MIND: A Large-scale Dataset for News Recommendation

Fangzhao Wu[†], **Ying Qiao**[‡], **Jiun-Hung Chen**[‡], **Chuhan Wu**[§], **Tao Qi**[§],
Jianxun Lian[†], **Danyang Liu**[†], **Xing Xie**[†], **Jianfeng Gao**[†], **Winnie Wu**[‡], **Ming Zhou**[†]
[†]Microsoft Research [‡]Microsoft [§]Tsinghua University
{fangzwu, yiqia, jiuche, jialia}@microsoft.com
{t-danliu, xingx, jfgao, winniew, mingzhou}@microsoft.com
{wu-ch19, qit16}@mails.tsinghua.edu.cn

Abstract

News recommendation is an important technique for personalized news service. Compared with product and movie recommendations which have been comprehensively studied, the research on news recommendation is much more limited, mainly due to the lack of a high-quality benchmark dataset. In this paper, we present a large-scale dataset named MIND for news recommendation. Constructed from the user click logs of Microsoft News, MIND contains 1 million users and more than 160k English news articles, each of which has rich textual content such as title, abstract and body. We demonstrate MIND a good testbed for news recommendation through a comparative study of several state-of-the-art news recommendation methods which are originally developed on different proprietary datasets. Our results show the performance of news recommendation highly relies on the quality of news content understanding and user interest modeling. Many natural language processing techniques such as effective text representation methods and pre-trained language models can effectively improve the performance of news recommendation. The MIND dataset will be available at https://msnews.github.io.

1 Introduction

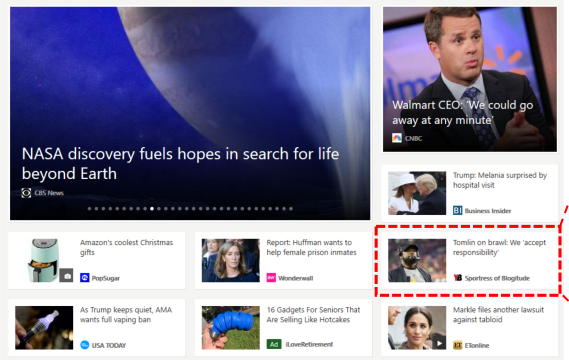
Online news services such as Google News and Microsoft News have become important platforms for a large population of users to obtain news information (Das et al., 2007; Wu et al., 2019a). Massive news articles are generated and posted online every day, making it difficult for users to find interested news quickly (Okura et al., 2017). Personalized news recommendation can help users alleviate information overload and improve news reading experience (Wu et al., 2019b). Thus, it is widely used in many online news platforms (Li et al., 2011; Okura et al., 2017; An et al., 2019).

In traditional recommender systems, users and items are usually represented using IDs, and their interactions such as rating scores are used to learn ID representations via methods like collaborative filtering (Koren, 2008). However, news recommendation has some special challenges. First, news articles on news websites update very quickly. New news articles are posted continuously, and existing news articles will expire in short time (Das et al., 2007). Thus, the cold-start problem is very severe in news recommendation. Second, news articles contain rich textual information such as title and body. It is not appropriate to simply representing them using IDs, and it is important to understand their content from their texts (Kompan and Bieliková, 2010). Third, there is no explicit rating of news articles posted by users on news platforms. Thus, in news recommendation users' interest in news is usually inferred from their click behaviors in an implicit way (Ilievski and Roy, 2013).

A large-scale and high-quality dataset can significantly facilitate the research in an area, such as ImageNet for image classification (Deng et al., 2009) and SQuAD for machine reading comprehension (Rajpurkar et al., 2016). There are several public datasets for traditional recommendation tasks, such as Amazon dataset[1] for product recommendation and MovieLens dataset[2] for movie recommendation. Based on these datasets, many well-known recommendation methods have been developed. However, existing studies on news recommendation are much fewer, and many of them are conducted on proprietary datasets (Okura et al., 2017; Wang et al., 2018; Wu et al., 2019a). Although there are a few public datasets for news recommendation, they are usually in small size and most of them are not in English. Thus, a public

[1]http://jmcauley.ucsd.edu/data/amazon/
[2]https://grouplens.org/datasets/movielens/

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 3597–3606



(a) An example Microsoft News homepage

Title	Mike Tomlin: Steelers 'accept responsibility' for role in brawl with Browns
Category	Sports
Abstract	Mike Tomlin has admitted that the Pittsburgh Steelers played a role in the brawl with the Cleveland Browns last week, and on Tuesday he accepted responsibility for it on behalf of the organization.
Body	Tomlin opened his weekly news conference by addressing the issue head on. "It was ugly," said Tomlin, who had refused to take any questions about the incident directly after the game, per Brooke Pryor of ESPN. "It was ugly for the game of football. I think all of us that are involved in the game, particularly at this level, ...

