MIND News Recommender

CSE 158 FA 23 Assignment 2

Abstract

Confronted with an overwhelming flood of news via websites, smartphones, and social media, individuals increasingly experience a sense of inundation and overload. In response, news recommendation systems (NRS) have emerged as crucial tools to provide personalized, relevant, and engaging news content. This project delves into the evaluation of NRS, blending foundational RS models from CSE 158 with cutting-edge algorithms to explore their unique performance attributes. Our investigation extends beyond simple performance metrics, aiming to explore current trends and developments in NRS. By juxtaposing classroom-acquired knowledge with the latest algorithmic advances, our study offers detailed insights into the evolving landscape of news recommendation technologies.

1. Dataset

After examining various datasets, we identified the Microsoft News Dataset (MIND) [1] as a suitable resource for our news recommendation research. MIND is a comprehensive English dataset developed by the Microsoft Research team, specifically tailored for advanced news recommendation studies.

The MIND dataset, encompassing approximately 160,000 English news articles and over 15 million impression logs from 1 million users, offers a rich resource for news recommendation research. Each article in the dataset includes comprehensive textual content such as title, abstract, body, category, and entities. The impression logs detail user interactions, including click events, non-click events, and the user's historical news click behavior prior to the impression.

Additionally, MIND provides a smaller sample dataset, ideal for initial testing of model performance before scaling up to the full training set. Upon downloading the sample training and validation sets, an initial step involves performing Exploratory Data Analysis (EDA) to understand the dataset's characteristics.

The dataset comprises four key files:

- behaviors.tsv: each test user's click history and the record of news read during the experiment timeframe.
- news.tsv: details like the news category, subcategory, title, and abstract.
- entity_embedding.vec: embedded representations of entities.
- Relation_embedding.vec: embedded representations of relations between entities.

As introduced by the dataset paper, news recommendation poses unique challenges distinct from other types of content recommendation. Firstly, the rapid turnover of news articles on websites means they are continuously updated and quickly become outdated, leading to a pronounced cold-start problem. Secondly, news articles contain rich textual information like titles and bodies, as shown in Figure 1, making it inadequate to represent them merely by IDs; understanding their content is crucial. Thirdly, unlike other platforms, news websites typically lack explicit user ratings for articles. Therefore, users' interests in news are often inferred implicitly from their click behaviors.

# users	50000	# news	51282
Median # history	20	# categories	17
Median # test	24	# subcategories	264
# total interactions		3522653	

Table 1: Detailed statistics of the MIND training dataset

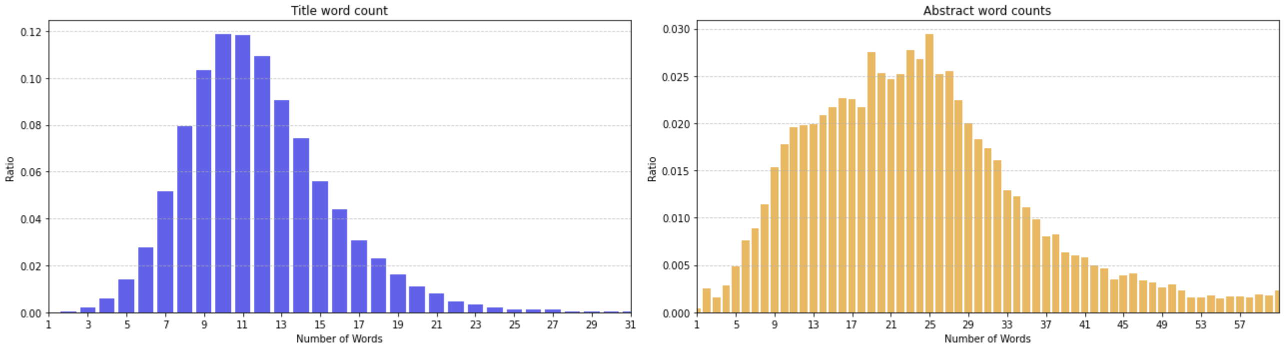


Figure 1: Word count of title and abstract in training set

An analysis of the training and validation sets reveals that they include data for 50,000 users, with only a 6.3% overlap between them. This significant lack of overlap indicates a severe cold start problem in the dataset. Furthermore, there are around 51,000 different news items in the dataset, with less than 1% overlap, further emphasizing the challenge of new item recommendations. Addressing these issues will be critical to developing an effective news recommendation system using the MIND dataset.

The observation highlighted in Figure 2, indicating a skewed distribution of news categories in the MIND dataset, is a significant aspect to consider for your news recommendation system. The predominance of articles classified under the "news" category suggests that there is a substantial imbalance in the dataset regarding the variety of news topics.

