich gemeinsam alle Herausforderungen dieses Masterstudiengangs gemeistert habe.

Acknowledgements

First, I would like to extend a big thanks to my supervisor Julia Neidhardt for her constant support and feedback. Many thanks also to Thomas E. Kolb for his continued assistance throughout the entirety of the project. I am also grateful for the opportunity to work with the amazing team at the Christian Doppler Lab for Recommender Systems and all partners of the lab. The whole team was a pleasure to work with.

Finally, I would like to extend a special thanks to Nora Pauelsen, with whom I have faced every challenge of this master's program together.

Kurzfassung

Neben den Genauigkeitsmetriken sind Fairness und Diversität zu weithin untersuchten Themen in Empfehlungssystemen geworden. Die Verbesserung dieser Metriken ist nicht nur aus ethischer und rechtlicher Sicht wichtig, sondern kann auch die allgemeine Zufriedenheit der Nutzer und Nutzerinnen verbessern. Obwohl Fairness- und Diversitätsmetriken weithin diskutiert werden, gibt es nur sehr wenige empirische Untersuchungen, insbesondere zum Vergleich verschiedener Algorithmen mit unterschiedlichen Metriken. Diese Arbeit untersucht die Rolle von Fairness und Diversität in Empfehlungssystemen für Nachrichten, insbesondere im Kontext der österreichischen Medienlandschaft. Ziel dieser Studie ist es, die effektivsten Ansätze zur Generierung von fairen und vielfältigen Nachrichtenempfehlungen zu identifizieren und gleichzeitig die möglichen negativen Folgen von einseitigen Empfehlungen und Filterblasen, wie z.B. gesellschaftliche Polarisierung und die Unterdrückung von Informationen, zu berücksichtigen. Die Forschungsmethoden umfassen eine umfassende Literaturrecherche über relevante Metriken zur Gruppenungerechtigkeit und modernste Algorithmen, die auf Fairness ausgerichtet sind. Darüber hinaus wurde ein Datensatz von Artikeln einer österreichischen Zeitung für die empirische Untersuchung verwendet, wobei die politische Ausrichtung, Fairness und Vielfalt der Empfehlungen analysiert wurden. Die Kernergebnisse dieser Studie sind, dass Genauigkeit und Fairness mit dem richtigen Modellierungsansatz gleichzeitig erreicht werden können, während die Vielfalt mit diesen Modellierungstechniken konstant gehalten werden kann. Die Studie empfiehlt die Verwendung von personalisierten Fairness-Modellen auf der Grundlage von Kausalvorstellungen für die Genauigkeit und die Verringerung bestimmter Unfairness-Kennzahlen und stellt fest, dass Fairness-Ziele für kollaborative Filtermodelle bei der Verringerung anderer Arten von Unfairness effektiver sind. Die Ergebnisse leisten einen Beitrag zum Fachgebiet, indem sie zeigen, wie wichtig es ist, Fairness- und Diversitätsmetriken in die Entwicklung und Bewertung von Empfehlungssystemen einzubeziehen und indem sie Hinweise auf die effektivsten Ansätze zur Erreichung dieser Ziele geben. Die Studie gibt weiters interessante Einblicke in das Leseverhalten und die politische Ausrichtung von Nachrichtenartikeln, welche von Österreicherinnen und Österreichern gelesen werden und zeigt den Bedarf an weiterer Forschung in diesem Bereich auf.

Abstract

Beyond accuracy metrics, such as fairness and diversity, have become widely studied topics in recommender systems. Improving these metrics is important not only from an ethical and legal perspective, but can also improve overall user satisfaction. Although fairness and diversity metrics are widely discussed, very little empirical research has been done, especially comparing multiple algorithms across different metrics. This thesis explores the role of fairness and diversity in news recommender systems, specifically in the context of the Austrian media landscape. This study aims to identify the most effective approaches for generating fair and diverse news recommendations, while addressing the potential negative consequences of biased recommendations and filter bubbles, such as societal polarization and the suppression of information. The research methods include an extensive literature review of relevant group unfairness metrics and state-of-the-art fairness-aware algorithms. In addition, a dataset of articles from an Austrian newspaper was used for empirical research, with analysis performed on political labeling, fairness, and diversity of recommendations. The key message of the study is that accuracy and fairness can be achieved simultaneously with the right modeling approach, while diversity can be held constant using these modeling techniques. The study recommends the use of Personalized Fairness based on Causal Notion models for accuracy and reducing certain unfairness metrics, and finds Fairness Objectives for Collaborative Filtering models more effective at reducing other types of unfairness. The findings contribute to the field by demonstrating the importance of incorporating fairness and diversity metrics into the design and evaluation of recommender systems, and by providing guidance on the most effective approaches to achieve these goals. The study also reveals interesting insights into the reading behaviors and political lean of news articles read by Austrians, and suggests the need for further research in this area.

Contents

Kurzfassung **xi**

Abstract **xiii**

Contents **xv**

1 Introduction **1**
1.1 Preface . . . 1
1.2 Motivation . . . 1
1.3 Considerations . . . 3
1.4 Research Questions . . . 4
1.5 Methodology . . . 4

2 Background **7**
2.1 Imbalanced Data . . . 7
2.2 Metrics . . . 8
2.3 Algorithms . . . 12

3 Empirical Research **19**
3.1 Data Description . . . 19
3.2 Gender . . . 21
3.3 Geography . . . 24

4 Comparing Fairness-Aware NRS Across Groups **29**
4.1 Preprocessing . . . 29
4.2 Results . . . 31

5 Discussion **39**
5.1 Political Labeling . . . 39
5.2 FOCF . . . 40
5.3 PFCN . . . 41

6 Conclusion **43**
6.1 Summary . . . 43

6.2 Contribution 44
6.3 Future Work 45

List of Figures **46**

List of Tables **47**

Bibliography **49**

CHAPTER 1

Introduction

1.1 Preface

Recommender systems (RS) have gained such vast popularity during the internet age, that they can be seen in all facets of our daily lives. We are surrounded by recommendations of what movies to watch, what clothes to buy, and what news to read. Not only are these systems instrumental in helping us make quicker decisions, but they have also become essential as the amount of information available on the internet far exceeds our processing capacity [Jac84]. However, as recommender systems have risen in popularity, so has our realization of possible negative aspects that come along with these recommendations. Negative effects such as filter bubbles, bias, and unfair recommendations have come into the spotlight recently. To combat these unintended consequences, many authors have focused on beyond-accuracy measures, such as diversity, serendipity, and fairness. In fact, in the past five years, more than 60 fairness-related recommender system papers have been published in top conferences and journals [WMZ+22].

1.2 Motivation

According to research performed by the Pew Research Center in 2020, approximately 86% of adults in the US consume news online through digital newspapers, social media, news apps, etc [She21]. With the majority of adults receiving their news online, there is an obligation to create recommendations that are fair and unbiased. At an individual level, overly personalized news recommendations may cause filter bubbles by not exposing the reader to other points of view, which over the long term can cause them to avoid counter-attitudinal information [Hel19]. At the societal level, this behavior can create more polarization in an already polarized world and poses a threat to democracy in a sense through the unintended suppression of information.

Fairness Definition. This thesis will focus on fairness of news recommender systems (NRS) between groups using the *consistent fairness* (CF) definition. Specifically, we define fairness as:

> *"The lack of discrimination against a certain group, ie the absence of a differential impact on the outcomes created for them."* [Meh22]

CF indicates that groups of users should be treated similarly, and therefore measurements of CF measure the inconsistency of the utility distribution [WMZ+22].

Most research involving group fairness revolves around fairness of accuracy [ETA+18, SHFT15, MZS20], and it is easy to see the implications of unfairness of group accuracy. On the surface level, it is important to ensure user satisfaction by providing recommendations of equal quality to all users using a service. Despite this importance, inequalities still exist, such as female users and older users receiving worse movie and music recommendations [ETA+18]. On a deeper level, unfair recommendations may have major societal impacts, such as the Google study which showed that female users are less likely to see job postings for highly compensated executive positions [DTD15].

Although fairness of accuracy is an important topic, existing research has shown unfairness and bias also exist in beyond-accuracy recommendation metrics like diversity and novelty [WC21]. Increasing diversity in particular has been linked to user satisfaction and better quality perception by users, therefore it is an important metric to consider in RS. [WMZ+22, EHWK14, PCH11, ZMKL05].

In the news domain in particular, diversity is crucial not only to keep users interested and engaged but also to expose users to counter-attitudinal behavior and keep readers from becoming trapped in filter bubbles [RD20]. For this reason, in addition to focusing on fairness of accuracy, this thesis also includes fairness of diversity in the analysis.

User Grouping. There are a variety of ways to divide groups, with common divisions including gender, age, and race. This research will split the data by two sensitive attributes: first by gender and then by geographic location.

Geographic location grouping is particularly interesting when it comes to news recommendations with the current state of the political divide in the United States. In the 2016 presidential election, Hillary Clinton won eight of the ten largest metropolitan areas and the majority of metropolitan areas of more than one million residents; Donald Trump won all other types of areas [Max19]. In the 2018 midterm election, Democrats won every congressional district in the largest urban areas, while Republicans won 87% of rural districts [Max19]. This difference in voting preference is attributed in part to the polarization between individuals living in large cities and small towns. City dwellers tend to be more diverse, better educated, and more likely to work in white-collar jobs [LTDSL22]. They also tend to have more progressive values around gender rights, homosexuality, immigration, and the family dynamic [LTDSL22].

This voting trend can also be seen in the 2019 Austrian parliament elections, where Vienna was the only Austrian state to vote differently than the others. Accordingly, this thesis explores the concepts of fairness and bias in the context of Austrian news media, with the motivation to ensure recommender systems do not create an even larger divide in society through selective recommendations and filter bubbles. Anonymized data is provided by the Austrian newspaper *STANDARD Verlagsgesellschaft m.b.H.*[1](*Der Standard*).

1.3 Considerations

In order to ensure the efficacy of recommender systems, it is crucial to mitigate unfairness phenomena. There are several reasons why fairness is of utmost importance.

1. From an ethics perspective, Aristotle listed fairness as one of the crucial virtues to make people live well, making it an important ethical consideration that remains a fundamental requirement for a just society [ARB09, Raw99].

2. From a legal perspective, anti-discrimination laws [Hol05] require that public services, employment, admissions, housing, etc., do not discriminate against different groups based on age, race, gender, and other such factors. For instance, in a job recommendation scenario, minority-owned companies should receive recommendations at a rate equivalent to white-owned companies.

3. From an individual perspective, a fair recommender system helps expose users to diverse information in the recommendations, including niche information, which can help break filter bubbles, alleviate societal polarization, broaden horizons, and enhance the value of recommendations.

4. From a product perspective, fairness is crucial for the success of recommender systems in the long run. When a system is unfair, it may lead to negative user experiences by recommending popular content to users with niche interests, or provide insufficient exposure to niche providers, thus limiting the diversity of content and users on the platform, and ultimately impacting its growth [MCBP+20].

[1] `https://www.derstandard.at/`

1.4 Research Questions

The main objective of this thesis is to implement multiple fairness-aware news recommender systems and compare the results across various metrics. The data is grouped by two sensitive attributes: gender and geographical location. Specifically, the main research questions of this thesis are defined as:

RQ1: To what extent do different reading behaviors of different groups impact recommender systems and recommendations in the domain of news media?

RQ2: When is it appropriate to use different fairness metrics for NRS?

RQ3: To what extent do different recommender system algorithms affect these fairness metrics, in terms of improvement made, cost to accuracy, affect to diversity etc.?

1.5 Methodology

The methodological approach follows a common data science methodology called Cross-Industry Standard Process for Data Mining (CRISP-DM) [She00], in addition to a Literature Review. Figure 1.1 shows the CRISP-DM process. The steps in the context of this thesis are:

1. **Literature Review**

   To establish the current state of group fairness research in recommender systems, the literature review in Chapter 2 provides the background on various fairness metrics and algorithms to improve group fairness.

2. **Business (Problem) Understanding**

   At this stage, we worked with *Der Standard* to understand any needs the business might have.

3. **Data Understanding**

   At this point, time was spent to understand where the data comes from, how it is collected, and what is contained in the data. Chapter 3 outlines the background of the data and an exploratory data analysis of applicable fields.

4. **Data Preparation**

   This stage is covered in Chapter 3 when data needed to be prepared for modeling. Steps include:

   - Click data was filtered to relevant click events.
   - Articles were translated into English and political lean was predicted using a Natural Language Processing (NLP) algorithm.

- Users were grouped according to two sensitive attributes: gender and geographical location.

5. **Modeling**

   Chapter 4, Section 4.1 covers the modeling section, in which hyperparameters were tuned and the models were applied to the data.

6. **Evaluation**

   Chapter 4, Section 4.2 covers the evaluation of the applied algorithms.

7. **Deployment**

   Deployment is not covered as part of this thesis.

