MIND: MIcrosoft News Dataset

A Large-Scale English Dataset for News Recommendation Research

UNSW SYDNEY

COMP9727 Project Team

Linbo Zhang z5352294
Jinghan Wang z5286124
Junyu Li z5467278

Date 02-08-2024

Problem Statement

- Personalized news recommendations
- Aim to deliver relevant articles to users based on their reading history and preferences.
- The primary challenge is to accurately capture user interests and manage the dynamic nature of news content.

Dataset

- Date Range:
  Covers user behavior logs from October 12 to November 22, 2019.

- Rich Users:
  Encompasses data from 1 million users, each with at least five news click records.

- Rich News Content:
  Includes news articles with IDs, titles, abstracts, bodies, and category labels, along with entity information linked to WikiData for knowledge-aware recommendations.

Dataset

- Train, validation, test split ready

	From	To
Train	Nov 9	Nov 14
Validation	Nov 15	Nov 15
Test	Nov 16	Nov 22

Dataset

- Example row from the behaviors.tsv

User	Timestamp	History	Impressions
U87243	11/10/2019 11:30:54 AM	N8668 N39081 N65259 N79529 N73408 N43615 N29379 N32031 N110232 N101921 N12614 N129591 N105760 N60457 N1229 N64932	N78206-0 N26368-0 N7578-0 N58592-0 N19858-1 N58258-0 N18478-0 N2591-1 N97778-0 N32954-0 N94157-1

Dataset

- behaviors.tsv

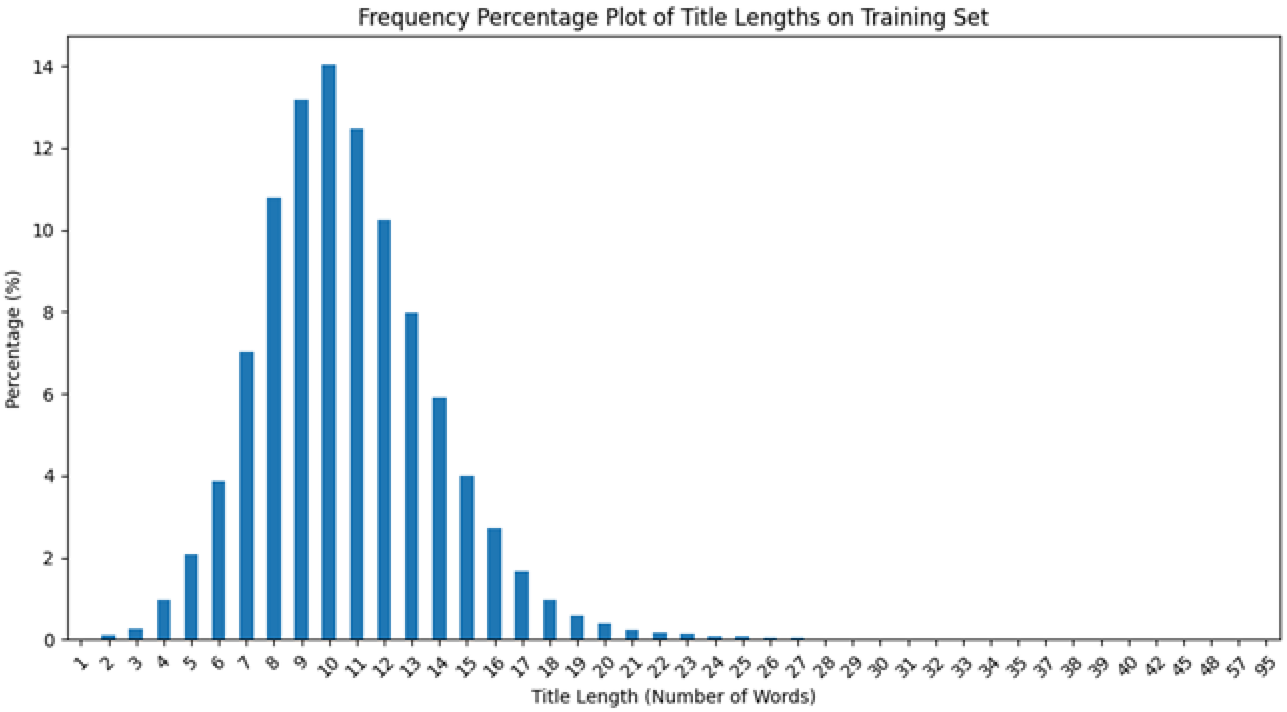
User	Timestamp	History	Impressions
U87243	11/10/2019 11:30:54 AM	N8668 N39081 N65259 N79529 N73408 N43615 N29379 N32031 N110232 N101921 N12614 N129591 N105760 N60457 N1229 N64932	N78206-0 N26368-0 N7578-0 N58592-0 N19858-1 N58258-0 N18478-0 N2591-1 N97778-0 N32954-0 N94157-1

