Recommending News Articles for Public Health Intelligence

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

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

Abstract

Public Health Intelligence (PHI) is the process of extracting useful information from vast amounts of data to help quickly identify and respond to health threats. Systems that perform PHI are used daily by different national and international organizations. One of the most prominent platforms is the Epidemic Intelligence from Open Sources Initiative (EIOS) platform, which continuously gathers health-related news items. However, the EIOS platform requires users to swift through unrelated information to their domain or work needs, even when using different filtering options. This inefficiency in assessing the relevance of each article creates the need to develop a recommender system that effectively positions each incoming article according to its significance. In this work, we present the first iteration of this system, making use of previous user interactions with the articles already available in the platform and the articles' content and metadata. We investigated various configurations to address the problem of data sparsity by conducting cluster-based harmonization. Our best-performing model reports an NDGC@K of 0.4108 and an F-measure@K of 0.7287, respectively, for $K = 100$ articles.

Keywords

Public Health Intelligence, Recommender Systems, Clustering, User Data, Health News Articles

1. Introduction

Every day, expert analysts swift through tens of thousands of health news articles to identify incoming health threats, such as an outbreak of a disease and other types of relevant health information regarding humans, animals, and plants. To do their work, the analysts use platforms that primarily aim to gather all news articles and reports on health topics. The Epidemic Intelligence from Open Sources (EIOS) platform is the most well-known Public Health Intelligence (PHI) resource. EIOS is an international initiative led by the World Health Organization (WHO) with a unified all-hazards One Health approach to early detection, verification, assessment and communication of public health threats using publicly available information[1].

The analysts working on identifying relevant health information for each of their purposes and domains have to carry out their day-to-day work and often prepare for large mass gatherings, e.g. sports championships or the Olympics games, which present an increased risk of disease outbreaks. Thus, analysts face the daily challenge of processing a high volume of information. EIOS collects 50,000 articles a day; as such, the possibility to organise information by relevance using a recommender system, a feature currently missing in EIOS, would improve analysts' experience by significantly alleviating the time spent identifying which articles are relevant for their purpose.

Health recommender systems are broad and encompass epidemic forecasting tools such as HealthMap [1] and EPIWATCH[2], which track disease spread by collecting information from various channels, including news and social media [2]. In crises, these recommender systems are pivotal for effectively allocating medical resources and guiding interventions. Moreover, they extend to environmental health monitoring, offering air and water quality advice, and are integrated into Personal Health Records

HealthRecSys'24: The 6th Workshop on Health Recommender Systems co-located with ACM RecSys 2024

*Corresponding author.

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

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

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

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

(PHRs) to suggest health actions [3, 4], such as vaccine recommendation features [5]. Lastly, health applications employ these systems to promote personalized health-related behaviour [6, 7]. Despite their potential, ensuring data privacy, system validation, multilingual adaptability, and ethical use is paramount for maintaining public trust and successfully deploying recommender systems in public health.

To address the need for more efficient identification of relevant articles coming to the EIOS platform, we created a content-based recommender system that is based on three data streams: (1) The content of the article, specifically the first 1000 characters, taking into account complete sentences; (2) The event type labels resulting from the application of a pandemics event classifier; (3) The user interactions with each article (i.e., relevance score), obtained using a scoring function that considers the type and number of interactions, augmented with a clustering procedure to tackle data sparsity. We tested XGBoost [8] with seven different data augmentation procedures.

The article's main contributions are:

- Usage of an event classifier labels to enrich the recommendation algorithm;
- Introducing a clustering-based approach for user activity harmonization to address data sparsity challenges;
- Development of a content-based system for recommending articles in real-world PHI scenarios.
- Error analysis conducted on example use cases to assess whether the recommender can flag relevant information missed by the users.

The data described and used in this paper was sourced from a live system. As a result, Intellectual Property and Privacy regulations apply, preventing dataset sharing. Nevertheless, the experiments detailed in this article are significant for health recommender systems. They offer valuable insights into implementing AI-based solutions using actual user data.

Section 2 describes the data, mainly the metadata used to train the recommender system. Section 3 describes the cluster-based procedure to perform data harmonization and tackle sparsity. Section 4 presents the recommender system, including model and evaluation metrics. Section 5 presents results, a discussion of the clustering plus recommendation pipeline, and an error analysis of the different clustering modalities. Finally, Section 6 presents the main conclusions and future work.

