NARPHI: News Articles Recommendation System for Public Health Intelligence

Diana F. Sousa[1,*], Luigi Spagnolo[1] and Nicolas Stefanovitch[1]

[1]European Commission Joint Research Centre, Ispra, Italy

Abstract

Several national and international institutions and organizations are committed to monitoring health news in order to derive Public Health Intelligence (PHI). Their primary objective, among others, is to identify emerging health threats. They achieve this by utilizing aggregator platforms such as the Epidemic Intelligence from Open Sources (EIOS) platform. However, their efforts are significantly hindered by the overwhelming amount of textual data, as they need to sift through more than 100 000 multilingual articles daily. To better tailor articles to different user interests (such as source and topics covered), we developed NARPHI, a recommendation system based on Graph Neural Networks (GNN). This system uses user and article features and past interactions (i.e., combined into ratings), recommending the most relevant articles for each user. Our initial findings with NARPHI show that the recommendations using the best combination of user community and article event type features align with user preferences.

Keywords

Recommender System, Health News Articles, Public Health Intelligence, Graph Neural Networks, Expert User Data

1. Introduction

National and international institutions and organizations dedicated to Public Health Intelligence (PHI) have a critical role in monitoring health news. Among their various tasks, they sift through thousands of articles to identify emerging health threats, such as disease outbreaks. They also monitor changes in vaccination policies, new medical developments, and different carriers of diseases, including humans, animals, and plants. As a result, various organizations may focus on the same or different topics and areas within public health.

To assist analysts in their work, aggregator platforms are designed to compile news articles, reports, and social media posts related to health topics. The most widely used platform globally is the Epidemic Intelligence from Open Sources (EIOS), created at the end of 2017. EIOS employs a comprehensive One Health approach to the early detection, verification, assessment, and communication of public health threats using publicly available information. This platform is led by the World Health Organization (WHO) and is the result of a long-standing collaboration between the WHO and the Joint Research Centre (JRC) of the European Commission (EC)[1].

The EIOS platform collects over 100 000 multilingual articles daily, making it extremely challenging to process such a vast amount of information for analysis. Users need to search for relevant information tailored to their specific domains and objectives. This situation creates a strong demand for automatic organization and prioritization based on individual needs. Implementing a user-based recommender system, a feature currently not available in EIOS, would significantly reduce the time analysts spend identifying articles of interest relevant to their goals.

To meet the demand for quicker identification of relevant articles on the EIOS platform, we developed NARPHI, a recommendation system for news articles focused on PHI. This system utilizes a Graph

Joint Proceedings of the ACM UMAP Workshops 2026, UMAP 2026, June 8–11, 2026, Gothenburg, Sweden

*Corresponding author.

diana.francisco-de-sousa@ec.europa.eu (D. F. Sousa); luigi.spagnolo@ec.europa.eu (L. Spagnolo); nicolas.stefanovitch@ec.europa.eu (N. Stefanovitch)

0000-0003-0597-9273 (D. F. Sousa); 0009-0008-0179-7468 (L. Spagnolo); 0009-0000-2061-3216 (N. Stefanovitch)

[1]https://www.who.int/initiatives/eios

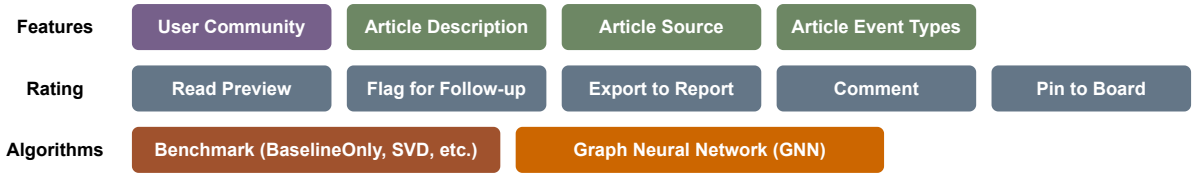

Figure 1: Main components of the tested implementations: user and article features, rating elements, and algorithms.

