# Being Diverse is Not Enough: Rethinking Diversity Evaluation to Meet Challenges of News Recommender Systems

Celina Treuillier, Sylvain Castagnos, Evan Dufraisse, Armelle Brun

# Being Diverse is Not Enough: Rethinking Diversity Evaluation to Meet Challenges of News Recommender Systems

Célina Treuillier
celina.treuillier@loria.fr
Université de Lorraine, CNRS, LORIA
Vandoeuvre-lès-Nancy, FRANCE

Sylvain Castagnos
sylvain.castagnos@loria.fr
Université de Lorraine, CNRS, LORIA
Vandoeuvre-lès-Nancy, FRANCE

Evan Dufraisse
evan.dufraisse@cea.fr
Université Paris-Saclay, CEA, LIST
Palaiseau, FRANCE

Armelle Brun
armelle.brun@loria.fr
Université de Lorraine, CNRS, LORIA
Vandoeuvre-lès-Nancy, FRANCE

## ABSTRACT

Modern societies face many challenges, one of them is the rise of affective polarization over the last 4 decades. In an attempt to understand its reasons, many researchers have questioned the role of Social Media in general, and Recommender Systems (RS) in particular, on the emergence of these extreme behaviors. Diversity in News Recommender Systems (NRS) was quickly perceived as a major issue for the preservation of a healthy democratic debate. However, after more than 15 years of research in Artificial Intelligence on the subject, the understanding of the real impact of diversity in recommendations remains limited. Through a case analysis on the well-known MIND dataset, we propose a critique of the diversity-aware recommendation and evaluation approaches, and provide some take-home messages related to the need of adapted datasets, diversity metrics and analytical methodologies.

## CCS CONCEPTS

• **Information systems** → **Recommender systems**; *Information retrieval diversity*; • **Human-centered computing** → Empirical studies in collaborative and social computing.

## KEYWORDS

News recommender systems, User modeling, Diversity impact, Dataset analysis, Polarization, Filter bubbles

**ACM Reference Format:**
Célina Treuillier, Sylvain Castagnos, Evan Dufraisse, and Armelle Brun. 2022. Being Diverse is Not Enough: Rethinking Diversity Evaluation to Meet Challenges of News Recommender Systems. In *Adjunct Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization (UMAP '22 Adjunct), July 4–7, 2022, Barcelona, Spain.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3511047.3538030


*UMAP '22 Adjunct, July 4–7, 2022, Barcelona, Spain*

ACM ISBN 978-1-4503-9232-7/22/07…$15.00
https://doi.org/10.1145/3511047.3538030

## 1 INTRODUCTION

Recommender Systems (RS) were first intended to direct users' attention towards resources that are more suitable to their needs or preferences [68]. This kind of approach was successful in many domains, such as e-commerce, VOD or search engines. However, various studies showed that user satisfaction is not restricted to the traditionally used precision measure, and McNee *et al.* have paved the way for a new generation of diversity-oriented RS [57]. Since then, the integration of diversity in machine learning models has shown many simultaneous benefits on user confidence, acceptance rate and user perceived qualities [35], or on the reduction of over-specialization [1]. Diversity has also been integrated into almost all RS evaluation frameworks [24, 61]. Nevertheless, diversity in its current form is beginning to show its limits in certain contexts.

Taking social media as an example, the literature still highlights plenty of scientific issues around diversity [67]. The shift toward online news sharing and consumption gave the opportunity to news readers to access a huge amount of news everyday. In that context, News Recommender Systems (NRS) have been developed to help them identify content they are interested in. Traditional RS transposed in news domain tend to provide recommendations that do not challenge users' prior beliefs, offering news similar to those previously consulted. This often results in the creation of filter bubbles [62]. An intuitive approach to this problem has been to increase as much as possible the diversity of the set of recommended news without altering too much the accuracy of the recommendations. The goals of such an approach are to provide a larger spectrum of opinions to all users, ensure ethical and fair recommendations [31, 72], and foster a healthy democratic debate [12, 53]. However, the role of NRS on user polarization is still disputed, and the benefits of diversity in this context may be overestimated. On the one hand, it is commonly accepted that the repeated consumption of news covering only a narrow range of opinions about a specific topic can admittedly lead to the adoption of extreme positions [75]. Based on this, recent studies highlight that social media companies like Facebook[1] and Twitter[2] intensify political sectarianism [26, 77]. On the other hand, social media is probably a key facilitator and an amplifier – rather than a cause – of polarization. A recent survey by Boxell *et al.* showed that political

[1] https://www.facebook.com/
[2] https://www.twitter.com/

polarization is a phenomenon that has been growing for 40 years, before the advent of the Internet and social networks [14]. They even point out that affective polarization – *i.e.* the tendency for partisans to dislike and distrust those from the other party [22] – has risen in USA, Canada, New Zealand and Switzerland, while it decreased in UK, Australia, Germany, Norway and Sweden where social media is also widely used.

As stated by Kubin *et al.* [46], one can regret the lack of research exploring ways social media can depolarize. In March 2020, an experiment in which subjects stopped using Facebook for a month highlighted a significant reduction of polarization on policy issues [5]. More surprisingly, it did not diminish divisiveness based on party identity. That can be explained by the fact that people tend to be more upset and angry at the other side when seeing political content on social media. That is consistent with the conclusions of Bail *et al.* according to which exposure to opposite views can increase political polarization [7].

In this context, it is questionable whether ensuring high diversity for all users is appropriate. Of course, this is a sensitive issue because it is vital to make certain that RS, by adapting their strategies to each individual, do not influence users and standardize opinions. Though, **we support the idea that current NRS have a lack of expert knowledge on user personality traits, leading to a poor understanding and attention on users' behavior classes**. Through this position paper, we would like to investigate this idea and raise the following research questions:

- **RQ1.** Does diversity of recommendations and content bring a systematic gain? In particular, we seek to determine the extent to which the diversity impacts the news consumption behaviors of users and contributes to broadening their spectrum of news access.
- **RQ2.** Is it sufficient to measure the influence of an NRS afterwards with single-number metrics, or does this influence occur with some variations over time?

Enlarging our vision, this problem can be extended this way:

- **RQ3.** Do NRS really influence all users equally? If not, we can imagine that all users are not receptive to bias and are not all exposed to filter bubbles.
- **RQ4.** Are there different classes of user behavior with respect to viewing diversity, independently of the recommender system? If so, do viewing habits prior to interactions with an NRS impact user trajectory during the recommendation phase?

We do not pretend to answer all these questions, but rather to open the debate on current practices in the field of social media and NRS. To contribute to this reflection, we study the well-known open MIcrosoft News Dataset (MIND)[3], and apply both commonly evaluation methods of recommendations diversity and a novel in-depth analysis taking into account the temporal aspect of that diversity. We aim to draw preliminary conclusions from this analysis and provide some take-home messages (TH) and perspectives for researchers in User Modeling, Adaptation and Personalization. In particular, we emphasize the need for (1) **enriched datasets** (including more information about users, opinions, context and triggering factors of news consumption, and topics of controversy) for a better understanding of users' behavior and personality traits, (2) **new analytical methodologies** to put the same effort into modeling the diversity trajectory of users with and without RS as was put into assessing the diversity provided by RS, (3) **new diversity metrics** suitable for NRS by taking the time dimension and opinions into account in addition to topics, and (4) **new human-computer interaction and recommendation models** to bring the appropriate and personalized amount of diversity, and guarantee a fair and depolarizing NRS.

The remainder of this paper is structured as follows: In Section 2, we first provide a literature review on how RS and NRS are commonly evaluated, with a particular emphasis on the notion of diversity and time dimension. We then compare, in Section 3, the existing open datasets and choose one to analyze. Section 4 is dedicated to the analysis of diversity within the MIND dataset as an attempt to measure its influence on users' news consumption. Finally, we provide some take-home messages for the community.

## 2 RECOMMENDER SYSTEM EVALUATION FRAMEWORKS

### 2.1 Single-Number Metrics

The main purpose of recommender systems is to provide users with content adapted to their preferences and/or needs, which they would not have consulted spontaneously [68]. To achieve this goal, many accuracy metrics have been designed and are used both to compute recommendations and to evaluate the quality of the models through offline datasets [4]. These metrics can either be error-based or ranking-based [69].

Error metrics are often preferred when evaluating single recommendations and/or predictions. They include for instance the *Mean Absolute Error* (MAE), the *Mean Squared Error* (MSE), and the *Root Mean Squared Error* (RMSE) [18]. Rank metrics are adapted to listwise or sequence-based recommenders (*precision*, *recall*, *nDCG*, *MAP*, *MRR*...) [4, 17, 36, 63, 73]. Their aim is to determine whether the ranking of the recommendations is appropriate, i.e. the items in the test set that correspond to the strongest user preferences are present in the top-N recommendations, and conversely the recommendations with the highest scores are highly valued items in the test set.

