Misinformation in video recommendations: an exploration of Top-N recommendation algorithms

Benedikt Hornig[1,*], Maria Soledad Pera[2] and Jan Scholtes[1]

[1]Maastricht University, Maastricht, The Netherlands

[2]Web Information Systems - Delft University of Technology, Delft, The Netherlands

Abstract

With this paper, we delve into the problem of misinformation propagation in the video recommendation domain, focusing on top-N recommendation algorithms (RAs). We evaluate a broad spectrum of RAs to probe their ability to minimize misinformation recommendations while optimizing the RAs for overall performance. The results of an empirical exploration conducted using a suite of Top-N RAs and a video recommendation dataset [1] show that certain RAs excel in both performance and misinformation handling, while others struggle in mitigating misinformation. Our findings emphasize the potential of neighbourhood-based, neural, and other advanced collaborative filtering (CF) approaches in combating misinformation and contributing to more responsible recommender systems. Inspired by our findings, we propose investigating hybrid RAs and exploring specific features influencing misinformation recommendations, to further enhance the understanding and effectiveness of mitigating misinformation in recommendation systems.

Keywords

Recommendation algorithms, Misinformation, Video recommendation

1. Introduction

Misinformation, which we define as in [1], as information which is incorrect or misleading, but is not deliberately deceptive, can have an impact on society that should not be underestimated. The 2016 US election for example has been impacted by the spreading of Fake News on social media [2]. This is exacerbated by the fact that misinformation spreads faster than truthful information on platforms like Twitter [3]. The presented misinformation is selected by a recommendation algorithm (**RA**) and then shown to the user. The selection of information can be filtered, e.g. by a misinformation detection system [4]. However, this approach is reactive and thus, with this study, we focused on the recommendations of the RAs themselves.

When a RA presents content to the user, misinformation may be included. Previous research on Top-N RAs, such as [5] and [6], has mainly focused either on a limited number of algorithms or on understanding misinformation diffusion. However, they did not extensively compare various

ROMCIR 2024: The 4th Workshop on Reducing Online Misinformation through Credible Information Retrieval (held as part of ECIR 2024: the 46th European Conference on Information Retrieval), March 24, 2024, Glasgow, UK

[*] Corresponding author.

b.hornig@student.maastrichtuniversity.nl (B. Hornig); m.s.pera@tudelft.nl (M. S. Pera); j.scholtes@maastrichtuniversity.nl (J. Scholtes)

https://solepera.github.io (M. S. Pera)

0009-0005-2693-7322 (B. Hornig); 0000-0002-2008-9204 (M. S. Pera); 0000-0003-0534-2491 (J. Scholtes)

RAs regarding the prevalence of misinformation in their recommendations. Our objective is to provide a more comprehensive analysis, specifically examining which types of RAs tend to recommend more content containing misinformation than others.

In this paper, we examine the prevalence of misinformation in the recommendations of RAs by exploring a broad spectrum of top-N RAs as to which of them are more prone to recommending misinformation content. In particular, we aim to answer this RQ: *Which types of top-N RAs are more prone to present items including misinformation in their recommendation results?*

In order to answer this question, we conducted an empirical exploration of multiple top-N RAs measuring their performance and the extent of misinformation they recommend. This exploration allowed us to identify RAs that exhibit a higher tendency to recommend misinformation, offering valuable insights into the accuracy-misinformation tradeoff. Furthermore, we examined various features of the recommendations of each RA, aiming to determine which factors play the most significant role in recommending content containing misinformation.

A summary of our contributions is the following:

- An assessment of the performance based on NDCG of multiple state-of-the-art RAs on a new YouTube video recommendation dataset.
- A comprehensive analysis of the amplification of misinformation by different state-of-the-art RAs commonly used in both industry and academia. This analysis is based on Normalized Score and SERP-MS, two metrics designed to account for such amplification.
- An exploratory analysis of features (views, likes, comments, and video duration) associated with the top-recommended items by several RAs and discussing their importance concerning the prevalence of misinformation in the recommendations of each RA.

With our results, a trend can be recognized in which top-N RAs are more prone to spreading misinformation, contributing to more responsible recommender systems. The code for our experimental setup can be found in the accompanying GitHub repository[1].

2. Related work

In recent years, the spread of misinformation has become a pressing issue with far-reaching consequences for society. In this Section, we present background and related literature pertaining both recommender systems and misinformation, that inform our work.

Impact on society. Misinformation recommendation can have several far-reaching effects on our society. In the political scope, politicians deliberately use misinformation to discredit their opponents, and we have already seen this impacting the presidential election in the US in 2016 as shown by Zhang et al.[2]. Furthermore, in contrast to true information, misinformation even tends to resurface as for one it has become a political instrument to bring back misinformation after some time, e.g. during election days, to heat up old rumours [7]. However, harming effects on society can be mitigated, if misinformation propagation can be reduced and corrective information is only shown to relevant users. Even though displaying corrective information

[1] https://github.com/AlmightyPeanut/RecSysInternship

intends to have a positive impact, it can inadvertently lead to outcomes contrary to the intended ones [8]. Because of the impactful effect misinformation can have, several studies have analysed this issue and investigated recommendations given by RAs.

