Audience-based news source characterization using probabilistic matrix factorization

Junita Sirait
Department of Computer Science
Princeton University
Princeton, NJ 08544
jsirait@princeton.edu

Abstract

In an era dominated by digital news outlets, understanding the landscape of online news sources is critical since news directly affects how people view the world and therefore make decisions. However, there is no one database that catalogues and characterizes the entire set of news sources, and current methods for characterizing news sources often rely on expert knowledge and resource-intensive natural language processing (NLP) techniques, limiting scalability and replicability. This paper introduces an alternative approach: characterizing news sources based on their audience, specifically the communities (subreddits) within the social media platform Reddit. By associating news sources with similar audiences, this project aims to enhance our understanding of their characteristics. This project employs probabilistic matrix factorization algorithms to extract latent features of news sources from Reddit data, focusing on share counts across various subreddits. The results demonstrate the utility of audience-based metrics for news characterization, including reliability classification and identification of similar news sources.

1 Introduction

In this digital age, online news is abundant as there is a lower barrier of entry for online news producer to establish digital presence and spread news. These online news sources reach people of all backgrounds and locations, and in turn affect how they see the world and eventually how they make decisions. Thus, it is important to have a good understanding of the online news source ecosystem due to their role in our society.

Despite the importance of news sources, no single database exists to track and analyze the vast number of them. Current efforts to understand and categorize news sources often rely on natural language processing (NLP) techniques that analyze content directly. While effective, these methods are resource-intensive, language-dependent, and difficult to replicate at scale. This study explores an alternative approach: characterizing news sources based on their audiences rather than their content. Here, "audience" refers to communities on the largely anonymous social media platform Reddit, known as subreddits. News sources are considered similar if they share similar audiences. For news sources with little prior analysis or characterization, identifying others with similar audiences can provide valuable insights based on previous evaluations of comparable sources.

2 Related Work

Efforts to characterize online news sources have included projects like MediaRank (Ye and Skiena [2019]), which ranks thousands of news outlets based on factors such as reputation, popularity, and political bias. MediaRank relies heavily on computational signals and article analysis through natural

Preprint. Under review.

language processing (NLP), which demands significant computational resources, storage, and domain knowledge. Other studies, such as Shin et al. (2022), employ human expertise for characterization, but this approach faces scalability issues.

For-profit organizations, like NewsGuard Technologies Inc.[1], which offers a paid browser extension to rate the credibility of news outlets, and Media Bias/Fact Check[2], have also entered this space. These commercial platforms rely on assessments by journalists, making their ratings potentially more accurate. However, due to their commercial nature, these products may introduce bias and are not easily reproducible or auditable.

This work, by contrast, focuses solely on audience-based metrics for characterizing news sources, an approach not commonly seen in previous studies. Although some projects, like MediaRank, incorporate audience data alongside NLP methods, this project uniquely emphasizes audience metrics.

A key inspiration for this project is the Netflix Prize Challenge, particularly the matrix factorization approach used by the winning team "BellKor's Pragmatic Chaos" (Töscher and Jahrer [2009]). In this project, news sources are analogous to movies, subreddits to users, and sharing counts to user ratings. While the Netflix competition aimed to recommend movies, this project seeks to learn the latent vectors of news sources using probabilistic matrix factorization.

3 Dataset

3.1 Source of dataset

One preliminary and obvious challenge is to define what counts as a news source. For the purposes of this project, and as I did in my previous project (Sirait [2022]), I use the definition used by the Global Database of Events, Language and Tone (GDELT) Project[3] and Muck Rack[4]. An online web page is defined as a news source in this project, if it exists in both the GDELT and Muck Rack databases. GDELT is a project that monitors world's broadcast, print, and web news from nearly everywhere on earth to codify human society's events, while Muck Rack is a Public Relations Management (PRM) platform that has a database list on their website containing the news sources that they monitor.