Error-based or rank-based metrics have been proven to assess the relevance of recommendations, but they suffer from two limitations. First, as stated by Scheidt and Beel, they are based on the entire dataset and hold no information if models perform the same over the whole time period or if they improved or worsened over time [71]. Second, accuracy has been proven to be insufficient to evaluate and explain user satisfaction [40, 55, 57, 91]. It is to address this second limitation and go beyond-accuracy metrics that evaluation frameworks have been enriched with many other complementary dimensions, such as: *Coverage* [30], *Confidence* [56], *Trust* [59], *Acceptance and Adoption Rates* [6], *Novelty* [89], *Serendipity* and *Unexpectedness* [2], and *Diversity* [47]. The latter dimension has allowed recommender systems to progress considerably over the last 15 years, as we will see.

Diversity in RS was first introduced by Smyth and McClave [74]. At that time, it was considered as the opposite of similarity (1 –

[3] https://msnews.github.io/

*similarity*), the similarity being another metric to measure the proximity of the recommendations with the known preferences of users. Smyth and McClave foregrounded the lack of diversity in recommendations produced by similarity-based approaches and the problems that it entails. For instance in e-commerce, the recommendations should all be similar to the product currently consulted by the active user. However, if they are also too similar to each other, it will be difficult for the user to make a decision among this set of alternatives. Diversity was therefore a means of modifying the list of recommendations without compromising accuracy, before becoming a full-fledged evaluation objective for RS. Subsequently, several diversity metrics have been proposed to take into account different needs and application contexts. The *Intra-List Similarity* (ILS) [92] and, in contrast, the *Intra-List Distance* (ILD) [90] aim to measure the average level of diversity within a set of recommendations in a permutation-insensitive way: rearranging positions of recommendations does not affect the measure. With this new definition, it becomes possible to measure diversity within a group, and not simply between pairs of items. A very similar and still widely used formula for $Diversity$ within a set of $n$ items was also proposed by Smyth and McClave [15]. The diversity of a set of items $(i_1, i_2, ...i_n)$ is seen as the average dissimilarity between all pairs of items, and computed as follows:

$$Diversity(i_1, i_2, ..., i_n) = \frac{\sum_{k=1}^{n} \sum_{j=1}^{n} (1 - Similarity(i_k, i_j))}{\frac{n}{2} * (n-1)} \quad (1)$$

Other metrics complete the picture such as the *Relative diversity* (RD) which measures the diversity brought by an item relatively to a set of items [15], the *Expected Intra-List Diversity* (EILD) which is rank-sensitive and rank-aware [78], the *aggregate diversity* used to measure the diversity of items across the recommendation lists of all users and prevent the long-tail problem [3], and the *inverse of Gini index* to measure of distributional inequality [27].

Many machine learning models have been developed to increase diversity in RS [41, 47, 80, 84, 88] and have been designed for different application domains such as e-commerce [35], online music services [8, 49], and social networks [70]. In parallel, many user studies have been conducted to prove the added value of diversity in RS. Zhang and Hurley [90] and Lathia [48] showed that users may suffer frustration when facing a lack of diversity. Research has even revealed that, in the case of equivalent accuracy between different models, diversity is not only perceived by users but also plays a prominent role in the acceptance of recommendations [25, 38]. Castagnos *et al.* highlighted the need for diversity so as to increase user confidence during the purchase decision process [16]. At last, Parapar *et al.* demonstrated that diversity can help to elicitate users' preferences, thus reducing the popularity bias in RS [60].

Despite these scientific advances, we note that all these works use single-number metrics and therefore only partially address the challenges of beyond-accuracy RS. This type of evaluation can potentially show significant improvement, especially in terms of diversity and impact, on average on the whole sets of users, items and sessions. Nevertheless, it does not guarantee that the improvement is consistent over time, nor does it guarantee that it applies to all users. In the case of social media and news recommendation, interactions are built over a much longer duration than most application domains, ranging from several weeks to several years, and this problem becomes even more prevalent. The next subsection will illustrate the specificities of NRS.

## 2.2 Evaluation of News Recommender Systems

NRS are intended to select information of interest for news readers [67]. They face a major news-related characteristic that makes the recommendation process more complex [53]: news items have a very short lifespan and are constantly replaced by fresh releases. Considering this, it is obvious that traditional RS should be adapted to the news recommendation domain, as news quickly becomes obsolete. This adaptation of models is accompanied by the need to revise NRS evaluation. The widespread *a posteriori* evaluation based on a complete dataset could face major issues by recommending a given user and at a specific time news that no longer exist or that have not yet been written. Besides, considering online shopping, music listening, or movie viewing, users have short-term needs (at the session level). In the field of news, diversity is of course still desirable, but preference elicitation and recommendations spread over a longer period (several days, weeks, months...). Moreover, user satisfaction is not quantifiable on the basis of a single consumed item but rather on the numerous news accessed on their topics of interest. In this regard, the inclusion of diversity and the evaluation frameworks must take into account the time dimension.

Let us first have a look to diversity, which is equally important, if not more, than accuracy when recommending news because of the potential role of NRS in polarisation. Even though diversity is a key principle in news recommendation, there is no consensus on its definition, nor on the associated metrics. For example, Bauer and Werthner distinguish two types of diversity: content diversity and provider diversity, based on different concepts and objectives [9]. These diversities do not influence the recommendations in the same way: by diversifying the content, the recommended news will cover different topics, but will potentially come from the same source, whereas by diversifying the providers, the news will represent diverse providers, but may address similar topics. According to Joris *et al.*, diversity represents a broad concept, relative to several fields (computer science, communication science, law and computational linguistics), and covers a large specter of aspects (news, media, democracy) [39]. Vrinjenhoek *et al.* affirm that diversity can not be defined with a unique and absolute value, but rather with an aggregate value of many aspects, especially in the news domain [79]. They explain that what defines a "good" diversity in a NRS depends mainly of the goal of the system, and the normative framework it aims to follow. Bringing diversity in NRS is a complex task, as well as evaluating its potential impact on users' news consumption. The democratic role of NRS [34] strengthens even more the importance of controlled evaluation of diversity [12]. NRS impact goes beyond acceptance or satisfaction issues, as it may play a role in the enclosure of users in filter bubbles [19, 62]. There is some evidence that providing diversity can have a depolarizing capacity [33]: it may encourage users to consume news related to opposing viewpoints, and even encourage a higher tolerance toward them. However, diversity can sometimes have deleterious impacts: Bail *et al.* show that diversity can potentially lead to more extreme viewpoints [7]. At last, some researchers show that the potential impact of NRS,

especially on social media, are different from one system to another, but also from one user to another [20, 44]. Non-personalized diversity – *i.e.* the diversity produced globally by the NRS with the same process for all users without taking into account their specificities – thus does not seem to have universal consequences on news consumption behavior. This encourages the modeling of the individual users' diversity needs, which may potentially reduce detrimental impacts [23, 85].

Let us now turn our attention to the temporal dynamics challenge. In the field of recommendation, it refers to the evolution of users' preferences over time and the need to adapt recommendations to this evolution [86, 87]. Joeran Beel fosters the need of a temporal evaluation, which can allow drawing more precise conclusions about RS and their actual performances over time [11]. Raza and Ding highlight that the dynamic updating of news is also accompanied by temporal dynamics of users' preferences, which may change in the short or long term, even seasonally [66]. Hence, authors propose a model that exploits temporal dynamics for news recommendations. Very few other studies address temporal dynamics in NRS: Li *et al.* integrated long-term preferences of users by building a time sensitive weighting scheme, as well as short-term preferences by analyzing recent users' reading history [50]. Other researchers propose to mine patterns of temporal changes in data streams, and then incorporate trends and temporal user habits in their NRS [52]. To the best of our knowledge, Lathia *et al.* is the first and unique work that explores the importance of temporal diversity and explain that RS must necessarily adapt to temporal changes, mainly because of data renewal, and that changes in temporal diversity affect users' preferences [48]. This work does not specifically focus on news and as far as we know, no work deal with temporal diversity in NRS.

From this literature review, we can conclude that the diversity approaches in NRS are based on different definitions, relate to many concepts, and there is no common framework, both for diversification processes and evaluation. Furthermore, standard evaluations of the impact of diversity on news consumption come to scattered conclusions: diversity can reduce user polarization, have no effect at all, or even exacerbate the adoption of extreme viewpoints. Thus, modeling users' need for diversity in a personalized way appears to be required to control more finely the impact of diversity according to different user profiles and needs. Moreover, the characteristics of the news induce strong temporal variations, and the evaluation of the individual dynamics, i.e. the trajectories specific to each user, are not currently applied. Together, these elements refer to the intuitions presented in the introduction, and we propose to confirm them on an existing dataset and highlight overarching perspectives responding to the challenges of NRS.