Dataset

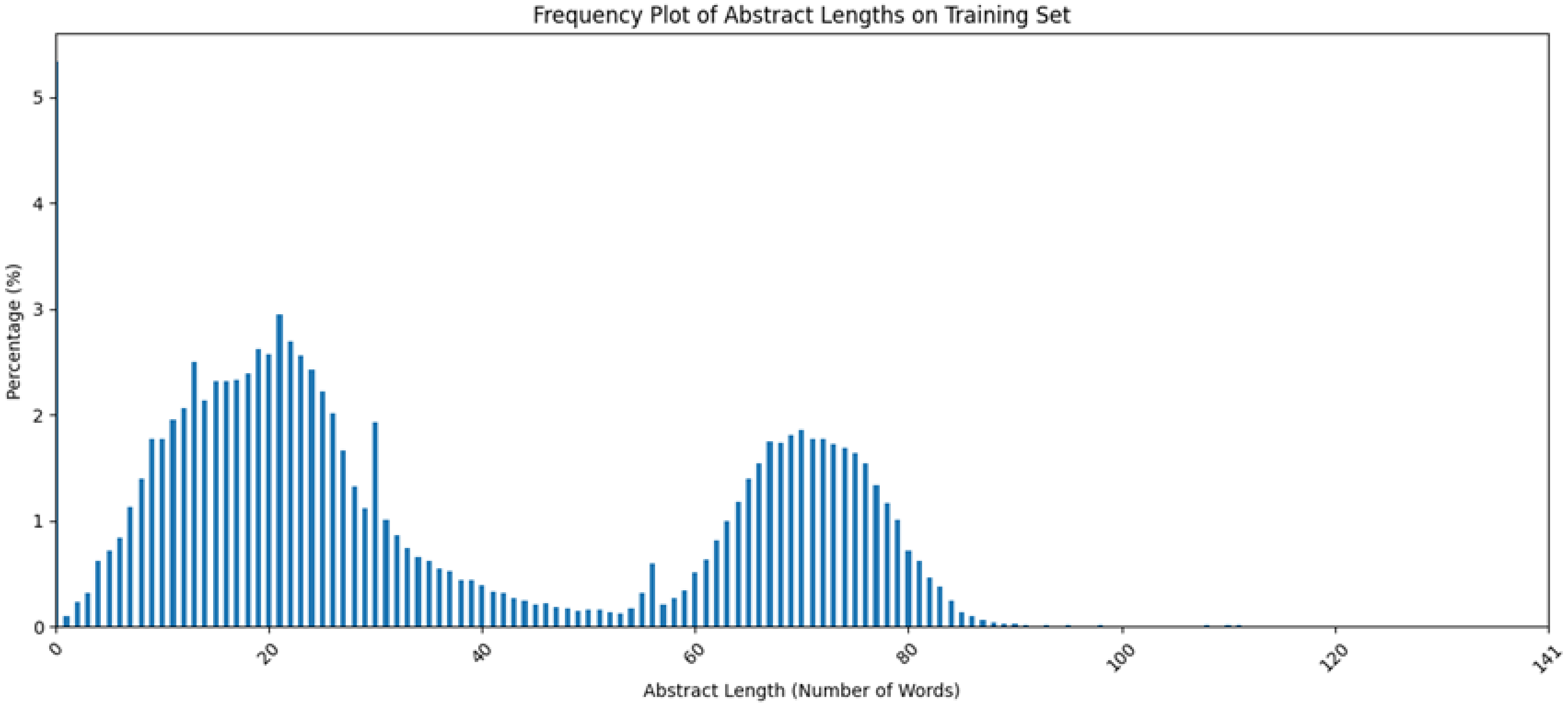
- news.tsv

News ID	Category	Subcategory	Title	Abstract	URL	Title Entity	Abstract Entity
N23144	health	weightloss	50 Worst Habits For Belly Fat	These seemingly harmless habits are holding you back and keeping you from shedding that unwanted belly fat for good.	https://assets.msn.com/labs/mind/AAB19MK.html	[{"Label": "Adipose tissue", "Type": "C", "WikidataId": "Q193583", "Confidence": 1.0, "OccurrenceOffsets": [20], "SurfaceForms": ["Belly Fat"]}]	[{"Label": "Adipose tissue", "Type": "C", "WikidataId": "Q193583", "Confidence": 1.0, "OccurrenceOffsets": [97], "SurfaceForms": ["belly fat"]}]