This work uses Reddit dataset obtained from PushShift (Baumgartner et al. [2020]) to find links of online news sources that are shared in various subreddits over the course of January - June 2021, which I have previously obtained and used for another related project that does not make use of recommender system (Sirait [2022]). Since we are looking at subreddits, the audience themselves are mostly and encouraged to be anonymous, which means that privacy of any individual is not compromised. Research also suggests that anonymous conditions encourage people to express their opinions more freely, which means that we have a higher confidence of faithful news source preferences of subreddits as audience (Luarn and Hsieh [2014]). Subreddits are a type of groups or dedicated communities in the social platform Reddit, where users are allowed to be or even encouraged to be anonymous. These communities are built based on shared interests of the members. Some examples of subreddits include `r/news`[5] which is self described as "the place for news articles about current events in the United States and the rest of the world," `r/sports`[6] described as "Sports News and Highlights from the NFL, NBA, NHL, MLB, MLS, and leagues around the world," and many others.

3.2 Dataset summary

This work focuses on 740 news sources shared across $36,727$ subreddits. There are only $296,440$ share counts, which means the data has a high sparsity of 98.9% (i.e., only 1.1% of the entries of the interaction matrix are nonzero). The range of the non-zero counts is between 1 and $39,381$ with mean 11.2 and median 1, indicating a long right tail. There is no further pre-processing done on this dataset,

[1] `https://www.newsguardtech.com/`
[2] `https://mediabiasfactcheck.com/`
[3] `https://www.gdeltproject.org/`
[4] `https://muckrack.com`
[5] `https://www.reddit.com/r/news/`
[6] `https://www.reddit.com/r/sports/`

since the goal is to attempt a simple news source characterization using subreddits-news sources interaction data focusing on the share counts, with probabilistic matrix factorization algorithms.

4 Method

This work uses probabilistic matrix factorization algorithms. Precisely, The data is represented as a sparse matrix Y of dimension $M \times N$, representing M news sources shared across N subreddits. Each of the cell y_{ij} in the matrix will record how many times news source i was shared in subreddit j. This problem can be viewed as matrix completion problem where we try to predict the missing values in Y. Following Murphy [2022] on Recommender Systems, specifically on Matrix Factorization, we can add some constraints and assume that Y is low rank and define $Z = VU^\top \approx Y$ where U is an $N \times K$ matrix, V is an $M \times K$ matrix, and K is the rank of the matrix. Then, the problem of the missing value prediction can be written as $\hat{y}_{ij} = v_j u_i^\top$, which is a matrix factorization problem statement. Note that we can view the matrix U can be viewed to be representations of the N subreddits preferences while the matrix V as the representations of the M news sources. The aim of this project is to extract the matrix V. We can then calculate similarities between news source a and news source b, for example by calculating the cosine distance between the entries V_a and V_b of the matrix V, or by visualizing them in a lower dimension. We can also use the extracted latent features of the news sources stored in V as features in classification task to characterize news sources, such as reliability classification task. Section 5 dives deeper into both use cases.

To solve this matrix factorization problem, we can set the conditional distribution over the observed counts $p(R|U, V)$ to be of a specific distribution, and set the priors on the subreddit and news source feature vectors $p(U)$ and $p(V)$ to also be of some specific distributions. Then, we can derive the log of posterior distribution, set our objectives, and use these to calculate gradients to update both U and V. Alternatively, in the case of a fully Bayesian approach, we can set hyperpriors on U and V, then use approximation techniques for the complex / intractable posterior, such as Markov Chain Monte Carlo (MCMC) or variational inference, then update the hyperparameters, U, and V accordingly. This project looks into three different approaches of probabilistic matrix factorization, which are summarized in the next subsection.

4.1 Gaussian PMF with point estimate

Inspired by Salakhutdinov and Mnih [2007], we can define the conditional distribution over the observed share counts as

$$p(Y|U, V, \sigma^2) = \prod_{i=1}^{N} \prod_{j=1}^{M} \left[\mathcal{N}(Y_{ij}|V_j U_i^\top, \sigma^2)\right]^{I_{ij}} \tag{1}$$

where $\mathcal{N}(x|\mu, \sigma^2)$ is the probability density function of the Normal distribution with mean μ and variance σ^2, and I_{ij} is the indicator function that is equal to 1 if subreddit i contains shared link of news source j and 0 otherwise. Inspired by the probabilistic matrix factorization method in the same paper, we can also then place Normal priors on the subreddit and news source feature vectors:

$$p(U|\sigma_U^2) = \prod_{i=1}^{N} \mathcal{N}(U_i|0, \sigma_U^2 \boldsymbol{I}) \quad \text{and} \quad p(V|\sigma_V^2) = \prod_{j=1}^{M} \mathcal{N}(V_j|0, \sigma_V^2 \boldsymbol{I}). \tag{2}$$