## 3 NEWS RECOMMENDER SYSTEMS DATASETS

The design of NRS strongly depends on the data that can be used: number of users, number of news, user and news attributes, user activities, recency, time span, etc.

Due to copyright and privacy issues, few online media provide free access to news content and user news consumption. The number of publicly available datasets, even for research purpose, is thus notably reduced. As a consequence, many studies explore proprietary and non-public datasets that can not be shared nor reused [42]. In most cases, these datasets are small and gather information about a limited number of users during a short period of time. To cope with this limit, some researchers perform evaluations on datasets that are not related to news, such as the well-known MovieLens dataset [37]. However, such datasets cannot be fully used to draw strong conclusions about news recommendation.

Yet, some datasets dedicated to news recommendation are freely available. A brief overview of these datasets is given in Table 1 [42, 67]. We would like to note that these datasets contain information about news and user news consumption. Other domains, such as natural language processing, are also interested in news datasets, but these datasets only contain news related data.

Among those datasets, the one that seems to be the most adequate to address the questions raised in introduction is the MIND dataset. Indeed, it covers the largest time frame (6 weeks) of English-written news consumption, and provides information about both news and recommendations (impressions) provided to users. Besides, MIND is widely used in the news recommender systems area [51, 64, 82].

In detail, MIND gathers 15M impression logs about 1M users, interacting with 160k English news articles. Data were collected by randomly selecting users who have interacted with the Microsoft News website during six weeks, between October 12 and November 22, 2019. Data are split into training, validation and test sets: the clicks made during the first four weeks (from 12 October to 8 November, 2019) were used to construct the news click history, while those made on the fifth (from 9 November to 15 November, 2019) and sixth weeks (from 16 November to 22 November, 2019) were used for training, validation and test sets. To ensure user privacy, data are anonymized and users are identified using unique identifiers. In line with the choice made in some studies [81], we work on the small version of the dataset, provided by MIND developers. This subset contains information about 50k users randomly extracted from the large dataset. The associated history concerns the first four weeks (from 12 October to 8 November, 2019), and the impression (recommendation) logs were recorded during the fifth week only (from 9 November to 15), along with the clicks performed by users during that fifth week, within the set of impressions. Data about the sixth week is not part of this subset. We assume that the news accessed during the first four weeks are also chosen in a set of recommended news, upon which no information is given. A graphical temporal representation of this data is presented in Fig. 1.

## 4 ANALYSIS OF THE IMPACT OF DIVERSITY ON MIND USERS' NEWS CONSUMPTION

To study the impact of diversification of the recommendations on users' news consumption and to evaluate the extent to which diversification is a solution to affective polarization, we conduct experiments on the MIND dataset.

### 4.1 Experimental settings

*4.1.1 Data Selection and pre-processing.* From Section 3, we can notice that MIND provides limited information about the news (Category, SubCategory, Title, Abstract, URL). The actual content

**Table 1: Summary of free datasets for the news recommendation task.**

| Dataset | Description | Data types | Size | Period |
|---|---|---|---|---|
| **Yahoo Webscope** | Several datasets avalaible<br>https://webscope.sandbox.yahoo.com/ | News articles information<br>Click data | Depends on the dataset | Depends on the dataset |
| **Plista [43]** | Plista & TU Berlin<br>13 German news portals<br>Accessible upon request for research<br>https://www.plista.com/ | Editor information<br>Readers information<br>(clicks and impressions)<br>Time-related information | 17M sessions<br>70k news<br>80M impressions<br>1M clicks | 30 days<br>(June 2013) |
| **Adressa [32]** | Adressavisen and the<br>Norwegian U. of<br>Sciences and Technology<br>Large and small versions available<br>https://reclab.idi.ntnu.no/dataset/ | News articles information<br>Click data | *Large version:*<br>3M users, 48k news<br>27M clicks<br>*Small version:*<br>15k users, 1k news<br>2.7M clicks | *Large version:*<br>10 weeks<br>(2017)<br>*Small version:*<br>1 week<br>(2017) |
| **MIND [83]** | Microsoft News Website<br>Large and small versions available<br>https://msnews.github.io/ | News articles information<br>Impressions logs | *Large version:*<br>1M users, 200k news<br>15M impression logs<br>*Small version:*<br>50k users, 60k news<br>1M impression logs | *Large version:*<br>6 weeks<br>(Oct.-Nov. 2019)<br>*Small version:*<br>5 weeks<br>(Oct.-Nov. 2019) |
| **Globo.com [21]** | Brazilian news portal<br>https://www.globo.com/ | News articles information<br>Click data | 3M clicks<br>1.2M sessions<br>330k users, 50k news | 2 weeks<br>(October 2017) |
| **Outbrain** | Outbrain Click Prediction challenge<br>https://www.kaggle.com/c/outbrain-click-prediction | News articles information<br>Click data | 2 billion page visits<br>700M users, 17M clicks | 2 weeks<br>(June 2016) |

(body) of the news is not provided. The same for the publication date, which is crucial to study user behavior. Fortunately, by processing the URL, the body of a news and its publication date can be retrieved. On the later, the lifespan of the majority of news is less than two days in the Microsoft News Service, as mentioned in [83]. The delay between publication and user consultation is thus reduced. We propose to estimate users' consumption date by the news publication date.

Filter bubbles are not related to the interest in a unique topic, but in the interest in a unique opinion. So, to conduct an analysis that specifically focus on this latter aspect, we select a subset of news that deal with one category, specifically the "news" category. Not only this category is predominant in the dataset, but it is also the closest category to the political domain, which is of particular interest as we are interested in political polarization.

Finally, **RQ2** being related to the temporal analysis of users news consumption, observations need to be made on users who consumed news throughout the complete dataset time span. We thus keep only users who accessed news each of the five weeks, and at least 3 news each week to have a valid interpretation of their consumption.

*4.1.2 News Representation.* In line with the literature [28, 54, 76], we used an unsupervised approach to represent the topic distributions of news items, namely LDA models[4] [13]. After concatenating the original MIND train/validation/test news items, we performed near-deduplication using MinHashLsh, resulting in 126,649 articles. Using the Spacy library[5], we removed stop-words, digits, and we lemmatized words, forming bag-of-words representations. News containing less than 20 words were filtered out. Thus, post-processed news items have between 21 and 4,351 words, with an average of 261 words. The model hyperparameters were tuned to optimize performance.

*4.1.3 Diversity Metric.* The news representation serves as a basis to measure their diversity. We select a widely used measure, i.e. the average dissimilarity between each pair of news (see equation (1)), that ranges between 0 (no diversity) and 1 (maximal diversity). Similarity was computed using cosine similarity [65], as commonly done in NRS field [45, 58].

The data structure used in the experiments are summarized in Table 2. The resulting dataset, which we call MIND5w, is made up of 1,475 users and 20,541 news.

## 4.2 Experimental Analysis

The experiments conducted below are designed to contribute to answer both research questions raised in Introduction. For **RQ1**, we adopt a holistic perspective of diversity. For **RQ2**, we adopt a temporal perspective to provide additional information about the temporal dynamics of users' news consumption.

As a prerequisite of the experiments, we ascertain that the news from the dataset actually offer users the possibility to get engaged with content more or less diverse. To this aim, we evaluate the average diversity of each pair of news in MIND5w. Herewith, the

[4]https://bab2min.github.io/tomotopy/v0.12.2/en/

[5]https://spacy.io/

| Column name | Description |
|---|---|
| News represent. | Vectorial represent. of news |
| Date | Date of the event (hist./reco./acc.) |
| UserID | Unique ID of the users |
| History | Boolean, true if news acc. in hist. |
| Recommended | Boolean, true if news reco. |
| Accessed | Boolean, true if news accessed |

**Table 2: Data in MIND5w**

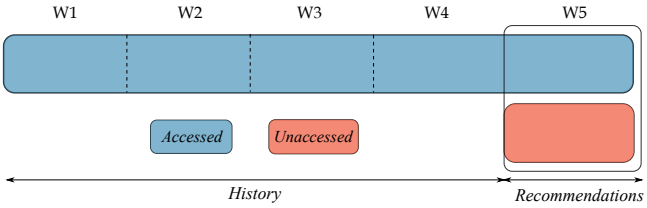



**Figure 1: Graphical representation of MIND5w**

average diversity is 0.79, which shows that news items are globally diversified. The standard deviation (0.27) affirms that really high, as well as lower diversity values actually exist in MIND5w. This scattered diversity distribution affirms that users have the opportunity to access both highly and little diversified news. Refining this analysis by considering all news pairs individually, a large proportion (53.5%) have a diversity greater than 0.9.

