

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[†]

{fangzwu, yiqia, jiuche, jialia}@microsoft.com {t-danliu, xingx, jfgao, winniew,
mingzhou}@microsoft.com

{wu-ch19, qit16}@mails.tsinghua.edu.cn

Remember.

- When reading technical papers, always identify:
 - the problem it wants to solve
 - What has been done to solve it
 - What are the limitations of existing solutions
 - Its proposed solution
 - Its experiments/analysis/findings
 - Its conclusion
 - Its recommendations

Pair up and fill out this table. In this session, we will focus our attention on the MIND

Problem	
Existing solution/s	
Limitations	
Proposed solution	

The Problem

- There is no large-scale dataset in the English language. This prevents the creation of state of the art methods to personalize news recommendation.
- NOTE: The unique issues to news recommendation were also noted:
 - Cold start. News articles on websites are updated quickly, and posted continuously, so news articles expire in a short time (Remember you want “fresh” news all the time.)
 - Representation issue. CF approaches do not work, the content needs to be understood.
 - Implicit rating. There is no explicit way to rate the news.

Existing solution/Limitations (Dataset)

- Large scale data are proprietary datasets exist (Microsoft/Bing News)
- Small, non-English datasets

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.

Existing solution/Limitations (News Recommendation)

- How to represent news articles (item profile)
 - URLs, categories,
- How to model users' interests (user profile)
 - Demographics, location, behavior from other sites
- Deep Learning
 - End-to-end approaches: learn representations of news articles (and user interest (using an autoencoder >> similar to RBMs) by Okura et al
 - Used CNN to create word embeddings and entity embeddings (content + user preference) by Wang et al
 - Used title, body, and category using LSTM by Wu et al]
- ISSUE: datasets are not publicly available, methods cannot be validated, and cannot create their own state of the art methods

Proposed solution (Dataset)

- Create a large-scale dataset called Microsoft News Dataset (MIND)
- Conduct experiments to show that MIND is a good testbed for state of the art methods

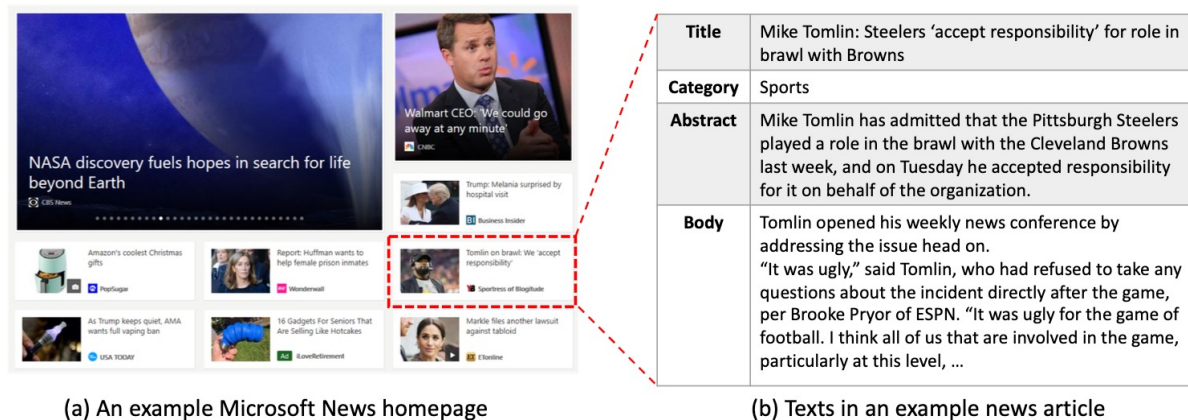


Figure 1: An example homepage of Microsoft News and an example news article on it.

- Click behavior (implicit feedback) of 1M users
- 160,000 English news articles

Proposed solution (Dataset)

- Microsoft News Dataset from Microsoft News
- 1 million users who had at least 5 news click records for 6 weeks (October 12 – November 22, 2019), anonymized
- Impression log: records the news articles displayed to a user when she visits MSN at a specific time, and click behaviors on articles.
- Data includes news click history that in turn became labeled samples for training and verification.

Data format/training/validation/test

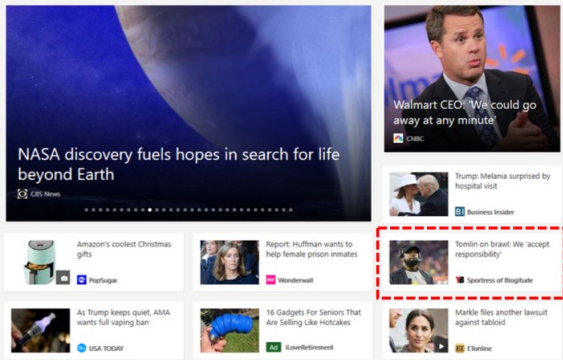
$[uID, t, ClickHist, ImpLog]$

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., $[(nID_1, label_1), (nID_2, label_2), \dots]$, where nID is news article ID and $label$ is the click label (1 for click and 0 for non-click). We used

- Remember: data for 6 weeks.
 - Weeks 1 - 4: click history
 - Week 5: training
 - Last Day of Week 5: validation
 - Week 6: testing

Data pre-processing

- Extraction of "rich entities" using named entity recognition (NER) and linking tool (via WikiData) and TransE. >> Look for these in your dataset!



(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.

EDA: The MIND

# 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.

Training	2,186,683
Validation	365,200
Test	2,341,619

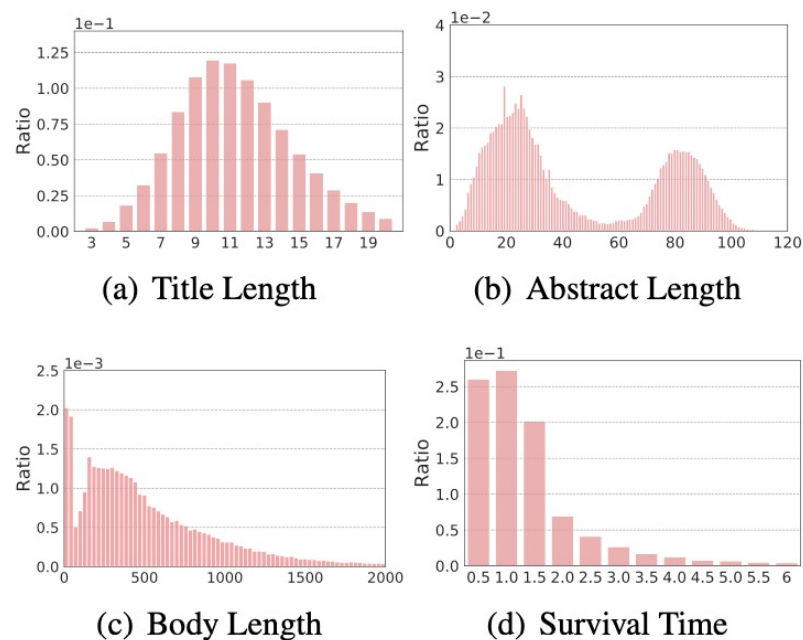


Figure 2: Key statistics of the MIND dataset.

Proposed solution (Personalized News Recommendation)

- Use NLP techniques such as CNN and transformers to represent news articles (item profile)
- Learn user interest representation by learning document representation from its sentences. (user profile)
- Formulate news recommendation as a special text matching problem

END for today.