Then we can derive the log of the posterior distribution over the user and movie features $\ln p(U, V|Y, \sigma_U^2, \sigma_V^2)$ as the sum of the log of priors and conditional likelihood and some constants that don't depend on the parameters. Then, as in the implementation for this project, we can use coordinate ascent on the log posterior to optimize each vector of U and V while holding other variables fixed. For brevity, in this paper we will refer to this Gaussian PMF with point estimate simply by "GPMF".

4.2 Bayesian Gaussian PMF with MCMC

As an extension to the previous method, Salakhutdinov and Mnih [2008] also consider a fully Bayesian approach to Gaussian matrix factorization. Similarly to GPMF with MAP point estimate,

the likelihood of the observed share counts is given by Eq. 1, but now the Gaussian prior distributions over the subreddit and news source feature vectors are given by:

$$p(U|\mu_U, \Lambda_U) = \prod_{i=1}^{N} \mathcal{N}(U_i|\mu_U, \Lambda_U^{-1}) \quad \text{and} \quad p(V|\mu_V, \Lambda_V) = \prod_{i=1}^{M} \mathcal{N}(V_i|\mu_V, \Lambda_V^{-1}). \tag{3}$$

Following the probabilistic matrix factorization process described in the same paper, we then further place Gaussian-Wishart priors on the subreddit and news source hyperparameters $\Theta_U = \{\mu_U, \Lambda_U\}$ and $\Theta_V = \{\mu_V, \Lambda_V\}$. Then, due to the complexity of the posterior, MCMC-based methods are used to approximate the predictive distribution. Since conjugate priors are used for the parameters and hyperparameters, the conditional distributions derived from the posterior distribution are easy to sample from, so the implementation uses Gibbs sampling algorithm for inference, to update the hyperparameters as well as the subreddit and news sources features. This project uses the implementation of Salakhutdinov and Mnih [2008] by Lory Pack[7]. For brevity, in this paper we will refer to this fully bayesian Gaussian PMF with point estimate by "BPMF".

4.3 Poisson PMF with variational inference

Unlike Gaussian matrix factorization, Poisson matrix factorization uses Poisson distribution in defining the conditional distribution over the share counts and places Gamma priors on the subreddit and news source feature vectors. This makes sense in this project since the share counts are positive and more closely represented using Poisson distribution than Gaussian distribution. Gopalan et al. [2015] developed hierarchical Poisson matrix factorization (HPF) which captures long-tailed distribution, therefore is relevant to this project. Following Gopalan et al. [2015], we now define the conditional distribution over the observed share counts as

$$p(Y|U, V, \sigma^2) = \prod_{i=1}^{N} \prod_{j=1}^{M} \left[Poisson(Y_{ij}|U_i^\top V_j)\right]^{I_{ij}}. \tag{4}$$

Inspired by Gopalan et al. [2015], we then place Gamma priors on the latent attributes of news sources and subreddits, as well as additional priors on the subreddit-specific and news source-specific rate parameter of those Gammas, capturing the diversity of subreddits and news sources. So, for each subreddit i, $U_i \sim Gamma(a, \xi_i)$ where $\xi_i \sim Gamma(a', a'/b')$ and for each news source j, $V_j \sim Gamma(c, \eta_j)$ where $\eta_j \sim Gamma(c', c'/d')$. Similar to Gopalan et al. [2015], the exact posterior is intractable, so we use approximation method called mean-field variational inference, where we posit a family of distributions over the hidden variables and solve for a member of that family that minimizes the Kullback-Liebler (KL) divergence to the true posterior. This project uses the implementation of Gopalan et al. [2015] by David Cortes[8]. For brevity, in this paper we will refer to this hierarchical Poisson matrix factorization implementation by "HPF".

