Analysing the Effect of Recommendation Algorithms on the Spread of Misinformation

Miriam Fernandez
miriam.fernandez@open.ac.uk
The Open University, Knowledge Media Institute
United Kingdom

Alejandro Bellogín
alejandro.bellogin@uam.es
Universidad Autónoma de Madrid, Ingeniería Informática
Spain

Iván Cantador
ivan.cantador@uam.es
Universidad Autónoma de Madrid, Ingeniería Informática
Spain

ABSTRACT

Recommendation algorithms (RAs) have been pointed out as one of the major culprits of misinformation spreading in the digital sphere.[1] However, it is still unclear how these algorithms propagate misinformation, e.g., which particular recommendation approaches are more prone to suggest misinforming items, or which internal parameters of the algorithms could be influencing more on their misinformation propagation capacity. Motivated by this fact, in this work, we present an analysis of the effect of some of the most popular recommendation algorithms on the spread of misinformation on Twitter (X). A set of guidelines on how to adapt these algorithms is provided based on such analysis and a comprehensive review of the research literature.

KEYWORDS

recommender systems, misinformation, social networks

1 INTRODUCTION

Online misinformation[2] is a high-dimensional socio-technical problem with multiple influencing factors, including: (i) the ways in which **information** is constructed and presented [18, 71], (ii) the **users**' personality, values, emotions and susceptibility [37, 67] as well as the presence of bots and malicious accounts [23, 61], (iii) the architectural characteristics of the **digital platforms** where such information is propagated (i.e., the structure of the social networks, constraints on the type of messages and sharing permissions, etc.) [4], and (iv) the **algorithms** that power the recommendation of information within those platforms [22].

Recommendation algorithms (RAs) have been criticised for providing recommendations without taking into account their potential negative consequences or ethical implications [19]. These criticisms include filtering the information observed by users, who may be placed into filter bubbles where the only content they access is the type of content they like, generated by people with similar opinions [48, 65]. This comes as a consequence of the fact that RAs are part of the so-called *feedback loop*, i.e., systems that aim to reinforce a cycle that attempts to optimise user retention and interactions.

Additionally, these algorithms tend to rely on engagement signals for the recommendation of information (such as user preferences on topics, social connections between users, and the relatedness between the topics in social networks [75]), and are therefore affected by popularity and homogeneity biases [7, 34]. In this context, filter bubbles and biases may limit the exposure of users to diverse points of view [52] and reduce the quality of the information the users access [16], potentially making them vulnerable to misinformation. E.g., YouTube has been criticised for amplifying videos that are divisive and conspiratorial.[3]

Despite these criticisms, there is an important gap in the research literature when it comes to understanding the impact that RAs have on the spread of false and misleading information [22, 31]. Some works have studied the effect that RAs may have on the creation of filter bubbles [6, 8, 48], and others have created models to understand the effect that common popularity biases in RAs may have on the quality of items consumed by users [16]. However, a more in-depth investigation is needed to better understand which of these algorithms are more prone or susceptible of spreading misinformation, under which circumstances, and how the internal functioning of such algorithms could be modified or adapted to counter their misinformation recommendation behaviour. This is a very complex issue, and previous attempts have resulted in harmful effects. E.g., Twitter (X) modified its RA to recommend popular tweets into the feeds of people who did not subscribe to the accounts that posted those tweets. This change, which provides popular opposing views, was heavily criticised for amplifying inflammatory political rhetoric and misinformation.[4]

Misinformation is thus a problem with a high number of dimensions that interrelate to one another [72, 77], affecting what RAs learn and how they behave. For this reason, adapting RAs to counter their misinformation spreading behaviour requires an in-depth understanding not only of the internal mechanisms of such algorithms, but also of the data they manipulate, the users they serve, and the platforms they operate in.

Starting from this position, this paper investigates the effect of some of the most popular RAs on the spread of misinformation. A set of guidelines on how to adapt these algorithms is provided

[1] Please note that amplifying, spreading and propagating are indistinctly used in this paper to refer to the amplifying effect that recommendation algorithms may have on misinformation when recommending or suggesting items to users.

[2] In this paper we use the term *misinformation* to refer to misleading information, hoaxes, conspiracy theories, hyper-partisan content, click-bate headlines, pseudo-science and false news.

[3] Fiction is outperforming reality': how YouTube's algorithm distorts truth https://www.theguardian.com/technology/2018/feb/02/how-youtubes-algorithm-distorts-truth

[4] How Twitter's algorithm is amplifying extreme political rhetoric https://edition.cnn.com/2019/03/22/tech/twitter-algorithm-political-rhetoric/index.html

Websci '24, May 21–24, 2024, Stuttgart, Germany

ACM ISBN 978-1-4503-XXXX-X/18/06
https://doi.org/10.1145/3614419.3644003

based on the performed analysis and a deep review of the research literature. In our investigation, a dataset is created and released to the scientific community to stimulate discussions on the future design and development of RAs to counter misinformation.[5]

Our contributions can be summarised as:

- An analysis of previous work studying (i) the dimensions of the misinformation ecosystem that may affect the performance, results, and biases of RAs, (ii) the role of RAs on the amplification of misinformation, and (iii) effective strategies to counter misinformation and correct misperceptions.
- The creation of a dataset containing Twitter (X) user profiles, items, ratings, and misinformation labels that enables studying the effect of RAs on the amplification of misinformation.
- An analysis of different state-of-the-art RAs frequently used in industry and academia (including collaborative filtering techniques like nearest neighbours and matrix factorisation) on the amplification of misinformation, by means of three evaluation metrics proposed to account for such amplification: misinformation count, ratio difference, and Gini.
- A set of guidelines on how to modify and adapt RAs based on the conducted analysis as well as on the review of the research literature.

We note that the focus of this work is on recommendation algorithms and not on recommender systems. The latter include other aspects aside from the algorithm (e.g., user interface) that are not considered in our study. The remainder of the paper is structured as follows. Section 2 discusses related work on the field. Section 3 describes the dataset that we have generated for experimentation and that we are making available to the research community. Section 4 presents the different RAs that have been assessed, as well as the metrics and methods used to assess them. Sections 5 and 6 present our results and recommendations on how to adapt RAs to palliate the misinformation amplification effect. Discussions and conclusions are presented in Section 7.

2 RELATED WORK

Misinformation is a multifaceted (human, sociological, and technological) problem, and it has been the focus of investigation in several research fields including social sciences, journalism, computer science, psychology, and education. In this section, we aim to provide a summary of the literature based on four dimensions of the misinformation problem related with the design and building of RAs: *content*, *users*, *platform characteristics*, and *algorithms*. We complement this analysis with a summary of some of the strategies that have been found effective in correcting misperceptions.

Although our contributions are more in line with those works that have attempted to understand the impact of algorithms on the spread of misinformation, a multidimensional review of the literature considering other dimensions like content, users, and platforms is needed for the design of comprehensive RAs adaptation guidelines (see Section Adaptation Guidelines). We note that ours does not aim to be an exhaustive literature review on misinformation. For a comprehensive overview of the problem, the reader is referred to the following recent survey [2].

2.1 Misinformation Dimensions

2.1.1 Content. Content is an important factor of the misinformation problem, and also a key aspect to consider in the design, evaluation, and adaptation of RAs. Items to be recommended can be present in various *forms* (as news articles, research papers, blog entries, and social media posts) and discuss a wide range of *topics*, such as health and elections, to name a few. These items are not only textual, but sometimes include information in different *formats*, like images or videos. Moreover, combinations of these formats are frequently used to propagate misinformation (e.g., a news title linked with an image from a different place, or from a different time). The *framing* of misinforming items also varies between false news, rumours, conspiracy theories, and misleading content [71]. Other relevant elements about content are their *emotional tone*, their *origin* (news outlets, social contacts, public figures, etc.) as well as the *time* when they are posted. Note that recency is particularly relevant to the recommendation of news items.

Works in the literature have attempted to understand the characteristic of such content, and to develop algorithms that could automatically detect it. E.g., [12, 13, 28] studied information credibility on Twitter mainly based on content features, and created supervised machine learning classifiers to detect credibility. Their studies concluded that credible tweets tend to include more URLs, and are longer than non-credible tweets. Question and exclamation marks tend to concentrate on non-credible tweets, frequently using first and third-person pronouns. [55] also studied content features for misinformation detection. They concluded that lexical and Part of Speech (POS) patterns are key for correctly identifying rumours. Hashtags can result in high precision but lead to low recall. [29] showed how messages in news with negative sentiment became more viral. In terms of topics, [67] showed how false political news have a more pronounce cascading effect than false news about terrorism, natural disasters, science, or financial information.