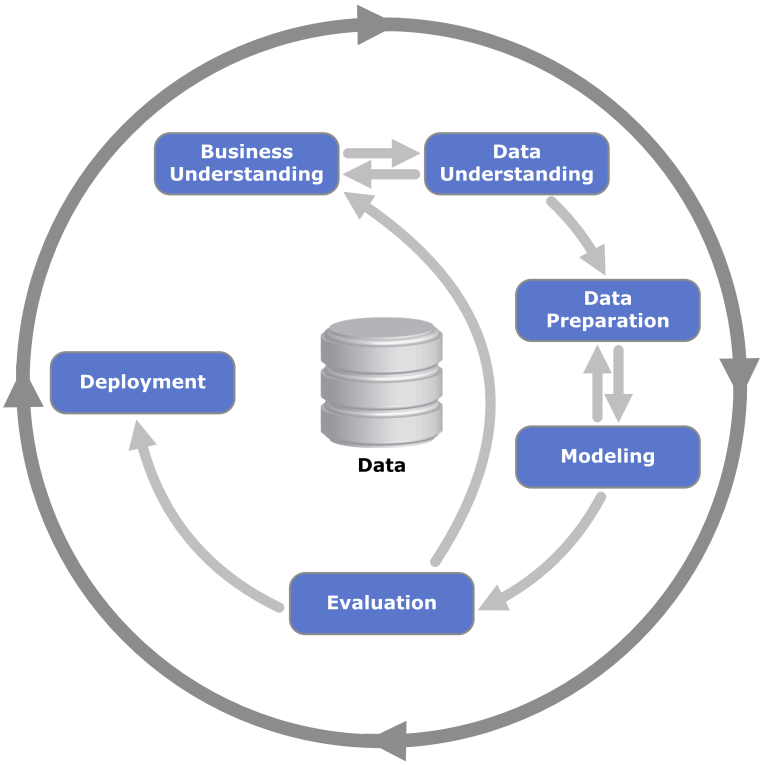

Figure 1.1: CRISP-DM process diagram [Jen12].

CHAPTER 2

Background

In order to fully understand the importance of fairness within NRS, this chapter starts with the main causes of unfairness. It then moves into how fairness can be measured via various metrics, and provides an understanding of algorithms including both traditional and fairness-aware algorithms. This will serve as the theoretical background before moving into the application of these topics.

2.1 Imbalanced Data

Bias in recommender systems can have a significant impact on the recommendations provided to users. There are two main forms of bias that can be inherited from data: observation bias and bias that comes from imbalanced data [YH17].

Observation bias is caused by a feedback loop in the data collection and recommendation system. An item displayed by the recommender system may result in user feedback for that item, which is then fed back into the model [FKT+18]. This feedback loop reinforces the ranking algorithm used by the recommender system to make more recommendations that are similar to those made previously. As a result, if a user is never exposed to a product, they are unable to provide feedback on it, which results in a lack of diversity in the suggestions made. For example, if a user on a movie streaming website was never given movie recommendations from a certain genre, the system will never learn the user's preference of that genre, leading to a bias.

On the other hand, an *imbalance in data* is a result when systematic bias is present due to societal, historical, or other ambient biases. It is more difficult to address this type of bias because the model is not aware of it. For example, in the context of women in STEM (science, technology, engineering, mathematics), due to societal problems, there may be a smaller proportion of women who succeed in STEM as compared to men [YH17]. This bias is not known to the model, which may only know the gender of users, but not their

proclivity towards STEM [YH17]. These biases have been researched and addressed in different contexts with use of multi-arm bandits and diversity-based recommendations [FKT+18]. Although these approaches tend to handle bias by increasing the diversity of a recommender system, they do not directly address the issue of fairness. More recently, fairness in recommender systems has been explored through the use of fairness metrics, which will be discussed in the next section.

2.2 Metrics

First, we provide the notations and explanations of commonly used variables in the chapter (see in Table 2.1).

Notation	Explanation
n	number of users
m	number of items
$\hat{r}_{u,i}$	prediction for user u on item i
$r_{u,i}$	ground truth feedback of user u on item i
$U = u_1, ..., u_n$	the whole set of users
$I = i_1, ..., i_m$	the whole set of items
$\hat{L}(u)$	a ranked list of items that a model produces for user u
$L(u)$	a ground-truth set of items that user u has interacted with
K	the length of the recommendation list
G_j	the j-th group of individuals
$E_j[\hat{r}]_i$	the average rating predicted rating of group j on item i
$E_j[r]_i$	the average rating of group j on item i

Table 2.1: Notations and explanations of common variables.

2.2.1 Common Accuracy Metrics

This section provides the definitions for common accuracy metrics of recommender systems which will be used as part of the analysis section. In all metrics, we use the Top-K recommendation list.

Normalized Discounted Cumulative Gain (NDCG@K). *NDCG@K* [BYRNo99] is a measure of ranking quality in which positions are discounted logarithmically. This gives higher weights to correctly recommended items in higher ranks. It is defined as the Discounted Cumulative Gain (DCG) divided by the Ideal Discounted Cumulative Gain (IDCG). Consider a ranked list where i_j represents the item at rank j. DCG is defined as:

$$DCG@K = \sum_{j=1}^{K} \frac{r_{u,i_j}}{log(j+1)}$$

and IDCG is the same equation as DCG, but with an ideal ranking (in which the highest ranked items take the highest ranking positions, perfectly). In this aspect, the metric is normalized by dividing by IDCG. Formally, NDCG is defined as:

$$NDCG@K = \frac{DCG@K}{IDCG@K}$$

Recall@K. This measure computes the proportion of relevant items recommended divided by the total number of relevant items [Wan21]. Higher recall means the system has successfully recommended a high proportion of relevant items. In the context of news recommender systems, it is important to capture the highest proportion of relevant articles as possible for the user. The measure is defined as:

$$Recall@K = \frac{1}{|U|} \sum_{u \in U} \frac{|\hat{L}(u) \cap L(u)|}{|L(u)|}$$

Hit@K. Also known as Hit Ratio (HR), this measure calculates the percentage of users with at least one relevant article in their recommendation list [AHD20]. It is defined as:

$$HR@K = \frac{1}{|U|} \sum_{u \in U} \delta \left(\hat{L}(u) \cap L(u) \neq \emptyset \right)$$

where δ represents an indicator functions that returns 1 when the condition is true and 0 when the condition is false. $\emptyset$ denotes an empty set.

Mean Reciprocal Rank (MRR@K). This metric computes the reciprocal rank of the first relevant item in the recommendation list [Wan21]. High MRR represents the algorithm finding a relevant item quickly in the ranking. It is defined as:

$$MRR@K = \frac{1}{|U|} \sum_{u \in U} \frac{1}{rank_u^*}$$

where $rank_u^*$ is defined as the rank position of the first relevant item found in the recommendation list.

2.2.2 Common Fairness Metrics

Fairness is an interesting concept in recommender systems because its definitions and corresponding metrics can vary widely. As stated previously, this paper will focus on group fairness using the definition of consistent fairness. Even with this narrower focus on fairness, there exist at least nine different metrics found in the literature. These nine metrics are displayed in Table 2.2. Since this thesis will focus only on comparing binary groups (ex. Male vs Female), only metrics that compare binary groups are covered.

In order to help with the understanding of some metrics below, we will use the example introduced above regarding the proportions of males vs females in STEM. In this context, our example is concerning a recommender system recommending (or not recommending) university students to take STEM courses.

Metric Name	Def	Target		
		Two Groups	More Groups	Individual
Absolute Difference	CO	✓	X	X
Value Unfairness	CO	✓	X	X
Absolute Unfairness	CO	✓	X	X
Underestimation Unfairness	CO	✓	X	X
Overestimation Unfairness	CO	✓	X	X
Non-Parity	CO	✓	X	X
Variance	CO	-	✓	✓
Min-Max Difference	CO	-	✓	✓
F-Statistic of ANOVA	CO	-	✓	X

Table 2.2: List of group fairness metrics. Table inspired by [WMZ+22].

Absolute Difference Unfairness (AD). *Absolute difference unfairness* revolves around the concept of fairness in the quality of recommendations between two groups. It is defined as the absolute difference of the average recommendation performance between group G_0 and group G_1. Different recommendation performance metrics may be used, such as $NDCG@K$ in [LCF+21], [FXG+20]. Some papers use this metric especially to check for fairness when group data is unbalanced [LCF+21], [FXG+20]. Lower values indicate fairer recommendations. For a function f computing average recommendation performance:

$$AD = |f(G_0) - f(G_1)|$$

Value Unfairness. *Value unfairness* [YH17] was created to measure the inconsistency in signed error between two user groups. Value unfairness is maximized when one group is consistently given recommendations above their true preference, and the other group is consistently given recommendations below their true preference. An example would be female students being under-recommended STEM courses, and male students being

over-recommended STEM courses. The equation for value unfairness is given as:

$$U_{val} = \frac{1}{m}\sum_{i=1} m|(E_0[\hat{r}]_i - E_0[r]_i) - (E_1[\hat{r}]_i - E_1[r]_i)|$$

Absolute Unfairness. *Absolute unfairness* [YH17], on the other hand, measures the inconsistency of absolute prediction error. Since absolute unfairness is unsigned, it captures solely the difference in reconstruction error between the two groups. Absolute unfairness is maximized when one group is given recommendations perfectly matching their preferences, and the other is given recommendations that are very different than their preferences. Absolute unfairness is defined as:

$$U_{abs} = \frac{1}{m}\sum_{i=1} m||E_0[\hat{r}]_i - E_0[r]_i| - |E_1[\hat{r}]_i - E_1[r]_i||$$

Underestimation Unfairness. *Underestimation unfairness* [YH17] measures how much the predictions underestimate the true ratings. Underestimation unfairness is important in situations where missing recommendations are weighed more heavily than extra recommendations. For example, underestimation unfairness could lead to students missing out on courses they would excel in.

$$U_{under} = \frac{1}{m}\sum_{i=1} m|max(0, E_0[r]_i - E_0[\hat{r}]_i) - max(0, E_1[r]_i - E_1[\hat{r}]_i)|$$

Overestimation Unfairness. *Overestimation unfairness* [YH17] measures the degree to which the predictions overestimate the true ratings. Overestimation unfairness can cause users to invest large amounts of time in order to sift through recommendations.

$$U_{over} = \frac{1}{m}\sum_{i=1} m|max(0, E_0[\hat{r}]_i - E_0[r]_i) - max(0, E_1[\hat{r}]_i - E_1[r]_i)|$$

Non-Parity Unfairness. *Non-parity unfairness* [YH17] is defined as the absolute difference in overall predictions between the two groups.

$$U_{nonpar} = |E_0[y] - E_1[y]|$$

2.2.3 Fairness of Diversity

This paper includes fairness of diversity in its main scope. Diversity in recommender systems refers to the diversity of items given in a recommendation list. When discussing

diversity fairness, most research refers to item fairness, in which recommender systems unfairly promote popular items, therefore creating a feedback loop. However, fairness of diversity can also be examined from the side of the users [WC21]. This unfairness occurs when certain users are receiving a more diverse recommendation list than other users. Some studies have found that existing methods to improve recommendation diversity can worsen user unfairness [LAK18], so it is important to add this metric to the model comparison. We will use two metrics in this analysis.

Entropy. *Entropy* is a common measure of diversity in recommender systems [MAP+20, MAP+22, PBG+20]. It measures the uniformity of a distribution; in the case of recommender systems, it measures the uniformity of a distribution of recommended items. A uniform distribution has the highest entropy or information gain, therefore high entropy is desired to increase diversity. Where $p(i)$ represents the probability of event i, entropy is defined as:

$$Entropy = -\sum_{i \in I} p(i) * \log p(i)$$

Dissimilarity. Entropy is focused more on measuring the exposure of items in a recommendation list, but we are also interested in the diversity of content of the items. For this, we introduce item *dissimilarity* as a measure [KP17, SAHN22]. The intra-list diversity can be measured by computing the cosine similarity of each pair of items (i,j) in the recommendation list $\hat{L}$, and calculating the average. The dissimilarity measure is defined as:

$$Dissimilarity = \frac{2}{K(K-1)} \sum_{i=1}^{K} \sum_{j=1}^{K} (1 - cos(i,j))$$

2.3 Algorithms

2.3.1 Traditional Algorithms

In the news domain, recommender systems are used to help readers discover and consume relevant news articles. News recommender systems, in particular, face a number of unique challenges, including the dynamic nature of the news environment, the need to consider the relevance of news articles, and the fact that users' interests may change over time.

Content-based filtering (CBF) algorithms are often used in news recommender systems because they can take into account the content and attributes of the news articles being recommended. These algorithms build recommendation lists by comparing the user profile and item profile based on the content of a shared attribute space [RD22]. CBF algorithms can be effective in recommending news articles that are relevant to the user's evolving interests, but they may struggle to handle large numbers of temporary or anonymous users, and may not be able to fully capture the semantics and context of the news content.

Collaborative filtering (CF) algorithms, on the other hand, are content-free and rely on the interactions between users and news articles to make recommendations [RD22]. CF algorithms can be effective in recommending news articles that other users with similar interests have enjoyed, but they require a sufficient amount of data about users' interactions with news articles to be effective. This can be a challenge in the fast-paced news environment, where the value of news articles can decay quickly.

In order to combine the strengths of both CF and CBF algorithms, hybrid approaches use both collaborative and content-based information to make recommendations [RD22]. These approaches can be particularly effective in situations where both types of information are available. Figure 2.1 shows the difference between the two methods.