(b) Texts in an example news article

Figure 1: An example homepage of Microsoft News and an example news article on it.

large-scale English news recommendation dataset is of great value for the research in this area.

In this paper we present a large-scale MIcrosoft News Dataset (MIND) for news recommendation research, which is collected from the user behavior logs of Microsoft News[3]. It contains 1 million users and their click behaviors on more than 160k English news articles. We implement many state-of-the-art news recommendation methods originally developed on different proprietary datasets, and compare their performance on the MIND dataset to provide a benchmark for news recommendation research. The experimental results show that a deep understanding of news articles through NLP techniques is very important for news recommendation. Both effective text representation methods and pre-trained language models can contribute to the performance improvement of news recommendation. In addition, appropriate modeling of user interest is also useful. We hope MIND can serve as a benchmark dataset for news recommendation and facilitate the research in this area.

2 Related Work

2.1 News Recommendation

News recommendation aims to find news articles that users have interest to read from the massive candidate news (Das et al., 2007). There are two important problems in news recommendation, i.e., how to represent news articles which have rich textual content and how to model users' interest in news from their previous behaviors (Okura et al., 2017). Traditional news recommendation methods usually rely on feature engineering to represent news articles and user interest (Liu et al., 2010; Son et al., 2013; Karkali et al., 2013; Garcin et al., 2013; Bansal et al., 2015; Chen et al., 2017). For example, Li et al. (2010) represented news articles using their URLs and categories, and represented users using their demographics, geographic information and behavior categories inferred from their consumption records on Yahoo!.

In recent years, several deep learning based news recommendation methods have been proposed to learn representations of news articles and user interest in an end-to-end manner (Okura et al., 2017; Wu et al., 2019a; An et al., 2019). For example, Okura et al. (2017) represented news articles from news content using denoising autoencoder model, and represented user interest from historical clicked news articles with GRU model. Their experiments on Yahoo! Japan platform show that the news and user representations learned with deep learning models are promising for news recommendation. Wang et al. (2018) proposed to learn knowledge-aware news representations from news titles using CNN network by incorporating both word embeddings and the entity embeddings inferred from knowledge graphs. Wu et al. (2019a) proposed an attentive multi-view learning framework to represent news articles from different news texts such as title, body and category. They used an attention model to infer the interest of users from their clicked news articles by selecting informative ones. These works are usually developed and validated on proprietary datasets which are not publicly available, making it difficult for other researchers to verify these methods and develop their own methods.

News recommendation has rich inherent relatedness with natural language processing. First, news is a common form of texts, and text modeling

[3]https://microsoftnews.msn.com/

Dataset	Language	# Users	# News	# Clicks	News information
Plista	German	Unknown	70,353	1,095,323	title, body
Adressa	Norwegian	3,083,438	48,486	27,223,576	title, body, category
Globo	Portuguese	314,000	46,000	3,000,000	no original text, only word embeddings
Yahoo!	English	Unknown	14,180	34,022	no original text, only word IDs
MIND	English	1,000,000	161,013	24,155,470	title, abstract, body, category

Table 1: Comparisons of the MIND dataset and the existing public news recommendation datasets.

techniques such as CNN and Transformer can be naturally applied to represent news articles (Wu et al., 2019a; Ge et al., 2020). Second, learning user interest representation from previously clicked news articles has similarity with learning document representation from its sentences. Third, news recommendation can be formulated as a special text matching problem, i.e., the matching between a candidate news article and a set of previously clicked news articles in some news reading interest space. Thus, news recommendation has attracted increasing attentions from the NLP community (An et al., 2019; Wu et al., 2019c).

2.2 Existing Datasets

There are only a few public datasets for news recommendation, which are summarized in Table 1. Kille et al. (2013) constructed the *Plista*[4] dataset by collecting news articles published on 13 German news portals and users' click logs on them. It contains 70,353 news articles and 1,095,323 click events. The news articles in this dataset are in German and the users are mainly from the German-speaking world. Gulla et al. (2017) released the *Adressa* dataset[5], which was constructed from the logs of the Adresseavisen website in ten weeks. It has 48,486 news articles, 3,083,438 users and 27,223,576 click events. Each click event contains several features, such as session time, news title, news category and user ID. Each news article is associated with some detailed information such as authors, entities and body. The news articles in this dataset are in Norwegian. Moreira et al. (2018) constructed a news recommendation dataset[6] from Globo.com, a popular news portal in Brazil. This dataset contains about 314,000 users, 46,000 news articles and 3 million click records. Each click record contains fields like user ID, news ID and session time. Each news article has ID, category, publisher, creation time, and the embeddings of its words generated by a neural model pre-trained on a news metadata classification task (de Souza Pereira Moreira et al., 2018). However, the original texts of news articles are not provided. In addition, this dataset is in Portuguese. There is a Yahoo! dataset[7] for session-based news recommendation. It contains 14,180 news articles and 34,022 click events. Each news article is represented by word IDs, and the original news text is not provided. The number of users in this dataset is unknown since there is no user ID. In summary, most existing public datasets for news recommendation are non-English, and some of them are small in size and lack original news texts. Thus, a high-quality English news recommendation dataset is of great value to the news recommendation community.