Initiatives by the journalism research field have also attempted to identify key features of misinforming content. Credibility Coalition published an article in 2018 [76] listing credibility indicators for news articles. These include content indicators, such as the use of clickbait titles, or the use of emotionally charged tone, and context indicators, such as the representation of sources cited in the article.

These studies have derived on the creation of tools that attempt to identify misinforming content automatically. Examples include: TweetCred,[6] ClaimBuster,[7] or the Global Disinformation Index.[8] These tools rely on the characterisation of misleading content, and on manually compiled lists of misleading articles and websites. Media literacy projects[9] and games (e.g., Fakey,[10] Go Viral[11]) have also emerged to teach users how to identify misleading content.

2.1.2 Users. Users are a key dimension of the misinformation problem, and a core aspect of the functioning of RAs. Researchers have

[5] https://github.com/abellogin/RecSysMisInfo

[6] https://chrome.google.com/webstore/detail/tweetcred/fbokljinlogeihdnkikeeneiankdgikg?hl=en

[7] https://idir.uta.edu/claimbuster/

[8] https://www.disinformationindex.org/

[9] https://digital-strategy.ec.europa.eu/en/news/winners-european-media-literacy-awards

[10] https://fakey.iuni.iu.edu/

[11] https://www.cam.ac.uk/stories/goviral

therefore studied the effect of different *motivations* [15], *personalities* [79], *values* [54], *emotions* [67], and *knowledge and literacy* [47], on the acceptance misinformation, as well as the *susceptibility* of users to spread misinformation [58] and to interact with malicious actors [69]. For example, extroverts and individuals with high cooperativeness and high reward dependence are found more prone to share misinformation [15]. Psychology also shows that individuals with higher anxiety levels are more likely to spread misinformation [33]. The attention that users pay also plays an important role. Information overload and limited attention seem to contribute to the spread of misinformation [56].

Malicious actors, such as bots and sock puppets accounts, also appear within the information ecosystem. Social bots play a disproportionate role in spreading articles from low credibility sources. They also target users with many followers through replies and mentions, manipulating them to reshare misinformation [23, 61]. Users are also sometimes hired to support and propagate arguments or claims simulating grassroots social movements [41]. This phenomenon is known as crowdturfing. A prominent example is the campaign uncovered by The Washington Post, which was enlisting teens to spam a pro-Trump agenda for the 2020 US elections.[12]

As with content detectors, tools exist to identify malicious actors within social networks. Examples include Botometer,[13] which checks the activity of a Twitter account, and gives a score indicating how likely it is for the account to be a bot, and BotSlayer,[14] which supports stakeholders to discover coordinated campaigns in Twitter. Other tools like misinfo.me [44] encourage users to self-reflect by providing them with an assessment of how they have been interacting with misinforming content and accounts.

2.1.3 *Platforms and Social Networks.*

Platforms are designed differently and, therefore, facilitate the spread of misinformation in different ways. *Content limitations* (e.g., maximum length for posts), *ability to share information* and select the subsets of users with whom such information is shared (*sharing permissions*), the ability to *vote* (e.g., Reddit) or to *express emotions* for content (e.g. Facebook), are important aspects of platform design that may shape the content, the way information spreads, and the *social network structure*. The typology and topology of the social network are indeed key factors of misinformation dynamics [21] and also important in the design of RAs. Note that RAs take into consideration not only similarity between items but also between users, as well as social connections, when generating recommendations.

Multiple works have focused on understanding how misinformation flows across different social networks. [61] analysed the spread of 400K articles on Twitter for ten months in 2016 and 2017. They concluded that low-credibility sources spread through original posts and reposting, while few are shared in replies (i.e., the spreading patterns of low-credibility content are less conversational). They also observed that some accounts in the network acted as "super-spreaders" posting a low credibility article hundreds or even thousands of times, suggesting that the spread is amplified through automated means. [67] analysed a dataset of rumour cascades on Twitter and concluded that false news diffused significantly farther, faster, deeper and more broadly than the true ones. When looking at the structure of user connections, [66] focused on recommending friends from outside the echo chamber and tested their approach on Twitter data. Their approach showed an increase in the diversity and novelty of recommendations.

Automatic tools like Hoaxy,[15] and the Fact-checking Observatory[16] have also emerged to help monitoring the spread of misinformation, particularly on Twitter (X).

2.1.4 *Algorithms.*

The full details of the algorithms that social networking sites have developed to personalise and recommend information to users are not known to the public. Their primary goal is, however, to increase user engagement and time spent on the platform, as a way of maximising revenue from ads shown. Economic interest behind advertising ecosystems, and their effect on misinformation, have been at the core of recent studies [63].

Critics of the RAs behind social networks [52] have emphasised that users do not decide what they see, but are exposed only to the information that those algorithms select for them, introducing users in so-called filter bubbles. Since RAs are designed to provide us with information that we like, based on our past interactions, and on people who are similar to us, we risk ending up in bubbles where we only receive information that is pleasant, familiar and confirms our beliefs. We may not see the diverse set of opinions and information potentially available in the network. Additionally, since past interests determine what we are exposed to in the future, this may be leaving less room for the unexpected encounters that spark creativity, innovation, and the democratic exchange of ideas. Furthermore, users may not even be aware of this information filtering process. A 2015 study conducted with 40 Facebook users indicated that 62% of those users were entirely unaware of any curation, believing instead that every single story from their friends and followed pages appeared in their news feed [20].

RAs are known to suffer from popularity bias (i.e., the algorithm promotes information that is trending on the platform [7, 34]). In addition to the potential effect of item popularity, the information that users consume in social networks is also influenced by two other types of biases: (i) social biases (information that users are exposed to mainly comes from friends or accounts they follow) and (ii) cognitive biases, particularly confirmation bias (users are more likely to consume information that agrees with their own beliefs).

Works have reported empirical studies and network simulations to understand whether filter bubbles do indeed exist in social media and whether these are the effect of RAs. In a 2015 study, Nikolov and colleagues [48] confirmed the presence of social bubbles on Twitter. They showed that collectively, people access information from a significantly narrower spectrum of sources through Twitter compared to a search baseline. A similar study [6] showed how Facebook's three filters (the social network, the algorithm, and a user's own content selection) decrease exposure to ideologically challenging news. The article concludes that the composition of the users' social network is the most important factor affecting the mix of content encountered on social media with individual choice also playing a large role. The news feed (i.e., the algorithm effect) has

[12] https://www.washingtonpost.com/politics/turning-point-teens-disinformation-trump/2020/09/15/c84091ae-f20a-11ea-b796-2dd09962649c_story.html

[13] https://botometer.osome.iu.edu/

[14] https://osome.iu.edu/tools/botslayer/

[15] https://hoaxy.iuni.iu.edu/

[16] https://fcobservatory.org/about/

a smaller impact on the diversity of information according to this study. A 2020 study confirms this effect claiming that, under the presence of homophily (i.e., users preferring interactions with individuals that are similar to them), echo chambers and fragmentation are system-immanent phenomena [8]. The effect of algorithmic popularity bias on the quality of information that users consume has also been investigated. Ciampaglia and colleagues [16] concluded that popularity bias hinders average quality when users are capable of exploring many items, as well as when they only consider very few top items due to scarce attention.

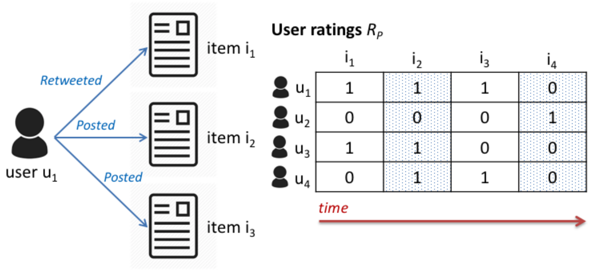
These works have aimed at understanding the effect that RAs may have on the creation of filter bubbles [6, 8, 48], as well as the effect that common popularity biases may have on the quality of items consumed by users [16]. The studies show how filter bubbles and popularity biases may make users more vulnerable to misinformation by reducing the diversity and quality of the information they are exposed to. *Our work aims to advance the state of the art by analysing how different RAs may influence the amplification of misinformation, and under which conditions. We do not account here for social or cognitive biases, just algorithmic effects. We hypothesise that, by better understanding these algorithms, and how they behave under the presence of misinformation, we can propose more informed adaptations to counteract the effect of false and misleading content.*

2.2 Strategies to Correct Misperceptions

Adaptations of RAs can be broadly focused on: (i) reducing the number of misinforming items they recommend and, (ii) adapting them to recommend information that could potentially help correct misperceptions. Understanding successful and unsuccessful strategies to counter misperceptions is therefore key to proposing more informed adaptations of RAs.

