# MIND: A Large-scale Dataset for News Recommendation

## 摘要：

本文提出一个新闻推荐大型数据集MIND。MIND由微软新闻的用户点击日志构建而成，拥有**100万**用户和**16万**多篇英文新闻文章，每篇文章都有丰富的标题、摘要、正文等文本内容。研究结果表明，**新闻内容的理解质量**和**用户兴趣模型的构建**在很大程度上决定着新闻报道的效果。**有效的文本表示方法**和预先训练好的**语言模型**等自然语言处理技术可以有效地提高新闻推荐的性能。

## 介绍：

1.新闻推荐的特殊性

更新快，冷启动问题非常重要；含有丰富的文本信息，需要对文本信息加以利用（标题，正文）；缺少明确的评分，用户兴趣从用户点击中含蓄表达。

## 相关工作：

1.新闻推荐中的2大问题：

如何通过丰富的文本信息表示新闻/如何从历史行为中建模用户偏好

2.新闻推荐：
特征工程：(Liu et al., 2010; Son et al., 2013; Karkali et al., 2013; Garcin et al., 2013; Bansal et al., 2015; Chen et al., 2017[1]）；

深度模型：新闻：去噪自动编码机+用户：GRU建模历史行为[2]

knowledge-aware新闻表示[3]

多视图学习，从新闻的title, body , category构建特征表示[4]

3.现存数据集

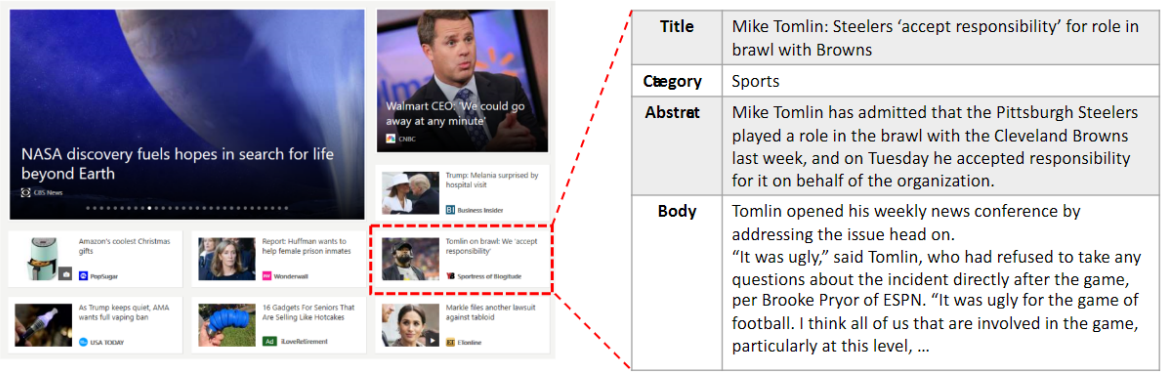
| Dataset | Language | # Users | # News | # Clicks | News information |
|---|---|---|---|---|---|
| Plista | German | Unknown | 70,353 | 1,095,323 | title, body |
| Adressa | Norwegian | 3,083,438 | 48,486 | 27,223,576 | title, body, category |
| Globo | Portuguese | 314,000 | 46,000 | 3,000,000 | no 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.

## MIND Dataset :

1.基本构成

- 每个用户至少含有5条点击信息，6周
- *[uID; t; ClickHist; ImpLog]* 用户ID；时间；历史点击ID list；新闻集合（ID，label是否点击）通过时间排序。
- test：最后一周数据；train：前五周数据
- 对于训练集中的样本，使用前四周的点击行为来构建新闻的点击历史。对于测试集中的示例，新闻点击历史提取的时间段是前五周，只保留了非空新闻点击history的样本。在训练数据中，我们使用第五周最后一天的样本作为验证集。
- 新闻结构：(ID, title, abstract, body, category)




(a) An example Microsoft News homepage (b) Texts in an example news article

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

为了便于知识型新闻推荐的搜索，我们将新闻文章的标题、摘要和正文中的实体提取到MIND dataset中，并将它们与内部的NER和实体链接工具WikiData 11中的实体链接起来。还从WikiData中提取了这些实体关系的知识三元组，并使用TransE (Bordes et al.， 2013)方法学习实体和关系的嵌入。这些实体、知识三元组以及实体和关系嵌入也包含在数据集中。