3 MIND Dataset

3.1 Dataset Construction

In order to facilitate the research in news recommendation, we built the MIcrosoft News Dataset (MIND)[8]. It was collected from the user behavior logs of Microsoft News[9]. We randomly sampled 1 million users who had at least 5 news click records during 6 weeks from October 12 to November 22, 2019. In order to protect user privacy, each user was de-linked from the production system when securely hashed into an anonymized ID using one-time salt[10] mapping. We collected the behavior logs of these users in this period, which are formatted into impression logs. An impression log records the news articles displayed to a user when she visits the news website homepage at a specific time, and her click behaviors on these news articles. Since in news recommendation we usually predict whether a user will click a candidate news

[4]http://www.newsreelchallenge.org/dataset/

[5]http://reclab.idi.ntnu.no/dataset/

[6]https://www.kaggle.com/gspmoreira/news-portal-user-interactions-by-globocom

[7]https://webscope.sandbox.yahoo.com/catalog.php?datatype=l

[8]It is public available at https://msnews.github.io for research purpose. Any question about this dataset can be sent to mind@microsoft.com.

[9]https://microsoftnews.msn.com

[10]https://en.wikipedia.org/wiki/Salt_(cryptography)

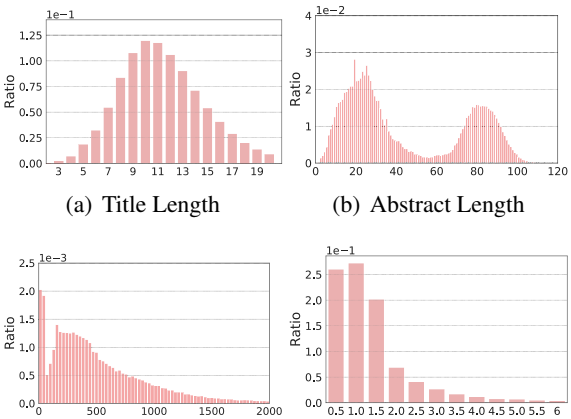
article or not based on her personal interest inferred from her previous behaviors, we add the news click histories of users to their impression logs to construct labeled samples for training and verifying news recommendation models. The format of each labeled sample is $[uID, t, ClickHist, ImpLog]$, where uID is the anonymous ID of a user, and t is the timestamp of this impression. $ClickHist$ is an ID list of the news articles previously clicked by this user (sorted by click time). $ImpLog$ contains the IDs of the news articles displayed in this impression and the labels indicating whether they are clicked, i.e., $[(nID_1, label_1), (nID_2, label_2), ...]$, where nID is news article ID and $label$ is the click label (1 for click and 0 for non-click). We used the samples in the last week for test, and the samples in the fifth week for training. For samples in training set, we used the click behaviors in the first four weeks to construct the news click history. For samples in test set, the time period for news click history extraction is the first five weeks. We only kept the samples with non-empty news click history. Among the training data, we used the samples in the last day of the fifth week as validation set.

Each news article in the MIND dataset contains a news ID, a title, an abstract, a body, and a category label such as "Sports" which is manually tagged by the editors. In addition, we found that these news texts contain rich entities. For example, in the title of the news article shown in Fig. 1 "Mike Tomlin: Steelers 'accept responsibility' for role in brawl with Browns", "Mike Tomlin" is a person entity, and "Steelers" and "Browns" are entities of American football team. In order to facilitate the research of knowledge-aware news recommendation, we extracted the entities in the titles, abstracts and bodies of the news articles in the MIND dataset, and linked them to the entities in WikiData[11] using an internal NER and entity linking tool. We also extracted the knowledge triples of these entities from WikiData and used TransE (Bordes et al., 2013) method to learn the embeddings of entities and relations. These entities, knowledge triples, as well as entity and relation embeddings are also included in the MIND dataset.

3.2 Dataset Analysis

The detailed statistics of the MIND dataset are summarized in Table 2 and Fig. 2. This dataset contains 1,000,000 users and 161,013 news articles. There

[11] https://www.wikidata.org/wiki/Wikidata:MainPage



(a) Title Length (b) Abstract Length

(c) Body Length (d) Survival Time

Figure 2: Key statistics of the MIND dataset.