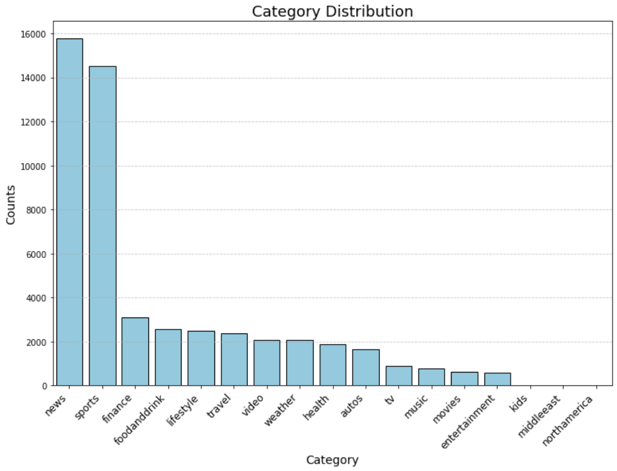


Figure 2: Category distribution on training set

2. Task

In our recommendation system, we aim to predict user engagement, specifically whether users will click on news items displayed on the homepage. The model's effectiveness will be gauged using the Area Under the Curve (AUC) metric. AUC is widely used in binary classification tasks, particularly effective for probabilistic predictions and in contexts with imbalanced datasets. Following the model's training on the designated training dataset, its predictive accuracy will be assessed on a separate validation set, using AUC as the benchmark. The effectiveness of our model will be evaluated by comparing its result with the LibFM [3] solution provided by the dataset, which we will use as our baseline.

Our approach, informed by insights from exploratory data analysis, includes testing a combination of user features, item features, and their interactions. To model user behavior, we aggregate the news information of all news articles in a user’s history. For news features, we analyze each item's category and subcategory and represent them through one-hot encoding. Additionally, we represent each news title and abstract as a Term Frequency-Inverse Document Frequency (TF-IDF) vector. This technique enables a nuanced comparison of news items based on their textual content, enriching the recommendation process.

3. Models

3.1. Similarity model

As an intuitive approach, we decided to first use a basic model based on categories. With the categories and subcategories given, it was not hard to construct a similarity function based on it, equation as shown below.

$$\text{sim} = \left(\frac{\text{sum of matched categories with item}}{\text{total history new viewed by user}}\right) \times \left(1 + \frac{\text{sum of matched categories}}{\text{sum of matched categories with item}}\right)$$

We tested our initial model on the validation set, yielding an AUC of 51. This result indicates that the traditional model, reliant on an interaction similarity function, is inadequate for our purposes, primarily due to the limited range of features and the imbalance of labels in our dataset. Moreover, relying solely on categories for prediction proves insufficient, as the dataset's 40 categories are too broad to effectively cater to a diverse user base of 50,000.

In light of these findings, we have opted to employ a Factorization Machine (FM) approach. Factorization Machines are particularly adept at handling sparse data, which aligns well with the nature of our dataset. By converting our features into a sparse matrix format, we can effectively train the FM, potentially overcoming the limitations encountered with our initial model and improving the accuracy of our news recommendation system.

3.2. Factorization Machine

Recognizing the suitability of factorization machines (FMs) for datasets like MIND, particularly in addressing issues of data sparsity, we redirected our focus towards this methodology. FMs are renowned for their efficacy in handling sparse datasets, a primary challenge in our case.

During our exploration, we encountered a range of pre-built FM libraries, such as LibFM, FastFM, and LightFM. However, when attempting to install these libraries in our local environments, we faced significant obstacles due to their lack of ongoing maintenance and support. This issue prevented us from successfully implementing these libraries in our project. Ultimately, we chose to utilize a manually built factorization machine algorithm available on GitHub [2]. This approach also offered the opportunity to delve deeper into the inner workings of the FM algorithm, potentially leading to more tailored and effective solutions for our dataset's unique challenges.

Initially, we attempted to represent user-news interactions using a large sparse matrix. However, its immense size, comprising 50 thousand rows and columns, proved impractical. Consequently, we reverted to utilizing users' click histories, encoding them based on the frequency of categories and subcategories of news they read. Our hypothesis is that similar clicking patterns in news categories among users indicate parallel preferences and recommendations. This approach was then compared with the use of TF-IDF vectors to represent user history.

For representing news content features, we employed the TF-IDF vectorizer from scikit-learn to create vectors for the top 1000 words from each news title. We tested using solely news titles for the TF-IDF vectorization, despite the MIND dataset offering news abstracts as well. This decision stems from our belief that users are primarily influenced by news titles when deciding whether to explore further details of a news item, thus making the abstract less significant in influencing user behavior. Our results later confirmed this hypothesis.

We also explored the use of Singular Value Decomposition (SVD) as an alternative for modeling user-item interactions, reducing the dimensionality to 50 embeddings. However, this approach did not integrate well with other features in our model. We hypothesize that these embeddings may not be sufficiently representative for our purposes.