*4.2.1 Single-Number Analysis.* In this section we adopt a holistic perspective of diversity. The first four weeks of MIND5w are associated with the users' news consumption. The last week ($5^{th}$ week) is also associated with users' news consumption, supplemented by the news recommended to users (see Figure 1). To make a thorough analysis of these sets of news over all users, we propose to evaluate four diversity measures.

• *Diversity of history* is the users' average diversity of the news they accessed during the first four weeks. Users accessed on average 39 news.

• *Diversity of recommended news* corresponds to the average diversity of news recommended during the $5^{th}$ week. The number of recommended news that week is quite high, 140 news on average, which is significantly higher than the average number of news accessed in the history.

For that $5^{th}$ week, the dataset also provides information about the news the users chose to access among these recommendations and those they did not access. This allows to measure:

• *Diversity of accessed news* corresponding to the average user diversity of the news accessed during the $5^{th}$ week (among the recommendations). It has a similar meaning than the diversity of history, but over a shorter period. Users access on average 9 news during the $5^{th}$ week, which is similar to the weekly number of news accessed in the history.

• *Diversity of unaccessed news* corresponding to the average diversity of the recommended but not accessed news. The number of accessed news being relatively limited compared to the number of recommendations, these unaccessed news represent a large part of the recommended news, 131 news on average. This indicates that diversity of unaccessed and recommended news will be highly similar.

The distribution of the number of users over the diversity values for each of the four previously introduced diversity measures, is displayed in Figure 2.

Considering the distribution of the recommendations, we can see that the average diversity is 0.75, which is quite high and the standard deviation is 0.04, which is rather small. We would like to point out that the average diversity of recommendations and the average diversity of news are not directly comparable, although both diversities have been evaluated similarly. Indeed, the average diversity of recommendations is calculated over all pairs of news items in the recommendation sets, that we suppose to be made up pieces of news that are rather similar to each other (at least in terms of topics), even though the sets are diversified. The average diversity of the news in MIND5w is calculated over all pairs in the set of 20k news. As previously mentioned, the majority of the pairs have a really high diversity (>0.9) as most of them are totally unrelated, which naturally increases the average diversity.

Besides, the high diversity and small standard deviation in recommended news leads us to think that a diversification process actually exists in the recommender system of the Microsoft News Website, that guarantees that recommendations provided to users are diverse and meet a predefined diversity level.

As expected, the distribution of the diversity of the set of unaccessed news is similar to the one of the recommended news (Kolmogorov-Smirnov test, $p - value = 0.148 > \alpha = 5\%$). Recall that this similar distribution is more than probably due to the fact that the unaccessed news is a subset (93%) of the recommended ones.

If we focus on the news accessed during the $5^{th}$ week, the distribution significantly differs from the distribution of the recommended news. On the one hand, the average diversity of accessed news is significantly lower than for recommended news. So, users access news items that are on average less diversified than those recommended. To be precise, 67.1% of the users have a diversity of accessed news lower than the average diversity of recommendations. On the other hand, the standard deviation (0.17) is significantly higher, and reflects a greater variability in the average diversity of news accessed by users during the $5^{th}$ week.

The distribution of the accessed news in history (first four weeks) seems to be as spread as the one of the accessed news during the $5^{th}$ week. Yet, the average diversity in history is significantly smaller ($p - value \approx 0$). This leads us to question about the impact of a recommender system.

To refine this analysis, we study these diversities (recommended, accessed and history) for each user, displayed in Figure 3. If we focus on the diversity of recommended news and the diversity of history (Figure 3a), we note that the average diversity of the recommendations is consistent across all users, regardless of the number of news in the history. Beyond the previous finding about the high diversity of the recommended news in MIND5w, we confirm here that this diversity is not personalized, especially in terms of diversity in the history. Comparing the average diversity of accessed

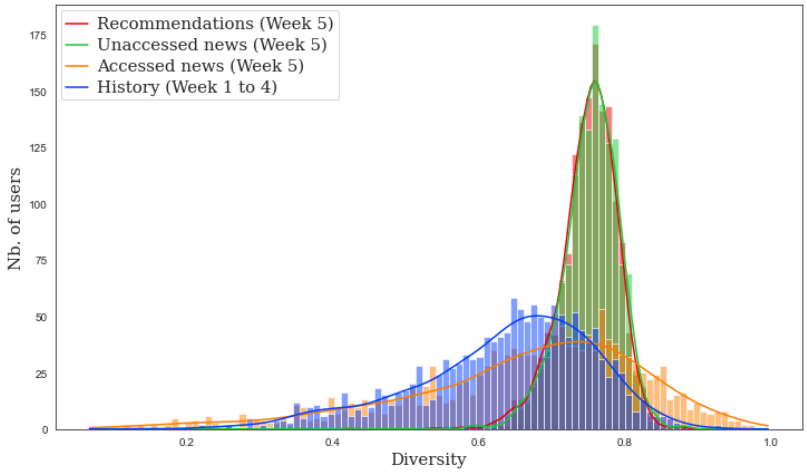



**Figure 2: Distribution of diversity among users.**

news and the diversity of history (Figure 3b), we see that the vast majority of the users are close to the diagonal, i.e. most of them have a diversity of history close to the diversity of accessed news. The prominence of users upper (but close to) the diagonal confirms the previously identified higher diversity during the $5^{th}$ week.

With the goal to understand why some users do not have similar diversity in history and in $5^{th}$ week, we conducted additional experiments, not presented here, that provide additional findings. First, about the impact of the number of news accessed in the history on the diversity of the accessed news, we confirm that this number does not explain the difference in diversity between history and accessed news. Second, during the $5^{th}$ week, 67.7% of users access news less diversified than the recommendations they get.

From these experiments, we can conclude that a high diversity of recommendations is far from leading to a systematic diverse news consumption. Altogether, these results motivate us to further analyze the extent to which RS impact users' news consumption over time.

*4.2.2 Temporal Analysis.* Let us now focus on the evolution of the diversity of users through time, namely on the complete time span of MIND5w. To this aim, we adopt a weekly perspective.

The distribution of users' average diversity for each week is presented in Figure 4. A first look at these distributions shows that they are close to each other. However, a Kolmogorov-Smirnov test attests that the distributions are not similar between weeks ($p-values \approx 0$), except for the week 1 that is similar to weeks 3 and 4 ($p-value = 0.124$ and $p-value = 1.160$, respectively). These results were confirmed with the Anderson-Darling test for $k$-samples. However, considering the evolution of the average diversity, it does not systematically increase or decrease.

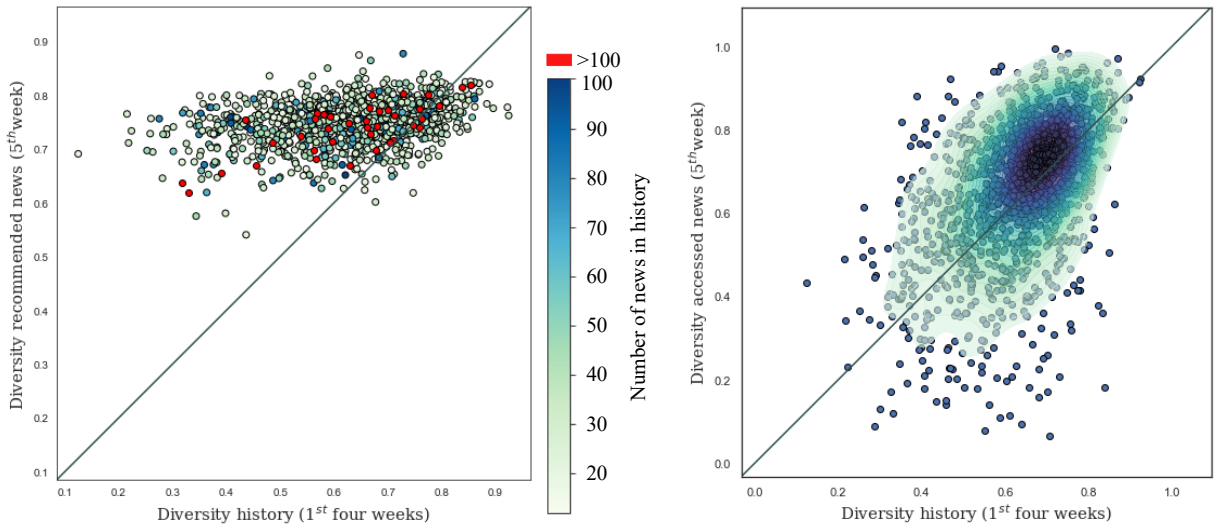
To analyze more precisely the evolution of user diversity through time, we choose to form four equally sized groups of users. These groups are defined from the quartile values of their weekly average diversity, and the first quarter contains users with lower diversity values. A study of the values of the three quartiles highlight that they remain stable across the five weeks, which confirms that even if distribution of diversity between weeks are different, they are close.