# News	161,013	# Users	1,000,000
# News category	20	# Impression	15,777,377
# Entity	3,299,687	# Click behavior	24,155,470
Avg. title len.	11.52	Avg. abstract len.	43.00
Avg. body len.	585.05		

Table 2: Detailed statistics of the MIND dataset.

are 2,186,683 samples in the training set, 365,200 samples in the validation set, and 2,341,619 samples in the test set, which can empower the training of data-intensive news recommendation models. Figs. 2(a), 2(b) and 2(c) show the length distributions of news title, abstract and body. We can see that news titles are usually very short and the average length is only 11.52 words. In comparison, news abstracts and bodies are much longer and may contain richer information of news content. Thus, incorporating different kinds of news information such as title, abstract and body may help understand news articles better.

Fig. 2(d) shows the survival time distribution of news articles. The survival time of a news article is estimated here using the time interval between its first and last appearance time in the dataset. We find that the survival time of more than 84.5% news articles is less than two days. This is due to the nature of news information, since news media always pursue the latest news and the exiting news articles get out-of-date quickly. Thus, cold-start problem is a common phenomenon in news recommendation, and the traditional ID-based recommender systems (Koren, 2008) are not suitable for this task. Representing news articles using their textual content is critical for news recommendation.

4 Method

In this section, we briefly introduce several methods for news recommendation, including general recommendation methods and news-specific recommendation methods. These methods were developed in different settings and on different datasets. Some of their implementations can be found in Microsoft Recommenders open source repository[12]. We will compare them on the MIND dataset.

4.1 General Recommendation Methods

LibFM (Rendle, 2012), a classic recommendation method based on factorization machine. Besides the user ID and news ID, we also use the content features[13] extracted from previously clicked news and candidate news as the additional features to represent users and candidate news.

DSSM (Huang et al., 2013), deep structured semantic model, which uses tri-gram hashes and multiple feed-forward neural networks for query-document matching. We use the content features extracted from previous clicked news as query, and those from candidate news as document.

Wide&Deep (Cheng et al., 2016), a two-channel neural recommendation method, which has a wide linear transformation channel and a deep neural network channel. We use the same content features of users and candidate news for both channels.

DeepFM (Guo et al., 2017), another popular neural recommendation method which synthesizes deep neural networks and factorization machines. The same content features of users and candidate news are fed to both components.

4.2 News Recommendation Methods

DFM (Lian et al., 2018), deep fusion model, a news recommendation method which uses an inception network to combine neural networks with different depths to capture the complex interactions between features. We use the same features of users and candidate news with aforementioned methods.

GRU (Okura et al., 2017), a neural news recommendation method which uses autoencoder to learn latent news representations from news content, and uses a GRU network to learn user representations from the sequence of clicked news.

DKN (Wang et al., 2018), a knowledge-aware news recommendation method. It uses CNN to learn news representations from news titles with both word embeddings and entity embeddings (inferred from knowledge graph), and learns user representations based on the similarity between candidate news and previously clicked news.

NPA (Wu et al., 2019b), a neural news recommendation method with personalized attention mechanism to select important words and news articles based on user preferences to learn more informative news and user representations.

NAML (Wu et al., 2019a), a neural news recommendation method with attentive multi-view learning to incorporate different kinds of news information into the representations of news articles.

LSTUR (An et al., 2019), a neural news recommendation method with long- and short-term user interests. It models short-term user interest from recently clicked news with GRU and models long-term user interest from the whole click history.

NRMS (Wu et al., 2019c), a neural news recommendation method which uses multi-head self-attention to learn news representations from the words in news text and learn user representations from previously clicked news articles.

5 Experiments

5.1 Experimental Settings

In our experiments, we verify and compare the methods introduced in Section 4 on the MIND dataset. Since most of these news recommendation methods are based on news titles, for fair comparison, we only used news titles in experiments unless otherwise mentioned. We will explore the usefulness of different news texts such as body in Section 5.3.3. In order to simulate the practical news recommendation scenario where we always have unseen users not included in training data, we randomly sampled half of the users for training, and used all the users for test. For those methods that need word embeddings, we used the Glove (Pennington et al., 2014) as initialization. Adam was used as the optimizer. Since the non-clicked news are usually much more than the clicked news in each impression log, following (Wu et al., 2019b) we applied negative sampling technique to model training. All hyper-parameters were selected according to the results on the validation set. The metrics used in our experiments are AUC, MRR, nDCG@5 and nDCG@10, which are standard metrics for recommendation result evaluation. Each experiment was repeated 10 times.

[12] https://github.com/microsoft/recommenders

[13] The content features used in our experiments are TF-IDF features extracted from news texts.