2. Data

To train and test our model, we used a dataset of approximately 3.5 million articles from the EIOS platform from 01/01/2018 to 09/06/2022 (about four years and six months). This dataset contains all articles and information about user interactions with those articles in all the different languages captured by the platform. For this work, which constitutes the first iteration to create a recommendation solution for PHI systems, the features we focused on are the text of the article, the event labels generated through an event classifier, and the user activity for each article (i.e., relevance score). Figure 1 illustrates the high-level pipeline involving three input data streams in the recommender system.

2.1. Text

The dataset has the full text for each article. However, due to memory limitations and to keep the focus on the core information of the article, we decided to consider only the first few sentence(s), up to 1000 characters.

To preprocess this truncated-article text, we only removed stop words from English articles. In order to vectorise the articles, we used the TfidfVectorizer function from the scikit-learn[3] using the maximum document frequency set to ignore terms that have a document frequency strictly higher than 1.

[3] https://scikit-learn.org/

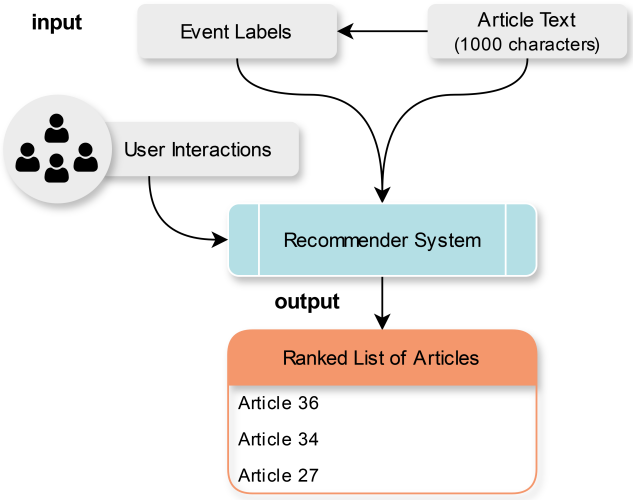

Figure 1: Recommender system high-level pipeline with the three data streams and expected output.

2.2. Event Labels

We assigned event labels to the articles to boost the system's performance and better characterize and differentiate between articles. We ran an event classifier for each article within the dataset to classify them into one or more of 27 events following a taxonomy and pipeline created and developed by Piskorski et al. [9]. Some of the most frequent labels are (1) Reporting Cases (i.e., reporting on cases of infections, hospitalizations, deaths, recoveries of single persons and groups, provision of updates thereon, which covers a short time span and specific location), (2) Reporting Situation (i.e., provision of updates on the overall situation of the outbreak, current total figures, observed trends, forecast, which spans longer period of time, and also covers cross-regional and cross-country comparisons), (3) Measuring Vaccine/Medicine Roll-out (i.e., covers events revolving around the roll-out of vaccines, medicines, equipment to combat the disease or mitigate the consequences, and includes also events related to sharing experience, measure hesitancy, anti-vax movements, etc.). Other coarse-grain labels are Impact, Violation, Research & Development, Communication, Support, and Miscellaneous.

To preprocess these event labels, we applied the MultiLabelBinarizer function, given that each article can have more than one label wrapped to work with ColumnTransformer, both from the scikit-learn library.

2.3. User Activity

The user activity for each article is pre-determined by the weighted sum of user interactions, which we express as a relevance score. Different types of interactions yield different weights. The platform computes the user activity using the weights presented in Table 1.

Table 1
Types of user activities and their corresponding weights.

User Activity	Weight
Read Preview	0 or 1
Read Detail	0 or 2
Flag for Follow Up	3
Export to Report	5
Attach to Team Communication	5
Comment	5
Pin to Board	Variable

When it comes to the "Read Preview" interaction, the weight assigned will be zero if there are no other user interactions on the article, and one otherwise (excluding "Read Detail"). For the "Read Detail" interaction, the weight assigned will be zero if there are no other user interactions on the article, and two otherwise (excluding "Read Preview"). As for the "Pin to Board" activity, the weight assigned is five or ten, based on whether the board is private or public, respectively. The weights assigned to each activity are proportional to the complexity of the activity being performed.

One of the issues we had to address before the application of our system was the low proportion of articles with user interactions (2.03%). The news feeds presented to users are ordered by time and user preference settings (i.e., pre-determined keywords, languages, etc.). When a new story emerges, EIOS users often interact with the first article reporting on the story, with the article they deemed to be from the most reliable source, or even with the article that reports the story in their language, among other preferences.