Presenting people with corrective information is likely to fail in changing their salient beliefs and opinions, or may, even, reinforce them [25, 50]. Human beings strive for internal psychological consistency. We tend to favour information that confirms and supports our previous beliefs and values (confirmation bias). Inconsistency, on the other hand, tends to become psychologically uncomfortable (cognitive dissonance) and we tend to reject it [24]. Recent works explored the development of RS to present users with corrective information based on their reading history [70], however, the effectiveness of the proposed recommendations on users has not been evaluated.

Nevertheless, some strategies have been found to be effective in correcting misperceptions [40], such as exposing users to related but disconfirming stories [9], or revealing the demographic similarity of the opposing group [25]. In the context of health misinformation, [68] argue that "observational correction" is an accurate strategy for changing misconceptions. i.e., those who witness a correction on social media, but are not directly engaged with the misinformation item, are less affected by cognitive dissonance, and thus more amenable to correction. Vraga and Bode also suggested: (i) citing highly credible factual information with links to expert sources, (ii) offering a coherent alternative explanation for the misinformation, (iii) using multiple corrections to reinforce the message, and (iv) trying to correct misinformation early, before misperceptions are entrenched. In the context of more polarised topics, such as political misinformation, it is however unclear whether corrections work, or even worsen the problem for users who are unwilling to revise their beliefs [35, 50]. We have considered these suggestions when proposing adaptations to RAs (see Section Adaptation Guidelines).



	i_1	i_2	i_3	i_4
u_1	1	1	1	0
u_2	0	0	0	1
u_3	1	1	0	0
u_4	0	1	1	0

Figure 1: User-ratings indicate which items the user has interacted with (i.e., posted or shared). A user profile contains 1 for item i_x if the user has interacted with the item and 0 otherwise. The grey colour in the matrix (items i_2 and i_4) indicate that such items are misinformation.

3 DATASET CREATION

Generating datasets that enable studying how different RAs amplify misinformation constitutes a significant challenge. These datasets should contain information about users, items, ratings (i.e., user-item interactions) and labels about which of those items are misinformation. An illustration of these components is presented in Figure 1. We consider explicit user interactions, such as posting or sharing data, as ratings.

Existing datasets in the literature either: (i) provide a set of labelled misinforming items (e.g., datasets generated by fact-checker organisations) – without providing information about users or user-item interactions (ratings) or, (ii) provide social media data collections (e.g., Twitter datasets, which contain information about users and items) but they either do not provide labels about which of those items are misinformation, or do not provide comprehensive information about user-item interactions (ratings). See the Media Futures project for examples of such datasets.[17]

In some cases, a handful of interactions per user are provided (e.g., the COAID[18]) or the Fakenewsnet [62][19] datasets. However, the timelines (user-interactions) of these users are not collected and analysed to investigate additional interactions of the same users with other misinforming items. A recent work (June 2023) [59] claimed the generation of a dataset containing all necessary components for the development of trust-based RS for fake news mitigation. However, to the best of our knowledge, this dataset has not been made publicly available. We also do not consider datasets with automatically generated labelled data [42, 43], but focus on those manually labelled, preferable by fact-checkers.

Our dataset creation process involved selecting Twitter (X) due to its API accessibility for collecting user, item, and rating data at the time of generating this dataset in November 2020. Note that in

[17] https://mediafutureseu.github.io/datasets.html
[18] https://github.com/cuilimeng/CoAID
[19] https://github.com/KaiDMML/FakeNewsNet

Algorithm 1: Ratio-based user profile generator

```
Function generate user u, ratio r
    neg ⟵ { i ∈ u: i is misinformative } ;
     // Negative claims
    neu ⟵ u \ neg ;            // Neutral claims
    desNeg ⟵ r · |u| ;         // Desired negative ratio
    desNeu ⟵ (1 - r) · |u| ;   // Desired neutral ratio
    while (desNeg > |neg|) OR (desNeu > |neu|) do
        if desNeg > |neg| ;    // Downsampling negative
         then
            desNeg ⟵ desNeg - 1;
        end
        if desNeu > |neu| ;    // Downsampling neutral
         then
            desNeu ⟵ desNeu - 1;
        end
        newTotal ⟵ desNeg + desNeu;
        desNeg ⟵ r · newTotal;
        desNeu ⟵ (1 - r) · newTotal;
    end
    userProfile(u) ⟵ sample(neg, desNeg) ∪ sample(neu, desNeu);
end
```

March 2023, Twitter (now named X) officially ended their free API access. Misinforming claims were gathered by merging datasets from the CoronaVirusFacts Alliance [20], Misinfo.me [44], Covid-19 two myths [42], and CMU-MisCov19 [43], considering tweet IDs, URLs, and text for matching. A list of 39,525 false claims was obtained, represented by sets of URLs, texts, and tweet IDs. Items were defined as pieces of information (news), and tweets sharing the same information were considered the same item. Users and their ratings were collected by identifying users who posted or retweeted these items, resulting in a dataset of 2,921 users, 1,014,004 items, and 1,116,658 interactions. Sparsity reduction measures included limiting items to English and selecting those with URLs. The final dataset was compiled to facilitate research on misinformation detection and contains comprehensive user-item interactions.

4 EVALUATION SETUP

In this section, we describe how we have exploited the dataset described in the previous section to simulate possible scenarios under different proportions of misinformative items in the system and/or shared by each user (Section User Profiles). Besides, we describe the evaluated recommendation algorithms (Section Recommendation Algorithms), and explain how we propose to account for the presence of misinformation in the recommendations (Section Evaluation Metrics).

4.1 User Profiles

The first step when collecting information to build the dataset was to identify users who originally shared (i.e., *tweeted*) claims that were explicitly labelled as misinformative items. It is important to highlight that due to the limited amount of tweets identified by fact-checkers, the generated dataset may not be fully representative of the misinformation status within the social network [26]. Note that all users have between 1 and 15 ratings associated with misinformative items. To have some control over the amount of information exploited by the algorithms, we included the following constraints. First, we imposed a ratio r that every user should satisfy regarding the amount of misinformative vs. neutral (either non-informative or unknown) items; for instance, $r = 0.5$ means that every user should have as many misinformative as neutral items. Second, we allowed sub-sampling to match the desired level of misinformation ratio, both in terms of misinformative or neutral items. An algorithm to achieve this is presented in Algorithm 1, where its main idea is to obtain the maximum number of either types of items that satisfy a given ratio.

Table 1: Statistics from the obtained datasets according to the methodology presented in Section 4.1. *Density* accounts for the number of cells with information in the user-item matrix, that is, $R/(U \cdot I)$, considering R the number of interactions and U and I the number of users and items.

Ratio	Users	Items	Interactions	Density (%)
∅	2,921	1,014,004	1,116,658	0.038
0.2	2,919	28,378	33,065	0.040
0.5	2,921	5,761	10,084	0.060
0.8	1,999	914	3,909	0.214

Table 2: Parameters of evaluated recommendation algorithms. Values in bold denote the *typical* parameterisation that will be referenced later. For MF, k denotes the number of factors, λ controls the overfitting, and n is the number of iterations. For UB and IB, k denotes the number of neighbours, *sim* is the similarity, and q is the exponent of similarity value.

Rec	Parameters
MF	$k = \{20, \mathbf{50}, 100\}$, $\lambda = \{\mathbf{0.1}, 0.01\}$, n=$\{\mathbf{20}, 100\}$
IB	$k = \{10, \mathbf{50}, 100\}$, sim = {jac, cos, **pearson**}, q=$\{\mathbf{1}, 2, 3\}$
UB	$k = \{10, \mathbf{50}, 100\}$, sim = {jac, cos, **pearson**}, q=$\{\mathbf{1}, 2, 3\}$

Finally, to control against the base scenario where no constraints are imposed, we considered the special value $r = \emptyset$ as the situation where no filter is applied, that is, all the users in the dataset are transformed into user profiles and considered for training the recommendation algorithms. In the experiments, we tested three values of ratio r that may fit a wide range of real-world situations available in actual social networks: a conservative $r = 0.2$ (all users share more neutral than misinformative items), an unbiased $r = 0.5$, and an extreme $r = 0.8$. Table 1 shows statistics about the generated datasets according to these ratios. We note the especially low density values, in particular compared against standard datasets in recommendation, whose density ranges between 4 and 6% [30]. We leave as future work to experiment with users of different ratios co-existing in the same simulation of the system.