	Overall				Overlap Users				Unseen Users			
	AUC	MRR	nDCG@5	nDCG@10	AUC	MRR	nDCG@5	nDCG@10	AUC	MRR	nDCG@5	nDCG@10
LibFM	59.93	28.23	30.05	35.74	60.23	28.08	29.94	35.66	59.72	28.35	30.14	35.81
DSSM	64.31	30.47	33.86	38.61	64.70	30.39	32.84	38.62	64.02	30.53	33.88	38.61
Wide&Deep	62.16	29.31	31.38	37.12	62.53	29.22	31.33	37.11	61.89	29.38	31.41	37.13
DeepFM	60.30	28.19	30.02	35.71	60.58	28.05	29.91	35.62	60.10	28.31	30.10	35.77
DFM	62.28	29.42	31.52	37.22	62.62	29.30	31.45	37.18	62.03	29.50	31.57	37.25
GRU	65.42	31.24	33.76	39.47	65.80	31.15	33.73	39.47	65.14	31.31	33.78	39.46
DKN	64.60	31.32	33.84	39.48	64.88	31.19	33.76	39.43	64.40	31.42	33.89	39.52
NPA	66.69	32.24	34.98	40.68	67.10	32.18	35.00	40.72	66.39	32.29	34.97	40.65
NAML	66.86	32.49	35.24	40.91	67.15	32.36	35.17	40.88	66.65	32.58	35.28	40.94
LSTUR	67.73	32.77	35.59	41.34	68.13	32.70	35.59	41.38	67.43	32.82	35.58	41.31
NRMS	67.76	33.05	35.94	41.63	68.23	33.05	36.03	41.74	67.41	33.05	35.88	41.55

Table 3: Results on the test set of the MIND dataset. Overlap users mean the users included in training set.

5.2 Performance Comparison

The experimental results of different methods on the MIND dataset are summarized in Table 3. We have several findings from the results.

First, news-specific recommendation methods such as *NAML*, *LSTUR* and *NRMS* usually perform better than general recommendation methods like *Wide&Deep*, *LibFM* and *DeepFM*. This is because in these news-specific recommendation methods the representations of news articles and user interest are learned in an end-to-end manner, while in the general recommendation methods they are usually represented using handcrafted features. This result validates that learning representations of news content and user interest from raw data using neural networks is more effective than feature engineering. The only exception is *DFM*, which is designed for news recommendation but cannot outperform some general recommendation methods such as *DSSM*. This is because in *DFM* the features of news and users are also manually designed.

Second, among the neural news recommendation methods, *NRMS* can achieve the best performance. *NRMS* uses multi-head self-attention to capture the relatedness between words to learn news representations, and capture the interactions between previously clicked news articles to learn user representations. This result shows that advanced NLP models such as multi-head self-attention can effectively improve the understanding of news content and modeling of user interest. The performance of *LSTUR* is also strong. *LSTUR* can model users' short-term interest from their recently clicked news through a GRU network, and model users' long-term interest from the whole news click history. The result shows appropriate modeling of user interest is also important for news recommendation.

Third, in terms of the AUC metric, the performance of news recommendation methods on unseen users is slightly lower than that on overlap users which are included in training data. However, the performance on both kinds of users in terms of MRR and nDCG metrics has no significant difference. This result indicates that by inferring user interest from the content of their previously clicked news, the news recommendation models trained on part of users can be effectively applied to the remaining users and new users coming in future.

	NAML		LSTUR		NRMS	
	AUC	nDCG@10	AUC	nDCG@10	AUC	nDCG@10
LDA	54.29	31.88	53.27	30.41	52.93	30.50
TF-IDF	56.07	33.06	55.53	32.32	55.43	32.31
Avg-Emb	57.97	34.29	61.06	36.10	61.10	36.49
Attention	60.76	36.80	64.95	39.06	65.31	39.66
CNN	63.10	38.07	64.76	39.04	64.77	39.10
CNN+Att	65.10	39.53	65.86	39.93	66.05	40.10
Self-Att.	65.46	39.89	65.64	39.81	65.91	40.02
Self-Att+Att	65.60	40.05	65.91	39.91	66.22	40.23
LSTM	65.20	39.66	65.88	39.87	66.27	40.21
LSTM+Att	66.17	40.23	66.37	40.31	66.91	40.85

Table 4: Different news representation methods. *Att* means attention mechanism.

5.3 News Content Understanding

Next, we explore how to learn accurate news representations from textual content. Since the MIND dataset is quite large-scale, we randomly sampled 100k samples from both training and test sets for the experiments in this and the following sections.