Table 1
EIOS dataset overview, All and with Dimensionality Setups (DS) presented as a combination of (*user*, *article*)

	All	DS 1 (2, 2)	DS 2 (2, 5)	DS 3 (2, 10)	DS 4 (5, 2)	DS 5 (10, 2)
Users	1 350	1 116	1 014	681	981	858
Items	1 752 791	380 931	45 224	1 799	380 931	380 930
Ratings	3 243 371	1 526 923	316 231	23 383	1 526 611	1 526 035
Average Ratings	1.0900	1.1410	1.3258	1.9555	1.1407	1.1404
Sparsity	99.86 %	99.64 %	99.31 %	98.12 %	99.59%	99.53 %

Neural Network (GNN) built on various factors, including article and user features, past user interactions with the articles, the types of interactions (e.g., comments in an article), and user communities—reflecting the domains of interest for new users (i.e., addressing the cold start issue).

The article's main contributions are:

- Establishing a rating procedure that considers various types of user interactions.
- Tackling the cold start problem in recommendation systems by utilizing insights from user communities.
- Developing NARPHI, a GNN-based recommendation system for news articles focusing on PHI.

The data used in this paper was sourced from a live system. As a result, intellectual property and privacy regulations apply, which prevent the sharing of the dataset. Nevertheless, the experiments described in this article are significant for health recommender systems, providing valuable insights into implementing AI-based solutions using real user data.

Section 2 provides an overview of related work, focusing on previous implementations of health recommender systems and filtering applications designed for health news aggregator platforms. Section 3 outlines the experimental setup, detailing the data used and the specifications and adjustments made during training. In Section 4, NARPHI's first results are presented along with a discussion of the recommendation pipeline and of the various system configurations. Finally, Section 5 summarizes the main conclusions and outlines directions for future work.

2. Related Work

There are several health news aggregator platforms, both open-access and subscription-based, that focus on epidemic intelligence and early warning of public health threats, in addition to EIOS. Some of the resources from these platforms are already integrated into EIOS. For instance, the Global Public Health Intelligence Network (GPHIN)[2] and ProMED-mail[3] provide data on global news reports to identify potential public health threats and information regarding infectious disease outbreaks. HealthMap[4] [1]

[2]https://gphin.canada.ca/cepr/articles.jsp?language=en_CA
[3]https://promedmail.org/
[4]https://www.healthmap.org

and EPIWATCH[5] are other global disease surveillance platforms that collect information from multiple sources to provide real-time updates on emerging public health threats.

Many of these platforms provide resources for data analytics and various filtering options. However, they all face a common challenge similar to EIOS: an overwhelming amount of information. This issue is worsened by the increasing number of sources being added daily. Although different strategies are being developed to manage this influx, recommender systems emerge as a natural solution for these platforms. They offer the flexibility to tailor content to individual user interests based on a range of features. The initial implementation of a recommender system in EIOS was limited because it only took into account aggregated user interactions for each article and relied solely on content-based methods [2].

3. Experimental Setup

This section presents the data and training details for the tested recommender system implementations. Figure 1 showcases the main components, including user and article features, rating elements, and algorithms.

3.1. Data

The EIOS dataset contains over 1.75 million articles, 1 350 users, and more than 3.2 million ratings from the EIOS platform from September 2023 to March 2024 ($\approx$ 7 months). This dataset contains all articles and information related to user interactions with those articles in various languages as captured by the platform. The data features analyzed include user and article attributes, such as community (users) and event type labels, descriptions, and sources (articles). Detailed descriptions of all features are provided in the following sections.

Among the available articles, 1 371 860 were only interacted with once by a single user. To reduce data sparsity and avoid memory errors, we filtered out articles with very few ratings and users who rarely rated articles in five configurations where we set the minimum number of ratings per article or per user to be 2, 5, or 10. Table 1 showcases the six different setups considering *All* of the data and the five *Dimensionality Setups (DS)*.

3.1.1. Users

The users mainly consist of expert analysts in the PHI domain who are part of organizations such as the European Centre for Disease Prevention and Control (ECDC), the Food and Agriculture Organization of the United Nations (FAO), and the WHO.