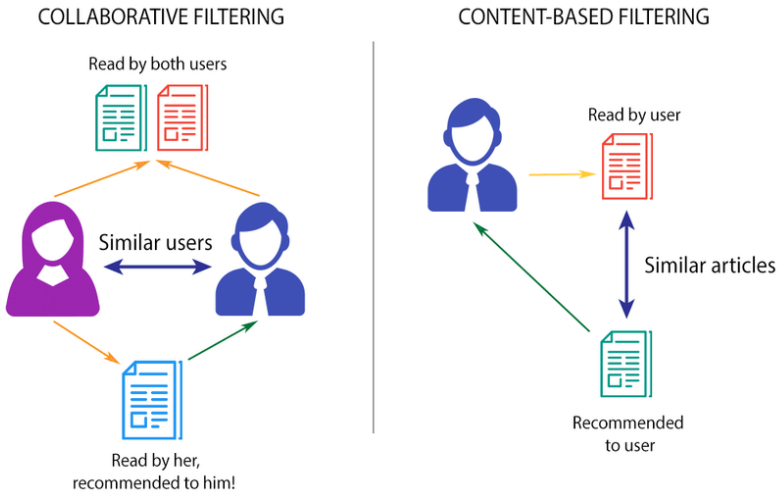

Figure 2.1: Content-Based Filtering vs Collaborative Filtering by [Ton18]

2.3.2 Matrix Factorization

Building on the collaborative filtering algorithm is *matrix factorization* (MF), which is a popular model-based method that first gained recognition in the Netflix competition [KBV09]. Matrix factorization models map users and items to a joint latent factor space, such that user-item interactions are modeled as inner products in that space [KBV09]. These methods have become popular in recent years due to their combination of scalability and predictive accuracy while also offering flexibility for modeling various real-life situations [KBV09].

The rating of user u on item i can be represented by the following equation:

$$r_{ui} \approx \mathbf{p}_u^T \mathbf{q}_i + w_u + v_i$$

where p_u is the matrix representing the uth user and q_i is the matrix representing the ith item, and w_u and v_i are scalar bias terms for user and item [YH17]. The goal is to learn the matrices $\mathbf{P}$ and $\mathbf{Q}$ by minimizing the Mean Square Error, plus a regularization

term to prevent overfitting (where λ is the coefficient of regularization), making the total cost function:

$$J = \frac{1}{|R|} \sum_{r_{u,i} \in R} (y_{ui} - r_{ui})^2 + \frac{\lambda}{2}(||\mathbf{P}^2|| + ||\mathbf{Q}^2||)$$

where R represents the observed ratings.

Matrix factorization will help serve as a baseline when comparing modeling results in later chapters.

2.3.3 Fairness Objectives for Collaborative Filtering (FOCF)

The problem with these traditional algorithms is that they are susceptible to bias as discussed previously, which may result in unfairness to certain protected groups. Recently, a number of techniques have emerged with the intention of focusing on improving fairness as it relates to the previously discussed metrics.

The initial strategy to protect algorithms from unfair and biased decisions was to remove sensitive features. This strategy can alleviate a certain amount of unfairness, but it is often incomplete. Oftentimes, sensitive features are correlated with unprotected features, and as a result, a model may still make biased decisions [KAS11]. In addition, methods such as collaborative filtering may be able to infer latent user attributes from a user's behavior [YH17].

Another common strategy is to enforce demographic parity, which aims to guarantee the overall proportion of members in a protected group has the same positive or negative outcomes as the rest of the population [YH17]. For example, where we have a binary decision $\hat{Y} \in \{0, 1\}$ and a binary protected attribute $A \in \{0, 1\}$, the constraint is expressed as [HPS16]:

$$Pr\{\hat{Y} = 1 | A = 0\} = Pr\{\hat{Y} = 1 | A = 1\}$$

Methods [KAS11] have been created to solve this parity issue by adding a regularization term to control for demographic parity. However, this approach is limited in that it doesn't take into account the fact that sensitive features are commonly correlated with user preferences. Features such as gender, race, and age typically affect user ratings, and therefore this method may make the recommender system less effective.

To address the issue of dependence on preferences and sensitive features, Hardt, *et al.* [HPS16] propose a new method. Given the same binary decision and sensitive feature in the previous example, and the true label $Y \in \{0, 1\}$, they propose the constraint [HPS16]:

$$Pr\{\hat{Y} = 1 | A = 0, Y = y\} = P\{\hat{Y} = 1 | A = 1, Y = y\}, y \in \{0, 1\}$$

This constraint encourages fairness which also respects differences in group preferences, and is the basis of the fairness metrics proposed by Yao & Huang [YH17].

This brings us to the algorithms created by Yao & Huang [YH17], which are based on the previously discussed metrics such as value unfairness, absolute unfairness, etc. They

use standard matrix factorization and collaborative filtering methods to learn latent representations of users and items; however, to improve model fairness, they augment the objectives with these unfairness metrics as penalty functions and minimize these new cost functions [YH17].

The authors found that not only could each metric be optimized separately, but also optimizing any of the metrics typically decreased the other unfairness metrics [YH17]. They also found that optimizing these metrics leads to no significant increase in reconstruction error [YH17].

2.3.4 Personalized Fairness based on Causal Notion (PFCN)

Causal Fairness. The FOCF method is an *associated-based* (or correlation-based) fairness notion, which mainly focuses on discovering differences in statistical metrics between groups. Contrary to association-based concepts, *causal-based notions* use causal models, which take advantage of prior information about the structure of the world to explain how variable changes in the system propagate [LCX+21]. Causal-based fairness notions have become more researched in recent years to address unfairness in machine learning methods [KLFH19, KLRS18, ZB18a, ZB18b]. The authors of *Towards Personalized Fairness based on Causal Notion* [LCX+21] build their fairness-aware technique around causal notions.

In this method, in order to ensure fairness, the probability distributions of the model outcomes must be equivalent in the factual and counterfactual world in relation to each individual [LCX+21]. In other words, for a recommender system to be fair in terms of causal fairness, if a user's sensitive features are changed there should be no changes in the recommendations results, given the other features that are not dependent on the sensitive features remain unchanged. In some cases, this method may be more reasonable than enforcing equality of metrics such as in association-based notions, as users of different groups may have different preferences.

The authors [LCX+21] define fairness (specifically counterfactual fairness), as a recommender model in which for all possible users u with sensitive features $Z = z$ and features that are not causally dependent on Z, $X = x$:

$$P(L_z|X = x, Z = z) = P(L_{z^{\text{‘}}}|X = x, Z = z)$$

for all L and for any value $z^{\text{‘}}$ attainable by Z, where L denotes the Top-N recommendation list for user u.

Framework. In classification tasks, the most straightforward way to ensure independence between sensitive features and predictions is to avoid using sensitive features in the candidate input feature set [KLRS18]. However, it is not so easy with recommender systems. Typical recommender systems take advantage of the user-item interaction matrix

and the model may capture the relevance between user features and user behaviors if there is a causal impact present. Instead, counterfactual fairness in recommender systems must be realized in a nontrivial way.

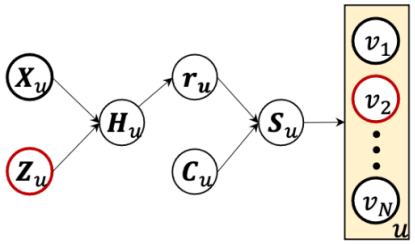

Figure 2.2: Causal path from [LCX+21].

The causal relations for recommendation systems are represented in Figure 2.2. For any given user u, X_u and Z_u represent the insensitive and sensitive user features, respectively. H_u is the user-item interaction history, r_u is the user embedding, C_u is the candidate item set, and S_u is the predicted scores for candidate items.

For any given user u, the scoring function S_u creates the recommendation list by taking the user embedding r_u and the candidate item embedding C_u as inputs [LCX+21]. As shown in 2.2, the user embedding r_u is learned from the user history matrix H_u, which depends on the user features X_u (non-sensitive features) and Z_u (sensitive features) [LCX+21]. As a result, to meet the counterfactual fairness requirement, the causal path between Z_u and the recommendation output shows that only the independence between r_u and Z_u must be ensured [LCX+21].

The Model. The main idea of the authors' method is to use adversary learning; specifically, the authors train a predictor that learns informative representations for the recommendation task, while simultaneously training an adversarial classifier that seeks to minimize the predictor's ability to predict the sensitive features from the representation [LCX+21]. This strategy thereby removes the sensitive feature information from the user embedding.

The architecture of the model is shown in Figure 2.3. The first step is to introduce a set of filter functions that filter the sensitive feature information from the user embeddings r_u. The authors discuss two methods of training filter functions, however, for our purposes only one will be discussed. That method is called the *separate method* (SM), and in this method, one filter function is trained for each potential combination of sensitive features. For example, if **Q** is the set of sensitive features containing age and gender, three filter functions are trained: $f_A, f_G, f_{A,G}$.

To learn the filter functions, adversary learning is used to train a set of discriminators; where k represents the number of sensitive features, each sensitive feature Z_k, a classifier C_k is trained which aims to predict Z_k from the user embeddings [LCX+21]. Therefore,

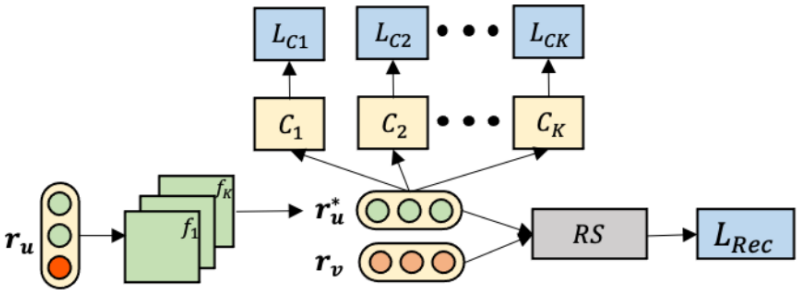

Figure 2.3: Framework from [LCX+21].

the objective of the filter functions is to make it difficult to predict the sensitive features from the user embeddings, and the objective of the discriminator is to defeat the filter functions. Since our analysis trains separate models for each feature, we will adapt the model accordingly.

The loss of the recommendation task L_{Rec} and the loss of the discriminators L_C (to predict the sensitive features) are optimized together. Therefore, the adversary learning loss L is defined as:

$$L = \sum_{u,v} \left(L_{Rec}(u, v) - \lambda \sum_{z \in Z} L_C(r_u^*, z) \right)$$

where λ denotes the adversarial coefficient and controls the trade-off between performance and fairness.

In the original paper [LCX+21], the authors apply their technique to four different methods, to demonstrate its effectiveness on both deep and shallow learning algorithms. This paper will apply their technique to those same methods, and therefore, a short introduction to each method is provided here.

Probabilistic matrix factorization (PMF). PMF is a class of MF models having intuition coming from Bayesian learning for parameter estimation, in which it adds Gaussian observation noise into the user and item latent factor distributions [MS07]. The model scales well with large numbers of observations, and sparse, imbalanced datasets [MS07]. In a social recommender system, PMF is utilized to integrate the user-item rating matrix and social network structure, and a similar concept is applied in an NRS [LXL+12], which addresses the data sparsity issue by including news content, user interactions, and social network information into the PMF model.

PMF runs under the assumption that the entries of the user-item interaction matrix are normally distributed with mean equal to the inner product of the user and item vectors with constant variance. Under this model, the user-factor and item-factor matrices are learned by maximizing the log-likelihood of the observed ratings under the

probabilistic model. PMF is typically more effective than standard matrix factorization in recommendation settings because the model can effectively handle missing values and uncertainty in the rating matrix.

Biased Matrix Factorization (BiasedMF). This algorithm is the same as the matrix factorization method (subsection 2.3.2) discussed previously but adds a global bias term [KBV09]. The bias term takes into account that some users may have a higher tendency to rate items positively or negatively, and that some items may be inherently more popular or less popular than others. BiasedMF aims to model these biases along with the latent factors that represent user and item preferences.

Wide & Deep model (WideDeep). The Wide & Deep model, shown in Figure 2.4, attempts to achieve both memorization and generalization in recommender systems by combining a generalized linear model (wide model) with a neural network (deep model) [CKH+16]. The linear model consists of a wide set of cross-product feature transformations and is able to memorize feature interactions, but generalization requires a large amount of effort at the feature engineering stage. The deep neural network requires less feature engineering and can generalize more effectively to new feature combinations using low-dimensional dense embeddings learned for sparse features. However, deep neural networks can over-generalize and are less effective with sparse user-item interactions. Recommendations involving memorization are typically more topical and relevant to the items which users have already rated, whereas generalization helps improve the diversity of recommendations [CKH+16].

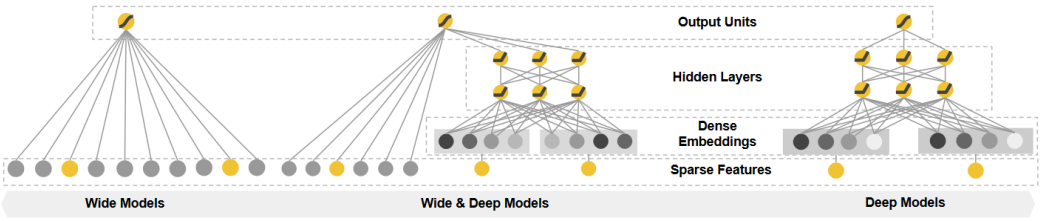

Figure 2.4: Visualization of the Wide and Deep model [CKH+16].