Analysis of RAs. Several studies ([1], [9], [10]) have analysed RAs and found that the recommendations they present are heavily influenced by the user's viewed content history. However, the content distribution can be influenced by selecting different RAs and recommending different items to the user, as different RAs can recommend different items[11]. To investigate this, Fernández et al. [5] analysed some RAs in this preliminary study on a limited number of RAs to identify which ones are more likely to recommend misinformation and how they can be modified. They put emphasis on the modifications of the RAs and issues that lead to more recommendations incorporating misinformation. Still, this shows that choosing different RAs has a big impact on which content is recommended and therefore on how much misinformation is spread.

Not only selecting different RAs, but also altering existing algorithms can have an effect on misinformation recommendation. There have been many suggestions on how to alter RAs like query expansion, data fusion ([12], [13]) or proposing new algorithms like Badami et al. [14] who addressed anti-polarization by proposing an algorithm that promotes diverse perspectives. The user themselves are also able to influence their recommendations by deliberately choosing or searching for their desired content [15]. However, if they are not conscious of this fact, they may be more likely exposed to misinformation just by the selection of a specific RA.

Perspective for this study. Srba et al. [1] conducted an external audit of YouTube's recommendation system, assessing its effectiveness in mitigating misinformation promotion and recommending debunking items. Although improvements were observed for certain topics compared to a prior study [10], opportunities for enhancement remain. The study also introduced a new dataset of YouTube recommendations. However, no existing study has comprehensively compared the misinformation predictions of various RAs on YouTube data. Leveraging this dataset and its misinformation classifier, we trained and evaluated a selection of RAs, assessing their performance and the prevalence of misinformation in their recommendations.

Fernández et al. [16] analysed misinformation recommendations from different RAs using Twitter data. They compared matrix factorization (MF), user and item nearest neighbour (NN) algorithms to random and popular algorithms, highlighting popularity bias in RAs. While they suggested strategies to mitigate this bias, the study mainly focused on pre- and post-recommendation strategies, revealing a research gap. Our study briefly compares the performance of selected RAs, but primarily delves into analysing how various RAs recommend misinformation content.

In conclusion, there are several studies on the analysis and extension of RAs to reduce the recommendation of misinformation. These studies suggest different strategies for modifying RAs to promote diversity, reduce polarization and misinformation. However, there is still a need for further research in this area, particularly on comparing the models themselves.

3. Experimental Setup

In this Section, we describe the experiments and their prerequisites we conducted to answer the research questions mentioned in Section 1.

Dataset. Dataset recreation for our empirical exploration relies on the dataset from [9], collected via YouTube agents. This dataset was annotated for misinformation categories (debunking, promoting, neutral) and used as seed data for training a classifier. Video data was obtained using the YouTube API[2] and youtube-dl[3] for transcripts. For this, the repositories from [9][4] [5] and [10][6] were used to build upon. The dataset covers recommendations from the YouTube home page and when a user clicks on a video, spanning five topics: 9/11 conspiracies, moon landing conspiracies, chemtrails conspiracies, flat earth conspiracies, and vaccines conspiracies.

Each video is labelled using a model much like in [9], first introduced in [10], from here on called Papadamou's model. Papadamou's model is indicating whether the video promotes, debunks, or remains neutral towards misinformation by utilizing snippet (video title and description), transcript, and top-200 comments, encoded into word embeddings with fastText. Since the data originally used for training this model does not contain video tags, this part was omitted. The model, featuring 4 layers (256, 128, 64, 32 units) with ReLU activation and a Softmax output layer, classifies videos with a threshold of $\geq$0.7. Dropout (d=0.5) is used for regularization, and oversampling addresses class imbalance. For training, 10-fold cross validation was used.

Our dataset comprises 15,070 videos, with 82.6% automatically annotated by the aforementioned classifier, which was trained on seed data manually annotated in [9]. The dataset includes 1.3% promoting, 2.6% debunking, and 96% neutral videos. Retrieval challenges for video data limited the use of some videos from the original dataset [9]. Specifically, 12% lacked comments due to disabled or unavailable comment sections.

Selected RAs. In our exploration, we considered a broad range of RAs from popularity-based to state-of-the-art, exemplifying different popular approaches in the literature.

As **baseline models**, we employed random and most popular RAs. They served as a basic benchmark to compare the other models to.

To incorporate **Nearest Neighbour (NN)** methods, we examined the item- and user-based NN algorithms. They identify the k most similar items or users to a given item or user. These algorithms operate on the assumption that users with similar preferences exhibit comparable interactions or favour items with similar characteristics. By leveraging the concept of similarity, I-k NN [17] and U-k NN [18] can provide personalized recommendations based on similar user-item interactions.