User engagement varies significantly; some are highly active, logging over 100 000 interactions with articles over the years, while others have interacted only with a single article. These different levels of engagement can be attributed to varying onboarding procedures for stakeholders. Some users are automatically integrated into the platform, even if they do not actively use it, while others join on an ad hoc basis.

As user features, we have information about the community to which each user belongs (e.g., FAO). This information is particularly important for addressing the cold-start issue. By knowing a user's community, we can make educated assumptions about their interests and tailor our recommendations accordingly.

3.1.2. Items

The dataset encompasses a total of 1 752 791 articles in 80 languages with metadata regarding description, source, and event types:

[5] https://www.epiwatch.org/

- **Description**: The article description coincides with the initial 250 tokens of an article, focusing on complete sentences. This excerpt typically encapsulates the article's main ideas. The description is provided in the article's original language.
- **Source**: The article's source refers to its origin, such as a specific news outlet, social media platform, or organizational website.
- **Event Types**: The articles' event types refer to the application of an event type classifier that assigns labels such as *Reporting Cases* and *Displacement of People*, along with 27 other detailed options [3]. Each article can be assigned no label or multiple labels. Users can then filter by their preferred event types before browsing the content.

We utilized all three data streams as features for the articles. We experimented with different combinations, such as omitting one or two features, to evaluate the influence of various features on the system's recommendations.

3.1.3. Ratings

The rating that each user assigns to an item (such as an article) is calculated using a weighted sum of their interactions with that particular item. The EIOS platform allows for various types of interactions, including:

- **Read Preview**: The user reads the preview information of the article.
- **Flag for Follow-up**: The user marks the article to revisit it at a later time.
- **Export to Report**: The user finds the article relevant for an ongoing report and adds it to the material to be considered.
- **Comment**: The user provides comments on the article.
- **Pin to Board**: The user pins the article to a private or public board related to the article's topic.

Table 2
Scoring System for Each Type of User Interaction

User Interaction	Score	Repeatable
Read Preview	1	No
Flag for Follow-up	3	No
Export to Report	5	Yes
Comment	5	No
Pin to Board	5 or 10	Yes

Table 2 illustrates each type of user interaction and the corresponding score based on their level of engagement. Interactions requiring higher levels of engagement receive higher scores. Specifically, the *Pin to Board* interaction can score either 5 or 10, depending on whether the user pins the article to a private or public board, respectively.

The repeatable actions *Export to Report* and *Pin to Board* are interactions that users may perform multiple times for the same item. This repeatability suggests that the item holds increased significance for the user; for instance, the more frequently a user exports an item to various locations, the more important that item becomes to them. Repeatability is calculated as the sum of a user's repeatable interactions with a specific item. If s_{ui} represents the score associated with a particular user interaction, then the repeatable action score r_{ui} can be defined as $x \times s_{ui}$ where x stands for the number of repeats.

For the user interaction labelled as *Comment*, we did not consider the action to be repeatable, even though it is possible for a user to write multiple comments on the same item. This is due to the diverse ways in which users choose to comment. For instance, some users may comment once with all the relevant information in a single text block, while others may prefer to share that same information across multiple comments for clarity or emphasis. Given these varying styles, we decided to count the act of commenting only once for any given user on any specific item.

3.2. Training

This section presents the models tested, the evaluation metrics used, and the details of the overall implementation of the NARPHI system.

3.2.1. Models

As benchmark collaborative filtering algorithms (vanilla models), we tested basic algorithms (NormalPredictor and BaselineOnly), k-NN inspired algorithms (KNNBasic, KNNWithMeans, kNNWithZScore, and KNNBaseline), matrix factorization-based algorithms (SVD, SVDpp, and NMF), Slope One, and Co-clustering (all available in the surprise package[6]).