4.2 Recommendation Algorithms

In this work we study how different families of algorithms may amplify misinformation under different initial constraints. For this,

[20] https://www.poynter.org/ifcn-covid-19-misinformation/

we focus on the most common algorithmic techniques to produce recommendations, namely collaborative filtering approaches [39].

These techniques have the main advantage that do not depend on user or item metadata or attributes, since they only require the user-item interactions to model the user preferences and, based on that, to produce suggestions. Because of this, they are widely used in several domains, ranging from movie or music recommendation to the travel domain [5, 60, 78]. However, they are well-known to suffer from popularity bias, or the *rich gets richer* effect [7, 34]. Therefore, they are good candidates to analyse if popularity bias translates into a potential misinformation spreading, or under which conditions this is more likely to occur. To properly understand this behaviour, we selected three classical methods that are widely used in both academia and industry:

- A **matrix factorisation algorithm (MF)** [32] that uses Alternate Least Squares in its minimisation formula. This method learns latent factors for users and items, and tries to reconstruct the original user-item interaction matrix by minimising the distance between the original and reconstructed matrices.
- A **nearest neighbour algorithm based on users (UB)**, which exploits similar-minded users from the community[21] to produce the recommendations [49]. We use a non-normalised version as follows since it has shown better ranking performance [3]:

$$s(u,i) = \sum_{v \in N(u;k)} s(v,i) w(u,v)^q \qquad (1)$$

  $N(u;k)$ denotes the k closest users (neighbours) in terms of similarity to user u, $w(u,v)$ is a similarity function, and q is a weight to emphasise the value of such similarity.
- A **nearest neighbour algorithm based on items (IB)**, which, similarly to UB, generates recommendations according to a neighbourhood, but in this case by exploiting similar items to those previously interacted by the user [49]. We also use a non-normalised version.

Besides, we also included a random (Rnd) and a most-popular (Pop) baselines that provide non-personalised recommendations while controlling biases in the data: whereas Rnd will produce completely unbiased suggestions, Pop will be guided purely by the items with more interactions.

Since there is no training-test split in our experiments, we cannot optimise any accuracy metric to select the hyperparameters. To overcome this limitation, we experimented with some typical parameterisations of these approaches, together with other variations, all of them presented in Table 2.

4.3 Evaluation Metrics

We propose the following three evaluation metrics that measure how many misinformative items are present in the recommendation lists provided by the RAs, either in an absolute way (count), compared against the user prior distribution (ratio difference), and from a global perspective (Gini):

- **Misinformation Count (MC)** measures how many of the recommended items to a user are misinformation. Its value is normalised by a cutoff, which corresponds to the ranking size, so that comparable measurements could be produced at different sizes. The higher the MC value, the more misinformation included in the recommendations; its range is in $[0, 1]$.
- **Misinformation Ratio Difference (MRD)** computes how much the ratio of misinformation has changed in a user basis with respect to what is observed in training. In particular, we calculate the ratio of misinformation in training for each user (let us call it m_t^u), and compare it with the observed ratio of misinformation in the recommendation list (m_r^u). This metric is the average of the differences $m_t^u - m_r^u$. Hence, the larger the MRD value in absolute terms, the higher the change with respect to training, whereas its sign indicates the direction of such change: a positive value one would show that the ratio is larger in training. The range of this metric is $[-1, 1]$.
- **Misinformation Gini (MG)** measures the dispersion over a distribution, as it is done to account for diversity in recommendation [11]. We compute the distribution over the misinformative items by considering the number of times each item was recommended, and add another item that represents the rest of the items in the collection, i.e., the neutral or informative ones. In this way, when the distribution is uniform (all the misinformative items have been recommended a similar number of times), MG would produce a higher value than when the above distribution is highly skewed. Note that, in contrast to the other two metrics, this one is not computed in a user basis, but for the entire set of recommendations. Its range is $[0, 1]$.

5 RESULTS

In this section, we report and discuss the results of the evaluation metrics presented in Section 4.3 on different scenarios where the ratio of misinformation has been configured as explained in Section 4.1. First, in Section 5.1, we explore the effect on misinformation spread of the most common instantiations of the analysed recommendation algorithms; then, in Section 5.2, we perform a sensitivity analysis on the effect of the parameters of such algorithms for misinformation spreading. Some limitations of these experiments are discussed in Section 7, after guidelines for adaptation are proposed in Section 6.

5.1 Misinformation Spread of RAs

Table 3 shows our proposed misinformation evaluation metrics (count, ratio difference, and Gini) when testing different configurations of the misinformation ratio to create the user profiles. For these experiments, we tested the RAs presented in Section 4.2 using their most typical configuration of parameters (as shown in Table 2), which have demonstrated to be very effective on different domains.

Our first observation is that, except when ratio is ∅ (that is, when no control on the amount of misinformation interacted by the users is imposed), the popularity-based Pop algorithm is the most effective method in spreading misinformation, both in terms of MC (where all the items presented to the user are labelled as misinformative) and

[21] We note that the term *neighbourhood* is used as in classical recommendation, to denote users selected according to the similarity, and it has no relation with the structure of the Twitter (X) network.

Table 3: Misinformation metrics for the typical configurations of CF RAs at different ratios of misinformation present in the user profiles. MC, MG, and MRD denote count, Gini, and ratio difference of misinformation, see Section Evaluation Metrics.

Ratio	Rec	MC@5	MC@10	MC@20	MRD@5	MRD@10	MRD@20	MG@5	MG@10	MG@20
∅	Rnd	0.001	0.001	0.001	0.031	0.031	0.031	0.000	0.000	0.000
∅	Pop	0.000	0.002	0.098	0.032	0.030	−0.065	0.000	0.000	0.000
∅	MF	0.053	0.040	0.032	−0.021	−0.007	0.000	0.001	0.001	0.001
∅	IB	0.018	0.013	0.008	0.014	0.020	0.024	0.000	0.000	0.000
∅	UB	0.068	0.054	0.044	−0.036	−0.022	−0.012	0.006	0.006	0.006
0.2	Rnd	0.027	0.026	0.026	0.109	0.110	0.110	0.009	0.012	0.016
0.2	Pop	1.000	1.000	1.000	−0.864	−0.864	−0.864	0.006	0.012	0.026
0.2	MF	0.995	0.984	0.919	−0.859	−0.848	−0.783	0.189	0.222	0.237
0.2	IB	0.091	0.063	0.047	0.049	0.077	0.093	0.005	0.004	0.005
0.2	UB	0.327	0.213	0.131	−0.188	−0.073	0.009	0.054	0.041	0.028
0.5	Rnd	0.133	0.131	0.131	0.367	0.369	0.369	0.088	0.098	0.106
0.5	Pop	1.000	1.000	1.000	−0.500	−0.500	−0.500	0.006	0.012	0.026
0.5	MF	1.000	0.998	0.970	−0.500	−0.498	−0.470	0.203	0.246	0.266
0.5	IB	0.132	0.112	0.100	0.368	0.387	0.396	0.012	0.017	0.027
0.5	UB	0.340	0.235	0.217	0.160	0.264	0.279	0.059	0.051	0.064
0.8	Rnd	0.759	0.757	0.755	0.226	0.228	0.230	0.599	0.631	0.650
0.8	Pop	1.000	1.000	1.000	−0.015	−0.015	−0.015	0.006	0.012	0.026
0.8	MF	1.000	0.995	0.969	−0.015	−0.010	0.016	0.221	0.265	0.300
0.8	IB	0.667	0.627	0.515	0.252	0.194	0.154	0.280	0.348	0.375
0.8	UB	0.897	0.766	0.586	0.022	0.054	0.082	0.286	0.334	0.356

MRD (where the method produces the most negative differences with respect to the misinformation ratio in training, meaning that it increases such ratio for all users consistently). The reason for this might be obvious: once we force all users to have at least 20% of their items to be misinformative, it is more likely that the most popular items in the system are, at the same time, misinformative.