In the **Collaborative Filtering (CF)** family, we analysed various RAs: Matrix Factorisation (MF) [19], Bayesian Personalized Ranking MF (BPRMF) [20], Collaborative Metric Learning (CML) [21], Funk's Singular Value Decomposition algorithm (FunkSVD) [22], Logistic MF

[2] https://developers.google.com/youtube/v3
[3] https://youtube-dl.org
[4] https://github.com/kinit-sk/yaudit-recsys-2021
[5] https://github.com/kinit-sk/yaudit-papadamou-model
[6] https://github.com/kostantinos-papadamou/pseudoscience-paper

(LogMF) [23], Non-Negative MF (NNMF) [24], Probabilistic MF (PMF) [25], and SVD++ [26]. CF models learn latent factors for users and items, aiming to minimize the discrepancy between the original and reconstructed user-item interaction matrix.

Among **Factorisation Machines (FM)** algorithms, we evaluated a Field-aware FM (FFM) [27] algorithm. FMs combine linear and pairwise feature interactions, enabling them to capture complex relationships between features. By incorporating pairwise feature interactions, FMs enhance the recommendation performance compared to traditional linear models.

Finally, in the **Neural Algorithms** family, we investigated the following algorithms: Deep MF (DMF) [28], Generalised MF (GMF) [29], and Neural MF (NeuMF) [29]. These models use neural network architectures to capture complex patterns and user-item interactions.

Metrics. While the primary focus of this study is not to evaluate the performance of RAs, we still consider it to provide context for the tradeoff between accurate recommendations and misinformation. To quantify performance and misinformation, we considered the **Normalized Discounted Cumulative Gain (NDCG)** which measures the ranking quality of the recommended items. It takes into account both the relevance of the items and their positions in the recommendation list. Specifically, we used NDCG@10 for our evaluation[7].

The **Normalized Score (NS)** metric quantifies the popularity of misinformation in a list of items (videos). It considers labels of promoting (1), debunking (-1), or neutral (0) assigned to the items. NS is the average of individual annotations without considering the item's rank. The NS value ranges from -1 to 1, where values close to -1 indicate less misinformation for lists with mostly debunking content, close to 1 for promoting content, and close to 0 for balanced or neutral content.

To also incorporate the ranking of the items in each recommendation, **Search Engine Result Page Misinformation Score (SERP-MS)** was used as it works similarly to NS, but considers the rank of the items in the list. It is computed as follows:

$$SERP\text{-}MS = \frac{\sum_{r=1}^{n}(x_i * (n - r + 1))}{\frac{n*(n+1)}{2}}$$

where x_i is the label value, n the number of items in the list, and r the rank of the item. Its value is interpreted similarly as NS with a range of $[-1, 1]$.

Experiments. We conducted three experiments to answer our RQ posted in Section 1. To offer context to the misinformation analysis, we first examine the performance of the RAs under study, based on NDCG@10. Then, we assess the degree of misinformation recommendations of each RA with the metrics SERP-MS and NS. Finally, we explore some features (views, likes, comments, duration) of the recommended videos to gain a deeper understanding of the items recommended by each RA.

For experiment purposes, we used the algorithms and experimental platform of Elliot [30]. The RAs were trained using 5-fold cross-validation with a 90/10 train-test split. To find the best hyperparameters to optimize each algorithm for NDCG, grid search was used for each RA. In

[7] Note that in addition to NDCG, we considered MRR, Precision and Recall. Results showed similar trends to those observed for NDCG, which is why, due to the page limitation, we only report NDCG@10 results.

the first two experiments, we employed pairwise Student's paired t-test ($p<0.05$) to determine the significance of the recommendation results, systematically evaluating each combination of RA pairs. The results of the described experiments are reported in Section 4.

4. Results

Here, we present the results of the misinformation classifier evaluation, the performance, and misinformation evaluation of the RAs and the feature exploration as described in Section 3.

Misinformation classifier. First, we examined the performance of the misinformation classifier used to label videos presented by the top-N RAs to mitigate errors that could be introduced as a result of needing to train the classifier used in [1]. When training the misinformation classifier as described in Section 2, we achieved similar results compared to the original classifier as used in [1], which is shown in Table 1. We attribute performance differences to the fact that for some of the videos we were not able to collect associated data.

Table 1
Performance comparison across the results reported in [1] on the original misinformation classifier versus the counterpart version we reproduced in our study.