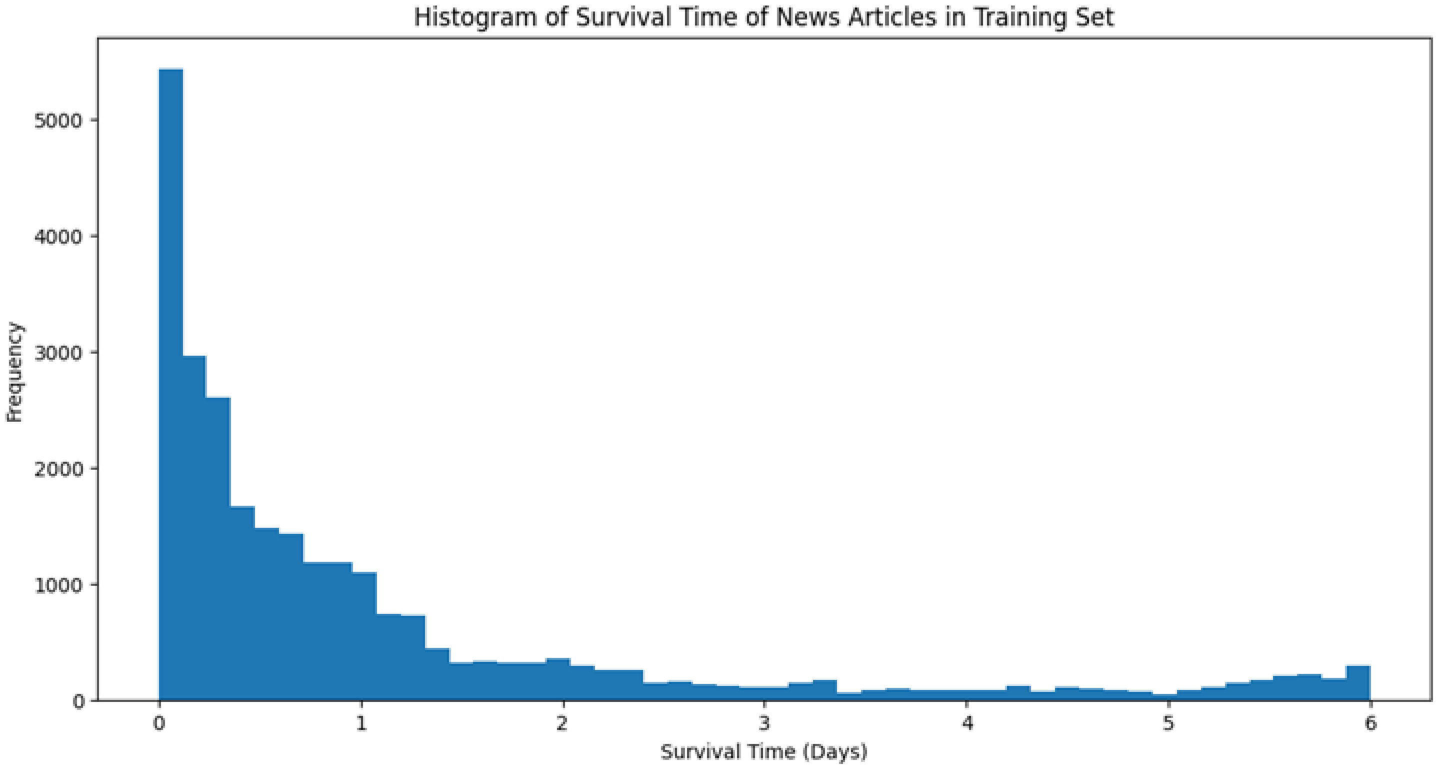
Data Analysis



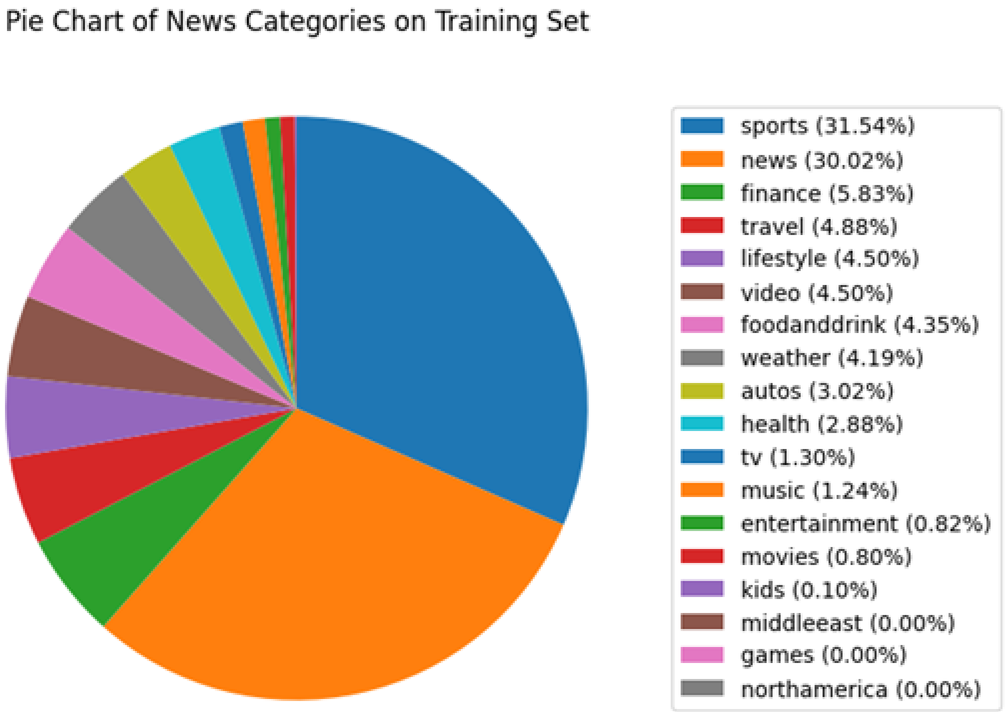
Data Analysis



Data Analysis



Data Analysis



Methods: Evaluation Metrics

- AUC: Measures the area under the ROC curve, indicating the model's ability to distinguish between clicked and unclicked news.

- MRR: Mean Reciprocal Rank assesses the ranking quality of the first relevant news article.

- nDCG@5: Normalized Discounted Cumulative Gain at rank 5 evaluates the ranking quality of the top 5 recommended news.

- nDCG@10: Normalized Discounted Cumulative Gain at rank 10 evaluates the ranking quality of the top 10 recommended news.

Methods: Smaller Set

- Subset the dataset using the provided MINDsmall set.
- 50,000 users across training, validation and testing set.

Model: LibFM (Factorization Machine)

- Model Type: LightFM Hybrid Recommender (LINK)

- Data Preparation: Combines TF-IDF vectors of user's history and news articles

- Dataset Initialization: Encodes user and news IDs, and builds interaction and item feature matrices

- Training: Utilizes WARP loss function for ranking quality

- Evaluation: Assesses model performance using AUC score

Model: Neural Collaborative Filtering (NCF)

- User and Item Embeddings: Captures latent features from interactions

- Content-Based Features: Incorporates TF-IDF vectors for user's history and current news

- Combined Features: Merges collaborative and content-based data

- Output: Predicts click probability for news articles

Dataset further justification

The MIND dataset was proposed by Microsoft, contains about 160k English news articles and more than 15 million impression logs generated by 1 million users.

We use the MIND-small sub dataset which randomly sampling 50,000 users and their behavior logs.