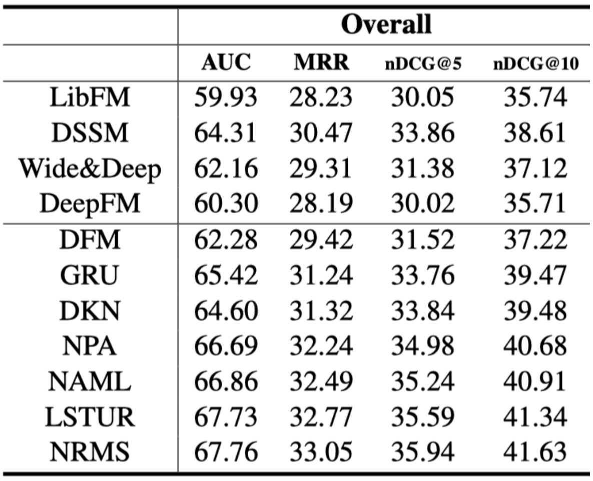
Further insights and detailed analysis of our model's performance will be elaborated on in the results section of our study.

4. Literature

As mentioned in the first section, our research utilizes the MIND dataset, a rich collection of data sourced from Microsoft News user click logs, encompassing over 1 million users and 160,000 English news articles. MIND has served as a testing ground for various state-of-the-art news recommendation techniques, originally developed using private datasets. Significant insights have been gained by applying these algorithms to MIND and comparing their effectiveness.

The dataset is segmented into three parts: a training set for building initial models, a validation set for model refinement and parameter adjustment, and a test set for evaluating the final performance of the recommender system. Previous studies on MIND have focused on metrics such as AUC, MRR (Mean Reciprocal Rank, indicating the rank of the first relevant item), and nDCG@5 or nDCG@10 (Normalized Discounted Cumulative Gain, measuring the relevance of the top 5 or 10 recommended items).

In comparison, other datasets for news recommendation have been limited by language, scale, or content representation. For instance, the Plista dataset, comprising user click logs from 13 German news portals, provides insights into German-language news interactions. The Adressa dataset, offering extensive data and categories, is used for studying user preferences in Norwegian news contexts. The Yahoo! dataset, though employed for session-based news recommendations, is limited in size (only 14,180 news items) and represents content using word IDs instead of original text. Many eligible datasets remain proprietary, restricted to internal use by large companies or institutions, and are not accessible for broader research.



	Overall			
	AUC	MRR	nDCG@5	nDCG@10
LibFM	59.93	28.23	30.05	35.74
DSSM	64.31	30.47	33.86	38.61
Wide&Deep	62.16	29.31	31.38	37.12
DeepFM	60.30	28.19	30.02	35.71
DFM	62.28	29.42	31.52	37.22
GRU	65.42	31.24	33.76	39.47
DKN	64.60	31.32	33.84	39.48
NPA	66.69	32.24	34.98	40.68
NAML	66.86	32.49	35.24	40.91
LSTUR	67.73	32.77	35.59	41.34
NRMS	67.76	33.05	35.94	41.63

Figure 3: Performance leaderboard of popular models

In the realm of news recommendation, state-of-the-art approaches like NRMS (Neural News Recommendation with Multi-Head Self-Attention), LSTUR, and NAML have surpassed traditional methods, such as those employing factorization machine algorithms. These advanced methods leverage neural networks to directly learn from raw data about news content and user interests in an end-to-end fashion, moving away from the handcrafted feature engineering typical in general recommendation methods.

NRMS (Neural News Recommendation with Multi-Head Self-Attention) [5] is a key advancement in news recommendation. It employs a multi-head self-attention mechanism to learn news content, especially titles, and user preferences for precise recommendations. This model analyzes the relationship between words and their context, focusing on different semantic aspects, and models user preferences by evaluating their news browsing history. NRMS has achieved impressive metrics, including an AUC of 67.76, establishing it as a leader in news recommendations.

LSTUR (Long and Short-Term User Representation) [6] closely follows NRMS in performance. It distinguishes between users' long-term and short-term interests, using an embedding vector for long-term preferences and a GRU network [7] for short-term interests. LSTUR adeptly balances these preferences in its recommendations, achieving an AUC of 67.73.

NAML (Neural News Recommendation with Attentive Multi-View Learning) [4] ranks third, excelling at modeling diverse news content sections like titles and bodies. It uses CNNs for content analysis and an attention network for aggregating information, ensuring a nuanced understanding of user preferences. With an AUC of 66.86, NAML effectively matches news with user interests, highlighting significant content and behavioral patterns.