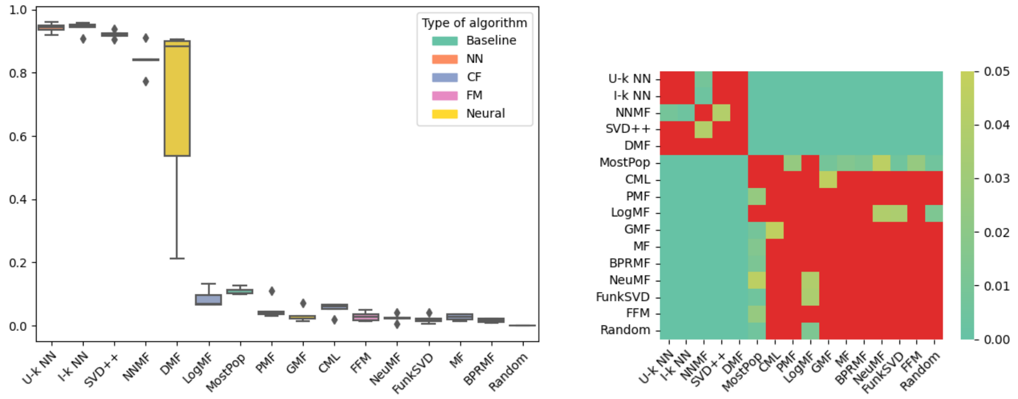
	Precision	Recall	F1-Score	Accuracy
original model	0.81	0.82	0.82	0.82
our model	0.84	0.84	0.83	0.83

RA performance. To provide context to the evaluation of misinformation prevalence in the recommendations provided by the top-N RAs, we briefly present the results of the evaluation of the performance of each algorithm. For the evaluation of the RAs, each Figure shows each of the algorithms in sorted order from highest to lowest. Additionally, the colour of each bar represents the type of algorithm that is shown as categorized in Section 2. All algorithms in the graphs are sorted by their NDCG@10 evaluation.

Figure 1 illustrates the NDCG@10 scores, accompanied by the visualisation of statistical significance, where the red colour indicates that the difference between the scores of this RA pair is not statistically significant, whereas the green colour shows that they are.

When comparing performance of the top-N RAs with NDCG@10, we observed a splitting of the algorithms in two large groups with NNMF laying between them. The first group includes SVD++, U-k NN, I-k NN and DMF. They are individually outperforming all the algorithms of the second group, which includes all other algorithms except NNMF. Analysing the group containing SVD++, the NN algorithms and DMF in more detail, achieved NDCG@10 scores greater than 0.9. NNMF performs slightly worse, with a NDCG@10 score of above 0.8.

The group, which is individually significantly worse than the algorithms just discussed, includes PMF, GMF, FunkSVD, FFM, LogMF, CML, MF, NeuMF and Random. They all evaluate below 0.2 NDCG@10.



(a) NDCG@10

(b) NDCG@10 paired t-test

Figure 1: NDCG@10 scores of each algorithm and their Student's paired t-test for comparison with counterparts. Green indicates significant (p<0.05) results; red for non-significant.

Interestingly, the neural algorithms are completely separated from each other. DMF is in the group of the best performing algorithms, whereas GMF and NeuMF are in the group of the worst performing algorithms. In contrast, the NN algorithms are both giving results that place them in the top group.

The baseline algorithm, choosing only the most popular item, is performing better than FFM, NeuMF and Random in both performance evaluations, but is not significantly better than the other algorithms of the group. These algorithms were unable to give higher quality results than just recommending the most popular item.

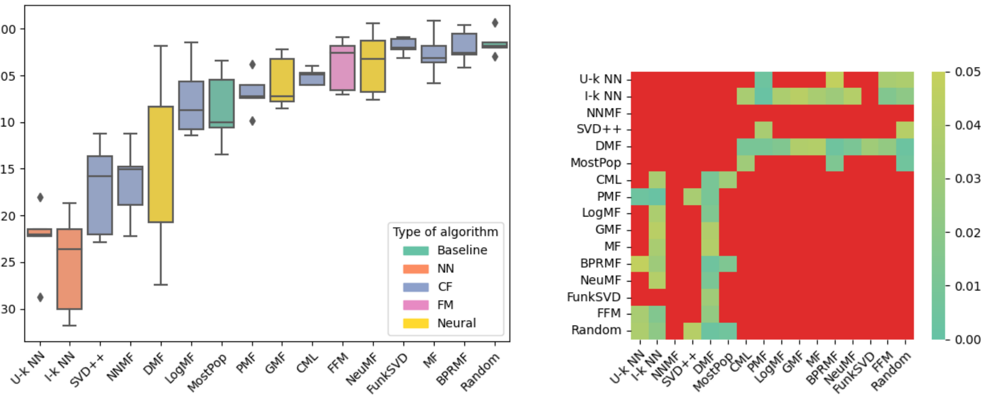
RA misinformation. Having contextualized the algorithm performance, we now probe the presence of misinformation among their suggestions. Our results are shown in Figures 2 and 3.

None of the RAs evaluated with a positive score for SERP-MS and neither for NS. This observation can be attributed to the dataset's distribution, which contains more than twice the number of debunking videos compared to promoting ones.

Certain RAs continue to perform well compared to the other RAs in the context of misinformation recommendation. I-k NN and DMF remain consistently good compared to the other algorithms and significantly outperform most other RAs with scores lower than -0.15 for both SERP-MS and NS. U-k NN, SVD++ and NNMF also perform similarly, recommending more content debunking misinformation, although their superiority is statistically significant compared to a smaller number of RAs.