Figure 5 represents the flow between quarters from one week to the next one in a Sankey Diagram. The thickness of each flow is proportional to the number of users in that flow. Let us start by highlighting that the transition patterns remain stable over the weeks: it confirms a similar global impact of the recommendations on users diversity of accessed news through weeks. Each possible transition (from any quarter to any quarter) actually occurs, although almost 80% of users remain in the same or adjacent quarter between two consecutive weeks. So, few users change significantly their consumption habits between two weeks.

This study of users' behavior between adjacent weeks remains limited and we would like to highlight user specific temporal behavior along the five weeks according to these quarter shifts. A variation ranges in $[-3, +3]$, where a variation is equal to 3 when a user moves from the $1^{st}$ to the $4^{th}$ quarter, i.e. her diversity greatly increases. A variation equal to -3 represents a user who moves from the $4^{th}$ quarter to the $1^{st}$ one, *i.e.* her diversity dramatically decreases. If the variation equals to 0, the user remains in the same quarter.

These variations are visually represented with a heatmap in Figure 6. A red variation represents an increase in the diversity, the darker the larger. A blue variation represents a decrease in the diversity. The Figure represents the diversity variation of each week relatively to week 1 to study the impact of the recommender system through time. This provides an overall picture of the trajectories taken by users in relation to their average diversity during the first week. We define three types of users, depending on their receptiveness to recommendations' diversity.

- **Positively receptive users**: users accessing more diverse news after 5 weeks of use of the system.
- **Negatively receptive users**: users accessing less diverse news at the end of the 5 weeks.
- **Resistant users**: users accessing equally diverse news after the 5 weeks.




**(a) Diversity of recommended news *vs* history.**

**(b) Density of the diversity of accessed news *vs* history.**

**Figure 3: Comparison of diversity of history, recommendations and accessed news. Each dot in (a) and (b) represents a user.**

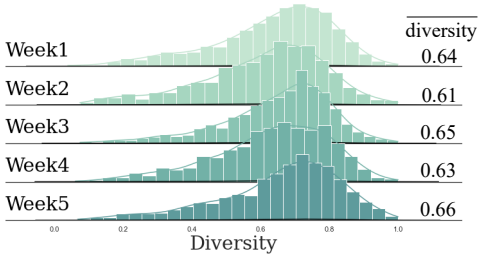



**Figure 4: User diversity distribution for the 5 weeks**

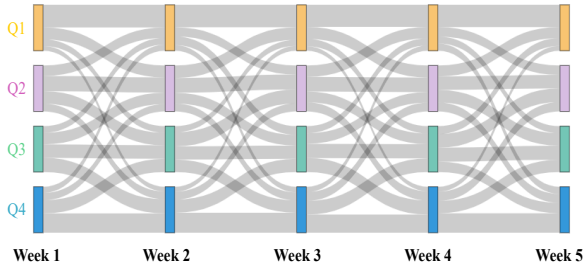



**Figure 5: Variation flow over weeks (*Q1 = quarter 1, ...*)**

On Figure 6, users are sorted according to their global evolution between weeks 1 and 5 (last column), then according to their evolution between weeks 1 and 2 (first column). Looking at the right-most column, we observe that previously presented types of users are divided into thirds: 32.1% of the users increase their diversity between week 1 and week 5, 32.5% of the users consume less diverse news in week 5 than in week 1. For both cases, the extrema represent users with the biggest change in their diversity. The white band in the center of the last column represents users whose average diversity did not change between weeks 1 and 5 (35.4% of the users).

If we now take a closer look at both positively and negatively receptive users, we observe that almost 90% of them observe a similar or null variation between week 1 and 2. Besides, more than 60% of the users have a variation direction in the first column similar to the one in the last column (red/red or blue/blue). For each of the three types of users, 25% of users have a consistent variation direction over the entire time span, equally distributed among the user types. In addition, only 15% of the users have temporary opposite variations (the others going through transient states with no variation). We can conclude that the variation between weeks 1 and 2 is an accurate predictor of the one at week 5.

Finally, the most striking finding is that a constant high diversity of recommendations, as used in MIND5w, has a negative impact on as many users as there are users with a positive impact. This tends to confirm findings of Bail *et al.* according to which exposing potentially polarized users to diversity can be counterproductive [7].

With regard to resistant users, 50% of them do not have any variation in the first week either. The other users evenly split between positive and negative variation. For about 30% of resistant users, the diversity of news accessed is not impacted along the five weeks. Looking more closely, the vast majority of users whose average diversity remains stable are those who belong to extreme quarters in the first week.

We can conclude that users who have either a really high diversity or those who have a very low diversity during the $1^{st}$ week

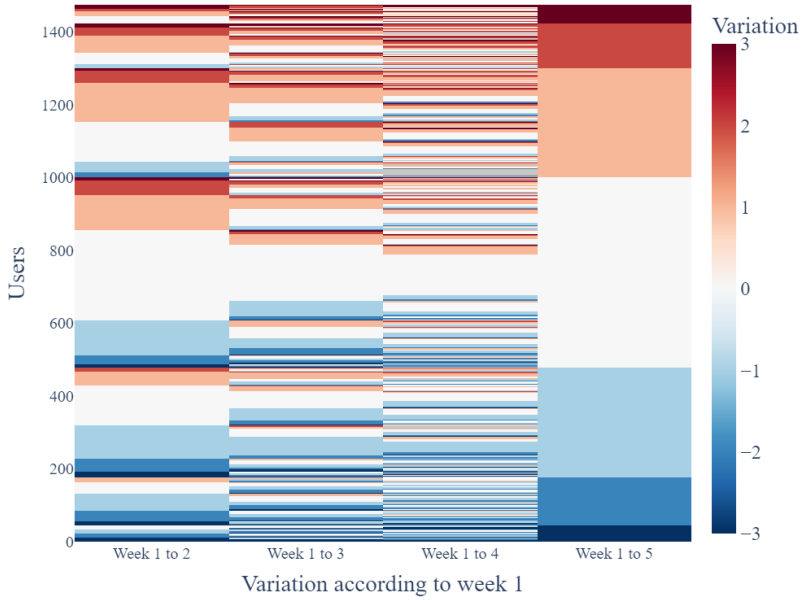



**Figure 6: Heatmap of quarter changes between week 1 and every other week. Each pixel line represents a user.**

are the most resistant users to the recommendations in terms of further news consumption.

To go further, we wonder whether these variations are influenced by the number of news accessed by the users over the five weeks, or by their initial diversity ($1^{st}$ week). Additional experiments provide us with arguments to state that all variations exist, whatever is the initial quarter of the users. In addition, the number of accessed news throughout the five weeks only impacts this variation if it is quite high: an extreme variation is never observed for such users.

### 4.3 Discussion

The Single-Number and temporal analyses allowed us to put forward important elements. First, new information has been highlighted about the dataset. The recommender system used in the Microsoft News Website is active all along the five weeks. The system ensures that the recommended news items are diverse, and fit a predefined high level of similarity close to 0.75. We are now able to answer **RQ1**: diversification does not lead to a systematic increase in the diversity of the news accessed, users significantly vary in terms of diversity of the news they access, ranging in the complete span of possible diversity values. More important, users' average diversity is lower (12%) than the average diversity of the recommendations, and 66% of users access less diverse news than recommendations.

From a temporal perspective, this varying impact is confirmed, and seems to operate every week, including at the user level. Each week, a significant proportion of users observe a variation in their average diversity and we highlighted three types of users: positively receptive, negatively receptive and resistant users. This temporal variation contributes to answer **RQ2**: single-number evaluations are insufficient, and there is a critical need for an NRS evaluation framework that takes into account both the differences between users and the temporal aspect.

## 5 CONCLUSION AND TAKE-HOME MESSAGES (TH)

The preliminary conclusions drawn above on the impact of the recommendations on user diversity, especially over time, deserve to be further discussed to highlight some needs and future research directions for the UMAP community.

First, we have seen that providing users with equally diverse sets of recommendations neither impacts similarly users, nor increases systematically the diversity of the news they consume. This leads us to focus on two elements.