5.3.1 News Representation Model

First, we compare different text representation methods for learning news representation. We select three news recommendation methods which have strong performance, i.e., *NAML*, *LSTUR* and *NRMS*, and replace their original news representation module with different text representation methods, such as *LDA*, *TF-IDF*, average of word embed-

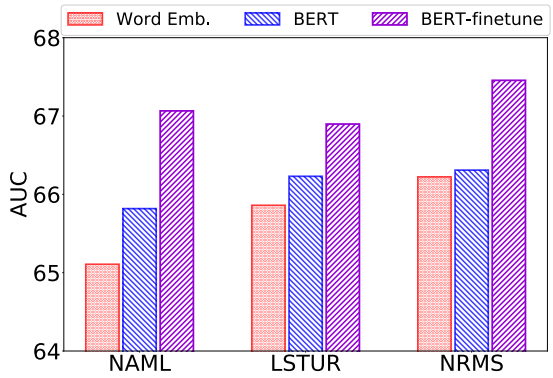


Figure 3: BERT for news representation.

	AUC	MRR	nDCG@5	nDCG@10
Title	66.22	31.92	34.53	40.23
Abs.	64.17	30.49	32.81	38.57
Body	66.32	31.88	34.42	40.22
Title + Abs. + Body (Con)	67.07	32.34	34.98	40.74
Title + Abs. + Body + Cat. (Con)	67.09	32.40	35.03	40.80
Title + Abs. + Body + Cat. + Ent. (Con)	67.23	32.41	35.04	40.83
Title + Abs. + Body (AMV)	67.38	32.37	35.12	40.79
Title + Abs. + Body + Cat. (AMV)	67.50	32.43	35.21	40.96
Title + Abs. + Body + Cat. + Ent. (AMV)	67.60	32.51	35.24	41.03

Table 5: News representation with different news information. "Abs.", "Cat." and "Ent." mean abstract, category and entity, respectively.

ding (*Avg-Emb*), *CNN*, *LSTM* and multi-head self-attention (*Self-Att*). Since attention mechanism is an important technique in NLP (Yang et al., 2016), we also apply it to the aforementioned neural text representation methods. The results are in Table 4.

We have several findings from the results. First, neural text representation methods such as *CNN*, *Self-Att* and *LSTM* can outperform traditional text representation methods like *TF-IDF* and *LDA*. This is because the neural text representation models can be learned with the news recommendation task, and they can capture the contexts of texts to generate better news representations. Second, *Self-Att* and *LSTM* outperform *CNN* in news representation. This is because multi-head self-attention and LSTM can capture long-range contexts of words, while CNN can only model local contexts. Third, the attention mechanism can effectively improve the performance of different neural text representation methods such as *CNN* and *LSTM* for news recommendation. It shows that selecting important words in news texts using attention can help learn more informative news representations. Another interesting finding is that the combination of LSTM and attention can achieve the best performance. However, to our best knowledge, it is not used in existing news recommendation methods.

5.3.2 Pre-trained Language Models

Next, we explore whether the quality of news representation can be further improved by the pre-trained language models such as BERT (Devlin et al., 2019), which have achieved huge success in different NLP tasks. We applied BERT to the news representation module of three state-of-the-art news recommendation methods, i.e., *NAML*, *LSTUR* and *NRMS*. The results are summarized in Fig. 3. We find that by replacing the original word embedding module with the pre-trained BERT model, the performance of different news recommendation methods can be improved. It shows the BERT model pre-trained on large-scale corpus like Wikipedia can provide useful semantic information for news representation. We also find that fine-tuning the pre-trained BERT model with the news recommendation task can further improve the performance. These results validate that the pre-trained language models are very helpful for understanding news articles.

5.3.3 Different News Information

Next, we explore whether we can learn better news representation by incorporating more news information such as abstract and body. We try two methods for news text combination. The first one is direct concatenation (denoted as *Con*), where we combine different news texts into a long document. The second one is attentive multi-view learning (denoted as *AMV*) (Wu et al., 2019a) which models each kind of news text independently and combines them with an attention network. The results are shown in Table 5. We find that news bodies are more effective than news titles and abstracts in news representation. This is because news bodies are much longer and contain richer information of news content. Incorporating different kinds of news texts such as title, body and abstract can effectively improve the performance of news recommendation, indicating different news texts contain complementary information for news representation. Incorporating the category label and the entities in news texts can further improve the performance. This is because category labels can provide general topic information, and the entities are keywords to understand the content of news. Another finding is that the attentive multi-view learning method is better than direct text combination in incorporating different news texts. This is because different news texts usually has different characteristics, and it is better

	AUC	MRR	nDCG@5	nDCG@10
Average	65.22	31.22	33.66	39.39
Attention	66.17	31.94	34.52	40.24
Candidate-Att	66.01	31.62	34.20	39.87
GRU	66.37	31.99	34.59	40.33
LSTUR	66.44	32.00	34.57	40.31
Self-Att	66.91	32.48	35.12	40.85

Table 6: Different user modeling methods.