SVD++, DMF and NNMF achieve slightly lower SERP-MS scores compared to its NS scores, indicating that videos debunking misinformation were generally higher up in their recommendation lists. Conversely, U-k NN and I-k NN show opposite results.

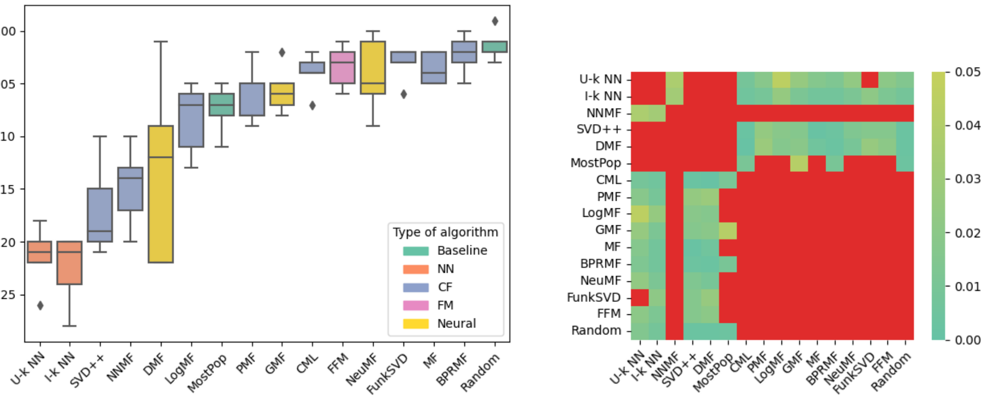
Surprisingly, the baseline algorithm, which simply recommends the most popular item, performs significantly better than algorithms such as GMF and MF in both misinformation evaluations. This suggests that the simplistic approach of relying solely on popularity can sometimes be more effective at minimizing recommended content containing misinformation



(a) SERP-MS

(b) SERP-MS paired t-test

Figure 2: SERP-MS scores of each algorithm and their Student's paired t-test for comparison with counterparts. Green indicates significant (p<0.05) results; red for non-significant.



(a) NS

(b) NS paired t-test

Figure 3: NS scores of each algorithm and their Student's paired t-test for comparison with counterparts. Green indicates significant (p<0.05) results; red for non-significant.

than more complex algorithms.

Another algorithm that shows lower scores for SERP-MS and NS compared to RAs with similar NDCG@10 Performance evaluations is FunkSVD. It achieved a score of -0.096 for SERP-MS and -0.1 for NS. Consequently, FunkSVD is giving lower scores than each of the worst performing algorithms, which are namely GMF, LogMF, MF, and Random. They cannot achieve scores lower than -0.05 for either misinformation metric. When only looking at NS, FFM and NeuMF join this group of worst performing algorithms regarding misinformation. These algorithms consistently underperform, indicating their limitations in effectively recommending less misinformation.

Neighbourhood-based recommendation algorithms stand out as the only category consistently excelling in minimizing misinformation items and favouring items debunking misinformation. Unlike this category, other recommendation algorithm types lack a clear advantage in reducing misinformation items. In the family of FMs, only FFM was examined and did not achieve favourable SERP-MS and NS scores compared to top-performing RAs, leaving room for potential better performance by other FM-based RAs.

In summary, neighbourhood-based RAs, DMF, SVD++ and NNMF continue to perform well and stand out among the RAs evaluated. Additionally, MostPop and FunkSVD have shown better misinformation recommendations compared to other RAs with similar performance evaluations. Conversely, algorithms such as GMF, LogMF, MF, and Random consistently demonstrate poorer performance in recommending fewer items containing misinformation.

Exploratory analysis. To gain a deeper understanding of the recommendations made by each RA, we analysed features associated with the top-10 recommended items. The distribution of views, likes, comments, and video duration is visualized in Figure 4.

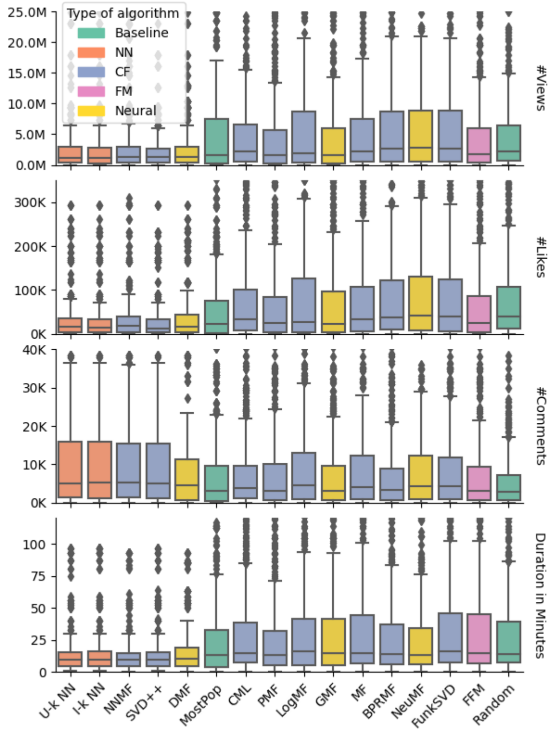


Figure 4: Distribution of the number of views, likes, and comments and the duration of the videos recommended across each of the analysed RAs sorted by their NDCG@10 scores.