- About the time span covered by the dataset and the span of use of the system. Recall that users that have been selected from the original MIND dataset are those who used the system during the five weeks. We saw that the diversity of the news consumed by users after five weeks remains as spread as the one in the first week and that the average diversity slightly evolves. It is more than probable that these users were already using the system before the data were captured. We can ask whether the impact of the diversity of the news recommended on the diversity of the news consumed occurred ahead the first week, which could explain the small average impact during the 5 weeks span. However, no information about the prior use of the system before these five weeks, which could have helped us to understand this limited impact and model more precisely the actual impact.
- About the individual impact of the recommendations. Although the average diversity of the news consumed by users are close between weeks 1 and 5, a specific focus shows that the impact differs fundamentally between users. Some users

totally polarize, other users totally open their interests and some do not modify their consumption habits. However, no information in the dataset could contribute to explain the individual impact.

**TH1: There is a cruel lack of open datasets that cover a significantly longer period to study the impact of the recommender systems in the long run, as well as richer datasets that can include additional information about the users (including a distinction between recommended and non-recommended news in user history), the contexts, and even the news to provide a more accurate model of the impact and adoption of the recommendations.**

Second, considering again the individual impact of diversified recommendations on users, that is not dependent on the users' prior diversity, we wonder who are the users who partially, even totally, reject the diversity of the recommendations and polarize progressively (**RQ3**). Why do they reject this diversity? We put forward the idea of the need of personalized diversification strategies, that could:

- consider user personality traits and user behavior classes to adapt the diversity level of the recommendations to these personal features to increase the adoption of diversified sets of recommendations.
- redefine the traditional diversity metrics, even propose and use different diversity metrics. Such new metrics could consider specific elements that users may be sensitive to, or specific elements of the news: recency, title, abstract, author, source, etc.

Besides, beyond the need to model opinions conveyed in the news in order to ensure an opinion diversification, we wonder to what extent the recommendation strategy should consider or not the topic of interest. Given that the topics discussed in the news do not all have the same polarizing effect – they present higher or lower levels of controversy [29], which impacts differently on polarizing behavior [10] –, should this controversy be considered to define the diversification strategy?

**TH2: There is a crucial need to define new diversity measures and personalized recommendation strategies that consider the users and the topics.**

Third, going into details of the news consumption along the weeks, even though the initial change in news consumption (first weeks) is an accurate indicator of the global change, the news consumption of some users significantly fluctuates, despite the constant diversity of the recommendations. Few elements have been highlighted in the experiments conducted, but some questions are still pending. What factors impact the actual news consumption from one week to another? What does a change in the diversity of news consumption between two or several weeks mean? Are there time-aware user behavior classes? Although these questions are naturally raised, methodologies for a temporal analysis of the news consumption are cruelly lacking in the literature to thoroughly understand the reasons for the evolution of users (**RQ4**). This is even more important as these answers should be used to provide personalized diverse recommendations, with the goal to control the impact on users news consumption.

**TH3: The literature is lacking of well-established methodologies to model the diversity trajectory of users.**

## ACKNOWLEDGMENTS

This work has been funded by the BOOM ANR Project - ANR-20-CE23-0024.

## REFERENCES

[1] Panagiotis Adamopoulos and Alexander Tuzhilin. 2014. On Over-Specialization and Concentration Bias of Recommendations: Probabilistic Neighborhood Selection in Collaborative Filtering Systems. In *8th ACM Conference on Recommender Systems (RecSys'14) (RecSys '14)*. ACM, Foster City, Silicon Valley, California, USA, 153–160.

[2] Panagiotis Adamopoulos and Alexander Tuzhilin. 2014. On Unexpectedness in Recommender Systems: Or How to Better Expect the Unexpected. *ACM Trans. Intell. Syst. Technol.* 5, 4 (dec 2014), 32 pages.

[3] Gediminas Adomavicius and YoungOk Kwon. 2011. Maximizing Aggregate Recommendation Diversity: A Graph-Theoretic Approach. In *Workshop on Novelty and Diversity in Recommender Systems (DiveRS 2011), held in conjunction with ACM RecSys'11*. ACM, Chicago, USA, 3–10.

[4] Charu C. Aggarwal. 2016. *Recommender Systems - The Textbook*. Springer, Berlin, Germany. 1–498 pages.

[5] Hunt Allcott, Luca Braghieri, Sarah Eichmeyer, and Matthew Gentzkow. 2020. The Welfare Effects of Social Media. *American Economic Review* 110, 3 (2020), 629–676.

[6] Marcelo G. Armentano, Ingrid Christensen, and Silvia Schiaffino. 2015. Applying the Technology Acceptance Model to Evaluation of Recommender Systems. *Polibits* 51 (01 2015), 73–79. https://doi.org/10.17562/PB-51-10

[7] Christopher A. Bail, Lisa P. Argyle, Taylor W. Brown, John P. Bumpus, Haohan Chen, M. B. Fallin Hunzaker, Jaemin Lee, Marcus Mann, Friedolin Merhout, and Alexander Volfovsky. 2018. Exposure to opposing views on social media can increase political polarization. *Proceedings of the National Academy of Sciences* 115, 37 (2018), 9216–9221.

[8] Ian Baracskay, Donald J Baracskay III, Mehtab Iqbal, and Bart Piet Knijnenburg. 2022. The Diversity of Music Recommender Systems. In *Proc. of the International conference IUI 2022* (Helsinki, Finland) *(IUI '22 Companion)*. ACM, Helsinki, Finland, 97–100.

[9] Michael D. Bauer. 2021. *Diversity in News Recommender Systems*. Master's thesis. echnische Universität Wien.

[10] Fabian Baumann, Philipp Lorenz-Spreen, Igor M Sokolov, and Michele Starnini. 2021. Emergence of polarized ideological opinions in multidimensional topic spaces. *Physical Review X* 11, 1 (2021), 011012.

[11] Joeran Beel. 2017. It's Time to Consider 'Time' when Evaluating Recommender-System Algorithms. https://arxiv.org/ftp/arxiv/papers/1708/1708.08447.pdf.

[12] Abraham Bernstein, Claes de Vreese, Natali Helberger, Wolfgang Schulz, Katharina Zweig, Christian Baden, Michael A Beam, Marc P Hauer, Lucien Heitz, Pascal Jürgens, et al. 2020. Diversity in news recommendations. In *erspectives Workshop: Diversity, Fairness, and Data-Driven Personalization in (News) Recommender*. Dagstuhl Publishing, Wadern, Germany, 43–61.

[13] David M Blei, Andrew Y Ng, and Michael I. Jordan. 2003. Latent Dirichlet allocation. *J. Mach. Learn. Res.* 3, 4-5 (2003), 993–1022. https://doi.org/10.1016/b978-0-12-411519-4.00006-9

[14] Levi Boxell, Matthew Gentzkow, and Jesse M Shapiro. 2020. *Cross-Country Trends in Affective Polarization*. Working Paper 26669. National Bureau of Economic Research.

[15] Keith Bradley and Barry Smyth. 2001. Improving recommendation diversity. In *Proceedings of the twelfth Irish conference on artificial intelligence and cognitive science*, Vol. 85. Citeseer, NUIM Department of Computer Science, Maynooth, Ireland, 141–152.

[16] Sylvain Castagnos, Nicolas Jones, and Pearl Pu. 2010. Eye-Tracking Product Recommenders' Usage. In *Proceedings of RecSys'10*. ACM, Barcelona, 29–36.

[17] Davide Chicco and Giuseppe Jurman. 2020. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics* 21, 6 (2020), 13 pages.

[18] Davide Chicco, Matthijs J. Warrens, and Giuseppe Jurman. 2021. The coefficient of determination R-squared is more informative than SMAP, MAE, MAPE, MSE and RMSE in regression analysis evaluation. *Peerj computer science* 7 (2021), 24 pages.