Deep Matrix Factorization (DMF). DMF [XDZ+17] extends and adapts the approach of matrix factorization by introducing a deep neural network to model the interactions between users and items. The model takes the user and item embeddings obtained from the factorization step as input and a fully connected layer learns the non-linear interactions between user and item embeddings. The authors train the model using a combination of stochastic gradient descent and backpropagation to minimize the MSE [XDZ+17]. The authors show the model performs well, especially on sparse datasets [XDZ+17].

CHAPTER 3

Empirical Research

In order to understand the recommendations that an NRS produces, it is important to understand the underlying reading behaviors of the database of users. This section will take a deeper dive into the data we will be using, and analyze the differences in reading behaviors of our specified groups.

3.1 Data Description

Data is provided by the Austrian newspaper *Der Standard*, and includes customer data collected from the year 2000 through 10 November 2021. All of the data have been completely anonymized. Relevant data includes gender, user creation date, time of visit, geolocation of visit, channel resort, object type, and article publishing date.

User Interactions. This analysis uses the last 30 days of user interaction data available from *Der Standard*; The dates are between 12 October 2021 and 10 November 2021. For this analysis, we consider a click on an article to be user interest in that article, and therefore consider it an interaction with that item. Only interactions with articles were used. There is also the case of interactions on a single article multiple times; the user may have read part of the article and came back to it later, or other situations in which the user returned to the same article. The amount of interactions by a user on each article is not considered in this analysis. In this event, only the first click is kept for the purposes of recording day, time, geolocation, etc.

Political Labels. The political leaning of articles is also of interest in this section, so the text from the articles was prepared. Articles of the categories: *Diverse, Inland, International, Panorama, Europe, Columns, Commentary, World Chronicle, Economic Policy, and Austria Chronicle* were first translated into English and then ran through a

semantic analysis NLP model[1]. The model was initially trained on a political bias dataset from Media Bias Fact Check[2]. The training dataset contains more than 52k articles from 100+ news outlets. Labels were created by human annotators with the following possible labels: *extreme right, right, right center, least biased, left center, left, and not labeled.* For the purposes of this analysis, articles predicted as *extreme right* were given the label *right* because of the very small number of predictions with that label.

Predicting political bias is already a difficult task, and first having to translate the articles makes the task of an NLP model even more difficult. After applying the model, unfortunately, most of the articles were scored as unknown (78%). Table 3.1 displays the political label broken down by article category. The analysis focuses only on articles with labeled predictions. Figure 3.1 illustrates the lean of each category.

Subcategory	Left	LeftCenter	LeastBiased	RightCenter	Right	Unknown
Diverse	0.000	0.065	0.261	0.032	0.279	0.363
Inland	0.000	0.065	0.012	0.000	0.080	0.842
International	0.005	0.130	0.014	0.000	0.116	0.735
Panorama	0.001	0.054	0.023	0.000	0.074	0.849
Europe	0.001	0.113	0.031	0.001	0.033	0.821
Columns	0.000	0.075	0.000	0.000	0.075	0.850
Commentary	0.005	0.119	0.014	0.003	0.031	0.828
World Chronicle	0.000	0.095	0.095	0.003	0.041	0.766
Economic Policy	0.001	0.126	0.051	0.003	0.072	0.747
Austria Chronicle	0.001	0.093	0.059	0.000	0.071	0.776
All	0.001	0.086	0.043	0.003	0.080	0.786

Table 3.1: Political lean prediction by subcategory.

It is interesting to note the category differences. *Diverse* seems to be the most right-leaning on average, and none of the categories sticks out as the most left. *Diverse* is also the most non-polarizing with the most articles labeled as *least biased*, whereas the proportion of unbiased articles in *Inland, International, Columns, and Commentary* is much lower, making them more polarizing categories. As an overall note, the vast majority of articles leaning right are labeled as *right* as opposed to *right center*. On the contrary, left-leaning articles are almost always labeled *left center*. It is unclear why this occurs in the data. It is possible the left-leaning articles tend to be more unbiased than the right, or it is possible the human annotators of the training set were more likely to label articles as *left center*.

[1] https://huggingface.co/valurank/distilroberta-mbfc-bias
[2] https://mediabiasfactcheck.com/

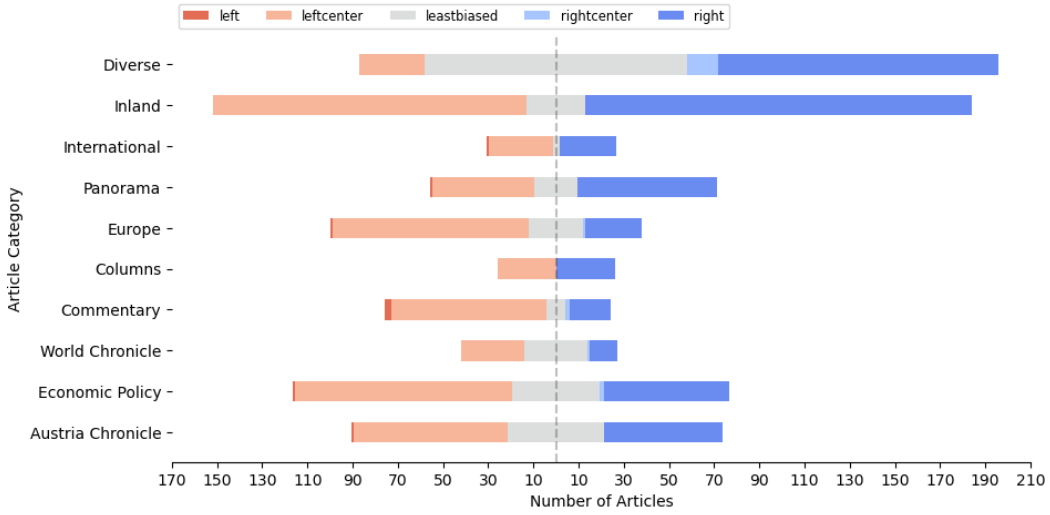

Figure 3.1: Article political leaning by category.

3.2 Gender

Our first sensitive attribute is gender. In this section, we will explore how gender is specified and analyze the different reading behaviors across genders.

3.2.1 Exploratory Data Analysis

After registering for a user account, a user is allowed to select a gender of either *male* or *female*, or leave the information blank. Before 2017, users were prompted to select a gender when signing up for an account, whereas in 2017 and after, there is no prompt to assign a gender. This is apparent in the data (see Figure 3.2), where the proportion of *Not Specified* gender sharply increases in 2017. For the purposes of this analysis, only users with assigned gender are included. Consequently, the data contains a higher proportion of users who created their account before 2017 than is actually representative. Nevertheless, this gender analysis makes use of the data available with the consideration that the proportion of users with gender is much lower for newer users.

After merging the user data with the click data, some interesting statistics arise in Table 3.2. Although females make up 36% of the overall user base, they make up less than 20% of the users who actively read an article in the last 30 days. Females were also active on less days as compared to their male counterparts, and on those active days read less articles on average. So, not only is the user base unbalanced, but also females tend to be less active customers. Is it possible there is a case of unfair recommendations from *Der Standard* recommendation algorithms? It is possible we are seeing a feedback loop here, where females make up a smaller proportion of the user base, and therefore receive less accurate recommendations. As a consequence of worse recommendations, females could be less active and read fewer articles on the website, consequently reducing

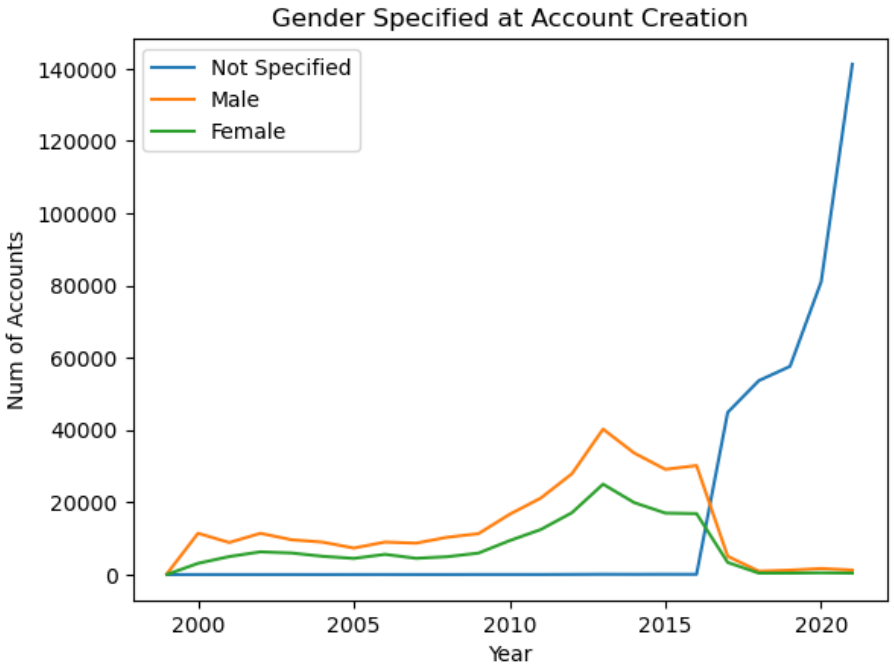

Figure 3.2: Account gender by year

their recommendation quality even further. This theory will be explored further in the modeling section.

	Male	Female
Proportion of User Base	0.638	0.362
Proportion of Readers in Last 30 Days	0.805	0.195
Avg Active Days	17.2	14.4
Avg # Articles Read	93.9	68.3
Avg Article Read per Active Day	4.4	3.7

Table 3.2: Reading statistics by gender.

3.2.2 Reading Behaviors.

Next, it is interesting to see if there exists a difference in reading preference between the two genders. Table 3.3 shows the top 10 categories read by each gender. The topics that are not contained in the other gender's top 10 are highlighted. The reading behaviors are fairly similar, with the most notable difference following typical gender stereotypes: men rank *Sports* higher than women rank *Sports*. Men also rank *Science* in the top 10, whereas women prefer the *Culture* and *Diverse* categories. We dig a bit deeper by also displaying the subcategories of each gender's preferences in Table 3.4. Here we also see

gender stereotypes involved, where men prefer *Football (Soccer)*, *Network Policy*, and *Innovations*, whereas women prefer *World Chronicles*, *Columns*, and *Health* articles.

Top Category Rank	Male	Female
1	Inland	Panorama
2	Panorama	Inland
3	Web	Opinion
4	Sport	Web
5	Opinion	International
6	International	Lifestyle
7	Economy	Economy
8	Budget	Budget
9	Lifestyle	Culture
10	Science	Diverse

Table 3.3: Reading category preference by gender.

Top Subcategory Rank	Male	Female
1	Inland	Inland
2	Panorama	Panorama
3	Football (Soccer)	Commentary
4	Commentary	Diverse
5	Europe	Austria Chronicle
6	Economic Policy	World Chronicle
7	Austria Chronicle	Columns
8	Network Policy	Europe
9	Innovations	Economic Policy
10	Diverse	Health

Table 3.4: Reading subcategory preference by gender.

3.3 Geography

Our second sensitive attribute is geographical location. In this section, we discuss the location of *Der Standard* users and how we apply an overall user location. We then provide reading statistics and reading preferences, and finally, apply the political lean labels to the user data.

3.3.1 Exploratory Data Analysis

Next, we start with a thorough exploratory data analysis of the geographical location attribute. When users interact with the website, the geolocation is recorded by the city and surrounding area. From here on out we will refer to it as a city-specific region. First, we display the last 30 days of interaction data on a country level (see Table 3.5) to see where most users are interacting with the website.

Country	Count	Percentage
Austria	41,855,028	0.878
Germany	2,433,502	0.051
United States	669,469	0.014
Switzerland	413,288	0.009
Italy	248,128	0.005
United Kingdom	202,623	0.004
Spain	127,587	0.003
Netherlands	121,464	0.003
Ukraine	111,172	0.002
France	102,262	0.002

Table 3.5: Interaction data: top 10 countries.

Not surprisingly, the top two countries with the most interactions are Austria and Germany, with 88% and 5% of the overall interactions, respectively. Somewhat surprisingly is the United States having the third most interactions. These interactions could come from German-speaking Americans, or possibly Austrian or German travelers accessing the website from the United States. If we zoom in on Austria, Figure 3.3 shows a heatmap of where the interactions are coming from. As expected, Vienna has the largest number of interactions, followed by other major cities such as Linz, Graz, Innsbruck, and Salzburg.

This map also shows another important point: although Vienna and other cities account for a large number of user interactions, there are also a substantial amount of interactions coming from outside the cities. Austrians from small villages are also accessing *Der Standard* to get their daily news, and they need to be represented in the RS as well.

As stated, these figures show where users are clicking on the website, but not where the users are actually located. Although user location is not provided in the data, we could infer it based on the interaction location. To do this inference, we split the data by user and determined the most frequent location of where users are reading articles.

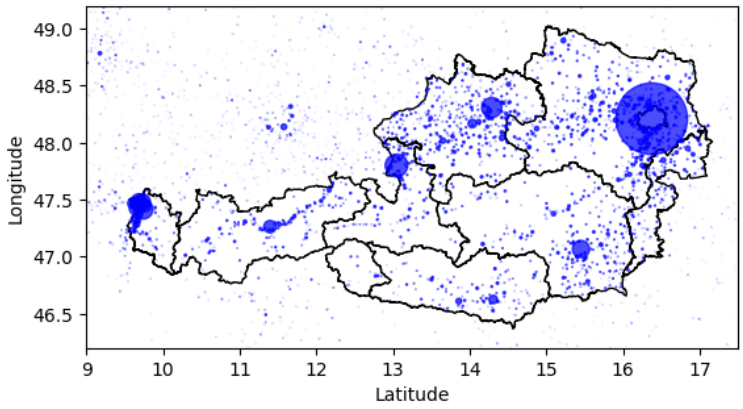

Figure 3.3: Interactions by geographical location.