2.数据分析：

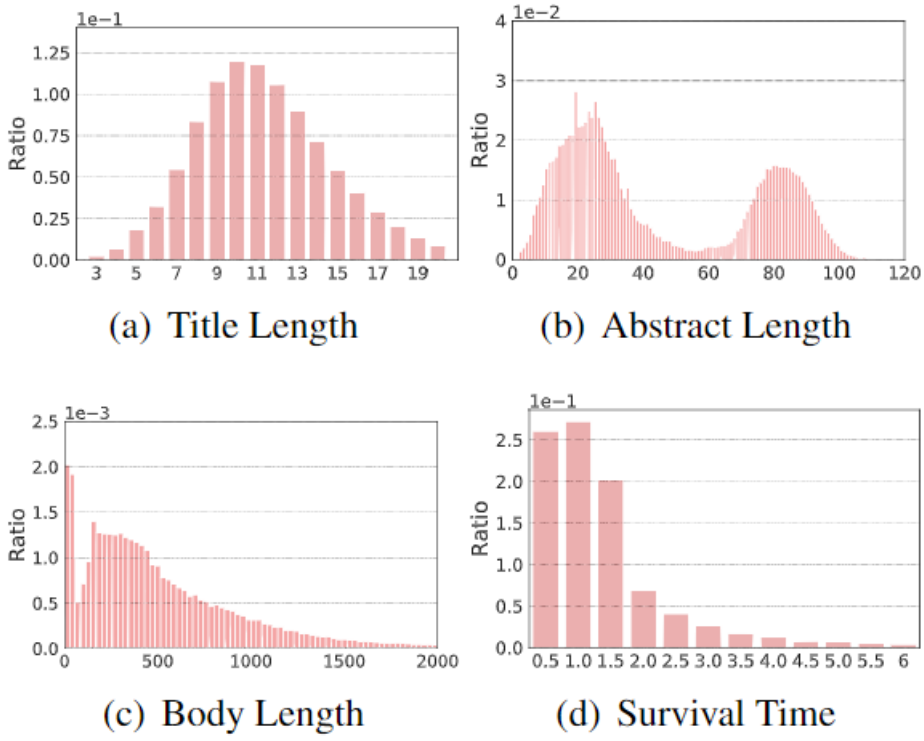



Figure 2: Key statistics of the MIND dataset.

| | | | |
|---|---|---|---|
| # 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.

Survival Time：使用新闻文章在数据集中首次出现和最后一次出现之间的时间间隔估计新闻文章的生存时间。我们发现，超过84.5%的新闻文章存活时间不到两天。这是由于新闻信息的不真实，新闻媒体总是追求最新的新闻，现有的新闻文章很快就会过时。

## 方法：

文章对比了一些推荐系统中的通用方法和新闻推荐中的常用方法，因为常用方法比较老，而且从结果上看，不适用于新闻推荐的问题。此处只列举新闻推荐中的方法。

- DFM [5]：将不同深度的网络结合起来，用于捕捉特征之间的交互
- GRU [6]：news：autoencoder（新闻内容）user：GRU（历史点击数据）
- **DKN** [7]：knowledge-aware，通过WordEmbedding和EntityEmbedding从新闻标题中学习新闻表示
- NPA [8]：个性化注意力机制，通过用户偏好，选择重要的word和news article，学习新闻和用户表示
- NAML [9]：多视图学习，结合了新闻的多种特征
- LSTUR [10]：对用户的兴趣进行长短期建模，short来自最近点击记录（GRU），long来自整个点击记录
- NRMS [11]：使用多头注意力机制来构建new表示

## 实验：

1.实验细节：

- 因为大多数新闻推荐任务只使用了新闻title，所以在测评中也只是用title。
- 为了模拟实际的新闻推荐场景，训练数据中总是有看不见的用户，我们**随机抽取一半用户进行训练**，并使用所有用户进行测试。对于那些需要单词嵌入的方法，本文使用Glove (Pen-nington et al.， 2014)作为初始化。
- 测评标准：AUC、MRR、nDCG@5和nDCG@10

| | 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.

实验结论：

- 总体来说，general方法比新闻推荐的方法效果差，原因是新闻推荐往往采用端到端的训练方式，而general则是采用手工特征。说明通过神经网络的方法获取特征相比于特征工程更加有效。
- NRMS效果较好说明，最新的语言模型比如多头自注意模型能有效地提高对新闻内容的理解和对用户兴趣的建模。
- LSTUR效果较好说明，用户建模阶段采用合适的方法也很重要。
- 冷启动问题的结果说明即便是对于没见过的用户和新闻，模型也可以进行较好的推荐。

2.新闻内容理解：

不同技术在text embedding中起到的作用：

| | 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.

实验结论：

- 基于神经网络的方法相比于传统方法在获取text表示方面效果更好。因为神经网络的方法可以跟着任务一起学习。
- self-att和LSTM相比于CNN有更好的效果，因为他们可以捕获长距离的语义特征。
- Attention可以显著提高方法性能，选择重要的单词，可以提高new表示

3.预训练模型探究：

探究更新的语言模型，比如BERT是否能够取得更好的效果。答案是肯定的，BERT基于WIKI提供更加丰富的语义，并且经过finetune后，性能进一步提升。

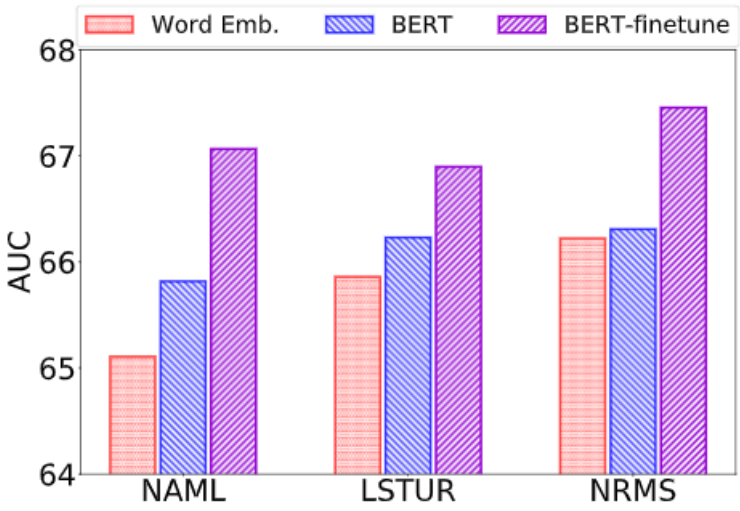



Figure 3: BERT for news representation.