[19] Uthsav Chitra and Christopher Musco. 2020. Analyzing the Impact of Filter Bubbles on Social Network Polarization. In *WSDM'20: Proceedings of the 13th International Conference on Web Search and Data Mining*. Houston, Texas, ACM, 115–123. https://doi.org/10.1145/3336191.3371825

[20] Matteo Cinelli, Gianmarco De Francisci Morales, Alessandro Galeazzi, Walter Quattrociocchi, and Michele Starnini. 2021. The echo chamber effect on social

media. *Proceedings of the National Academy of Sciences* 118, 9 (2021), e2023301118.
[21] Gabriel de Souza Pereira Moreira, Felipe Ferreira, and Adilson Marques da Cunha. 2018. News session-based recommendations using deep neural networks. In *Proceedings of the 3rd Workshop on Deep Learning for Recommender Systems*. ACM, Vancouver, Canada, 15–23.
[22] James N. Druckman, Samara Klar, Yanna Krupnikov, Matthew Levendusky, and John Barry Barry. 2021. Affective polarization, local contexts and public opinion in America. *Nature Human Behaviour* 5 (2021), 28–38.
[23] Yu Du, Sylvie Ranwez, Nicolas Sutton-Charani, and Vincent Ranwez. 2021. Is diversity optimization always suitable? Toward a better understanding of diversity within recommendation approaches. *Information Processing & Management* 58, 6 (2021), 102721.
[24] Tome Eftimov, Bibek Paudel, Gorjan Popovski, and Dragi Kocev. 2021. A Framework for Evaluating Personalized Ranking Systems by Fusing Different Evaluation Measures. *Big Data Research* 25 (2021), 100211.
[25] Michael D. Ekstrand, F. Maxwell Harper, Martijn C. Willemsen, and Joseph A. Konstan. 2014. User perception of differences in recommender algorithms. In *RecSys '14*. ACM, Foster City, USA, 161–168.
[26] Eli J. Finkel, Christopher A. Bail, Mina Cikara, Peter H. Ditto, Shanto Iyengar, Samara Klar, Lilliana Mason, Mary C. McGrath, Brendan Nyhan, David G. Rand, Linda J. Skitka, Joshua A. Tucker, Jay J. Van Bavel, Cynthia S. Wang, and James Druckman. 2020. Political sectarianism in America. *Science* 370, 6516 (2020), 533–536.
[27] Daniel Fleder and Kartik Hosanagar. 2008. Blockbuster culture's next rise or fall: The effect of recommender systems on sales diversity. *Management Science* 55, 5 (01 2008), 697–712.
[28] Florent Garcin, Kai Zhou, Boi Faltings, and Vincent Schickel. 2012. Personalized news recommendation based on collaborative filtering. In *Proc. - 2012 IEEE/WIC/ACM Int. Conf. Web Intell. WI 2012*. IEEE/ACM, Macau, China, 437–441.
[29] Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. 2018. Quantifying controversy on social media. *ACM Transactions on Social Computing* 1, 1 (2018), 1–27.
[30] Mouzhi Ge, Carla Delgado, and Dietmar Jannach. 2010. Beyond accuracy: Evaluating recommender systems by coverage and serendipity. In *RecSys'10 - Proceedings of the 4th ACM Conference on Recommender Systems*. ACM, Barcelona, Spain, 257–260. https://doi.org/10.1145/1864708.1864761
[31] Fausto Giunchiglia, Jahna Otterbacher, Styliani Kleanthous, Khuyagbaatar Batsuren, Veronika Bogina, Tsvi Kuflik, and Avital Shulner Tal. 2021. Towards Algorithmic Transparency: A Diversity Perspective. *ArXiv* abs/2104.05658 (2021), 11 pages.
[32] Jon Atle Gulla, Lemei Zhang, Peng Liu, Özlem Özgöbek, and Xiaomeng Su. 2017. The adressa dataset for news recommendation. In *Proceedings of the international conference on web intelligence*. ACM, Leipzig Germany, 1042–1048.
[33] Lucien Heitz, Juliane A Lischka, Alena Birrer, Bibek Paudel, Suzanne Tolmeijer, Laura Laugwitz, and Abraham Bernstein. 2022. Benefits of Diverse News Recommendations for Democracy: A User Study. *Digital Journalism* DOI: 10.1080/21670811.2021.2021804 (2022), 1–21.
[34] Natali Helberger. 2019. On the democratic role of news recommenders. *Digital Journalism* 7, 8 (2019), 993–1012.
[35] Rong Hu and Pearl Pu. 2011. Helping Users Perceive Recommendation Diversity. In *Workshop on Novelty and Diversity in Recommender Systems (DiveRS 2011), held in conjunction with RecSys'11*. ACM, Chicago, USA, 43–50.
[36] F.O. Isinkaye, Y.O. Folajimi, and B.A. Ojokoh. 2015. Recommendation systems: Principles, methods and evaluation. *Egyptian Informatics Journal* 16, 3 (2015), 261–273.
[37] Youchun Ji, Wenxing Hong, Yali Shangguan, Huan Wang, and Jing Ma. 2016. Regularized singular value decomposition in news recommendation system. In *2016 11th International Conference on Computer Science & Education (ICCSE)*. IEEE, Cambridge, UK, 621–626.
[38] Nicolas Jones. 2010. *User Perceived Qualities and Acceptance of Recommender Systems*. PhD Thesis. Ecole Polytechnique Fédérale De Lausanne.
[39] Glen Joris, Camiel Colruyt, Judith Vermeulen, Stefaan Vercoutere, Frederik Grove, Kristin Van Damme, Orphee de clercq, Cynthia Van Hee, Lieven Marez, Véronique Hoste, Eva Lievens, Toon De Pessemier, and Luc Martens. 2020. *News Diversity and Recommendation Systems: Setting the Interdisciplinary Scene*. Springer, Windisch, Switzerland, 90–105.
[40] Marius Kaminskas and Derek Bridge. 2016. Diversity, serendipity, novelty, and coverage: a survey and empirical analysis of beyond-accuracy objectives in recommender systems. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 7, 1 (2016), 1–42.
[41] Mahmut Karakaya and Tevfik Aytekin. 2018. Effective methods for increasing aggregate diversity in recommender systems. *Knowledge and Information Systems* 56 (08 2018). https://doi.org/10.1007/s10115-017-1135-0
[42] Mozhgan Karimi, Dietmar Jannach, and Michael Jugovac. 2018. News recommender systems–Survey and roads ahead. *Information Processing & Management* 54, 6 (2018), 1203–1227.
[43] Benjamin Kille, Frank Hopfgartner, Torben Brodt, and Tobias Heintz. 2013. The plista dataset. In *Proceedings of the 2013 international news recommender systems workshop and challenge*. ACM, Kowloon Hong Kong, 16–23.
[44] Brent Kitchens, Steven L. Johnson, and Peter Gray. 2020. Understanding Echo Chambers and Filter Bubbles: The Impact of Social Media on Diversification and Partisan Shifts in News Consumption. *MIS Quarterly* 44, 4 (2020), 1619–1650.
[45] Michal Kompan and Mária Bieliková. 2010. Content-based news recommendation. In *International conference on electronic commerce and web technologies*. Springer, Bilbao, Spain, 61–72.
[46] Emily Kubin and Christian von Sikorski. 2021. The role of (social) media in political polarization: a systematic review. *Annals of the International Communication Association* 45, 3 (2021), 188–206.
[47] Matevž Kunaver and Tomaž Požrl. 2017. Diversity in recommender systems–A survey. *Knowledge-based systems* 123 (2017), 154–162.
[48] Neal Lathia, Stephen Hailes, Licia Capra, and Xavier Amatriain. 2010. Temporal diversity in recommender systems. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*. ACM, Geneva, Switzerland, 210–217.
[49] Amaury L'Huillier, Sylvain Castagnos, and Anne Boyer. 2014. Understanding Usages by Modeling Diversity over Time. In *In extended proceedings of the 22nd Conference on User Modelling, Adaptation and Personalization (UMAP 2014)*. ACM, Aalborg, Denmark, 6 pages.
[50] Lei Li, Li Zheng, Fan Yang, and Tao Li. 2014. Modeling and broadening temporal user interest in personalized news recommendation. *Expert Systems with Applications* 41, 7 (2014), 3168–3177.
[51] Danyang Liu, Jianxun Lian, Shiyin Wang, Ying Qiao, Jiun-Hung Chen, Guangzhong Sun, and Xing Xie. 2020. KRED: Knowledge-aware document representation for news recommendations. In *14th ACM Conference on Recommender Systems (RecSys'20)*. ACM, Online, 200–209.
[52] Andreas Lommatzsch, Benjamin Kille, and Sahin Albayrak. 2017. Incorporating context and trends in news recommender systems. In *Proceedings of the international conference on web intelligence*. IEEE/ACM, Leipzig, Germany, 1062–1068.
[53] Gabriel Machado Lunardi, Guilherme Medeiros Machado, Vinicius Maran, and José Palazzo M. de Oliveira. 2020. A metric for Filter Bubble measurement in recommender algorithms considering the news domain. *Applied Soft Computing* 97 (2020), 106771.
[54] Tapio Luostarinen and Oskar Kohonen. 2013. Using Topic Models in Content-Based News Recommender Systems. *Proc. 19th Nord. Conf. Comput. Linguist. (NODALIDA 2013)* 2, 1 (2013), 239–251. https://aclanthology.org/W13-5622.pdf
[55] Andrii Maksai, Florent Garcin, and Boi Faltings. 2015. Predicting online performance of news recommender systems through richer evaluation metrics. In *Proceedings of the 9th ACM Conference on Recommender Systems*. ACM, Vienna, Austria, 179–186.
[56] Maciej A. Mazurowski. 2013. Estimating confidence of individual rating predictions in collaborative filtering recommender systems. *Expert Systems with Applications* 40, 10 (2013), 3847–3857.
[57] Sean M McNee, John Riedl, and Joseph A Konstan. 2006. Being accurate is not enough: how accuracy metrics have hurt recommender systems. In *CHI'06 extended abstracts on Human factors in computing systems*. ACM, Montreal, Canada, 1097–1101.
[58] Judith Möller, Damian Trilling, Natali Helberger, and Bram van Es. 2018. Do not blame it on the algorithm: an empirical assessment of multiple recommender systems and their impact on content diversity. *Information, Communication & Society* 21, 7 (2018), 959–977.
[59] John O'Donovan and Barry Smyth. 2005. Trust in recommender systems. In *International Conference on Intelligent User Interfaces, Proceedings IUI*. ACM, San Diego, California, 167–174. https://doi.org/10.1145/1040830.1040870
[60] Javier Parapar and Filip Radlinski. 2021. Diverse User Preference Elicitation with Multi-Armed Bandits. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*. ACM, Stanford CA, USA, 130—-138.
[61] Javier Parapar and Filip Radlinski. 2021. Towards Unified Metrics for Accuracy and Diversity for Recommender Systems. In *15th ACM Conference on Recommender Systems (RecSys'21)*. Association for Computing Machinery, New York, NY, USA, 75–84.
[62] Eli Pariser. 2011. *The filter bubble: what the Internet is hiding from you*. Penguin Press, New York.
[63] Denis Parra and Shaghayegh Sahebi. 2013. *Recommender Systems: Sources of Knowledge and Evaluation Metrics*. Springer Berlin Heidelberg, Berlin, Heidelberg, 149–175.
[64] Tao Qi, Fangzhao Wu, Chuhan Wu, and Yongfeng Huang. 2021. Personalized news recommendation with knowledge-aware interactive matching. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, Online, 61–70.
[65] Faisal Rahutomo, Teruaki Kitasuka, and Masayoshi Aritsugi. 2012. Semantic cosine similarity. In *The 7th International Student Conference on Advanced Science and Technology ICAST*. The University of Seoul, Seoul, South Korea, 1.
[66] Shaina Raza and Chen Ding. 2019. News recommender system considering temporal dynamics and news taxonomy. In *2019 IEEE International Conference on Big Data (Big Data)*. IEEE, Los Angeles, CA, USA, 920–929.