Note: locations that were not on a city-specific level were discarded. We recognize that users could be reading articles on their commute or at the workplace, but given the data to work with, this is the best solution for estimating user location. We also argue that although a user may commute and read articles in a location other than home, they are also part of the community in which they work. We will call this location *user location*, which simply means the location in which they access articles most frequently, not necessarily where they live. The top 10 user locations are given in Table 3.6 with the corresponding number of users who are located there.

User Location	Count	Percentage
Vienna	51,592	0.416
Salzburg	4,725	0.038
Dornbirn	4,235	0.034
Linz	3,956	0.032
Graz	3,063	0.025
Hard	2,664	0.022
Lauterach	2,292	0.019
Lustenau	2,260	0.018
Innsbruck	1,636	0.013
Schwarzach	1,628	0.013

Table 3.6: Top 10 user locations.

Next, in order to split our data into binary groups, we chose to categorize users living in the five largest Austrian cities by population as large-city users, and all others reside in smaller cities or the countryside. According to Wikipedia[3], the top five most populous cities in Austria are:

City	Population
Vienna	1,931,593
Graz	292,630
Linz	207,247
Salzburg	155,331
Innsbruck	130,585

Table 3.7: Five most populous cities in Austria.

After splitting the data into these two user groups, we then calculated basic reading statistics for each group (Table 3.8). As you can see, the user base is very closely split between users in big cities vs other locations. However, users located in big cities tend to be more active, with more active days on average, more total articles read, and more articles read per active reading day. In the results section we will see if these readers are also more represented in the recommendation accuracy.

	Big City	Other
Count of Readers in Last 30 Day	64,972	59,095
% of Readers in Last 30 Days	0.524	0.476
Avg Active Days	15.1	14.4
Avg # Artcles Read	71.82	66.5
Avg Article Read per Active Day	3.83	3.69

Table 3.8: Reading statistics by geographical location.

3.3.2 Reading Behaviors

Next, we look at the reading behaviors of the users based on user location. Table 3.9 lists the top 10 categories read by each user group. Interestingly, the user groups reading interests are not strongly affected by geography. Both user groups have almost identical top categories; the only difference is in rank 10, where Big City users prefer *Culture* whereas the Other users prefer *Science*. Looking at the top subcategories in Table 3.10, here there is even less of a difference between the two groups. The only change is a small switch in the positioning of the categories *Football (Soccer)* and *Commentary*. It seems Austrians everywhere enjoy the same types of articles regardless of location.

[3] https://en.wikipedia.org/wiki/List_of_cities_and_towns_in_Austria

Top Category Rank	Big City	Other
1	Panorama	Panorama
2	Inland	Inland
3	Opinion	Web
4	Web	Opinion
5	Sport	Sport
6	International	International
7	Economy	Economy
8	Budget	Budget
9	Lifestyle	Lifestyle
10	Culture	Science

Table 3.9: Reading category preference by geographical location.

Top Subcategory Rank	Big City	Other
1	Inland	Inland
2	Panorama	Panorama
3	Commentary	Football (Soccer)
4	Football (Soccer)	Commentary
5	Economic Policy	Europe
6	Europe	Economic Policy
7	Diverse	Diverse
8	Austria Chronicle	Austria Chronicle
9	Columns	Columns
10	World Chronicle	World Chronicle

Table 3.10: Reading subcategory preference by geographical location.

Political Labels. Next, we use the predicted political labels to analyze political reading behaviors between geographical locations in Austria. After removing articles that were not scored, we display the predicted political reading habits of the top 10 cities in Austria in Figure 3.4. Interestingly, this picture is not what we assumed to find. According to the political labels, the location of Austrians does not affect their political reading patterns. Even Vienna has a nearly 50/50 split in right vs left-leaning articles. When analyzing our groupings of Big City vs Other, there is no correlation in political reading patterns. In table 3.11 we group articles by either *left*, *right*, or *leastbiased* and display the lean split by geographical location.

With these findings, we will instead move forward with an analysis of diversity using the articles' categories as opposed to the articles' political content. More discussion on the political findings will be examined in the Discussion and Conclusion chapters.

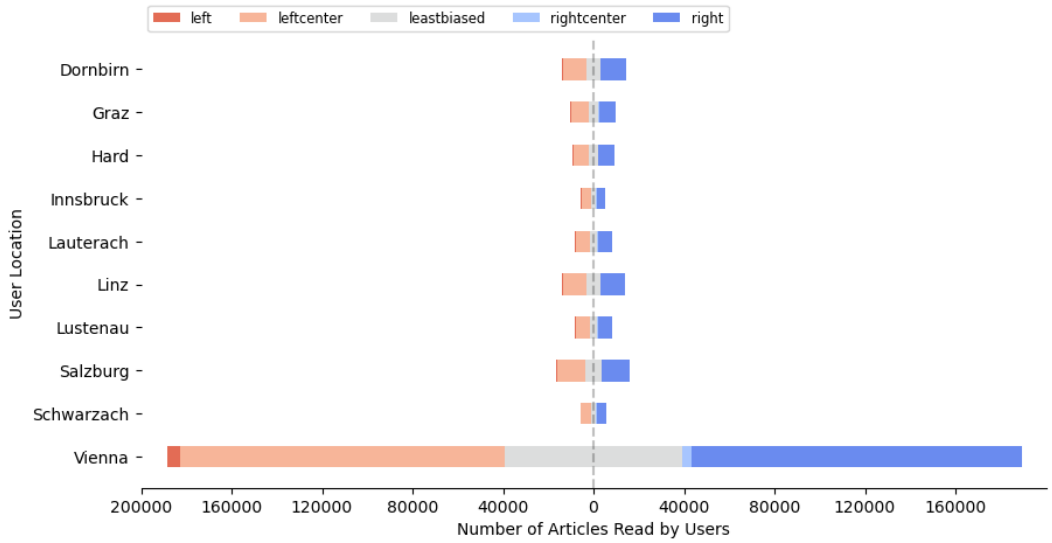

Figure 3.4: Political lean by city.

geo_location	left	leastbiased	right
Other	0.401	0.213	0.386
Big City	0.395	0.209	0.396

Table 3.11: Political lean by geographical location.

CHAPTER 4

Comparing Fairness-Aware NRS Across Groups

In this chapter, we first prepare the data and then perform hyperparameter tuning. Next, we apply the algorithms to our datasets, and display the results. Finally, we provide overall recommendations for future researchers.

4.1 Preprocessing

To perform the modeling part of the analysis, a publicly available library called RecBole [ZMH+21] was utilized. Specifically, an extension of the library title RecBole-FairRec [ZMH+21] was employed. Hyperparameters were tuned using a sample of 2,000 users due to time and resource constraints. All models have the following training setup. Hyperparameters specific to the models will be covered in the corresponding sections.

- Train/Test/Split was set at 80%/10%/10% of the data.
- Evaluation setting was *uni100*, which means during the evaluation step, 100 negative items are sampled with uniform distribution for each positive item in the testing set The model is then evaluated on these positive items with their sampled negative items.
- train-batch-size is set to the default, 2048 observations

4.1.1 FOCF

The hyperparameters from the original authors [YH17] were used for the start of the tuning process (see Table 4.1). However, after a few trials, it was quickly discovered the weight decay must be reduced and the fair-weight should be adjusted for some models.

The fair-weight refers to the coefficient of the fairness loss in proportion to the accuracy loss. Interestingly, the FOCF authors used an equal weighting for all models and achieved desirable results. In our experiments, we found that an equal weighting for the value and absolute models had a severe impact on accuracy. Figure 4.1 shows an example of this phenomenon with the value model. As you can see from the figure, the steep decrease in accuracy occurs when moving from fair-weight = 0.35 to 0.5. This is also the point where a noticeable increase in unfairness occurs. Based on these experiments, fair-weight was set to 0.3 for the value and absolute models, where there was a negligible decrease in accuracy and the lowest unfairness.

Model	Learning-Rate	Fair-Weight	Weight-Decay	Learner
Original	0.0001	1.0	0.0010	adam
Baseline	0.0002	0.0	0.0001	adam
Value	0.0002	0.3	0.0001	adam
Absolute	0.0002	0.3	0.0001	adam
Over	0.0002	1.0	0.0001	adam
Under	0.0002	1.0	0.0001	adam
NonParity	0.0002	1.0	0.0001	adam

Table 4.1: Hyperparameter settings for FOCF models.

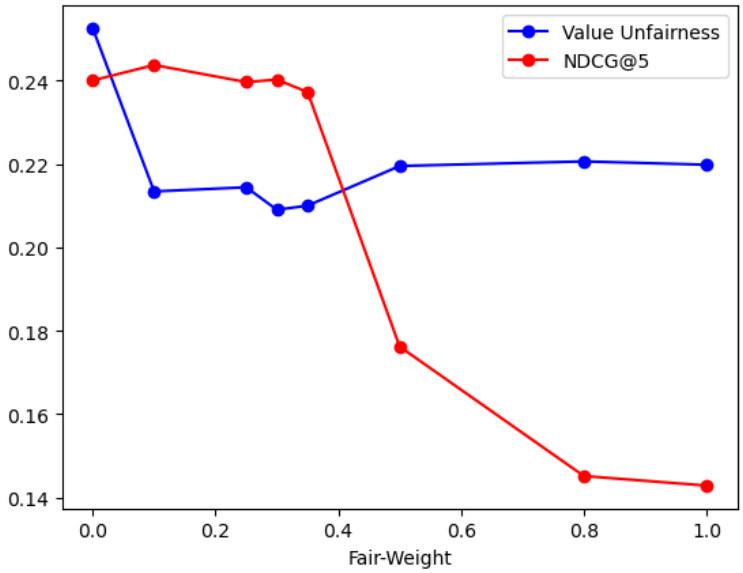

Figure 4.1: Fairness-accuracy trade-off by fair-weight setting.

4.1.2 PFCN

For the PFCN models, the original hyperparameters were used from the authors [LCX+21], with one exception: the DMF model. This model required more hyperparameter tuning

to report satisfactory results. The biggest change was to the learning rate. We found the validation results of each epoch to bounce around sporadically, a clear indication the learning rate was too high. We also needed to reduce the weight-decay from the suggested 0.001 to 0.0001, which is also in line with the other models. After reducing the weight-decay and learning-rate, we were able to achieve satisfactory results.

First, we list the hyperparameters in Table 4.2 that are consistent for every model, then list the parameters that are unique for each model.

All Models	dis-hidden-size=[128,256,128,128,64,32] filter-mode=sm dis-dropout=0.3 train-epoch-interval=5 dis-weight=10 neg-sample-num=uniform, 1 weight-decay=0.0001 activation=leakyrelu activation=leakyrelu
DMF	num-layers=3 dis-activation=leakyrelu mlp-activation=relu mlp-dropout=0.2
WideDeep	mlp-hidden-size=[64, 32, 16] dropout=0.2

Table 4.2: Hyperparameter settings for PFCN models.

Note: filter-mode = sm refers to the method of creating the filter function, where *sm* is the separate method. In our case, since we are training on only one sensitive attribute per model, one filter function is trained for that attribute.

To run the full experiment, 10,000 random users were sampled from users who have reported their gender. When running the models on the full sample, several fair-weights were also tested in order to corroborate the results of the smaller sample. The full modeling results are reported in the next section.

4.2 Results

Since there is a large number of models and metrics we are comparing, this section will first be broken down into three categories of metrics: *accuracy, unfairness,* and *diversity.* Each section will discuss the results separately, and at the end we provide a summary table bringing the results together (Table 4.11). For ease of viewing, the best model and any models that came in very close are highlighted for each metric. Models that also performed comparably well are highlighted with a paler shade of yellow.

Other table notes:

- the accuracy metrics are all ranking metrics, meaning they are typically set at a value of K, such that the metric includes K recommended items in the item list. For this analysis, the value of K is set at five for all metrics.

- The Baseline model refers to the FOCF method with no unfairness metric added to the loss function, therefore it is normal matrix factorization.

- The other FOCF methods are labeled by which unfairness metric is added to the cost function. We will also refer to the models in this manner, for example, *Under model* referring to the under-unfairness-reducing-model. Note: the metrics are measuring unfairness, therefore smaller values are desirable.

- $NDCG\_F$ and $NDCG\_M$ refer to NDCG of Females and Males, respectively. $NDCG\_O$ and $NDCG\_BC$ refer to the geolocations of Other and Big City, respectively.

4.2.1 Model Accuracy Comparison

Table 4.3 presents the results of our modeling experiment for the sensitive feature, gender. The performance of various models is evaluated based on the normalized discounted cumulative gain (NDCG), a comprehensive ranking metric. Our analysis reveals that the models employing the PFCN approach outperform the FOCF models, with the exception of the DMF model. Specifically, BiasedMF is identified as the top-performing model across all the accuracy metrics, including NDCG, recall, hit, and mean reciprocal rank. The PMF and WideDeep models also exhibit competitive performance, ranking second in accuracy metrics.