Why: We choose it because it contains the important features as in the large dataset, and it also have relatively large enough samples to generate when we using the content-based filtering in the

	News ID	Category	SubCategory	Title	Abstract	URL	Title Entities	Abstract Entities
0	N55528	lifestyle	lifestyleroyals	The Brands Queen Elizabeth, Prince Charles, an...	Shop the notebooks, jackets, and more that the...	https://assets.msn.com/labs/mind/AAGH0ET.html	[{"Label": "Prince Philip, Duke of Edinburgh",...	[]
1	N19639	health	weightloss	50 Worst Habits For Belly Fat	These seemingly harmless habits are holding yo...	https://assets.msn.com/labs/mind/AAB19MK.html	[{"Label": "Adipose tissue", "Type": "C", "Wik...	[{"Label": "Adipose tissue", "Type": "C", "Wik...
2	N61837	news	newsworld	The Cost of Trump's Aid Freeze in the Trenches...	Lt. Ivan Molchanets peeked over a parapet of s...	https://assets.msn.com/labs/mind/AAJgNsz.html	[]	[{"Label": "Ukraine", "Type": "G", "WikidataId...
3	N53526	health	voices	I Was An NBA Wife. Here's How It Affected My M...	I felt like I was a fraud, and being an NBA wi...	https://assets.msn.com/labs/mind/AACk2N6.html	[]	[{"Label": "National Basketball Association", ...
4	N38324	health	medical	How to Get Rid of Skin Tags, According to a De...	They seem harmless, but there's a very good re...	https://assets.msn.com/labs/mind/AAAKEkt.html	[{"Label": "Skin tag", "Type": "C", "WikidataI...	[{"Label": "Skin tag", "Type": "C", "WikidataI...

Dataset further justification

The two tsv that are important:

1. behaviors.tsv: #interactions 156,965

2. news.tsv

timestamp: calculate to --> epochhrs

click_history: before the impressionId happens, what news articals did the user view

impressions: for each certain impressions, Nxxx-1: clicked; Nxxx-0: not clicked

	impressionId	userId	timestamp	click_history	impressions
0	1	U13740	11/11/2019 9:05:58 AM	N55189 N42782 N34694 N45794 N18445 N63302 N104...	N55689-1 N35729-0
1	2	U91836	11/12/2019 6:11:30 PM	N31739 N6072 N63045 N23979 N35656 N43353 N8129...	N20678-0 N39317-0 N58114-0 N20495-0 N42977-0 N...
2	3	U73700	11/14/2019 7:01:48 AM	N10732 N25792 N7563 N21087 N41087 N5445 N60384...	N50014-0 N23877-0 N35389-0 N49712-0 N16844-0 N...
3	4	U34670	11/11/2019 5:28:05 AM	N45729 N2203 N871 N53880 N41375 N43142 N33013 ...	N35729-0 N33632-0 N49685-1 N27581-0
4	5	U8125	11/12/2019 4:11:21 PM	N10078 N56514 N14904 N33740	N39985-0 N36050-0 N16096-0 N8400-1 N22407-0 N6...

Dataset further justification

The two tsv that are important:

1. behaviors.tsv

extract features and generate embedding from the “title” and “abstract” (later used in model)

2. news.tsv: #articals 51282

	itemId	category	subcategory	title	abstract	url	tle_entities	abstract_entities
0	N55528	lifestyle	lifestyleroyals	The Brands Queen Elizabeth, Prince Charles, an...	Shop the notebooks, jackets, and more that the...	https://assets.msn.com/labs/mind/AAGH0ET.ht	"Label": 'rince hilip, Duke dinburgh",...	[]
1	N19639	health	weightloss	50 Worst Habits For Belly Fat	These seemingly harmless habits are holding yo...	https://assets.msn.com/labs/mind/AAB19MK.h	"Label": ıdipose ssue", ype": "C", Vik...	[{"Label": "Adipose tissue", "Type": "C", "Wik...
2	N61837	news	newsworld	The Cost of Trump's Aid Freeze in the Trenches...	Lt. Ivan Molchanets peeked over a parapet of s...	https://assets.msn.com/labs/mind/AAJgNsz.ht		[{"Label": "Ukraine", "Type": "G", "WikidataId...