Interestingly, these results evidence that our simulations with a positive misinformation ratio produce situations where a small number of misinformative items get popular very quickly. This is indeed quite realistic, as it often occurs in social media where fake news or other dubious pieces of information are spread rapidly. However, and according to our results, such spread can be slowed down with an appropriate use of recommendation algorithms, as we shall see next.

Besides the random Rnd recommender, which usually includes the lowest number of misinformative items in its suggestions due to its complete disregard of the interactions between users and items, we observe that the methods based on neighbours (UB and IB) spread less misinformation. We should note that the Rnd recommender is actually reflecting the distribution of the population, hence, in these cases, *most of the items are not misinformative*, which is true by design except when $r = 0.8$. The methods based on neighbours, which are expected to produce more relevant personalised recommendations than the Rnd recommender, are able to keep the spread of misinformation between 10 and 30%, as long as the original ratio of misinformation in the user profiles is not too high, that is, for $r = 0.2, 0.5$. This is in contrast with the Pop and the MF recommenders. The latter algorithm basically *follows* the Pop method, in the sense that in those configurations where the ratio is positive, it produces results very close to those obtained for Pop. This can be attributed to a strong popularity bias evidenced by this and other algorithms in the area, a well-studied problem by the community [34]: in general, good results are obtained when producing popular but slightly personal results for each user, even though the utility of such recommendations is very limited, and hence, a tradeoff between novel, diverse, and popular items is demanded by the users [11]. In this work, we can add another negative consequence of this behaviour: a larger presence of misinformative items in the recommendations.

Moreover, by analysing the misinformation Gini metric, we can better understand the differences between these two approaches. Recall that MG measures how uniform the distribution of misinformative items is, from a global perspective. Hence, since MF obtains higher values than Pop, these results show that Pop is always recommending the same (limited) set of the misinformative items. MF, in contrast, recommends a wider range of items, even though most of them turn out to also be misinformative. In this sense, we could infer that the spread of misinformation is different for these algorithms: Pop is very aggressive on suggesting the same items over and over again, as if it was a bot or a viral account in the system; on the other hand, MF distributes more evenly the misinformative items across the population, hence spreading out a larger number of distinct misinformative items.

Most of these observations dramatically change when there are no constraints on the misinformation ratio. In our Table 3, when ratio is ∅ we observe that the popularity-based recommender does not longer spread misinformative items in the same way. This is attributed to the items being less common among the entire population and, hence, not popular enough to be recommended. However, MF and, surprisingly, UB seem to be very effective in recommending a significantly large number of misinformative items (especially, if we compare against the random recommender) also in this situation. Our initial conclusion, hence, is that the MF algorithm, independently of the starting scenario, will increase the presence of misinformative items due to its recommendations, which is exacerbated in the long term if we consider recommendation algorithms as part of the feedback loop. Neighbour-based methods, and in particular, IB seem to be safer in this respect, since they tend to control the spread under some reasonable limits. In the next section, we continue our analysis exploring how sensitive the recommenders are to spreading misinformation when different values of their model parameters are used.

5.2 Effect of Recommendation Parameters

Table 4 shows a complementary analysis of the results presented before, but only for the metric MC@10, as it is the easiest to interpret. In this table, we aggregate the performance obtained by the RAs focusing on the amount of information exploited by each algorithm. This translates into the number of factors for MF and number of neighbours for UB and IB. A row where High appears in the Info column aggregates the values of all the recommenders of the same type whose number of factors or neighbours are above some predefined threshold. In particular, for this analysis and considering the parameters shown in Table 2, we consider 100 factors or neighbours as High, 50 factors or neighbours as Med, and 20 factors or 10 neighbours as Low.

We observe that under controlled conditions (i.e., a positive misinformation ratio), the number of factors or neighbours do not have a strong effect in changing the spread of misinformation for MF or IB. For UB, in contrast, a low number of neighbours drastically reduces the number of misinformative items being recommended. This observation might be linked to the previous discussion on popularity bias: as investigated in the Recommender Systems area [10], UB with large neighbourhoods tends to be closer to popularity, in this case, a lower value allows recommending less popular items which, in the controlled conditions, are more likely to not be misinformative (by design, as discussed in the previous section).

This effect, interestingly, is also observed when no constraints on misinformation ratios are imposed. Therefore, we conclude that a low number of neighbours in UB could be helpful in stopping the spread of misinformative items under all the conditions we have tested. Additionally, we also observe in Table 4 that a lower number of factors in MF limits the number of misinformative items recommended. This behaviour, however, is inconsistent in the rest of the simulated conditions, where a low number of factors ensures that almost all the recommended items will be misinformative.

Table 5 shows a summary of results similar to that of Table 4, but considering other parameters of the algorithms. In this case, for MF we analyse the number of iterations (High is used for 100, Low for 20) and for UB and IB the similarity weight q to determine how much each neighbour influences the final prediction.[22] From the results, similarly to the analysis shown in Table 4, we observe that neither MF nor IB seem to be affected by the above parameters in their abilities to increase the spread of misinformation. However, a large q in UB consistently reduces the amount of misinformative items recommended by this algorithm.

Table 4: Misinformation count measured at cutoff 10 aggregating the results according to the amount of information used by each algorithm: number of factors in MF and neighbours in IB and UB.

Rec	Info	∅	0.2	0.5	0.8
MF	High	0.077	0.907	0.953	0.959
MF	Med	0.062	0.988	0.999	0.995
MF	Low	0.015	0.997	1.000	1.000
IB	High	0.016	0.121	0.213	0.668
IB	Med	0.016	0.122	0.211	0.667
IB	Low	0.017	0.136	0.213	0.662
UB	High	0.057	0.242	0.272	0.772
UB	Med	0.055	0.213	0.235	0.766
UB	Low	0.038	0.048	0.072	0.469

Table 5: Misinformation count measured at cutoff 10 aggregating the results according to the additional information exploited by each algorithm: number of iterations in MF and similarity weight q in IB and UB.

Rec	Info	∅	0.2	0.5	0.8
MF	High	0.077	0.997	1.000	1.000
MF	Low	0.074	0.997	1.000	1.000
IB	q=1	0.016	0.136	0.213	0.668
IB	q=2	0.017	0.136	0.213	0.665
IB	q=3	0.017	0.136	0.213	0.664
UB	q=1	0.057	0.242	0.272	0.772
UB	q=2	0.042	0.098	0.123	0.764
UB	q=3	0.037	0.052	0.088	0.761

6 ADAPTATION GUIDELINES

In this section, we propose a series of adaptation guidelines for RAs based on our analysis, as well as on the review of the research literature. Inspired by the field of context-based RAs [1], we argue that RAs can be adapted at three different points: (i) pre recommendation, (ii) within the model, and (iii) post recommendation.

If we want to palliate the misinformation amplification effect of RAs, we need to reduce the popularity effect. This can be achieved by selecting RAs based on neighbours (like UB and IB), and by reducing the number of neighbours used within the models. The more neighbours the RA uses, the more it resembles popularity. Our recommendation for adaptation is consequently to reduce the number of neighbours. E.g., one could think about RAs that cluster the users' social contacts in different subgroups according to similarity (where the first cluster contains the most similar social contacts, and the last cluster the more dissimilar contacts). The algorithm could then recommend a ratio of items from each of those clusters. That will mean that (i) we will be reducing the popularity effect on one hand, since the neighbourhoods are smaller, and (ii) we could be introducing some diversity in the recommendations provided (since we could even recommend some items from the user's most dissimilar social contacts, taking into consideration not introducing a large cognitive dissonance).

Tools that allow for the detection of malicious actors (bots, sockpuppets) could also help us to adapt RAs by doing pre or post adaptations. E.g., one may use tools like Botometer to discard all user accounts that resemble bots, or that frequently spread misinformation, and do not take into account any of their content for recommendations. On the other hand, the adaptation could be done a posteriori, by re-ranking recommended items based on the "reliability" of the account from where those posts originated. This relates

[22] As noted in [3], a higher value of q will make smaller similarities drop to 0, while higher ones will be (relatively) emphasised.

to the notion of trust in RAs, although in a slightly different manner, since trust is normally considered among two users [51], while here we are referring to the reliability of accounts. The same could be thought about content. Content that ranks low on credibility could be either discarded (pre-adaptation) or re-ranked (post-adaptation). These elements could also be addressed at the model level, incorporating scores of reliability and credibility of users and content as part of the user and item profiles. This would enable more sensible adaptations, like diluting the weights of potentially misinforming users and items over cycles of recommendation.