Upon examining the number of views and likes of the recommended videos for each RA, a striking similarity emerges. As anticipated, higher views correspond to more user interactions, leading to increased likes. Notably, videos recommended by SVD++, the NN algorithms, and DMF tend to have fewer views and likes compared to those recommended by other algorithms. Following them, NNMF has recommended videos with slightly more views and likes, but still fewer than most of the other algorithms.

The baseline algorithm, which recommends popular videos, does not necessarily yield the highest view counts due to the audit's methodology, where sock-puppets clicked on the first recommended video rather than the most viewed. Thus, in our case, the most popular videos are not equivalent to the most watched ones on YouTube.

Examining comments, DMF, NeuMF, and the random baseline algorithm stand out for recommending videos with slightly fewer comments. Notably, videos recommended by SVD++, U-k NN, I-k NN, and DMF have similar comment counts despite lower views, indicating that videos debunking misinformation tend to receive more comments.

Regarding video duration, a distinct separation is evident. SVD++, NN algorithms, DMF, and NNMF tend to recommend shorter videos, mostly below 30 minutes. BPRMF, MostPop, LogMF, and MF suggest slightly longer videos, most up to one hour, while other algorithms generally recommend slightly longer videos.

The NN algorithms, DMF, SVD++, and NNMF also exhibit distinct patterns in the views, likes, and duration of recommended videos. In contrast, comment counts align more closely with other algorithms' recommendations.

5. Discussion and limitations

The results of our experiments provide valuable insights into the performance and effectiveness of RAs in the context of recommending misinformation. In this Section, we discuss the implications of our findings and acknowledge the limitations of our study.

Comparison of performance and misinformation recommendation. By comparing the performance and the popularity of misinformation in recommendations lists of different top-N RAs illustrated by Figures 1, 2 and 3, we can draw several implications.

Some algorithms struggled in terms of overall recommendation quality, but performed well when it came to filtering out misinformation. For instance, GMF had similar NDCG@10 evaluations to the baseline recommending the most popular items, but GMF was significantly worse in filtering misinformation than the baseline. Similarly, MF achieved a higher NDCG@10 score than NeuMF, but achieved worse scores than NeuMF when evaluated on misinformation.

In contrast, this phenomenon also exists the other way around. Algorithms like FunkSVD and CML achieved better SERP-MS and NS scores compared to algorithms with similar performance evaluations. This shows, that these algorithms were able to provide more recommendations debunking misinformation, when compared to RAs with similar recommendation quality.

Seeing this difference between performance and misinformation scores emphasizes the need to evaluate RAs specifically in the context of misinformation to ensure their reliability in providing the least misinformation.

This difference also highlights a crucial balance between accuracy and misinformation that needs to be carefully considered for different applications. When we apply these results to large social media platforms like YouTube, the choice of RA can have significant impacts. Since millions of users are using these platforms, a small change in the scores of the misinformation evaluations results in a change in the recommendations for all of these users. Getting this balance between the quality of the recommendations and the amount of content promoting misinformation right depends on what is the ultimate goal of the platform.

Overall, SVD++, the NN RA family, DMF and NNMF demonstrated better performance than most RAs in terms of recommendation quality and additionally were better than most in combating misinformation. This shows the potential of these algorithms in both maintaining high-quality recommendations whilst limiting videos promoting misinformation.

One baseline algorithm (MostPop) and FunkSVD showcased improvements compared to algorithms with similar performance in mitigating the recommendation of misinformation, despite their lower overall performance. This combined with the results of the neighbourhood-based algorithms and NNMF suggests that even algorithms with simpler recommendation strategies can contribute to reducing the spread of false information. This assertion is further supported by the work of Fernandez et al. [5], who suggested adaptations for certain algorithms to mitigate misinformation recommendations.

However, algorithms such as GMF, LogMF, MF consistently struggled in both overall recommendation performance and combating misinformation. These algorithms may require further improvements like DMF did with MF or alternative approaches to effectively address the challenges of misinformation in recommendations. Considering the number of users and content on social media websites, small improvements with regard to misinformation can already have big implications since recommending misinformation can have major consequences for our society as shown by multiple studies ([2], [7]). Improvements are imperative, as evidenced by the audit on YouTube conducted by Srba et al. [1], which did not reveal notable enhancements in mitigating the popularity of misinformation in recommendation outcomes.

Since misinformation spreads faster than other information [3], indicating a higher resonance with the users, we expected that videos promoting misinformation to experience a lively discussion resulting in more comments. Nonetheless, this was not the case, as they showcase more comments than expected on videos provided by RA recommending more items debunking misinformation. Analysing this user behaviour is worthy of future investigation.