[67] Shaina Raza and Chen Ding. 2022. News recommender system: a review of recent progress, challenges, and opportunities. *Artificial Intelligence Review* 55, 1 (2022), 1–52.

[68] Paul Resnick and Hal R Varian. 1997. Recommender systems. *Commun. ACM* 40, 3 (1997), 56–58.

[69] F. Ricci, L. Rokach, B. Shapira, and P. Kantor. 2011. *Recommender systems handbook.* Springer, Berlin, Germany.

[70] Javier Sanz-Cruzado and Pablo Castells. 2018. Enhancing Structural Diversity in Social Networks by Recommending Weak Ties. In *Proceedings of the 12th ACM Conference on Recommender Systems (RecSys '18).* ACM, Vancouver, British Columbia, Canada, 233–241.

[71] Teresa Scheidt and Joeran Beel. 2021. Time-dependent Evaluation of Recommender Systems. In *Perspectives on the Evaluation of Recommender Systems Workshop (PERSPECTIVES 2021), co-located with the 15th ACM Conference on Recommender System (RecSys'21)* (Amsterdam, The Netherlands). ACM, Amsterdam, Netherlands, 9 pages.

[72] Laura Schelenz. 2021. *Diversity-Aware Recommendations for Social Justice? Exploring User Diversity and Fairness in Recommender Systems.* Association for Computing Machinery, New York, NY, USA, 404–410.

[73] Thiago Silveira, Min Zhang, Xiao Lin, Yiqun Liu, and Shaoping Ma. 2019. How good your recommender system is? A survey on evaluations in recommendation. *International Journal of Machine Learning and Cybernetics* 10 (2019), 813–831.

[74] Barry Smyth and Paul McClave. 2001. Similarity vs. Diversity. In *Case-Based Reasoning Research and Development*, David W. Aha and Ian Watson (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 347–361.

[75] Cass R. Sunstein. 2009. *Going to extremes: how like minds unite and divide.* Oxford University Press, Oxford, New York.

[76] Nava Tintarev, Emily Sullivan, Dror Guldin, Sihang Qiu, and Daan Odjik. 2018. Same, Same, but Different: Algorithmic Diversification of Viewpoints in News. In *Adjunct Publication of the 26th Conference on User Modeling, Adaptation and Personalization* (Singapore Singapore). ACM, Singapore, 7–13.

[77] Jay J. Van Bavel, Steve Rathje, Elizabeth Harris, Claire Robertson, and Anni Sternisko. 2021. How social media shapes polarization. *Trends in Cognitive Sciences* 25, 11 (2021), 913–916.

[78] Saul Vargas and Pablo Castells. 2011. Rank and relevance in novelty and diversity metrics for recommender systems. In *RecSys'11.* ACM, Chicago, USA, 109–116.

[79] Sanne Vrijenhoek, Mesut Kaya, Nadia Metoui, Judith Möller, Daan Odijk, and Natali Helberger. 2021. Recommenders with a mission: assessing diversity in news recommendations. In *Proceedings of the 2021 Conference on Human Information Interaction and Retrieval.* ACM, Online, 173–183.

[80] Jacek Wasilewski and Neil J. Hurley. 2016. Incorporating Diversity in a Learning to Rank Recommender System. In *FLAIRS Conference.* AAAI Press, Key Largo, Florida, 572–577.

[81] Chuhan Wu, Fangzhao Wu, Tao Qi, and Yongfeng Huang. 2022. Feedrec: News feed recommendation with various user feedbacks. In *Proc. of the International Conference WWW '22.* ACM, Lyon, France, 9 pages.

[82] Chuhan Wu, Fangzhao Wu, Yang Yu, Tao Qi, Yongfeng Huang, and Qi Liu. 2021. NewsBERT: Distilling pre-trained language model for intelligent news application. *Findings of the Association for Computational Linguistics* EMNLP 2021 (2021), 3285–3295.

[83] Fangzhao Wu, Ying Qiao, Jiun-Hung Chen, Chuhan Wu, Tao Qi, Jianxun Lian, Danyang Liu, Xing Xie, Jianfeng Gao, Winnie Wu, et al. 2020. Mind: A large-scale dataset for news recommendation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics.* ACL, Online, 3597–3606.

[84] Qiong Wu, Yong Liu, Chunyan Miao, Yin Zhao, Lu Guan, and Haihong Tang. 2019. Recent Advances in Diversified Recommendation. *ArXiv* abs/1905.06589 (2019), 7 pages.

[85] Wen Wu and Yu Chen, Liand Zhao. 2018. Personalizing recommendation diversity based on user personality. *User Modeling and User-Adapted Interaction* 28 (2018), 237—-276.

[86] Liang Xiang, Quan Yuan, Shiwan Zhao, Li Chen, Xiatian Zhang, Qing Yang, and Jimeng Sun. 2010. Temporal recommendation on graphs via long-and short-term preference fusion. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining.* ACM, Washington, DC, USA, 723–732.

[87] Liang Xiong, Xi Chen, Tzu-Kuo Huang, Jeff Schneider, and Jaime G Carbonell. 2010. Temporal collaborative filtering with bayesian probabilistic tensor factorization. In *Proceedings of the 2010 SIAM international conference on data mining.* SIAM, Columbus, Ohio, USA, 211–222.

[88] Rui Ye, Y. Hou, Te Lei, Yunxing Zhang, Qing Zhang, Jiale Guo, Huaiwen Wu, and Hengliang Luo. 2021. Dynamic Graph Construction for Improving Diversity of Recommendation. In *Fifteenth ACM Conference on Recommender Systems.* ACM, Amsterdam, Netherlands, 651—-655.

[89] Liang Zhang. 2013. The Definition of Novelty in Recommendation System. *Journal of Engineering Science and Technology Review* 6, 3 (2013), 141–145.

[90] M. Zhang and N. Hurley. 2008. Avoiding monotony: Improving the diversity of recommendation lists. In *Proceedings of 2nd ACM International Conference on Recommender Systems.* ACM, Lausanne, Switzerland, 123—-130.

[91] Tao Zhou, Zoltán Kuscsik, Jian-Guo Liu, Matúš Medo, Joseph Rushton Wakeling, and Yi-Cheng Zhang. 2010. Solving the apparent diversity-accuracy dilemma of recommender systems. *Proceedings of the National Academy of Sciences* 107, 10 (2010), 4511–4515.

[92] Cai-Nicolas Ziegler, Sean M McNee, Joseph A Konstan, and Georg Lausen. 2005. Improving recommendation lists through topic diversification. In *Proceedings of the 14th international conference on World Wide Web.* ACM, Chiba, Japan, 22–32.