Further adaptations could be done on the modelling of users and items by incorporating some of the elements discussed in Sections Content and Users. Studies have shown how personality, values, emotions, and vulnerability of users affect their likelihood of propagating misinformation. Considering these aspects when profiling users and items, could help RAs to be more selective on their recommendations. While obtaining this information about users is not trivial, and automatic methods are not always correct, multiple approaches have emerged in recent years capable of estimating users' personalities, values, and emotions based on previous social media interactions [14, 57]

When thinking about adaptation of RAs it is also important to consider those strategies that have been proven effective when correcting misperceptions (see Section Strategies to Correct Misperceptions). RAs could be adapted to promote corrective information without introducing a high-degree of cognitive dissonance (e.g., by providing corrections that are "observational" −over topics where the user is not emotionally invested) [68] or providing corrections from users (social connections) that are similar, revealing the similarities of the opposing group [25].

7 DISCUSSION AND CONCLUSIONS

The goal of the presented study has been to analyse the impact of RAs on the spread of misinformation in social networks, and particularly Twitter (X). This is a novel and exciting research area. However, one where multiple challenges emerge.

Particularly challenging is the generation of datasets. For our study, users have been selected based on the Coronavirus Facts Alliance Dataset. Although this dataset covers misinformation (identified by fact-checkers) from 74 countries, it is specific to COVID-19. Users who do not spread COVID-19 misinformation are not covered in our data. Users in our dataset have very low numbers of misinforming items in their timeline, which may not be entirely representative of a social network. One may expect to see some users spreading a lot of misinformation, while many may spread none [26]. The creation of synthetic datasets could be a potential solution for future work, even though real effects would need to be measured via users studies, which opens up new challenges in terms of privacy and ethics. Despite these limitations, our dataset provides the research community with a unique opportunity to investigate responsible recommendations in the context of social media misinformation.

Our study has focused on the analysis of RAs based on Collaborative Filtering techniques. Analysing content-based and Hybrid methods requires capturing RAs dealing with natural language and its inherent subtleties (negation, sarcasm, etc.). An in-depth algorithmic survey is therefore required to better understand the impact of these techniques in the recommendation of misinformation. This includes classical and hybrid collaborative algorithms [36] and more recent methods aimed at understanding the natural language by, for instance, using Neural Networks [17, 73]. A special problem that deserves further analysis is the situation when there are not enough user interactions, i.e., the so-called cold-start users. While some works have explored this for general recommendation systems [38], the impact of misinformation spread on these types of users remains not addressed, and whether the lack of user information from the recommendation perspective increases or decreases the amount of misinformation presented to those users.

Table 6: Evolution of MC@10 depending on which recommender is used to present items to users. For $t = 2$, we simulate that all users accept their top-3 recommendations, train the recommendation algorithms again, and measure the misinformation of the items returned by each method.

Rec. cycle	HKV	UB
t=0	0.200	0.200
t=1	0.984	0.213
t=2 (after UB)	0.658	0.355
t=2 (after HKV)	0.988	0.762

Our results show that it is possible to limit the inherent spread of misinformation derived from RAs by configuring these techniques. However, it should be emphasised that no tradeoff with respect to the potential loss (or gain) in accuracy derived by such changes was measured. The Recommender Systems community has shown that several beyond-accuracy dimensions compete between each other and against accuracy when designing the perfect user experience, and it is extremely difficult to find an algorithm that is optimal for more than one dimension at the same time [11, 27]. Nonetheless, in this work we wanted to focus strictly on the spread of misinformation, so we decided to isolate the problem and study it independently.

Moreover, considering the difficulty of collecting the data for our study, as presented in Section 3, not needing to separate the data into training and test (for classical evaluation of the recommendation algorithms) allowed us to devote more data to the purpose of the study. We hope to conduct analyses with accuracy measurements in the future by enriching our current dataset and/or generating synthetic ones, as considered in recent studies [53, 64]. Note that accuracy and metrics that target user satisfaction, may not be the most effective ones when aiming to reduce the impact and spread of misinformation. Algorithms promoting a certain degree of cognitive dissonance, as suggested by existing literature on correcting misperceptions (see Section 2.2), and metrics that focus on computing a balanced degree of user satisfaction and discomfort, may be more suitable to assess and combat misperceptions.

It is also worth noting that the results presented so far only involve one cycle in the feedback loop. Some of the results that we obtained, such as reducing the spread to a range of 10-30%, instead of 90%, may not be enough if the users engage in several cycles of receiving recommendations and interacting with them. In fact, we have simulated another cycle of recommendation in Table 6, where we contrast how the misinformation evolves starting from the same data (user profiles built with a misinformation ratio of 0.2) and running two recommenders (MF and UB) after we assume that all users accept their top-3 recommendations, either those produced by MF or UB. As we observe, the number of misinformative items in the recommendations would increase steadily at each recommendation cycle, although this speed is much lower for UB than for MF. Since no parameter tuning was performed for this simulation, the actual results after another cycle of recommendation might be different depending on which dimension (e.g., spread or accuracy) is optimised.

An important point to make is the need for ethical guidelines [45]. We need to be careful when adapting existing algorithms to ensure that we do not introduce damaging effects. E.g., algorithmic adaptations that may reduce the recommendation of misinformation, but that tend to promote misinformation of a more harmful nature should not be considered successful. This research requires navigating the careful tension between privacy, security, economic interests, censorship and cultural differences, and requires to be addressed from multiple disciplines that can assess not only the technological aspect, but also the individual and the social one. As discussed, there is ample room for investigation in the proposed work, opening a novel, exciting and interdisciplinary line of research. Our initial findings have already paved the way for advancements in the field [53]. At the same time, the ethical implications that generative Artificial Intelligence may have (or is already having), as a source of misinformation in the social media ecosystem, may require to adapt or modify our presented framework, to further explore the effect of recommendation algorithms in such context [46, 74].

ACKNOWLEDGEMENTS

This work has been co-funded by H2020 Co-Inform (ID:770302) and HERoS (ID:101003606) projects and the Spanish Ministry of Science and Innovation (project references: PID2019-108965GB-I00 and PID2022-139131NB-I00).

REFERENCES

[1] Gediminas Adomavicius and Alexander Tuzhilin. 2015. Context-Aware Recommender Systems. In *Recommender Systems Handbook*, Francesco Ricci, Lior Rokach, and Bracha Shapira (Eds.). Springer, 191–226.

[2] Esma Aïmeur, Sabrine Amri, and Gilles Brassard. 2023. Fake news, disinformation and misinformation in social media: a review. *Social Network Analysis and Mining* 13, 1 (2023), 30.

[3] Fabio Aiolli. 2013. Efficient top-n recommendation for very large scale binary rated datasets. In *Proceedings of RecSys*. ACM, 273–280.

[4] Hunt Allcott, Matthew Gentzkow, and Chuan Yu. 2019. Trends in the diffusion of misinformation on social media. *Research & Politics* 6, 2 (2019), 2053168019848554.

[5] Xavier Amatriain and Justin Basilico. 2015. Recommender Systems in Industry: A Netflix Case Study. In *Recommender Systems Handbook*, Francesco Ricci, Lior Rokach, and Bracha Shapira (Eds.). Springer, 385–419.

[6] Eytan Bakshy, Solomon Messing, and Lada A Adamic. 2015. Exposure to ideologically diverse news and opinion on Facebook. *Science* 348, 6239 (2015), 1130–1132.

[7] Alejandro Bellogín, Pablo Castells, and Iván Cantador. 2017. Statistical biases in Information Retrieval metrics for recommender systems. *Information Retrieval Journal* 20, 6 (2017), 606–634.

[8] Chris Blex and Taha Yasseri. 2020. Positive algorithmic bias cannot stop fragmentation in homophilic networks. *The Journal of Mathematical Sociology* (2020), 1–18.

[9] Leticia Bode and Emily K Vraga. 2015. In related news, that was wrong: The correction of misinformation through related stories functionality in social media. *Journal of Communication* 65, 4 (2015), 619–638.

[10] Rocío Cañamares and Pablo Castells. 2018. Should I Follow the Crowd?: A Probabilistic Analysis of the Effectiveness of Popularity in Recommender Systems. In *Proceedings of SIGIR*. ACM, 415–424.