5 Results

In this section, I report the result of fitting each of the algorithms in Section 4. Then I analyze the usefulness of the latent features of the news sources, as used as features in reliability classification task, finding similar news sources using cosine distance.

5.1 Fitting three PMF algorithms

The three algorithms are fitted with grid search to set the value of K, the dimension of the latent features, to minimize RMSE. The RMSE on the test set achieved for the Gaussian matrix factorization with point estimate (GPMF in Section 4.1) is 188.2, while for the fully Bayesian matrix factorization (BPMF in Section 4.2) is 270.3, and for the Poisson matrix factorization (HPF in Section 4.3) is 184.4. As noted by Murphy [2022], placing Gaussian priors on the latent features of news sources and subreddits, as well as representing the share counts using Gaussian, still works arguably well for probabilistic matrix factorization. This is despite the fact that share counts are strictly non-negative

[7] `https://github.com/LoryPack/BPMF`

[8] `https://github.com/david-cortes/hpfrec`

and do not exhibit a Normal distribution. BPMF performs suprisingly poorly, despite the fact that this algorithm performs better than GPMF when fitted on Netflix dataset as reported by Salakhutdinov and Mnih [2008]. HPF, on the other hand, performs the best among the three algorithms as expected, since it was developed by Gopalan et al. [2015] to cater to count distributions that are long tailed.

Unlike the usual use cases of matrix factorization, such as to make recommendations, our aim is to extract the latent features of news sources V and use it to make characterizations of news sources by finding similar news sources, as discussed in Section 5.2.

5.2 Utilizing latent features

After fitting the three PMF algorithms summarized in Section 4, this project then extracts the latent features of the news sources V. We then use V as the features of the news sources to characterize them in terms of reliability classification. We will also use the latent features to find similar news sources for some specific news sources of interest using cosine similarity metric.

5.2.1 Reliability classification task

One aspect of news characterization is the investigation of their reliability. As a way to evaluate the usefulness of the latent attributes V of the news sources as extracted from a probabilistic matrix factorization algorithm, this project builds a classifier to classify news sources reliability ("reliable" or "unreliable" as defined by Media Bias Fact Check). The classifier itself is an ensemble of Nearest Neighbor and LightGBM, as found by AutoML[9], a Python package that performs hyperparameters tuning and model selection, to be the best classifier for the task. Table 1 summarizes the performance of this classifier.

	Predicted unreliable	Predicted reliable
True unreliable	0.75	0.25
True reliable	0.11	0.89

(a)

AUC	0.79
F1	0.86
Accuracy	0.83
Precision	0.84
Recall	0.89

(b)

Table 1: (a) confusion matrix of the reliability classification task on the held out test set with latent attributes of news sources from HPF as the features, (b) values of various evaluation metrics for the classifier performance on the held out test set

The above classifier performs reasonably well on the binary reliability classification task and, with more model selection and hyperparameter tuning, better classifiers could be found. As a baseline, when using directly their share counts in various subreddits as features, or using the principal components (from PCA) as features, the resulting classifiers of news sources reliability do not perform much better than chance (about 0.50 accuracy). This suggests we could reasonably be used to characterize news sources, such as their reliability, using only their latent attributes extracted using probabilistic matrix factorization algorithm on their share counts on Reddit.

5.2.2 Finding similar news sources

MediaRank, Media Bias Fact Check, Newsguard, and other fact-checking sites have catalogued and characterized a lot of news sources reasonably well. However, not all news sources are included in their database.

From Table 2(a), for example, we see that the three probabilistic matrix factorization algorithms are able to recognize what news sources are similar to `breitbart.com` based on various aspects, such as its far right political leaning, unreliability, and country of origin of the USA. These lists are generated by calculating the cosine distance between the latent feature vectors of `breitbart.com` and all the other news sources in our database, then ranking them based on this distance in ascending order. Specifically, the lists generated by using the latent features from GPMF and HPF have higher agreement with each other than with the list using the latent features from BPMF.