Due to the high density of item features and data sparsity, we chose GNNs to address our hypothesis following the testing of the benchmark algorithms. These networks have proven effective in identifying and capturing the relationships among users, items, and their associated features [4]. Additionally, GNNs have the potential to provide explainability, clarifying the decision-making process. Therefore, GNN-based methods are a suitable choice for this health-related task. We used a sentence BERT-based model *distiluse-base-multilingual-cased-v2* [5] to encode the articles' descriptions before applying the GNN. For the network to learn the connections between users and articles, it required mapping user features to article features and a representation of the edges between articles and users. The graph data structure is then passed to the neural network. In this case, it is a heterogeneous graph, as both users and articles are nodes.

To run the different GNNs combinations of features, all ratings were scaled to fit a predefined range of $[t_{min}, t_{max}] = [0, 10]$, based on the original range of ratings $[r_{min}, r_{max}] = [1, 101]$. The original distribution of these ratings did not follow a normal distribution, which hindered convergence during training. To address this, we employed the following equation to scale each rating:

$$rating_{scaled} = \frac{rating - r_{\min}}{r_{\max} - r_{\min}} \times (t_{\max} - t_{\min}) + t_{\min} \tag{1}$$

Table 3
Benchmark Algorithms tested over Non-normalized Data and the first two Dimensionality Setups (Table 1) using the RMSE Metric

Algorithm	DS 1	DS 2
BaselineOnly	**1.1730**	**1.8701**
CoClustering	1.4489	2.2466
SlopeOne	*	2.1238
KNNBaseline	1.3038	1.9820
KNNBasic	1.3770	2.0672
KNNWithMeans	1.3037	1.9843
KNNWithZScore	1.4276	2.1534
SVD	1.1939	1.8925
SVDpp	*	1.9118
NMF	1.3393	2.0745
NormalPredictor	1.6338	2.5150

* The algorithm could not run due to memory limitations.

3.2.2. Evaluation

The performance of the NARPHI system was assessed using two metrics: Root Mean Squared Error (RMSE) for benchmark algorithms and Mean Average Precision (MAP) for the GNN-based model. RMSE,

[6] https://surprise.readthedocs.io/

Table 4
NARPHI System tested over different combinations of Features (User Community (UC), Article Description (AD), Article Source (AS), and Article Event Types (AE)), Normalized Data, and five Dimensionality Setups (Table 1) using the MAP@10 Metric

Features	DS 1	DS 2	DS 3	DS 4	DS 5
UC	0.0996	0.1027	0.1070	0.0939	0.0963
AD	0.0960	0.1023	0.1264	0.1040	0.0977
AS	0.1029	0.1111	0.1219	0.1076	0.1086
AE	0.1136	0.1004	0.1503	0.1086	0.1147
UC + AD	0.0956	**0.1382**	0.1236	0.0981	0.1113
UC + AS	0.0916	0.1031	0.1303	0.1049	0.1119
UC + AE	0.0933	0.1015	0.1333	0.1137	**0.1215**
A(DS)	0.0957	0.0945	0.1230	0.1024	0.1177
A(DE)	**0.1152**	0.0964	0.1309	0.1057	0.0925
A(SE)	0.0984	0.1111	0.1160	0.1145	0.1124
UC + A(DS)	0.1037	0.1284	0.1073	0.1095	0.0953
UC + A(DE)	0.1022	0.1120	0.1162	**0.1147**	0.1011
UC + A(SE)	0.1082	0.1277	**0.1644**	0.0955	0.1017
A(DSE)	0.1065	0.1195	0.1214	0.0990	0.0827
UC + A(DSE)	0.1114	0.1102	0.1193	0.1093	0.0964

also known as root mean squared error, is a widely used measure for evaluating the quality of predictions. Using Euclidean distance indicates how far the predictions deviate from the actual measured values. On the other hand, MAP is a ranking quality metric that considers the number of relevant recommendations and their positions in the list. We calculated MAP at K as the Average Precision (AP) arithmetic mean at K across all users or queries. We did not use RMSE for the GNN-based models because we modified the original rating distribution for training, which made the comparison unreliable.