This interaction pattern means that if we have a single story reported in multiple articles from multiple sources, the user activity will vary widely among almost identical articles, with only a few articles getting interacted with. Thus, raw user activity does not directly equate to user interests. In the following section, we will outline how we intend to tackle this issue using clustering to make the relevance score a reliable measure of user interest.

3. Cluster-based Harmonization

We considered that articles with no interaction are articles for which the relevance is unknown rather than zero, transforming the problem into a semi-supervised learning one. We corrected the relevance score of articles in clusters to deal with this and fall back on a supervised learning problem.

The harmonization of user activity/relevance scores happens at the level of clusters of related articles, some of which have an interaction score and others potentially none. We intended that the clusters captured reports on the same event; as such, they were computed considering both the time and semantic aspects. The clustered article data corresponds to the text described in the Data section. The entire dataset was split into five-day chunks, capturing a story's average duration, as represented in Figure 2. Inside a chunk, all the pairs of articles were compared using sentence embeddings, and the pairs whose similarity was above a given threshold were put into a graph. The semantic similarity model used was `distiluse-base-multilingual-cased-v2`, with a threshold of 0.90. Finally, the graphs of all clusters were merged, and the set of connected components yielded the global set of clusters. This approach is designed to be adaptable, allowing it to pick up news stories that last longer than five days and preventing the merging of similar stories from widely different time spans.

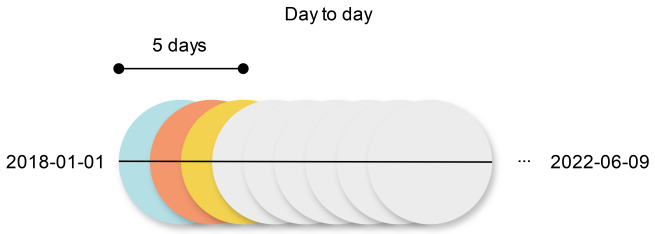

Figure 2: Representation of five day time-span local clusters in the timeframe considered.

Once the clusters were computed, the second step of our procedure was to harmonize the score of all the articles belonging to each cluster. To illustrate this, we will consider this example cluster of four identical articles and their corresponding user activities scores:

- **Cluster**: [Article 1, Article 2, Article 3, Article 4]
- **User Activities**: [0, 5, 17, 0]

Table 2
Counts of the number of articles per modality, their percentage with user activity, and the threshold used for predictions.

Modality	Number of Articles	User Activity	Threshold
Original	3 589 739	2.03	9
Sum	3 589 739	2.25	10
High	3 589 739	2.25	9
AVG	3 589 739	2.22	9
Low	3 589 739	2.20	9
Random	3 589 739	2.22	9
Discard	3 288 085	2.22	9
Null	3 287 754	2.20	9

Clusters containing articles with only zero relevance are left untouched, except for the Null configuration, detailed below. Clusters with mixed or only positive relevance were further processed to reassign the relevance score of every article within that cluster. We considered seven different modalities to perform the harmonization, which are illustrated in the following example:

- **Original**: Nothing changes → [0, 5, 17, 0].
- **Sum**: Application of the sum of all user activities in the cluster to all the articles in the cluster → [22, 22, 22, 22].
- **High**: Application of the highest user activity in the cluster to all the articles in the cluster → [17, 17, 17, 17].
- **Average**: Application of the average of all user activities computed by dividing the sum of all user activities by the number of articles in the cluster → [5.5, 5.5, 5.5, 5.5].
- **Low**: Application of the lowest user activity in the cluster to all the articles in the cluster → [5, 5, 5, 5].
- **Random**: To each cluster, application of a random configuration from the ones described above → [22, 22, 22, 22] or [17, 17, 17, 17] or [5.5, 5.5, 5.5, 5.5] or [5, 5, 5, 5].
- **Discard**: Keep only articles in the cluster that have user activity → [5, 17].
- **Null**: Remove clusters where there is no article with user activity → [0, 5, 17, 0].

The Discard and Null modalities constitute filtering options, not modifying the relevance score but excluding articles with no score, using different approaches. For Discard, all non-relevant articles are removed from the cluster for the clusters with at least one relevant article. For Null, all clusters where all the articles have a zero relevance score are removed.