Other features than the number of comments, namely the number of views, likes, and the duration of the videos displayed in Figure 4, show generally lower values for videos provided by RAs tending to recommend less misinformation debunking videos. Thus, indicating a connection between these features and the recommendation of items containing misinformation.

Our results indicate, that the neighbourhood-based family of RAs are the only type of RAs less prone to presenting items promoting misinformation compared to the other RAs. While our evaluation encompassed only one representative of the FM type, it is plausible that other RAs within the FM RA family might outperform FFM. However, FFM itself exhibited less promising outcomes in both recommendation quality and its capability to suggest fewer items containing misinformation. Regarding CF and neural-based RAs, we did not identify a clear pattern of inclination towards recommending misinformation items. Instead, these types comprised individual algorithms that either excelled or lagged in the aspect of recommending items incorporating misinformation.

Limitations. Our study has certain limitations. The evaluation was conducted on a specific dataset about video recommendations on YouTube, and the results may not be fully generalizable to other domains or platforms. The availability and quality of data, as well as the specific characteristics of the dataset, can impact algorithm performance and effectiveness.

The study did not consider hybrid recommendation approaches or user and item features, and extending the research to compare these results with hybrid algorithms is recommended.

The availability and quality of data can impact the training and evaluation of algorithms. In our case, the unavailability of certain videos in our dataset compared to the dataset used for training the original misinformation classifier may have influenced the classifier's performance. Ensuring an accurate labelling of the videos as misinformation promoting or not is crucial for the evaluation with SERP-MS and NS.

Our trained misinformation classifier achieved comparable results to the original model [1], but unavailability of some videos in our dataset may have influenced performance. This highlights the importance of comprehensive and up-to-date data for training accurate classifiers.

Despite these limitations, our study provides valuable insights into the performance and effectiveness of RAs in the context of misinformation. The findings contribute to the growing body of research aiming to develop robust recommendation systems that mitigate the spread of false information and improve user experiences.

6. Conclusions

In this work, we investigated which top-N RAs are more prone to recommending misinformation. Our study provides valuable insights into the performance and effectiveness of RAs in combating misinformation for video recommendation. SVD++, the nearest neighbour algorithms, DMF and NNMF showcased strong overall performance and effectiveness in mitigating misinformation. Further, algorithmic approaches, including simpler strategies, can also contribute to reducing the spread of false information. Continued research in RAs is necessary to enhance their performance, robustness, and reliability in delivering accurate and trustworthy content to users.

For future work, we recommend extending this study to further investigate more datasets from different domains to give complement our findings. Furthermore, examining which parts of the RAs increase or decrease the amount of misinformation, e.g. by deepening the data exploration with more features, to understand why certain RAs recommend more misinformation than others on a deeper level. To further analyse misinformation recommendation in top-N RAs, we propose to also investigate hybrid RAs and compare to our results. For this, we additionally recommend a deep analysis into which features tend to give an indication for misinformation content. Exploring this avenue will help enrich the understanding of the factors influencing misinformation recommendations and contribute to more effective and responsible recommender systems.

Overall, our findings contribute to developing more reliable and responsible recommender systems in the fight against false information.

References

[1] I. Srba, R. Moro, M. Tomlein, B. Pecher, J. Simko, E. Stefancova, M. Kompan, A. Hrckova, J. Podrouzek, A. Gavornik, et al., Auditing youtube's recommendation algorithm for misinformation filter bubbles, ACM Transactions on Recommender Systems 1 (2023) 1–33.

[2] X. Zhang, A. A. Ghorbani, An overview of online fake news: Characterization, detection, and discussion, Information Processing & Management 57 (2020) 102025.

[3] S. Vosoughi, D. Roy, S. Aral, The spread of true and false news online, Science 359 (2018) 1146–1151. doi:10.1126/science.aap9559.

[4] X. Zhou, R. Zafarani, A survey of fake news: Fundamental theories, detection methods, and opportunities, ACM Comput. Surv. 53 (2020). doi:10.1145/3395046.

[5] M. Fernandez, A. Bellogín, Recommender systems and misinformation: the problem or the solution?, OHARS Workshop. 14th ACM Conference on Recommender Systems (2020).

[6] R. Pathak, F. Spezzano, M. S. Pera, Understanding the contribution of recommendation algorithms on misinformation recommendation and misinformation dissemination on social networks, ACM Trans. Web 17 (2023). doi:10.1145/3616088.

[7] J. Shin, L. Jian, K. Driscoll, F. Bar, The diffusion of misinformation on social media: Temporal pattern, message, and source, Computers in Human Behavior 83 (2018) 278–287.

[8] R. Iizuka, F. Toriumi, M. Nishiguchi, M. Takano, M. Yoshida, Impact of correcting misinformation on social disruption, Plos one 17 (2022) e0265734.