	Model	NDCG	NDCG_F	NDCG_M	Recall	Hit	MRR
FOCF	Baseline	0.3517	0.3914	0.3425	0.2989	0.7448	0.4895
	Value	0.3502	0.3907	0.3408	0.2979	0.7428	0.4880
	Absolute	0.3504	0.3911	0.3409	0.2981	0.7431	0.4882
	NonParity	0.3519	0.3872	0.3437	0.2985	0.7435	0.4903
	Under	0.3403	0.3802	0.3311	0.2944	0.7296	0.4740
	Over	0.3517	0.3914	0.3425	0.2989	0.7448	0.4895
PFCN	BiasedMF	0.3838	0.4080	0.3781	0.3102	0.7734	0.5353
	PMF	0.3800	0.4016	0.3750	0.3070	0.7686	0.5329
	WideDeep	0.3760	0.4032	0.3697	0.3047	0.7663	0.5253
	DMF	0.3422	0.3628	0.3374	0.2792	0.7312	0.4846

Table 4.3: Accuracy comparison - gender.

In contrast, the DMF model demonstrates poor performance among all the PFCN models, ranking only above the FOCF Under model in the overall rankings. This outcome could be attributed to inadequate hyperparameter tuning or unsuitability of this method for the given dataset.

Regarding the FOCF models, the comparison reveals little variation in performance among most of the models. This outcome is desirable, as it implies that unfairness metrics can be incorporated into the loss function with minimal impact on accuracy. However, the FOCF Under model exhibits a decrease of 0.1 points in NDCG. Notably, the NonParity, Baseline, and Over models exhibit the highest accuracy among the FOCF models. The Baseline model is expected to produce the best accuracy as the loss function is solely focused on improving accuracy without considering unfairness metrics. Interestingly, the Overestimation Unfairness model results show no noteworthy difference from the baseline in terms of accuracy.

Finally, we observe that, despite female users' disproportionate representation in the dataset, they receive more accurate recommendations than their male counterparts, as evidenced by the NDCG scores for the Female and Male columns. This outcome is unexpected, given our initial hypothesis that imbalanced data would lead to inaccurate recommendations. One plausible explanation for this counter-intuitive result is that female users may exhibit less diverse reading patterns, making it easier to provide them with more accurate recommendations. We will explore this topic further in the diversity section.

Table 4.4 presents the results of our investigation based on geographical location. Consistent with the previous analysis, the PFCN models outperform the FOCF models, except for the DMF model. Among the PFCN models, the PMF model slightly outperforms the BiasedMF model on this dataset.

Similarly, the comparison of FOCF models with each other reveals that the incorporation of different unfairness metrics in the loss function has minimal impact on accuracy. Unlike the gender-based analysis, we do not observe a noteworthy decrease in accuracy for the FOCF Under model.

Notably, the data exhibits a balanced distribution with respect to geographical location as the sensitive attribute, and users from the Other regions receive only slightly better recommendations than their Big City counterparts.

	Model	NDCG	NDCG_O	NDCG_BC	Recall	Hit	MRR
FOCF	Baseline	0.3243	0.3290	0.3199	0.3140	0.6987	0.4406
	Value	0.3192	0.3285	0.3106	0.3104	0.6801	0.4292
	Absolute	0.3246	0.3298	0.3198	0.3139	0.6962	0.4415
	NonParity	0.3244	0.3288	0.3204	0.3141	0.6985	0.4410
	Under	0.3221	0.3276	0.3171	0.3150	0.6919	0.4361
	Over	0.3243	0.3290	0.3199	0.3140	0.6987	0.4406
PFCN	BiasedMF	0.3506	0.3553	0.3462	0.3250	0.7199	0.4758
	PMF	0.3512	0.3552	0.3474	0.3273	0.7238	0.4766
	WideDeep	0.3429	0.3505	0.3359	0.3213	0.7147	0.4666
	DMF	0.3233	0.3295	0.3176	0.3047	0.6966	0.4412

Table 4.4: Accuracy comparison - geographic location.

4.2.2 Model Unfairness Comparison

The comparison of unfairness metrics from the gender analysis can be seen in Table 4.5. Here, we see more varying results than the accuracy measurement results. First, when comparing the absolute difference of NDCG, the DMF model performs best, with BiasedMF and PMF also performing well. Moving to the value and absolute unfairness measures, the FOCF models perform well. Not surprisingly, the models that are built to minimize these metrics perform the best. Overestimation unfairness is also minimized by the Value and Absolute models. And important note here - the PFCN models drastically increase value, absolute, and overestimation unfairness when compared to the Baseline FOCF model. The Under model does somewhat reduce the underestimation unfairness metric; however, surprisingly the WideDeep, PMF, and BiasedMF models reduce it dramatically. Similarly, the NonParity model reduces non-parity unfairness, but using the BiasedMF and PMF models results in an even more drastic reduction.

	Model	Diff NDCG	Value	Absolute	Under	Over	NonPar
FOCF	Baseline	0.0489	0.1113	0.0878	0.0834	0.0279	0.0355
	Value	0.0499	0.1088	0.0869	0.0845	0.0243	0.0382
	Absolute	0.0502	0.1088	0.0869	0.0845	0.0243	0.0382
	NonParity	0.0435	0.1112	0.0865	0.0818	0.0293	0.0098
	Under	0.0491	0.1649	0.1394	0.0704	0.0945	0.0301
	Over	0.0489	0.1113	0.0878	0.0834	0.0279	0.0355
PFCN	BiasedMF	0.0299	0.1908	0.1620	0.0256	0.1652	0.0005
	PMF	0.0266	0.1958	0.1671	0.0260	0.1697	0.0009
	WideDeep	0.0335	0.2327	0.2187	0.0078	0.2249	0.0045
	DMF	0.0254	0.2283	0.1997	0.0396	0.1887	0.0028

Table 4.5: Unfairness comparison - gender.

The results from the geography analysis (Table 4.6) look similar in most aspects. The absolute difference in NDCG is lower overall for all models, possibly on account of the more balanced dataset and similar reading patterns of the two groups. Again, the Value and Absolute models perform best on the value, absolute, and overestimation unfairness metrics. Also similar to the gender analysis, underestimation unfairness is reduced most effectively with the WideDeep, PMF, and BiasedMF models, and non-partiy unfairness is best reduced with the PFCN and NonParity models.

4.2.3 Model Diversity Comparison

The dissimilarity metric, as defined in Chapter 2, was utilized to measure dissimilarity based on article attributes of category and subcategory. The attributes were one-hot encoded and cosine similarity was computed for each item pair in the recommendation list. The average dissimilarity score was then calculated for all users.

In this section, we present the results of the dissimilarity analysis for the entire dataset and then display the metrics for user groupings. The gender-based analysis results, shown

	Model	Diff NDCG	Value	Absolute	Under	Over	NonPar
FOCF	Baseline	0.0091	0.1014	0.0743	0.0758	0.0256	0.0121
	Value	0.0180	0.0997	0.0712	0.0700	0.0298	0.0111
	Absolute	0.0100	0.1001	0.0734	0.0754	0.0247	0.0125
	NonParity	0.0084	0.1014	0.0743	0.0758	0.0256	0.0081
	Under	0.0105	0.1273	0.0972	0.0648	0.0625	0.0112
	Over	0.0091	0.1014	0.0743	0.0758	0.0256	0.0121
PFCN	BiasedMF	0.0091	0.1111	0.0858	0.0180	0.0931	0.0035
	PMF	0.0078	0.1134	0.0896	0.0165	0.0969	0.0032
	WideDeep	0.0146	0.1195	0.1060	0.0101	0.1094	0.0025
	DMF	0.0119	0.1187	0.0973	0.0325	0.0861	0.0009

Table 4.6: Unfairness comparison - geographical location.

in Table 4.7, indicate that the FOCF models slightly outperform the PFCN models. This outcome implies that the FOCF models provide users with a slightly more uniform exposure to overall item listings. However, the PFCN models provide a greater variety of articles to users, with the BiasedMF model exhibiting the highest dissimilarity.

The geography-based analysis, presented in Table 4.8, yields similar results with slight variations. The entropy is higher for the FOCF models, and the WideDeep model also ranks highly. Once again, the PFCN models provide the highest dissimilarity, with the FOCF Value model also demonstrating high dissimilarity.

	Model	Entropy	Dissimilarity
FOCF	Baseline	0.0049	0.92487
	Value	0.0048	0.92435
	Absolute	0.0049	0.9244
	NonParity	0.0049	0.92491
	Under	0.0049	0.92317
	Over	0.0049	0.92487
PFCN	BiasedMF	0.0047	0.92853
	PMF	0.0047	0.92683
	WideDeep	0.0048	0.92779
	DMF	0.0046	0.92565

Table 4.7: Diversity comparison - gender.

Next, we examine the metrics split by user group in Tables 4.9 and 4.10. The analysis indicates that, on average, the entropy is slightly higher for male users compared to female users. Conversely, the dissimilarity is higher for female users. The BiasedMF and PMF models demonstrate the best absolute difference, but there is little variability overall between all the models. The PFCN models exhibit the highest absolute difference in dissimilarity, with the PMF model performing the best.

In the geography-based analysis, we observe that, on average, the entropy is higher for the

	Model	Entropy	Dissimilarity
FOCF	Baseline	0.00550	0.93434
	Value	0.00531	0.93851
	Absolute	0.00550	0.93409
	NonParity	0.00551	0.93422
	Under	0.00560	0.93425
	Over	0.00550	0.93434
PFCN	BiasedMF	0.00540	0.94009
	PMF	0.00540	0.93898
	WideDeep	0.00560	0.93956
	DMF	0.00540	0.94391

Table 4.8: Diversity comparison - geographical location.

Big City group. The dissimilarity varies by model, with some models exhibiting higher dissimilarity for the Other group and some for the Big City group. Notably, there is no clear winner in terms of absolute difference of entropy and dissimilarity on this dataset. The baseline FOCF and WideDeep models demonstrate the smallest absolute difference in entropy, while the baseline and over FOCF models exhibit the lowest absolute difference in dissimilarity.

Overall, the variability of dissimilarity across the models is minimal, indicating that the application of accuracy fairness techniques has minimal impact on diversity.

	Model	Ent_F	Ent_M	Diff Ent	Dissim_F	Dissim_M	Diff Dissim
FOCF	Baseline	0.00670	0.00700	0.00030	0.92946	0.92380	0.00566
	Value	0.00670	0.00700	0.00030	0.92968	0.92312	0.00656
	Absolute	0.00670	0.00700	0.00030	0.92968	0.92318	0.00650
	NonParity	0.00670	0.00700	0.00030	0.92931	0.92389	0.00542
	Under	0.00670	0.00700	0.00030	0.92909	0.92179	0.00731
	Over	0.00670	0.00700	0.00030	0.92946	0.92380	0.00566
PFCN	BiasedMF	0.00650	0.00670	0.00020	0.93119	0.92791	0.00328
	PMF	0.00680	0.00690	0.00010	0.92684	0.92682	0.00002
	WideDeep	0.00660	0.00700	0.00040	0.93055	0.92715	0.00340
	DMF	0.00650	0.00680	0.00030	0.92827	0.92504	0.00323

Table 4.9: Diversity comparison split by gender.

4.2.4 Overall Recommendations

Finally, we bring the full results together in Table 4.11 and provide recommendations for future researchers. Overall, regarding accuracy, the clear winner is PFCN. With the exception of DMF, these models perform much better on both analyses compared to the FOCF models. However, when comparing the unfairness across models, the results are more mixed. The results depend entirely on which metric is of interest. PFCN models tend to effectively reduce the difference in NDCG, under, and non-parity unfairness. On

	Model	Ent_O	Ent_BC	Diff Ent	Dissim_O	Dissim_BC	Diff Dissim
FOCF	Baseline	0.00620	0.00620	0.00000	0.93430	0.93438	0.00008
	Value	0.00593	0.00611	0.00018	0.93785	0.93911	0.00126
	Absolute	0.00610	0.00620	0.00010	0.93421	0.93398	0.00024
	NonParity	0.00617	0.00621	0.00004	0.93388	0.93453	0.00065
	Under	0.00620	0.00640	0.00020	0.93330	0.93513	0.00183
	Over	0.00617	0.00620	0.00004	0.93430	0.93438	0.00008
PFCN	BiasedMF	0.00600	0.00610	0.00010	0.94156	0.93874	0.00282
	PMF	0.00610	0.00600	0.00010	0.93984	0.93819	0.00165
	WideDeep	0.00620	0.00620	0.00000	0.93902	0.94006	0.00104
	DMF	0.00620	0.00610	0.00010	0.94277	0.94496	0.00219

Table 4.10: Diversity comparison split by geographical location.

the other hand, these models drastically increase value, absolute, and over unfairness when compared to the baseline. FOCF models are better suited for value, absolute, and over unfairness. In regards to diversity, neither model class had a big impact on these metrics. FOCF produced slightly better results on entropy, and PFCN produced slightly better results on dissimilarity.