Table 2 showcases the augmentation in general percentage for each modality compared to Original, reflecting our extremely conservative clustering procedure. The Threshold column is the user activity value considered at the recommendation level to decide if an article should be recommended. We obtained this value by considering the average of the positive (> 0) user activities for each modality. Figure 3 reports the histogram of the user activity/relevance score of articles comparing the distribution of all the original data and the clustered articles' distribution of the sum modality, presenting similar profiles.

4. Recommender System

The data available does not specify which users interacted with the articles; it only shows the overall user activity for each article. Therefore, recommendations are not based on individual user behaviour but on global preferences towards specific topics and domains, making adopting a collaborative filtering approach unfeasible.

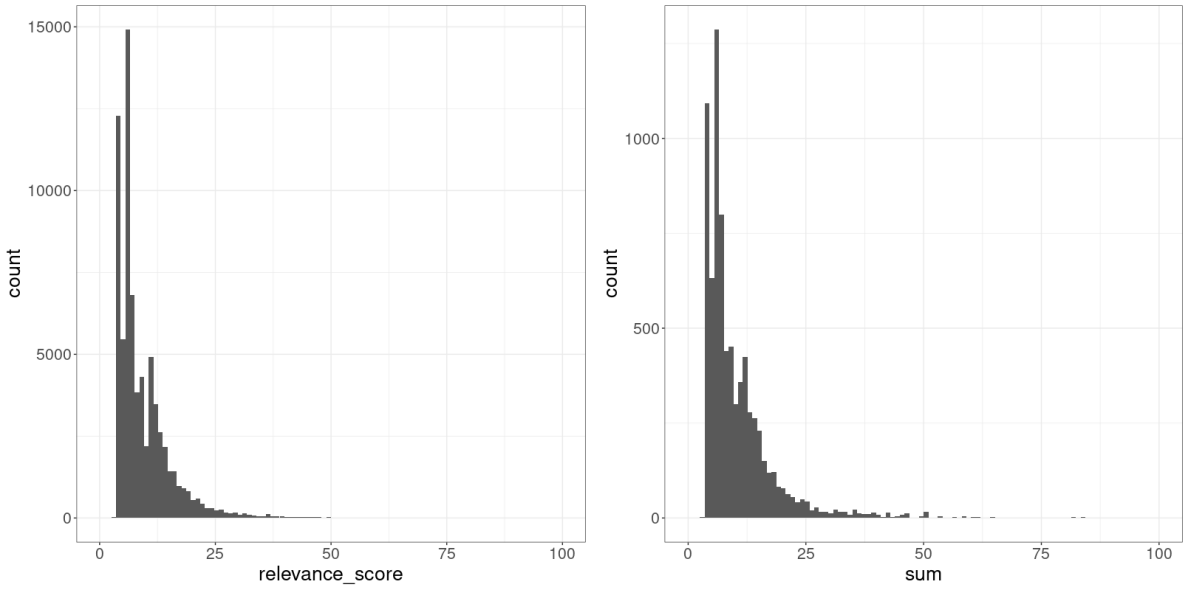

Figure 3: Histogram of relevance score distribution excluding non-relevant articles: original for all articles (left) and with sum harmonization (right).

4.1. Model

In this approach, each row of our data represented an article with a relevance score corresponding to the weighted sum of user interactions with the article. As stated in the previous sections, the features considered for training were the article attributes: a text section at the beginning of the article, the events labels that report on the article classification, and the relevance scores. Our goal was to recommend articles with higher engagement that are, therefore, more relevant.

We divided our data into training (80%) and testing (20%) with a 5-fold cross-validation. For the training data, we used an XGBoost regression model [8]. This model learns to predict each article's user engagement by building a series of decision trees sequentially, using gradient descent to minimize the loss. We did not do hyperparameter tuning, leaving the default parameters stated in the package documentation[4], to avoid overfitting the model to our data and maintain its generalizability to new data.

4.2. Evaluation Metrics

The evaluation metrics considered for the different settings were the following:

- **RMSE**: Root mean square error (RMSE) or root mean square deviation is one of the most commonly used measures for evaluating the quality of predictions. It shows how far predictions fall from measured true values using Euclidean distance.
- **NDGC@K**: Normalized Discounted Cumulative Gain (NDCG) considers both the relevance and the position of items in the ranked list in the top K items.
- **Precision@K**: Precision at K measures the proportion of relevant items among the top K items.
- **Recall@K**: Recall at K measures the coverage of relevant items in the top K items.
- **F-measure@K**: Harmonizes precision and recall to provide a balanced metric in the top K items.