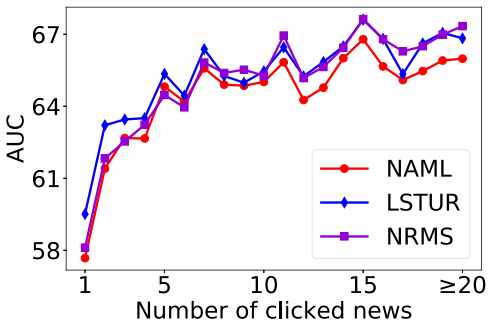


Figure 4: Users with different numbers of clicked news.

to learn their representations using different neural networks and model their different contributions using attention mechanisms.

5.4 User Interest Modeling

Most of the state-of-the-art news recommendation methods infer users' interest in news from their previously clicked news articles (Wu et al., 2019c; An et al., 2019). In this section we study the effectiveness of different user interest modeling methods. We compare 6 methods, including simple average of the representations of previously clicked news (*Average*), attention mechanism used in (Wu et al., 2019a) (*Attention*), candidate-aware attention used in (Wang et al., 2018) (*Candidate-Att*), gated recurrent unit used in (Okura et al., 2017) (*GRU*), long- and short-term user representation used in (An et al., 2019) (*LSTUR*) and multi-head self-attention used in (Wu et al., 2019c) (*Self-Att*). For fair comparison, the news representations in these methods are all generated using LSTM. The results are shown in Table 6.

We find that *Attention*, *Candidate-Att*, and *GRU* all perform better than *Average*. This is because *Attention* can select informative behaviors to form user representations, *Candidate-Att* can incorporate the candidate news information to select informative behaviors, and *GRU* can capture the sequential information of behaviors. *LSTUR* performs better than all above methods, because it can model both long-term and short-term user interest using behaviors in different time ranges. *Self-Att* can also achieve strong performance, because it can model the long-range relatedness between the historical behaviors of users for better user modeling.

We also study the influence of click history length on user interest modeling. In Fig. 4 we show the performance of three news recommendation methods, i.e., *LSTUR*, *NAML* and *NRMS*, on the users with different lengths of news click history. We find that their performance in general improves on the users with more news click records. This result is intuitive, because more news click records can provide more clues for user interest modeling. The results also show it is quite challenging to infer the interest of users whose behaviors on the news platform are scarce, i.e., the cold-start users.

6 Conclusion and Discussion

In this paper we present the MIND dataset for news recommendation research, which is constructed from user behavior logs of Microsoft News. It contains 1 million users and more than 160k English news articles with rich textual content such as title, abstract and body. We conduct extensive experiments on this dataset. The results show the importance of accurate news content understanding and user interest modeling for news recommendation. Many natural language processing and machine learning techniques such as text modeling, attention mechanism and pre-trained language models can contribute to the performance improvement of news recommendation.

In the future, we plan to extend the MIND dataset by incorporating image and video information in news as well as news in different languages, which can support the research of multi-modal and multi-lingual news recommendation. In addition, besides the click behaviors, we plan to incorporate other user behaviors such as read and engagement to support more accurate user modeling and performance evaluation. Many interesting researches can be conducted on the MIND dataset, such as designing better news and user modeling methods, improving the diversity, fairness and explainability of news recommendation results, and exploring privacy-preserving news recommendation. Besides news recommendation, the MIND dataset can also be used in other natural language processing tasks such as topic classification, text summarization and news headline generation.

Acknowledgments

We would like to thank Xi Chen, Virl Hill, Jesse Pannoni, Sally Salas and Ting Cai in the Microsoft News team for their great support in releasing this dataset and for their great help in preparing the data. We also want to thank Le Zhang, Miguel Gonzalez-Fierro and Tao Wu at Microsoft Azure for their support in Microsoft Recommenders repository and Azure resource. Finally, we thank Jingwei Yi and Ling Luo for their help on experiments.

References

Mingxiao An, Fangzhao Wu, Chuhan Wu, Kun Zhang, Zheng Liu, and Xing Xie. 2019. Neural news recommendation with long-and short-term user representations. In *ACL*, pages 336–345.

Trapit Bansal, Mrinal Das, and Chiranjib Bhattacharyya. 2015. Content driven user profiling for comment-worthy recommendations of news and blog articles. In *RecSys.*, pages 195–202. ACM.

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. In *NIPS*, pages 2787–2795.

Cheng Chen, Xiangwu Meng, Zhenghua Xu, and Thomas Lukasiewicz. 2017. Location-aware personalized news recommendation with deep semantic analysis. *IEEE Access*, 5:1624–1638.

Heng-Tze Cheng, Levent Koc, Jeremiah Harmsen, Tal Shaked, Tushar Chandra, Hrishi Aradhye, Glen Anderson, Greg Corrado, Wei Chai, Mustafa Ispir, et al. 2016. Wide & deep learning for recommender systems. In *DLRS*, pages 7–10. ACM.