The majority of contemporary algorithms in this field employ deep learning techniques, particularly neural networks, for modeling user-news interactions and news semantics. This area falls outside our expertise, limiting our ability to directly compare these methods with our model. The only comparable approach is one that utilizes LibFM, a classic factorization machine algorithm, closely mirroring our model's structure. According to its methodology, it models user history and news semantics using TF-IDF vectors, similar to the techniques we have implemented. However, it also models user-item interactions by including user-IDs and news-IDs. In our case, due to the dataset's large size, our attempt to integrate one-hot encoded user-IDs was unsuccessful, leading to null results. Consequently, we omitted user-IDs from our final model.

5. Results

After extensive testing with various parameters and feature combinations, we achieved our best result with an AUC of 58.11, closely aligning with the performance of LibFM. The absence of user-news interactions might account for the slight performance discrepancy compared to LibFM. Despite these considerations, our model demonstrates a notable improvement over our initial similarity-based model.

Our results reveal a few key findings. Firstly, we found that while including general categories doesn't enhance the model, incorporating subcategories substantially improves it. This can be attributed to the imbalanced distribution of categories that we previously observed in the dataset. For example, the “news” category, which is the largest in our dataset, comprises over 15,000 news items (Figure 2). Within this general “news” category, there are 30+ distinct subcategories (Figure 4). News items from different subcategories can vary significantly, even though they fall under the same general category. This variation makes the general category less informative for our model. Furthermore, our results indicate that modeling user history with TF-IDF vectors instead of mere categories led to a more notable performance boost. We found that using TF-IDF vectors for both the news title and the news abstract would yield a better result than using them only for the title. We also explored the impact of timestamps and user-item embeddings but observed only minimal effects on model performance.

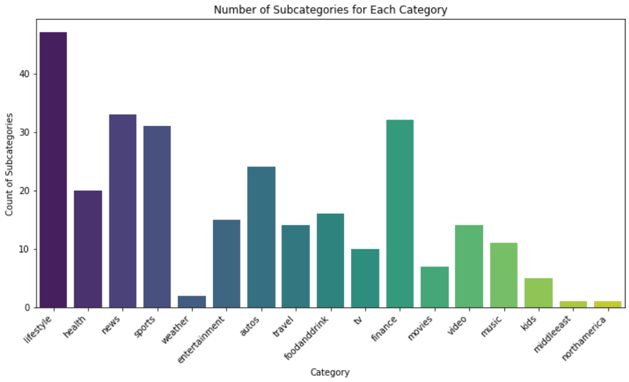


Figure 4: Subcategory distribution on training set

Considering these findings, our final model employs TF-IDF vectors for both the candidate news and the historic news, encompassing both news titles and abstracts. We found that the optimal number of maximum features for the TF-IDF vectors is 1000. Utilizing a smaller or larger number of features resulted in underfitting or overfitting, respectively. We also one-hot encoded the subcategories for both historical and candidate news, which boosted our model’s performance on the testing set by 0.08. For further optimization, we employed a grid search strategy to fine-tune our factorization model, which helped us identify the best hyperparameters (learning rate = 0.01, number of latent factors = 1, number of iterations = 10).

In conclusion, our research targets the complex field of news recommendation, where navigating large sparse matrices and addressing the cold-start problem present significant challenges. With careful feature selection and model adjustments, we attempted to handle the intricacies of news data and have made notable strides in improving recommendation accuracy. This study sets a foundation for us to further investigate deeper neural networks in this domain.

References

[1] Fangzhao Wu, Ying Qiao, Jiun-Hung Chen, Chuhan Wu, Tao Qi, Jianxun Lian, Danyang Liu, Xing Xie, Jianfeng Gao, Winnie Wu and Ming Zhou. MIND: A Large-scale Dataset for News Recommendation. ACL 2020.

[2] Ethen. Machine Learning documentary repository. GitHub 2018. Link to notebook.

[3] Steffen Rendle. 2012. Factorization Machines with libfm. TIST, 3(3):57:1–57:22.

[4] Chuhan Wu, Fangzhao Wu, Mingxiao An, Jianqiang Huang, Yongfeng Huang, and Xing Xie. 2019. Neural News Recommendation with Attentive Multiview Learning. IJCAI-19, pages 3863–3869

[5] Chuhan Wu, Fangzhao Wu, Suyu Ge, Tao Qi, Yongfeng Huang, and Xing Xie. 2019. Neural News Recommendation with Multi-Head Self-Attention. EMNLP-IJCNLP, pages 6390–6395.

[6] Mingxiao An, Fangzhao Wu, Chuhan Wu, Kun Zhang, Zheng Liu, and Xing Xie. 2019. Neural News Recommendation with Long-and Short-Term User Representations. ACL, pages 336–345.

[7] Shumpei Okura, Yukihiro Tagami, Shingo Ono, and Akira Tajima. 2017. Embedding-based News Recommendation for Millions of Users. KDD, pages 1933–1942. ACM.