We considered 5, 10, 15, and 100 items for K. For Precision, Recall, and F-measure, since the values considered are binary, we present only the $K = 100$ configuration to reflect better the real user needs in our setting.

[4] https://xgboost.readthedocs.io/en/stable/parameter.html

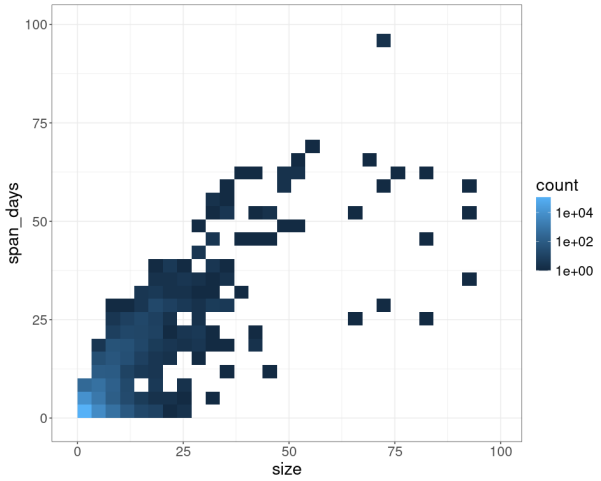

Figure 4: Heathmap of cluster size versus cluster span

Table 3
Statistics over clusters characteristics with different distributions.

	Only 0	Only *pos*	More 0	More *pos*	Eq. prop.
Count	124332	829	1188	182	3798
Max size	177	6	180	21	10
AVG size	2.3	2.1	5.3	3.3	2.0
Max span	133	7	244	28	12
AVG span	1.9	1.8	4.5	3.0	1.8
Max peak	25	6	11	4	4
% total rel.	0.00	0.24	0.18	0.06	0.52

5. Results and Discussion

This section presents the main results regarding all modalities and discusses the model's successes and potential limitations given the simplified approach.

5.1. Local Clusters Distribution

The settings used for clustering were conservative as it was performed on relatively long text with a high threshold. In total 8.7% of the articles were clustered. The data revealed a predominant pattern of small clusters, with 81% having a size of 2 and 99% under size 7. These clusters also tend to be short-lived, with 49% lasting a single day and 99% up to 8 days. The manual review confirms that articles in these clusters are remarkably similar, often being near-perfect duplicates. Notably, the clusters with the longest lifespan appear to be populated by automatically generated reporting articles.

In Figure 4, we plotted the distribution of cluster size and the distribution of the span of the cluster in days; some outliers fall outside the limits of the figure and are not shown. A cluster's median size was two articles, and the median span was two days. Table 3 reports several statistics over the clusters, grouping them based on whether the relevance of related articles contains only 0, only positive (*pos*), mostly 0, mostly *pos*, both 0 and *pos* in equal proportion. We report the mean and max cluster size and span, and the maximal peak article count, and the proportion of the total relevance. We can observe that clusters attracting most of the relevance tend to be relatively small and short.

5.2. Modality Performance

Table 4 presents the results of comparison of different clustering modalities for user data augmentation using the RMSE, NDGC@K, Precision@K, Recall@K and F-measure@K metrics, taking into account

Table 4
Comparison of different clustering modalities for user data augmentation using the RMSE, NDGC@K, Precision@K, Recall@K and F-measure@K metrics.

Modality	RMSE	NDGC@K				Precision@K	Recall@K	F-measure@K
		5	10	15	100	100	100	100
Original	1.5739	0.1903	0.1807	0.1606	0.1749	0.3466	1.0000	0.5136
Sum	1.7362	**0.4946**	**0.4382**	**0.4137**	**0.4108**	**0.5740**	1.0000	**0.7287**
High	1.6591	0.1722	0.2283	0.2318	0.2537	0.4320	1.0000	0.6015
AVG	**1.5313**	0.1612	0.1575	0.1559	0.1734	0.3478	0.9946	0.5137
Low	1.6188	0.1767	0.1950	0.1903	0.2201	0.4060	1.0000	0.5762
Random	1.6380	0.1622	0.2015	0.2023	0.2516	0.4440	1.0000	0.6139
Discard	1.6381	0.1516	0.1518	0.1377	0.1802	0.3880	1.0000	0.5575
Null	1.6318	0.1861	0.1884	0.1855	0.1834	0.3720	1.0000	0.5420

Table 5
Comparison of different clustering modalities with user data augmentation performance on the original test set (non-augmented) using the RMSE, NDGC@K, Precision@K, Recall@K, and F-measure@K metrics.