4.信息来源

探究使用更丰富的信息，是否可以提高实验结果，例如增加文本信息（abstract，body）。Con & attentive multi-view learning

| | 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.

实验结论：

- 整合标题、正文、摘要等不同类型的新闻文本可以有效提高新闻推荐的性能，说明不同的新闻文本包含了对新闻表示的补充信息。
- 在新闻文本中加入类别标签和实体可以进一步提高性能。这是因为类别标签可以提供一般的主题信息，而实体是理解新闻内容的关键字。
- 注意多视角学习方法在合并不同的新闻文本时优于直接的文本组合。这是因为不同的新闻文本通常具有不同的特征，最好使用不同的神经网络来学习它们的表征，并使用注意机制来建模它们的不同贡献。

5.用户偏好建模：

探究不同的用户兴趣建模方法的有效性。

| | 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.

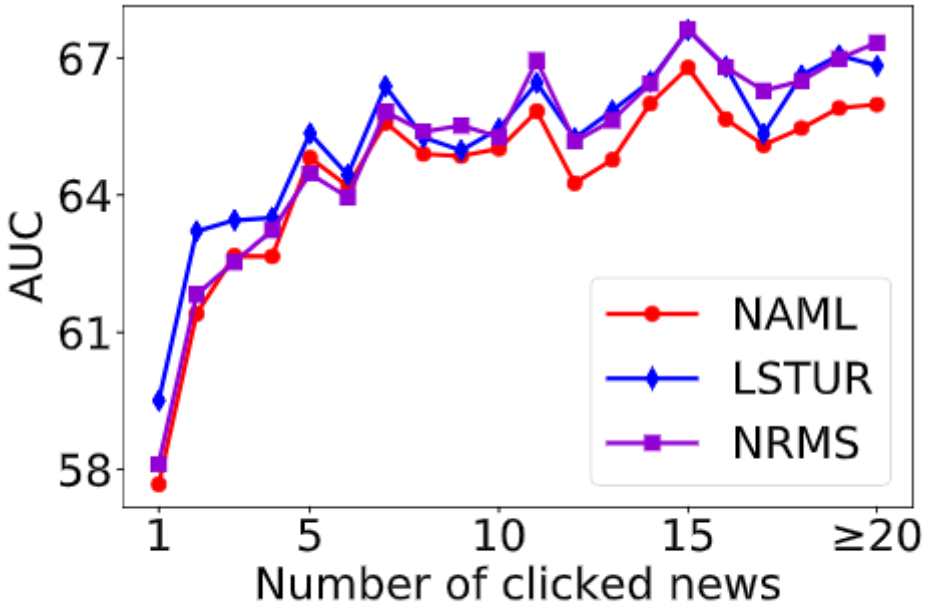



Figure 4: Users with different numbers of clicked news.

- Attention方法相较于其他方法更好，LSTUR和self-att表现良好的原因是因为建模了用户的长短期偏好。

**结论：**

在实验结果方面：对于新闻的表示，更先进的语言模型，预训练模型，Attention；更加丰富的信息（title，abstract，body）都有更好的效果，在用户表示方面，基于Attention的长短期建模的用户偏好效果较好。

在数据集方面，他们未来致力于添加图片来支持多模态任务，为用户添加更多的行为（阅读下载等行为）。


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

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

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

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

5. Jianxun Lian, Fuzheng Zhang, Xing Xie, and Guangzhong Sun. 2018. Towards better represen-tation learning for personalized news recommenda-tion: a multi-channel deep fusion approach. In IJ-CAI, pages 3805–3811. ↩

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

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

8. Chuhan Wu, Fangzhao Wu, Mingxiao An, Jianqiang Huang, Yongfeng Huang, and Xing Xie. 2019b. Npa: Neural news recommendation with personal-ized attention. InKDD, pages 2576–2584. ACM. ↩

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

10. Mingxiao An, Fangzhao Wu, Chuhan Wu, Kun Zhang, Zheng Liu, and Xing Xie. 2019. Neural news recom-mendation with long-and short-term user representa-tions. InACL, pages 336–345. ↩

11. Chuhan Wu, Fangzhao Wu, Suyu Ge, Tao Qi, Yongfeng Huang, and Xing Xie. 2019c. Neu-ral news recommendation with multi-head self-attention. InEMNLP-IJCNLP, pages 6390–6395. ↩