[11] Pablo Castells, Neil J. Hurley, and Saul Vargas. 2015. Novelty and Diversity in Recommender Systems. In *Recommender Systems Handbook*, Francesco Ricci, Lior Rokach, and Bracha Shapira (Eds.). Springer, 881–918.

[12] Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. 2011. Information credibility on twitter. In *Proceedings of the 20th WWW*. ACM, 675–684.

[13] Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. 2013. Predicting information credibility in time-sensitive social media. *Internet Research* 23, 5 (2013), 560–588.

[14] Jilin Chen, Gary Hsieh, Jalal U Mahmud, and Jeffrey Nichols. 2014. Understanding individuals' personal values from social media word use. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*. 405–414.

[15] Xinran Chen and Sei-Ching Joanna Sin. 2013. 'Misinformation? What of it?' Motivations and individual differences in misinformation sharing on social media. *Proceedings of the ASIS&T* 50, 1 (2013), 1–4.

[16] Giovanni Luca Ciampaglia, Azadeh Nematzadeh, Filippo Menczer, and Alessandro Flammini. 2018. How algorithmic popularity bias hinders or promotes quality. *Scientific Reports* 8, 1 (2018), 1–7.

[17] Gabriel de Souza Pereira Moreira, Dietmar Jannach, and Adilson Marques da Cunha. 2019. On the Importance of News Content Representation in Hybrid Neural Session-based Recommender Systems. In *Proceedings of INRA (CEUR Workshop Proceedings, Vol. 2554)*. 18–23.

[18] Michela Del Vicario, Alessandro Bessi, Fabiana Zollo, Fabio Petroni, Antonio Scala, Guido Caldarelli, H Eugene Stanley, and Walter Quattrociocchi. 2016. The spreading of misinformation online. *Proceedings of the National Academy of Sciences* 113, 3 (2016), 554–559.

[19] Mehdi Elahi, Dietmar Jannach, Lars Skjærven, Erik Knudsen, Helle Sjøvaag, Kristian Tolonen, Øyvind Holmstad, Igor Pipkin, Eivind Throndsen, Agnes Stenbom, et al. 2022. Towards responsible media recommendation. *AI and Ethics* (2022), 1–12.

[20] Motahareh Eslamimehdiabadi, Aimee Rickman, Kristen Vaccaro, Christian Sandvig, Kevin Hamilton, and Karrie Karahalios. 2015. " I always assumed that I wasn't really that close to [her]": Reasoning about invisible algorithms in the news feed. *Proceedings of the 33rd CHI* (2015).

[21] Miriam Fernandez and Harith Alani. 2018. Online misinformation: Challenges and future directions. In *Companion Proceedings of The Web Conference 2018*. 595–602.

[22] Miriam Fernandez and Alejandro Bellogín. 2020. Recommender Systems and Misinformation: The Problem or the Solution?. In *Proceedings of the OHARS Workshop, co-located with RecSys 2020*.

[23] Emilio Ferrara, Onur Varol, Clayton Davis, Filippo Menczer, and Alessandro Flammini. 2016. The rise of social bots. *Commun. ACM* 59, 7 (2016), 96–104.

[24] Leon Festinger. 1957. *A theory of cognitive dissonance*. Vol. 2. Stanford Univ. Press.

[25] R Kelly Garrett, Erik C Nisbet, and Emily K Lynch. 2013. Undermining the corrective effects of media-based political fact checking? The role of contextual cues and naïve theory. *Journal of Communication* 63, 4 (2013), 617–637.

[26] Andrew Guess, Jonathan Nagler, and Joshua Tucker. 2019. Less than you think: Prevalence and predictors of fake news dissemination on Facebook. *Science Advances* 5, 1 (2019).

[27] Asela Gunawardana and Guy Shani. 2015. Evaluating Recommender Systems. In *Recommender Systems Handbook*, Francesco Ricci, Lior Rokach, and Bracha Shapira (Eds.). Springer, 265–308.

[28] Aditi Gupta, Ponnurangam Kumaraguru, Carlos Castillo, and Patrick Meier. 2014. Tweetcred: Real-time credibility assessment of content on twitter. In *Proceedings of the SocInfo*. Springer, 228–243.

[29] Lars Kai Hansen, Adam Arvidsson, Finn Årup Nielsen, Elanor Colleoni, and Michael Etter. 2011. Good friends, bad news-affect and virality in twitter. In *Future information technology*. Springer, 34–43.

[30] F. Maxwell Harper and Joseph A. Konstan. 2016. The MovieLens Datasets: History and Context. *ACM Trans. Interact. Intell. Syst* 5, 4 (2016), 19:1–19:19.

[31] Taha Hassan. 2019. Trust and trustworthiness in social recommender systems. In *Companion Proceedings of The 2019 World Wide Web Conference*. 529–532.

[32] Yifan Hu, Yehuda Koren, and Chris Volinsky. 2008. Collaborative Filtering for Implicit Feedback Datasets. In *Proceedings of ICDM*. IEEE Computer Society, 263–272.

[33] Marianne E Jaeger, Susan Anthony, and Ralph L Rosnow. 1980. Who hears what from whom and with what effect: A study of rumor. *Personality and Social Psychology Bulletin* 6, 3 (1980), 473–478.

[34] Dietmar Jannach, Lukas Lerche, Iman Kamehkhosh, and Michael Jugovac. 2015. What recommenders recommend: an analysis of recommendation biases and possible countermeasures. *User Modeling and User-Adapted Interaction* 25, 5 (2015), 427–491.

[35] Jennifer Jerit and Yangzi Zhao. 2020. Political Misinformation. *Annual Review of Political Science* 23, 1 (2020), 77–94. https://doi.org/10.1146/annurev-polisci-050718-032814

[36] Mozhgan Karimi, Dietmar Jannach, and Michael Jugovac. 2018. News recommender systems - Survey and roads ahead. *Information Processing Management* 54, 6 (2018), 1203–1227.

[37] Natascha A Karlova and Karen E Fisher. 2013. A social diffusion model of misinformation and disinformation for understanding human information behaviour. (2013).

[38] Daniel Kluver and Joseph A. Konstan. 2014. Evaluating recommender behavior for new users. In *Eighth ACM Conference on Recommender Systems, RecSys '14, Foster City, Silicon Valley, CA, USA - October 06 - 10, 2014*, Alfred Kobsa, Michelle X. Zhou, Martin Ester, and Yehuda Koren (Eds.). ACM, 121–128. https://doi.org/10.1145/2645710.2645742

[39] Yehuda Koren and Robert M. Bell. 2015. Advances in Collaborative Filtering. In *Recommender Systems Handbook*, Francesco Ricci, Lior Rokach, and Bracha Shapira (Eds.). Springer, 77–118.

[40] David MJ Lazer, Matthew A Baum, Yochai Benkler, Adam J Berinsky, Kelly M Greenhill, Filippo Menczer, Miriam J Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, et al. 2018. The science of fake news. *Science* 359, 6380 (2018), 1094–1096.

[41] Kyumin Lee, Prithivi Tamilarasan, and James Caverlee. 2013. Crowdturfers, Campaigns, and Social Media: Tracking and Revealing Crowdsourced Manipulation of Social Media. In *Proceedings of the 7th ICWSM*.

[42] Joseph McGlynn, Maxim Baryshevtsev, and Zane A Dayton. 2020. Misinformation more likely to use non-specific authority references: Twitter analysis of two COVID-19 myths. *Harvard Kennedy School Misinformation Review* 1, 3 (2020).

[43] Shahan Ali Memon and Kathleen M Carley. 2020. Characterizing covid-19 misinformation communities using a novel twitter dataset. *arXiv preprint arXiv:2008.00791* (2020).

[44] Martino Mensio and Harith Alani. 2019. MisinfoMe: Who's Interacting with Misinformation? (2019).

[45] Silvia Milano, Mariarosaria Taddeo, and Luciano Floridi. 2020. Recommender systems and their ethical challenges. *AI & SOCIETY* (2020), 1–11.

[46] Scott Monteith, Tasha Glenn, John R. Geddes, Peter C. Whybrow, Eric Achtyes, and Michael Bauer. 2024. Artificial intelligence and increasing misinformation. *The British Journal of Psychiatry* 224, 2 (2024), 33–35. https://doi.org/10.1192/bjp.2023.136

[47] Xiaoli Nan, Yuan Wang, and Kathryn Thier. 2022. Why people believe health misinformation and who are at risk? A systematic review of individual differences in susceptibility to health misinformation. *Social Science & Medicine* (2022), 115398.