	Gender											
	Model	**NDCG**	**Diff NDCG**	**Value**	**Absolute**	**Under**	**Over**	**NonPar**	**Entropy**	**Dissimilarity**	**Diff Ent**	**Diff Dissim**
FOCF	Baseline	0.3517	0.0489	0.1113	0.0878	0.0834	0.0279	0.0355	0.0049	0.92487	0.00030	0.00566
	Value	0.3502	0.0499	0.1088	0.0869	0.0845	0.0243	0.0382	0.0048	0.92435	0.00030	0.00656
	Absolute	0.3504	0.0502	0.1088	0.0869	0.0845	0.0243	0.0382	0.0049	0.9244	0.00030	0.00650
	NonParity	0.3519	0.0435	0.1112	0.0865	0.0818	0.0293	0.0098	0.0049	0.92491	0.00030	0.00542
	Under	0.3403	0.0491	0.1649	0.1394	0.0704	0.0945	0.0301	0.0049	0.92317	0.00030	0.00731
	Over	0.3517	0.0489	0.1113	0.0878	0.0834	0.0279	0.0355	0.0049	0.92487	0.00030	0.00566
PFCN	BiasedMF	0.3838	0.0299	0.1908	0.1620	0.0256	0.1652	0.0005	0.0047	0.92853	0.00020	0.00328
	PMF	0.3800	0.0266	0.1958	0.1671	0.0260	0.1697	0.0009	0.0047	0.92683	0.00010	0.00002
	WideDeep	0.3760	0.0335	0.2327	0.2187	0.0078	0.2249	0.0045	0.0048	0.92779	0.00040	0.00340
	DMF	0.3422	0.0254	0.2283	0.1997	0.0396	0.1887	0.0028	0.0046	0.92565	0.00030	0.00323

	Geographical Location											
	Model	**NDCG**	**Diff NDCG**	**Value**	**Absolute**	**Under**	**Over**	**NonPar**	**Entropy**	**Dissimilarity**	**Diff Ent**	**Diff Dissim**
FOCF	Baseline	0.3243	0.0091	0.1014	0.0743	0.0758	0.0256	0.0121	0.00550	0.93434	0.00000	0.00008
	Value	0.3192	0.0180	0.0997	0.0712	0.0700	0.0298	0.0111	0.00531	0.93851	0.00018	0.00126
	Absolute	0.3246	0.0100	0.1001	0.0734	0.0754	0.0247	0.0125	0.00550	0.93409	0.00010	0.00024
	NonParity	0.3244	0.0084	0.1014	0.0743	0.0758	0.0256	0.0081	0.00551	0.93422	0.00004	0.00065
	Under	0.3221	0.0105	0.1273	0.0972	0.0648	0.0625	0.0112	0.00560	0.93425	0.00020	0.00183
	Over	0.3243	0.0091	0.1014	0.0743	0.0758	0.0256	0.0121	0.00550	0.93434	0.00004	0.00008
PFCN	BiasedMF	0.3506	0.0091	0.1111	0.0858	0.0180	0.0931	0.0035	0.00540	0.94009	0.00010	0.00282
	PMF	0.3512	0.0078	0.1134	0.0896	0.0165	0.0969	0.0032	0.00540	0.93898	0.00010	0.00165
	WideDeep	0.3429	0.0146	0.1195	0.1060	0.0101	0.1094	0.0025	0.00560	0.93956	0.00000	0.00104
	DMF	0.3233	0.0119	0.1187	0.0973	0.0325	0.0861	0.0009	0.00540	0.94391	0.00010	0.00219

Table 4.11: Full comparison of algorithms.

CHAPTER 5

Discussion

5.1 Political Labeling

The political labeling task was both an interesting and challenging piece of this thesis. Our initial assumption was that Austrians of different geographical locations would have dissimilar and distinct political reading patterns. This assumption was, however, not supported by the data, and no correlation was found between political reading behavior and geography. Here, we further elaborate on the limitations and potential theories that may explain the observed findings.

Our first discussion point is around the accuracy of the political labeling process. We acknowledge that the majority of articles (78%) were not able to be classified and were therefore labeled as *unknown*. For reference, only about 20% of the training data was labeled as *unknown*. This is an indication that there may be an underlying problem that is not allowing the algorithm to confidently predict political lean.

One possible limitation affecting accuracy is the translation of the articles from German to English. When translating between two languages, there is often some amount of information lost, as many complex ideas cannot be perfectly translated in a one-to-one relationship. Especially in the context of politics, word choice can be used to convey the same topic from two different perspectives. For example, words such as immigrant and expatriate have to same meaning but different connotations [Nas17]. When using the word expatriate, English speakers tend to think of someone coming from wealth and privilege, whereas the word immigrant is used for people who are in less privileged positions [Nas17]. Although both words have a similar meaning, they can be used to spark positive or negative emotions toward this class of people. Seeing as immigration is a crucial topic discussed by political parties, word choice and possible mistranslations can quickly change the tone of an article.

Another potential limitation that could have affected the accuracy of the political labeling task is the use of a binary political leaning classification system. Political beliefs and attitudes are not necessarily black and white, and individuals can have complex and nuanced perspectives on different issues. Especially in Austria where more than two political parties have influence in the government, we may have missed important variations and nuances in readers' political attitudes by labeling only left vs right. Future studies could consider incorporating a more nuanced classification system, such as a multi-dimensional approach that takes into account different dimensions of political ideology, such as economic, social, and foreign policy stances. This could potentially yield more accurate and comprehensive insights into readers' political reading behaviors.

Our second discussion point is around the data source. The limitation here is that we only had access to one Austrian newspaper, *Der Standard*. In general, *Der Standard* is known to be a more left-leaning news source. By only using one newspaper, our data is biased to only include readers of *Der Standard*. Our results could indicate that readers of *Der Standard* have homogenous reading patterns, regardless of geography. Since we do not have access to other newspapers' data, we cannot confidently extrapolate these results to Austrians who read other newspapers.

In conclusion, while our initial assumption of finding dissimilar and distinct political reading patterns based on geography was not supported by the data, there are still potential limitations and theories that could explain these findings. The accuracy of the political labeling process, the data source, and the binary classification system are all factors that could have influenced the observed results. Future studies could address these limitations and further explore the relationship between user location and political reading behaviors.

5.2 FOCF

The results of the FOCF model comparison yielded interesting findings, especially when compared to the original authors' [YH17] outcomes. The first contrasting finding is their results on unbalanced data. Their results from that analysis showed that underrepresented groups were treated unfairly by standard matrix factorization methods. In contrast, the results of our gender analysis show that although females are underrepresented in the data, they were actually receiving better recommendations across all models. There is agreement present in our findings with the original authors in that our unbalanced dataset (gender) had higher overall unfairness compared to our balanced dataset (geography).

Our other contrasting finding is that when optimizing for the value and absolute unfairness metrics, the original authors observed a larger drop than in our analysis. In addition, the Over model is specifically interesting in that we do not see a change in the over unfairness metric whatsoever. However, we do pick up on similar correlations between the metrics and models as the authors. For example, optimizing either the over or under unfairness metric causes the other to rise. We also have the positive outcome that optimizing for any unfairness metric does not have a substantial effect on the accuracy measurements.

The contrasting findings between our study and the original authors' emphasize the need for further research into the performance of FOCF models on unbalanced data. Future research can benefit from comparing the performance of FOCF models on different datasets and investigating the effectiveness of other fairness-aware algorithms. Moreover, it is essential to examine how the choice of fairness metric may affect the model's performance and the outcomes of the recommendations.

5.3 PFCN

The original authors [LCX+21] did not use the unfairness metrics outlined in this work, and instead defined unfairness as the ability of the attacking model to differentiate the sensitive features from user embeddings; this was measured using Area Under the Curve (AUC). Despite these differences, we found that the method proposed by the original authors can still be effective at reducing certain types of unfairness that we studied. This is an important finding, as it demonstrates the potential of using different fairness metrics in tandem to gain a more complete understanding of the performance of a recommender system.

Specifically, the PFCN method was particularly effective at reducing under and non-parity unfairness, as well as the absolute difference in NDCG, despite not being explicitly trained to reduce these metrics. This highlights the potential of exploring other fairness-aware methods that were not explicitly trained on these fairness metrics. It is possible that other machine learning models could also have a positive impact on reducing unfairness in recommender systems, and further research in this area is warranted. Ultimately, the ability to identify and address various types of unfairness in recommender systems will help to create fairer and more inclusive digital spaces, where users have access to diverse perspectives and information.

CHAPTER 6

Conclusion

6.1 Summary

In conclusion, this thesis provides valuable insights into the role of fairness and diversity in news recommender systems, specifically in the context of the Austrian media landscape. The research aimed to address the potential negative consequences of biased recommendations and filter bubbles, such as societal polarization and the suppression of information. By analyzing the performance of various models using accuracy, unfairness, and diversity metrics, this study aimed to identify the most effective approaches for generating fair and diverse news recommendations.

In order to identify relevant group unfairness metrics, we performed an extensive literature review and narrowed down the metrics into a suitable list. We described the evaluation metrics and techniques used to measure accuracy, fairness, and diversity. We also researched state-of-the-art fairness-aware algorithms to be applied to our data.

The empirical research section then revealed interesting findings in our data around reading behaviors and the political lean of news articles. Our gender analysis explored the topic of imbalanced data and how it might affect accuracy and unfairness. The geographical analysis was particularly interesting in that it showed no correlation between user location and political reading behaviors.

Finally, in the comparison of algorithms section we performed an extensive comparison of the methods and provided recommendations. The study found that PFCN models outperform FOCF models in terms of accuracy when recommending news articles, with the BiasedMF model performing the best across all accuracy metrics. The study also examined unfairness metrics and found that FOCF models performed well on these metrics The PFCN models worsened unfairness scores for some metrics, but vastly improved others. This finding highlights the importance of choosing the right metric based on the use case. Finally, the study explored the diversity of recommendations

using dissimilarity and entropy metrics, finding that neither method caused a drastic change in these metrics.

Challenges were encountered along the way, especially in terms of the application of the algorithms. Recbole-FairRec [ZMH+21] is a relatively new extension of Recbole, and with that comes a lack of functionality and first-user bugs. Overall, we were able to adapt the system to our needs and hope our usage and suggestions will further the use of this well-built package.

This study highlights the importance of incorporating fairness and diversity metrics into the design and evaluation of recommender systems. The findings demonstrate that accuracy and fairness can be achieved simultaneously with the right modeling approach, while diversity can be held constant using these modeling techniques. By developing more fair and diverse news recommender systems, we can help mitigate the negative consequences of biased recommendations and filter bubbles, and promote greater access to diverse information sources. This research has implications for both the news media industry and the wider public, as well as for researchers in the field of recommender systems. News media organizations can use the insights gained from this study to design more ethical and inclusive news recommender systems, ensuring that their recommendations do not unfairly prioritize certain groups or limit access to diverse viewpoints.

Additionally, the results of this study highlight the need for ongoing research into the relationship between user location and political reading behaviors. Although this analysis was limited to only one news source, this research can be extended to include more sources to get a more accurate picture of the reading choices of Austrians. This study can also act as a guideline for similar research performed in other countries around the world.

In future research, we recommend exploring the impact of other factors on the performance of recommendation systems, including user privacy concerns, the impact of recommendation systems on user behavior, and the use of other fairness techniques. Additionally, we recommend exploring the use of other datasets and evaluating the impact of different sensitive attributes on the performance of recommendation systems.

6.2 Contribution

Here we outline the research questions again that were answered by this work.

RQ1: To what degree do different reading behaviors of different groups impact recommender systems and recommendations in the domain of news media?

Here we had contrasting findings compared with our original hypothesis that readers of different geographical areas have different reading patterns. Although we could not perform an analysis of political reading preferences split by geography, this is an

interesting finding itself. We also observe a difference between the two analyses, in that the gender dataset with imbalanced data and slightly different reading patterns led to overall higher unfairness as compared to the geographical analysis with balanced data and similar reading patterns.

RQ2: When is it appropriate to use different fairness metrics for NRS?

For future researchers, we have provided the background of several unfairness metrics in Chapter 2, and guidance on the metrics in Results Section 4.2.

RQ3: To what extent do different recommender system algorithms affect these fairness metrics, in terms of improvement made, cost to accuracy, affect to diversity ect?

The results section of the analysis provided a thorough comparison of algorithms across accuracy, unfairness, and diversity metrics. The overall recommendation section (4.2.4) at the end of Chapter 4 provides guidance on which models are best in which situations.

6.3 Future Work

Future work, especially for the benefit of the political labeling section, could include more Austrian newspapers in the analysis. In this thesis, we only had access to a single Austrian newspaper, and therefore only have data from that paper. Of course, this creates an inherent bias in our data, and makes the reading behaviors not generalizable to all Austrians. We could also either (1) find a political labeling NLP algorithm that was built using the German language, or (2) create one ourselves. This would eliminate the potential information loss that comes with translating articles into another language.

As outlined previously, there is also grounds for further investigation into other fairness-aware algorithms. From this analysis, we showed that using algorithms outside their specific intended purpose may produce impressive results.