3.2.3. Implementation Details

In our experiments, beyond the benchmark algorithms described for which we did not do any hyperparameter tuning, we used a GNN model featuring SageConv layers for encoding graph features and linear layers for edge prediction, configured with 128 hidden units for a balance between complexity and efficiency. The model was trained with the Adam optimizer at a 0.01 learning rate for up to 300 epochs, employing early stopping to prevent overfitting. Computations were performed on an Nvidia Quadro RTX 5000 GPU to manage the model's complexity. This setup was chosen to optimize accuracy and efficiency while ensuring reproducibility, with room for further tuning to enhance future performance.

4. Results and Discussion

Table 3 presents the results of applying our first two data dimensionality setups (Table 1) to the benchmark collaborative filtering algorithms (vanilla models). It showcases the BaselineOnly algorithm [6] as being most suitable for non-normalized data; this algorithm predicts the baseline estimate for a given user and item. We interpret this performance as an objective indicator of the sparsity of our data that does not respond well to the other, more intricate benchmark algorithms, which informed our decision to test more dimensionality setups, perform data normalization, and use GNNs to capture the complex interactions between users and items.

Table 4 reports the results for our GNN-based approach, the NARPHI system, using MAP@10. We tested all possible combinations of features from the four available: User Community (UC), Article Description (AD), Article Source (AS), and Article Event Types (AE). These results represent the initial exploration of a GNN-based setup for the system. The optimal combination of features identified was UC + AS + AE for DS 3. Overall, the trend indicated that including the AS feature negatively impacted

performance, while the UC and AE features contributed positively. We believe this trend highlights the AS feature as an unbiased choice of sources by the analysts, while UC and AE are a more unambiguous indication of user preference.

5. Conclusion

Aggregator news health platforms for public health intelligence (PHI) are essential tools for hundreds of analysts worldwide. EIOS is the most commonly used example, as reflected in the volume of articles arriving on the platform daily and in multiple languages. This constant flow of information creates a need for effective filtering methods that can adapt to the requirements of the various communities using these platforms.

In this study, we introduced NARPHI, a graph neural network (GNN)-based system that considers four types of features for article recommendation. Our assessment identified that a combination of user community and article event type-specific features was the most effective.

There are still many avenues to explore to enhance NARPHI's performance. These include incorporating additional metadata, analyzing cross-community interests, experimenting with different model setups, and conducting a rigorous user-based evaluation of the system's implementation. Future work will focus on these areas.

Declaration of Generative AI During the preparation of this work, the authors used Grammarly to check grammar and spelling. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] C. C. Freifeld, K. D. Mandl, B. Y. Reis, J. S. Brownstein, Healthmap: global infectious disease monitoring through automated classification and visualization of internet media reports, Journal of the American Medical Informatics Association 15 (2008) 150–157. doi:10.1197/jamia.M2544.

[2] D. F. Sousa, N. Stefanovitch, L. Spagnolo, Recommending news articles for public health intelligence, in: Proceedings of the 6th International Workshop on Health Recommender Systems co-located with 18th ACM Conference on Recommender Systems (HealthRecSys'24), CEUR-WS, Bari, Italy, 2024, pp. 18–25.

[3] J. Piskorski, N. Stefanovitch, J. P. Linge, S. Kharazi, J. Mantero, G. Jacquet, A. Spadaro, G. Teodori, Multi-label infectious disease news event corpus, in: Proceedings of the Text2Story'23 Workshop, Elsevier, Dublin, Republic of Ireland, 2023, pp. 171–183.

[4] A. R. Mohammadi, Explainable graph neural network recommenders; challenges and opportunities, in: Proceedings of the 17th ACM Conference on Recommender Systems, RecSys '23, Association for Computing Machinery, New York, NY, USA, 2023, p. 1318–1324. URL: https://doi.org/10.1145/3604915.3608875. doi:10.1145/3604915.3608875.

[5] N. Reimers, I. Gurevych, Sentence-BERT: Sentence embeddings using Siamese BERT-networks, in: K. Inui, J. Jiang, V. Ng, X. Wan (Eds.), Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 3982–3992. URL: https://aclanthology.org/D19-1410/. doi:10.18653/v1/D19-1410.

[6] Y. Koren, Factor in the neighbors: Scalable and accurate collaborative filtering, ACM Transactions on Knowledge Discovery from Data (TKDD) 4 (2010) 1–24.