[48] Dimitar Nikolov, Diego FM Oliveira, Alessandro Flammini, and Filippo Menczer. 2015. Measuring online social bubbles. *PeerJ computer science* 1 (2015), e38.

[49] Xia Ning, Christian Desrosiers, and George Karypis. 2015. A Comprehensive Survey of Neighborhood-Based Recommendation Methods. In *Recommender Systems Handbook*, Francesco Ricci, Lior Rokach, and Bracha Shapira (Eds.). Springer, 37–76.

[50] Brendan Nyhan and Jason Reifler. 2010. When corrections fail: The persistence of political misperceptions. *Political Behavior* 32, 2 (2010), 303–330.

[51] John O'Donovan and Barry Smyth. 2005. Trust in recommender systems. In *Proceedings of IUI*. 167–174.

[52] Eli Pariser. 2011. *The filter bubble: What the Internet is hiding from you.* Penguin UK.

[53] Royal Pathak, Francesca Spezzano, and Maria Soledad Pera. 2023. Understanding the contribution of recommendation algorithms on misinformation recommendation and misinformation dissemination on social networks. *ACM Transactions on the Web* 17, 4 (2023), 1–26.

[54] Lara SG Piccolo, Alisson Puska, Roberto Pereira, and Tracie Farrell. 2020. Pathway to a Human-Values Based Approach to Tackle Misinformation Online. In *Proceedings of the 22nd HCI*. Springer, 510–522.

[55] Vahed Qazvinian, Emily Rosengren, Dragomir R Radev, and Qiaozhu Mei. 2011. Rumor has it: Identifying misinformation in microblogs. In *Proceedings of the EMNLP*. Association for Computational Linguistics, 1589–1599.

[56] Xiaoyan Qiu, Diego FM Oliveira, Alireza Sahami Shirazi, Alessandro Flammini, and Filippo Menczer. 2017. Limited individual attention and online virality of low-quality information. *Nature Human Behaviour* 1, 7 (2017), 0132.

[57] Daniele Quercia, Michal Kosinski, David Stillwell, and Jon Crowcroft. 2011. Our twitter profiles, our selves: Predicting personality with twitter. In *2011 IEEE third international conference on privacy, security, risk and trust and 2011 IEEE third international conference on social computing*. IEEE, 180–185.

[58] Jon Roozenbeek, Claudia R. Schneider, Sarah Dryhurst, John Kerr, Alexandra L. J. Freeman, Gabriel Recchia, Anne Marthe van der Bles, and Sander van der Linden. 2020. Susceptibility to misinformation about COVID-19 around the world. (2020).

[59] Dorsaf Sallami, Rim Ben Salem, and Esma Aïmeur. 2023. Trust-based Recommender System for Fake News Mitigation. In *Adjunct Proceedings of UMAP*. 104–109.

[60] Oguz Semerci, Alois Gruson, Catherinee Edwards, Ben Lacker, Clay Gibson, and Vladan Radosavljevic. 2019. Homepage personalization at Spotify. In *Proceedings of the 13th ACM Conference on Recommender Systems*, Toine Bogers, Alan Said, Peter Brusilovsky, and Domonkos Tikk (Eds.). ACM, 527.

[61] Chengcheng Shao, Giovanni Luca Ciampaglia, Onur Varol, Kai-Cheng Yang, Alessandro Flammini, and Filippo Menczer. 2018. The spread of low-credibility content by social bots. *Nature communications* 9, 1 (2018), 1–9.

[62] Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. 2020. FakeNewsNet: A Data Repository with News Content, Social Context, and Spatiotemporal Information for Studying Fake News on Social Media. *Big Data* 8, 3 (2020), 171–188.

[63] Emily Taylor, Lisa-Maria Neudert, Stacie Hoffmann, and Philip N Howard. 2020. Follow the Money: How the Online Advertising Ecosystem Funds COVID-19 Junk News and Disinformation.

[64] Antonela Tommasel and Ira Assent. 2023. Recommendation fairness and where to find it: An empirical study on fairness of user recommender systems. In *IEEE International Conference on Big Data, BigData 2023, Sorrento, Italy, December 15-18, 2023*, Jingrui He, Themis Palpanas, Xiaohua Hu, Alfredo Cuzzocrea, Dejing Dou, Dominik Slezak, Wei Wang, Aleksandra Gruca, Jerry Chun-Wei Lin, and Rakesh Agrawal (Eds.). IEEE, 4195–4204. https://doi.org/10.1109/BIGDATA59044.2023.10386616

[65] Antonela Tommasel and Filippo Menczer. 2022. Do Recommender Systems Make Social Media More Susceptible to Misinformation Spreaders?. In *Proceedings of RecSys*. 550–555.

[66] Antonela Tommasel, Juan Manuel Rodriguez, and Daniela Godoy. 2021. I Want to Break Free! Recommending Friends from Outside the Echo Chamber. In *Proceedings of RecSys*. 23–33.

[67] Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science* 359, 6380 (2018), 1146–1151.

[68] Emily K Vraga and Leticia Bode. 2020. Correction as a Solution for Health Misinformation on Social Media.

[69] Claudia Wagner, Silvia Mitter, Christian Körner, and Markus Strohmaier. 2012. When social bots attack: Modeling susceptibility of users in online social networks. In *Proceedings of the WWW'12 Workshop on 'Making Sense of Microposts'*, Vol. 2. 1951–1959.

[70] Shoujin Wang, Xiaofei Xu, Xiuzhen Zhang, Yan Wang, and Wenzhuo Song. 2022. Veracity-aware and event-driven personalized news recommendation for fake news mitigation. In *Proceedings of the ACM Web Conference*. 3673–3684.

[71] Claire Wardle and Hossein Derakhshan. 2017. Information disorder: Toward an interdisciplinary framework for research and policy making. *Council of Europe Report* 27 (2017).

[72] Liang Wu, Fred Morstatter, Kathleen M. Carley, and Huan Liu. 2019. Misinformation in Social Media: Definition, Manipulation, and Detection. *SIGKDD Explor.* 21, 2 (2019), 80–90.

[73] Canwen Xu and Julian J. McAuley. 2023. A Survey on Dynamic Neural Networks for Natural Language Processing. In *Findings of the ACL*, Andreas Vlachos and Isabelle Augenstein (Eds.). Association for Computational Linguistics, 2325–2336.

[74] Danni Xu, Shaojing Fan, and Mohan S. Kankanhalli. 2023. Combating Misinformation in the Era of Generative AI Models. In *Proceedings of the 31st ACM International Conference on Multimedia, MM 2023, Ottawa, ON, Canada, 29 October 2023- 3 November 2023*, Abdulmotaleb El-Saddik, Tao Mei, Rita Cucchiara, Marco Bertini, Diana Patricia Tobon Vallejo, Pradeep K. Atrey, and M. Shamim Hossain (Eds.). ACM, 9291–9298. https://doi.org/10.1145/3581783.3612704

[75] Fattane Zarrinkalam, Mohsen Kahani, and Ebrahim Bagheri. 2018. Mining user interests over active topics on social networks. *Inf. Process. Manag.* 54, 2 (2018), 339–357.

[76] Amy X Zhang, Aditya Ranganathan, Sarah Emlen Metz, Scott Appling, Connie Moon Sehat, Norman Gilmore, Nick B Adams, Emmanuel Vincent, Jennifer Lee, Martin Robbins, et al. 2018. A structured response to misinformation: Defining and annotating credibility indicators in news articles. In *Companion Proceedings of The Web Conference 2018*. 603–612.

[77] Xichen Zhang and Ali A. Ghorbani. 2020. An overview of online fake news: Characterization, detection, and discussion. *Inf. Process. Manag.* 57, 2 (2020), 102025.

[78] Fan Zhou, Ruiyang Yin, Kunpeng Zhang, Goce Trajcevski, Ting Zhong, and Jin Wu. 2019. Adversarial Point-of-Interest Recommendation. In *Proceedings of WWW*. ACM, 3462–34618.

[79] Bi Zhu, Chuansheng Chen, Elizabeth F Loftus, Chongde Lin, and Qinghua. He. 2010. Individual differences in false memory from misinformation: Personality characteristics and their interactions with cognitive abilities. *Personality and Individual Differences* 48, 8 (2010), 889–894.