[9] M. Tomlein, B. Pecher, J. Simko, I. Srba, R. Moro, E. Stefancova, M. Kompan, A. Hrckova, J. Podrouzek, M. Bielikova, An audit of misinformation filter bubbles on youtube: Bubble bursting and recent behavior changes, in: Proc. of the 15th ACM Conference on Recommender Systems, 2021, pp. 1–11.

[10] K. Papadamou, S. Zannettou, J. Blackburn, E. De Cristofaro, G. Stringhini, M. Sirivianos, "it is just a flu": Assessing the effect of watch history on youtube's pseudoscientific video recommendations, in: Proc. of the international AAAI conference on web and social media, volume 16, 2022, pp. 723–734.

[11] A. Tommasel, F. Menczer, Do recommender systems make social media more susceptible to misinformation spreaders?, in: Proc. of the 16th ACM Conference on Recommender Systems, 2022, pp. 550–555.

[12] M. Fröbe, S. Günther, A. Bondarenko, J. Huck, M. Hagen, Using keyqueries to reduce misinformation in health-related search results, in: Proc. of the 2nd Workshop Reducing Online Misinformation through Credible Information Retrieval, 2022, pp. 1–10.

[13] Y. Huang, Q. Xu, S. Wu, C. Nugent, A. Moore, Fight against covid-19 misinformation via clustering-based subset selection fusion methods, in: 2nd Workshop Reducing Online Misinformation through Credible Information Retrieval, ROMCIR 2022, 2022, pp. 11–26.

[14] M. Badami, O. Nasraoui, Paris: polarization-aware recommender interactive system, in: Proc. of the 2nd Workshop OHARS, 2021, pp. 1–13.

[15] W. Sun, O. Nasraoui, User polarization aware matrix factorization for recommendation systems, in: Proc. of the 2nd Workshop on Online Misinformation-and Harm-Aware Recommender Systems (OHARS 2021), Amsterdam, Netherlands, 2021, pp. 58–67.

[16] M. Fernández, A. Bellogín, I. Cantador, Analysing the effect of recommendation algorithms on the amplification of misinformation, 2021. URL: https://arxiv.org/abs/2103.14748.

[17] G. Linden, B. Smith, J. York, Amazon. com recommendations: Item-to-item collaborative filtering, IEEE Internet computing 7 (2003) 76–80.

[18] P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom, J. Riedl, Grouplens: An open architecture for collaborative filtering of netnews, in: Proc. of the 1994 ACM conference on Computer supported cooperative work, 1994, pp. 175–186.

[19] Y. Koren, R. Bell, C. Volinsky, Matrix factorization techniques for recommender systems, Computer 42 (2009) 30–37.

[20] S. Rendle, C. Freudenthaler, Z. Gantner, L. Schmidt-Thieme, Bpr: Bayesian personalized ranking from implicit feedback, arXiv preprint arXiv:1205.2618 (2012).

[21] C.-K. Hsieh, L. Yang, Y. Cui, T.-Y. Lin, S. Belongie, D. Estrin, Collaborative metric learning, in: Proc. of the 26th international conference on world wide web, 2017, pp. 193–201.

[22] S. Funk, Netflix update: Try this at home, 2006. URL: https://sifter.org/~simon/journal/20061211.html.

[23] C. C. Johnson, et al., Logistic matrix factorization for implicit feedback data, Advances in Neural Information Processing Systems 27 (2014) 1–9.

[24] X. Luo, M. Zhou, Y. Xia, Q. Zhu, An efficient non-negative matrix-factorization-based approach to collaborative filtering for recommender systems, IEEE Transactions on Industrial Informatics 10 (2014) 1273–1284.

[25] A. Mnih, R. R. Salakhutdinov, Probabilistic matrix factorization, Advances in neural information processing systems 20 (2007).

[26] Y. Koren, Factorization meets the neighborhood: a multifaceted collaborative filtering model, in: Proc. of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, 2008, pp. 426–434.

[27] Y. Juan, Y. Zhuang, W.-S. Chin, C.-J. Lin, Field-aware factorization machines for ctr prediction, in: Proc. of the 10th ACM conference on recommender systems, 2016, pp. 43–50.

[28] H.-J. Xue, X. Dai, J. Zhang, S. Huang, J. Chen, Deep matrix factorization models for recommender systems., in: IJCAI, volume 17, Melbourne, Australia, 2017, pp. 3203–3209.

[29] X. He, L. Liao, H. Zhang, L. Nie, X. Hu, T.-S. Chua, Neural collaborative filtering, in: Proc. of the 26th international conference on world wide web, 2017, pp. 173–182.

[30] V. W. Anelli, A. Bellogín, A. Ferrara, D. Malitesta, F. A. Merra, C. Pomo, F. M. Donini, T. Di Noia, Elliot: A comprehensive and rigorous framework for reproducible recommender systems evaluation, in: Proc. of the 44th international ACM SIGIR conference on research and development in information retrieval, 2021, pp. 2405–2414.