List of Figures

1.1 CRISP-DM process diagram [Jen12]. 5

2.1 Content-Based Filtering vs Collaborative Filtering by [Ton18] 13
2.2 Causal path from [LCX+21]. 16
2.3 Framework from [LCX+21]. 17
2.4 Visualization of the Wide and Deep model [CKH+16]. 18

3.1 Article political leaning by category. 21
3.2 Account gender by year 22
3.3 Interactions by geographical location. 25
3.4 Political lean by city. 28

4.1 Fairness-accuracy trade-off by fair-weight setting. 30

List of Tables

2.1 Notations and explanations of common variables. 8
2.2 List of group fairness metrics. Table inspired by [WMZ+22]. 10

3.1 Political lean prediction by subcategory. 20
3.2 Reading statistics by gender. 22
3.3 Reading category preference by gender. 23
3.4 Reading subcategory preference by gender. 23
3.5 Interaction data: top 10 countries. 24
3.6 Top 10 user locations. 25
3.7 Five most populous cities in Austria. 26
3.8 Reading statistics by geographical location. 26
3.9 Reading category preference by geographical location. 27
3.10 Reading subcategory preference by geographical location. 27
3.11 Political lean by geographical location. 28

4.1 Hyperparameter settings for FOCF models. 30
4.2 Hyperparameter settings for PFCN models. 31
4.3 Accuracy comparison - gender. 32
4.4 Accuracy comparison - geographic location. 33
4.5 Unfairness comparison - gender. 34
4.6 Unfairness comparison - geographical location. 35
4.7 Diversity comparison - gender. 35
4.8 Diversity comparison - geographical location. 36
4.9 Diversity comparison split by gender. 36
4.10 Diversity comparison split by geographical location. 37
4.11 Full comparison of algorithms. 38

Bibliography

[AHD20] Areej Alsini, Du Q. Huynh, and Amitava Datta. Hit ratio: An Evaluation Metric for Hashtag Recommendation, October 2020. arXiv:2010.01258 [cs].

[ARB09] Aristotle, W. D. Ross, and Lesley Brown. *The Nicomachean ethics.* Oxford world's classics. Oxford University Press, Oxford ; New York, 2009.

[BYRNo99] Ricardo Baeza-Yates, Berthier Ribeiro-Neto, and others. *Modern information retrieval*, volume 463. ACM press New York, 1999.

[CKH+16] Heng-Tze Cheng, Levent Koc, Jeremiah Harmsen, Tal Shaked, Tushar Chandra, Hrishi Aradhye, Glen Anderson, Greg Corrado, Wei Chai, Mustafa Ispir, Rohan Anil, Zakaria Haque, Lichan Hong, Vihan Jain, Xiaobing Liu, and Hemal Shah. Wide & Deep Learning for Recommender Systems, June 2016. arXiv:1606.07792 [cs, stat].

[DTD15] Amit Datta, Michael Carl Tschantz, and Anupam Datta. Automated Experiments on Ad Privacy Settings: A Tale of Opacity, Choice, and Discrimination, March 2015. arXiv:1408.6491 [cs].

[EHWK14] Michael D. Ekstrand, F. Maxwell Harper, Martijn C. Willemsen, and Joseph A. Konstan. User perception of differences in recommender algorithms. In *Proceedings of the 8th ACM Conference on Recommender systems*, RecSys '14, pages 161–168, New York, NY, USA, October 2014. Association for Computing Machinery.

[ETA+18] Michael D. Ekstrand, Mucun Tian, Ion Madrazo Azpiazu, Jennifer D. Ekstrand, Oghenemaro Anuyah, David McNeill, and Maria Soledad Pera. All The Cool Kids, How Do They Fit In?: Popularity and Demographic Biases in Recommender Evaluation and Effectiveness. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, pages 172–186. PMLR, January 2018. ISSN: 2640-3498.

[FKT+18] Golnoosh Farnadi, Pigi Kouki, Spencer K. Thompson, Sriram Srinivasan, and Lise Getoor. A Fairness-aware Hybrid Recommender System, September 2018. arXiv:1809.09030 [cs, stat].

[FXG+20] Zuohui Fu, Yikun Xian, Ruoyuan Gao, Jieyu Zhao, Qiaoying Huang, Yingqiang Ge, Shuyuan Xu, Shijie Geng, Chirag Shah, Yongfeng Zhang, and Gerard de Melo. Fairness-Aware Explainable Recommendation over Knowledge Graphs, June 2020. arXiv:2006.02046 [cs].

[Hel19] Natali Helberger. On the Democratic Role of News Recommenders. *Digital Journalism*, 7(8):993–1012, September 2019.

[Hol05] Elisa Holmes. Anti-Discrimination Rights without Equality. *The Modern Law Review*, 68(2):175–194, 2005. Publisher: [Modern Law Review, Wiley].

[HPS16] Moritz Hardt, Eric Price, and Nathan Srebro. Equality of Opportunity in Supervised Learning, October 2016. arXiv:1610.02413 [cs].

[Jac84] Jacob Jacoby. Perspectives on Information Overload. *Journal of Consumer Research*, 10(4):432–435, 1984. Publisher: Oxford University Press.

[Jen12] Kenneth Jensen. English: A diagram showing the relationship between the different phases of CRISP-DM and illustrates the recursive nature of a data mining project., April 2012.

[KAS11] Toshihiro Kamishima, Shotaro Akaho, and Jun Sakuma. Fairness-aware Learning through Regularization Approach. In *2011 IEEE 11th International Conference on Data Mining Workshops*, pages 643–650, Vancouver, BC, Canada, December 2011. IEEE.

[KBV09] Yehuda Koren, Robert Bell, and Chris Volinsky. Matrix Factorization Techniques for Recommender Systems. *Computer*, 42(8):30–37, August 2009. Conference Name: Computer.

[KLFH19] Aria Khademi, Sanghack Lee, David Foley, and Vasant Honavar. Fairness in Algorithmic Decision Making: An Excursion Through the Lens of Causality. In *The World Wide Web Conference*, pages 2907–2914, May 2019. arXiv:1903.11719 [cs, stat].

[KLRS18] Matt J. Kusner, Joshua R. Loftus, Chris Russell, and Ricardo Silva. Counterfactual Fairness, March 2018. arXiv:1703.06856 [cs, stat].

[KP17] Matevž Kunaver and Tomaž Požrl. Diversity in recommender systems – A survey. *Knowledge-Based Systems*, 123:154–162, May 2017.

[LAK18] Jurek Leonhardt, Avishek Anand, and Megha Khosla. User Fairness in Recommender Systems. In *Companion of the The Web Conference 2018 on The Web Conference 2018 - WWW '18*, pages 101–102, Lyon, France, 2018. ACM Press.

[LCF+21] Yunqi Li, Hanxiong Chen, Zuohui Fu, Yingqiang Ge, and Yongfeng Zhang. User-oriented Fairness in Recommendation. In *Proceedings of the Web Conference 2021*, pages 624–632, April 2021. arXiv:2104.10671 [cs].

[LCX+21] Yunqi Li, Hanxiong Chen, Shuyuan Xu, Yingqiang Ge, and Yongfeng Zhang. Towards Personalized Fairness based on Causal Notion. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, pages 1054–1063, New York, NY, USA, July 2021. Association for Computing Machinery.

[LTDSL22] Davide Luca, Javier Terrero-Davila, Jonas Stein, and Neil Lee. Progressive Cities: Urban-rural polarisation of social values and economic development around the world. page 40, 2022.

[LXL+12] Chen Lin, Runquan Xie, Lei Li, Zhenhua Huang, and Tao Li. PRemiSE: personalized news recommendation via implicit social experts. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 1607–1611, Maui Hawaii USA, October 2012. ACM.

[MAP+20] Masoud Mansoury, Himan Abdollahpouri, Mykola Pechenizkiy, Bamshad Mobasher, and Robin Burke. FairMatch: A Graph-based Approach for Improving Aggregate Diversity in Recommender Systems, May 2020. arXiv:2005.01148 [cs].

[MAP+22] Masoud Mansoury, Himan Abdollahpouri, Mykola Pechenizkiy, Bamshad Mobasher, and Robin Burke. A Graph-based Approach for Mitigating Multi-sided Exposure Bias in Recommender Systems. *ACM Transactions on Information Systems*, 40(2):1–31, April 2022. arXiv:2107.03415 [cs].

[Max19] Rahsaan Maxwell. Analysis | Why are urban and rural areas so politically divided? *Washington Post*, March 2019.

[MCBP+20] Martin Mladenov, Elliot Creager, Omer Ben-Porat, Kevin Swersky, R. Zemel, and Craig Boutilier. Optimizing Long-term Social Welfare in Recommender Systems: A Constrained Matching Approach. July 2020.

[Meh22] Sourabh Mehta. Why is the fairness in recommender systems required?, July 2022.

[MS07] Andriy Mnih and Russ R Salakhutdinov. Probabilistic Matrix Factorization. In *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc., 2007.

[MZS20] Alessandro B. Melchiorre, Eva Zangerle, and Markus Schedl. Personality Bias of Music Recommendation Algorithms. In *Fourteenth ACM Conference on Recommender Systems*, pages 533–538, Virtual Event Brazil, September 2020. ACM.

[Nas17] Kieran Nash. The difference between an expat and an immigrant? Semantics, January 2017.

[PBG+20] Gourab K. Patro, Arpita Biswas, Niloy Ganguly, Krishna P. Gummadi, and Abhijnan Chakraborty. FairRec: Two-Sided Fairness for Personalized Recommendations in Two-Sided Platforms. In *Proceedings of The Web Conference 2020*, pages 1194–1204, April 2020. arXiv:2002.10764 [cs].

[PCH11] Pearl Pu, Li Chen, and Rong Hu. A user-centric evaluation framework for recommender systems. In *Proceedings of the fifth ACM conference on Recommender systems*, RecSys '11, pages 157–164, New York, NY, USA, October 2011. Association for Computing Machinery.

[Raw99] John Rawls. *A Theory of Justice: Revised Edition*. Belknap Press, Cambridge, MA, September 1999.

[RD20] Shaina Raza and Chen Ding. A Regularized Model to Trade-off between Accuracy and Diversity in a News Recommender System. In *2020 IEEE International Conference on Big Data (Big Data)*, pages 551–560, December 2020.

[RD22] Shaina Raza and Chen Ding. News recommender system: a review of recent progress, challenges, and opportunities. *Artificial Intelligence Review*, 55(1):749–800, January 2022.

[SAHN22] Mete Sertkan, Sophia Althammer, Sebastian Hofstätter, and Julia Neidhardt. Diversifying Sentiments in News Recommendation. 2022.

[She00] Colin Shearer. The CRISP-DM model: the new blueprint for data mining. *Journal of data warehousing*, 5(4):13–22, 2000. Publisher: THE DATA WAREHOUSE INSTITUTE.

[She21] Elisa Shearer. More than eight-in-ten Americans get news from digital devices, 2021.

[SHFT15] Markus Schedl, David Hauger, Katayoun Farrahi, and Marko Tkalčič. On the Influence of User Characteristics on Music Recommendation Algorithms. In Allan Hanbury, Gabriella Kazai, Andreas Rauber, and Norbert Fuhr, editors, *Advances in Information Retrieval*, Lecture Notes in Computer Science, pages 339–345, Cham, 2015. Springer International Publishing.

[Ton18] Lionel Tondji. *Web Recommender System for Job Seeking and Recruiting*. PhD Thesis, February 2018.

[Wan21] Benjamin Wang. Ranking Evaluation Metrics for Recommender Systems, January 2021.

[WC21] Ningxia Wang and Li Chen. User Bias in Beyond-Accuracy Measurement of Recommendation Algorithms. In *Fifteenth ACM Conference on Recommender Systems*, pages 133–142, Amsterdam Netherlands, September 2021. ACM.

[WMZ+22] Yifan Wang, Weizhi Ma, Min Zhang, Yiqun Liu, and Shaoping Ma. A Survey on the Fairness of Recommender Systems. *ACM Transactions on Information Systems*, page 3547333, July 2022. arXiv:2206.03761 [cs].

[XDZ+17] Hong-Jian Xue, Xinyu Dai, Jianbing Zhang, Shujian Huang, and Jiajun Chen. Deep Matrix Factorization Models for Recommender Systems. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, pages 3203–3209, Melbourne, Australia, August 2017. International Joint Conferences on Artificial Intelligence Organization.

[YH17] Sirui Yao and Bert Huang. Beyond Parity: Fairness Objectives for Collaborative Filtering, November 2017. arXiv:1705.08804 [cs, stat].

[ZB18a] Junzhe Zhang and Elias Bareinboim. Equality of Opportunity in Classification: A Causal Approach. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[ZB18b] Junzhe Zhang and Elias Bareinboim. Fairness in decision-making — the causal explanation formula. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'18/IAAI'18/EAAI'18, pages 2037–2045, New Orleans, Louisiana, USA, February 2018. AAAI Press.

[ZMH+21] Wayne Xin Zhao, Shanlei Mu, Yupeng Hou, Zihan Lin, Yushuo Chen, Xingyu Pan, Kaiyuan Li, Yujie Lu, Hui Wang, Changxin Tian, Yingqian Min, Zhichao Feng, Xinyan Fan, Xu Chen, Pengfei Wang, Wendi Ji, Yaliang Li, Xiaoling Wang, and Ji-Rong Wen. RecBole: Towards a Unified, Comprehensive and Efficient Framework for Recommendation Algorithms, August 2021. arXiv:2011.01731 [cs].

[ZMKL05] Cai-Nicolas Ziegler, Sean M. McNee, Joseph A. Konstan, and Georg Lausen. Improving recommendation lists through topic diversification. In *Proceedings of the 14th international conference on World Wide Web*, WWW '05, pages 22–32, New York, NY, USA, May 2005. Association for Computing Machinery.