Modality	RMSE	NDGC@K				Precision@K	Recall@K	F-measure@K
		5	10	15	100	100	100	100
Sum	1.5794	**0.2259**	0.1878	0.1822	**0.2014**	0.3640	1.0000	0.5330
High	1.5718	0.1678	**0.2052**	**0.1828**	0.1864	0.3280	1.0000	0.4929
AVG	1.5732	0.1599	0.1542	0.1531	0.1660	0.3417	0.9944	0.5073
Low	**1.5715**	0.1535	0.1678	0.1627	0.1882	0.3520	1.0000	0.5195
Random	1.5719	0.1369	0.1627	0.1634	0.1966	**0.3880**	1.0000	**0.5578**
Discard	1.5752	0.1516	0.1518	0.1377	0.1548	0.3240	1.0000	0.4886
Null	1.5741	0.1861	0.1884	0.1788	0.1728	0.3460	1.0000	0.5139

5-fold cross validation.

Most modalities surpass the Original configuration. However, when considering NDGC@100, only Sum, High, Low, and Random perform distinctly better than the Original, with Sum being significantly better. The performance of Sum places the possibility that the actual user activity value represents the sum of all identical article interactions, performing twice as well as the Original.

Table 5 showcases the same procedure but using the Original modality test set. In this setting, the superior performance of the Sum modality is not as noticeable, but all modalities, except AVG, Discard, and Null, perform better than Original. A possible justification for this behaviour could be that our system performs better with more data regardless of how it is labelled, hindering the performance of Null and Discard modalities. Additionally, the AVG configuration could make stronger and weaker signals less noticeable, diluting their relative importance in a ranking setting.

5.3. Error Analysis

Table A1 (Appendix) showcases the false positives found across the five rounds of cross-validation for the different modalities at the top five ($K = 5$). All modalities introduce errors compared to the Original, with Sum and High introducing fewer wrong articles as also reflected in Table 4.

We analysed the articles for a fail rate of over or equal to 7/8 modalities to interpret what could have made most modalities assign relevance. We then analysed whether it was indeed a failure by our models or if it could have been a missed relevant article by the users and/or the clustering procedure for data augmentation. This selection resulted in six articles represented in Table 6 and marked with a asterisk ($*$) in Table A1 (Appendix). Table 7 reports on the details of these articles.

Even though Table 7 does not report on the sources for the articles, all of these are pieces that

Table 6
Most frequent articles and their scores across all modalities represented in the Top five ($K = 5$) false positives (fail rate $\geqslant 7/8$).

Article	Original	Sum	High	AVG	Low	Random	Discard	Null
1757749	51	46	34	41	46		51	41
210702	52	46	46	52	52	46	52	52
1177084	36		36	36	36	36	36	39
1642976	42		42	42	42	42	42	42
2083725	33		45	39	46	46	46	46
458168	68	64	64	68	68	64	68	68

Table 7
Most frequent articles across all modalities represented in the Top five ($K = 5$) false positives, their event labels, and the general topic they discuss (fail rate $\geqslant 7/8$).

Article	Event Labels	General Topic
1757749	REPORTING-CASES	First-case reporting on coronavirus in Africa
210702	MISCELLANEOUS-UNRELATED	Paediatric acute hepatitis reporting in the UK
1177084	IMPACT-OTHER	Coronavirus impact on the industry in India
1642976	MISCELLANEOUS-OTHER	Political landscape in Haiti
2083725	REPORTING-CASES	Coronavirus cases reporting in France
458168	REPORTING-SITUATION, REPORTING-CASES	Vaccination fears and chickenpox cases in Angola

primarily reflect the general opinion of an isolated expert of the respective fields and not official sources from health organisations, such as the WHO. So, even if the articles' domain and general topic might be relevant, analysts can avoid the article for not being factually about what is happening but more of a reflection on what has been happening throughout a specific outbreak, such as in article 458168. In this article, an expert demonstrates how vaccination fears are at fault for rising chickenpox cases in Angola. If other sources are already monitoring the number of cases, this piece can be overlooked because it is primarily about cause rather than consequence. Nevertheless, we believe this article and similar articles can indicate the worsening of ongoing outbreaks. As such, these shouldn't be ignored but used as indicators to flag future similar events pre-emptively.