[9] `https://supervised.mljar.com/api/`

GPMF	BPMF	HPF
theepochtimes.com	theblaze.com	newsbusters.org
newsbusters.org	14news.com	campusreform.org
thepostmillennial.com	rt.com	wnd.com
newsmax.com	newsmax.com	frontpagemag.com
wnd.com	whnt.com	theblaze.com
hotair.com	wsmv.com	theepochtimes.com
rt.com	kark.com	newsday.com
theblaze.com	wnd.com	newsmax.com
nationalinterest.org	lifesitenews.com	westernjournal.com
foxnews.com	lapresse.ca	hotair.com

(a)

GPMF	BPMF	HPF
eatthis.com	out.com	hopkinsmedicine.org
dailyprogress.com	daytondailynews.com	universetoday.com
californiaglobe.com	smithsonianmag.com	treehugger.com
francetvinfo.fr	commercialappeal.com	neurosciencenews.com
prageru.com	atlasobscura.com	quantamagazine.com
humanevents.com	nationalgeographic.com	medicalnewstoday.com
judicialwatch.org	mcall.com	discovermagazine.com
lawenforcementtoday.com	bleacherreport.com	mayoclinic.org
pressdemocrat.com	upworthy.com	the-scientist.com
foxbaltimore.com	wsaz.com	healthline.com

(b)

Table 2: (a) top ten most similar news sources to `breitbart.com` based on the three probabilistic matrix factorization algorithms, (b) top ten most similar news sources to `mindbodygreen.com` based on the three probabilistic matrix factorization algorithms

As another example, Table 2(b) shows three lists of news sources that are similar to `mindbodygreen.com`. There are not much information about this news site other than that it has a mix of pseudoscience[10]. However, looking at the list of similar news sources, we can see some life, nature, and health related news sources as captured by the HPF and BPMF algorithms correctly suggesting the general theme of `mindbodygreen.com`. Similarly, the list of similar news sources generated using the PMF algorithm contain some unreliable news sources such as `judicialwatch.org` and `prageru.com` correctly suggesting the unreliability of `mindbodygreen.com`.

In sum, we have seen that we can use the latent features of the news sources V to characterize news sources, such as by classifying their reliability, as well as to find similar news sources which allows for some reasonable inference of the characteristics of news sources that have not been thoroughly audited or characterized before.

6 Conclusion

In this study, I explore an audience-based approach to characterizing online news sources using three probabilistic matrix factorization algorithms: Gaussian matrix factorization with point estimate, fully bayesian Gaussian matrix factorization, and hierarchical Poisson matrix factorization. By building interaction data between news sources and subreddits using the share counts of news sources across Reddit communities, I extract latent features of news sources and demonstrate their effectiveness in reliability classification and identifying similar news sources. The findings of this project suggest that audience-based metrics offer a promising avenue for scalable and replicable news characterization, complementing existing content-based methods. This approach holds potential for enhancing our understanding of the online news ecosystem in an increasingly digital world.

References

J. Baumgartner, S. Zannettou, B. Keegan, M. Squire, and J. Blackburn. The pushshift reddit dataset, 2020.

[10]according to Media Bias Fact Check `https://mediabiasfactcheck.com/mindbodygreen/`

P. Gopalan, J. M. Hofman, and D. M. Blei. Scalable recommendation with hierarchical poisson factorization. In *Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence*, UAI'15, page 326–335, Arlington, Virginia, USA, 2015. AUAI Press. ISBN 9780996643108.

P. Luarn and A.-Y. Hsieh. Speech or silence: the effect of user anonymity and member familiarity on the willingness to express opinions in virtual communities. *Online Information Review*, 2014.

K. P. Murphy. *Probabilistic Machine Learning: An introduction*. MIT Press, 2022. URL `probml.ai`.

R. Salakhutdinov and A. Mnih. Probabilistic matrix factorization, 2007.

R. Salakhutdinov and A. Mnih. Bayesian probabilistic matrix factorization using markov chain monte carlo. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, page 880–887, New York, NY, USA, 2008. Association for Computing Machinery. ISBN 9781605582054. doi: 10.1145/1390156.1390267. URL `https://doi.org/10.1145/1390156.1390267`.

J. Sirait. Investigating news source characterizations using reddit audiencebased metrics, 2022.

A. Töscher and M. Jahrer. The bigchaos solution to the netflix grand prize, 2009.

J. Ye and S. Skiena. Mediarank: Computational ranking of online news sources, 2019.