Abhinandan S Das, Mayur Datar, Ashutosh Garg, and Shyam Rajaram. 2007. Google news personalization: scalable online collaborative filtering. In *WWW*, pages 271–280. ACM.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, pages 4171–4186.

Florent Garcin, Christos Dimitrakakis, and Boi Faltings. 2013. Personalized news recommendation with context trees. In *RecSys.*, pages 105–112.

Suyu Ge, Chuhan Wu, Fangzhao Wu, Tao Qi, and Yongfeng Huang. 2020. Graph enhanced representation learning for news recommendation. In *WWW*, pages 2863—2869.

Jon Atle Gulla, Lemei Zhang, Peng Liu, Özlem Özgöbek, and Xiaomeng Su. 2017. The adressa dataset for news recommendation. In *Proceedings of the international conference on web intelligence*, pages 1042–1048. ACM.

Huifeng Guo, Ruiming Tang, Yunming Ye, Zhenguo Li, and Xiuqiang He. 2017. Deepfm: a factorization-machine based neural network for ctr prediction. In *AAAI*, pages 1725–1731. AAAI Press.

Po-Sen Huang, Xiaodong He, Jianfeng Gao, Li Deng, Alex Acero, and Larry Heck. 2013. Learning deep structured semantic models for web search using clickthrough data. In *CIKM*, pages 2333–2338.

Ilija Ilievski and Sujoy Roy. 2013. Personalized news recommendation based on implicit feedback. In *Proceedings of the 2013 international news recommender systems workshop and challenge*, pages 10–15. ACM.

Margarita Karkali, Dimitris Pontikis, and Michalis Vazirgiannis. 2013. Match the news: A firefox extension for real-time news recommendation. In *SIGIR*, pages 1117–1118. ACM.

Benjamin Kille, Frank Hopfgartner, Torben Brodt, and Tobias Heintz. 2013. The plista dataset. In *Proceedings of the 2013 international news recommender systems workshop and challenge*, pages 16–23. ACM.

Michal Kompan and Mária Bieliková. 2010. Content-based news recommendation. In *International conference on electronic commerce and web technologies*, pages 61–72. Springer.

Yehuda Koren. 2008. Factorization meets the neighborhood: a multifaceted collaborative filtering model. In *KDD*, pages 426–434.

Lei Li, Ding-Ding Wang, Shun-Zhi Zhu, and Tao Li. 2011. Personalized news recommendation: a review and an experimental investigation. *Journal of computer science and technology*, 26(5):754.

Lihong Li, Wei Chu, John Langford, and Robert E Schapire. 2010. A contextual-bandit approach to personalized news article recommendation. In *WWW*, pages 661–670. ACM.

Jianxun Lian, Fuzheng Zhang, Xing Xie, and Guangzhong Sun. 2018. Towards better representation learning for personalized news recommendation: a multi-channel deep fusion approach. In *IJCAI*, pages 3805–3811.

Jiahui Liu, Peter Dolan, and Elin Rønby Pedersen. 2010. Personalized news recommendation based on click behavior. In *IUI*, pages 31–40. ACM.

Shumpei Okura, Yukihiro Tagami, Shingo Ono, and Akira Tajima. 2017. Embedding-based news recommendation for millions of users. In *KDD*, pages 1933–1942. ACM.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*, pages 1532–1543.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. In *EMNLP*, pages 2383–2392.

Steffen Rendle. 2012. Factorization machines with libfm. *TIST*, 3(3):57:1–57:22.

Jeong-Woo Son, A Kim, Seong-Bae Park, et al. 2013. A location-based news article recommendation with explicit localized semantic analysis. In *SIGIR*, pages 293–302. ACM.

Gabriel de Souza Pereira Moreira, Felipe Ferreira, and Adilson Marques da Cunha. 2018. News session-based recommendations using deep neural networks. In *DLRS*, pages 15–23. ACM.

Hongwei Wang, Fuzheng Zhang, Xing Xie, and Minyi Guo. 2018. Dkn: Deep knowledge-aware network for news recommendation. In *WWW*, pages 1835–1844.

Chuhan Wu, Fangzhao Wu, Mingxiao An, Jianqiang Huang, Yongfeng Huang, and Xing Xie. 2019a. Neural news recommendation with attentive multi-view learning. In *IJCAI-19*, pages 3863–3869.

Chuhan Wu, Fangzhao Wu, Mingxiao An, Jianqiang Huang, Yongfeng Huang, and Xing Xie. 2019b. Npa: Neural news recommendation with personalized attention. In *KDD*, pages 2576–2584. ACM.

Chuhan Wu, Fangzhao Wu, Suyu Ge, Tao Qi, Yongfeng Huang, and Xing Xie. 2019c. Neural news recommendation with multi-head self-attention. In *EMNLP-IJCNLP*, pages 6390–6395.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *NAACL*, pages 1480–1489.