6. Conclusion and Future Work

This article presented the first step in developing a recommendation system for a pre-existing platform, EIOS, developed for PHI. Therefore, the results and analysis still need to be completed. However, this work successfully showcases a pipeline for developing a content-based system recommending articles in real-world PHI scenarios. It introduces a clustering-based approach to tackle data sparsity and the use of event classifier labels to enrich the recommender algorithm. While more complex metadata and advanced models and approaches are available and will be used in the future, this first attempt successfully demonstrated a way of dealing with data sparsity for our case study, which in turn improved the model performance from an NDGC@K of 0.1749 to 0.4108, at $K = 100$, for the Sum cluster-based harmonization modality.

Looking ahead, we plan to further develop this approach by considering multiple users, article sources, other types of article metadata, and exploring the conjugation of clustering modalities and filters. Additionally, we aim to involve analysts in our approach to evaluate performance on actual end-users, thereby enhancing the robustness and applicability of our system.

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] J. S. Brownstein, C. C. Freifeld, L. C. Madoff, Digital disease detection—harnessing the web for public health surveillance, The New England journal of medicine 360 (2009) 2153. doi:10.1056/NEJMp0900702.

[3] J. M. Balbus, R. Barouki, L. S. Birnbaum, R. A. Etzel, P. D. Gluckman, P. Grandjean, C. Hancock, M. A. Hanson, J. J. Heindel, K. Hoffman, et al., Early-life prevention of non-communicable diseases, The Lancet 381 (2013) 3–4. doi:10.1016/S0140-6736(12)61609-2.

[4] H. Schäfer, S. Hors-Fraile, R. P. Karumur, A. Calero Valdez, A. Said, H. Torkamaan, T. Ulmer, C. Trattner, Towards health (aware) recommender systems, in: Proceedings of the 2017 International Conference on Digital Health, DH '17, Association for Computing Machinery, New York, NY, USA, 2017, p. 157–161. doi:10.1145/3079452.3079499.

[5] M. R. Pereira, Updated 2024 us vaccine recommendations from the advisory committee on immunization practices, American Journal of Transplantation 24 (2024) 514–516. doi:10.1016/j.ajt.2024.02.012.

[6] W. T. Riley, D. E. Rivera, A. A. Atienza, W. Nilsen, S. M. Allison, R. Mermelstein, Health behavior models in the age of mobile interventions: are our theories up to the task?, Translational behavioral medicine 1 (2011) 53–71. doi:10.1007/s13142-011-0021-7.

[7] H. Torkamaan, J. Ziegler, Recommendations as challenges: Estimating required effort and user ability for health behavior change recommendations, in: Proceedings of the 27th International Conference on Intelligent User Interfaces, IUI '22, Association for Computing Machinery, New York, NY, USA, 2022, p. 106–119. doi:10.1145/3490099.3511118.

[8] T. Chen, C. Guestrin, XGBoost: A scalable tree boosting system, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16, Association for Computing Machinery, New York, NY, USA, 2016, p. 785–794. doi:10.1145/2939672.2939785.

[9] 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.

A. Error Analysis with False Positives

Table A1
Top five ($K = 5$) false positives and their predicted scores across the five rounds of cross-validation for the different modalities.

Article	Original	Sum	High	AVG	Low	Random	Discard	Null
1757749*	51	46	34	41	46		51	41
2113787	42		37	50	38	37	45	
2355188	42		37	50	38	37	45	
1524234	34							
164429	55	49	55	42	42	41		
210702*	52	46	46	52	52	46	52	52
2713654	50			51		45		
1177084*	36		36	36	36	36	36	39
1642976*	42		42	42	42	42	42	42
1463715	36	48		32	48	36		
2083725*	33		45	39	46	46	46	46
462637	38			37	32			38
1388337	30			33				
458168*	68	64	64	68	68	64	68	68
2924180	40			40	40	40	40	40
1450025		44	32					
2671862		50	50		40	58		
455925		36						
857479		49						
1418055			32			32		
1152620			32			31		
185204			43	37	36	43		
3105707					30			
1255976						43		
471697							27	
390951							30	
2724292							30	
2274914							38	38
727629							42	
416240								66
2398504								33
979659								37
2163701								38
1133939								40