Dataset processing demonstration

	userId	timestamp	click	click_history	epochhrs
0	U1	2024-08-01 09:00:00	[A1, A2]	A0 A3	4768569
1	U2	2024-08-01 10:00:00	[B1]	B0	4768570

why?

expanding click

	userId	timestamp	click	click_history	epochhrs
0	U1	2024-08-01 09:00:00	A1	A0 A3	4768569
1	U1	2024-08-01 09:00:00	A2	A0 A3	4768569
2	U2	2024-08-01 10:00:00	B1	B0	4768570

Concatenating Historical Clicks with Raw Behaviour

	userId	timestamp	click	click_history	epochhrs	noclicks
0	U1	2024-08-01 09:00:00	A1	A0 A3	4768569	NaN
1	U1	2024-08-01 09:00:00	A2	A0 A3	4768569	NaN
2	U2	2024-08-01 10:00:00	B1	B0	4768570	NaN
3	U1	NaN	A0	NaN	4768569	[]
4	U1	NaN	A3	NaN	4768569	[]
5	U2	NaN	B0	NaN	4768569	[]

set the cutoff = 50;

filtering out click that less than 50 clicks

11.526% of the total.

Number of interactions in the behaviour dataset: 931302
Number of users in the behaviour dataset: 49949
Number of articles in the behaviour dataset: 4595

	epochhrs	userId	click	noclicks
0	437073.0	U13740	N55689	[N35729]
1	437106.0	U91836	N17059	[N20678, N39317, N58114, N20495, N42977, N2240...
2	437143.0	U73700	N23814	[N23877, N35389, N49712, N16844, N59685, N2344...
3	437069.0	U34670	N49685	[N35729, N33632, N27581]
4	437083.0	U19739	N33619	[N53696, N25722]

observation:
#items(news articles) << #users

MF Method

How to split the training set, validation set, test set?

```
# Split into 70% training set, 15% validation set, 15% test set
test_time_th = behaviour['epochhrs'].quantile(0.85)
valid_time_th = behaviour['epochhrs'].quantile(0.7)
train = behaviour[behaviour['epochhrs']< valid_time_th].copy()
```

70:15:15 **why?**

Temporal Relevance: Recommendations are typically more relevant if they are based on recent interactions.

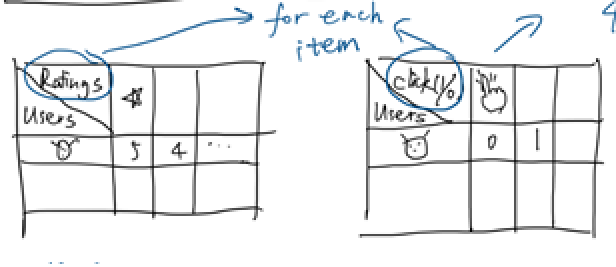
Avoid Data Leakage: Using future data to predict past interactions can lead to misleadingly high performance metrics.

Time leakage (e.g. splitting a time-series dataset randomly instead of newer data in test set using a TrainTest split or rolling-origin cross validation)

MF Method

MF: Matrix Factorization :

We has 2451 items (News)
49832 users



user	click	no-click
U_1	A_1	$N-U_1$
		$N-U_1$
U_1	A_2	
U_2	A_3	$N-U_2$
U_3	A_4	$N-U_3$

Negative Sampling : randomly assign some negative no-click items (News) to the user-click row, to represent the lack of preferences of users, to form the matrix.

— Because it's a binary (0/1) data.

≠

unlike the Movie rating (0 – 5) numbers.

MF Method

```python
# Build a matrix factorization model
class NewsMF(pl.LightningModule):
    def __init__(self, num_users, num_items, dim = 100, dropout_prob=0.2, reg=0.01): # add regularization
        super().__init__()
        self.dim=dim
        self.num_users = num_users
        self.num_items = num_items
        self.reg = reg
        self.useremb = nn.Embedding(num_embeddings=num_users, embedding_dim=dim)
        self.itememb = nn.Embedding(num_embeddings=num_items, embedding_dim=dim)

        self.dropout = nn.Dropout(p=dropout_prob) # the drop out probablity is set to 0.2


    def step(self, batch, batch_idx, phase="train"):
        batch_size = batch['userIdx'].size(0)
        uservec = self.useremb(batch['userIdx'])
        itemvec_click = self.itememb(batch['click'])

        # Apply dropout to embeddings
        uservec = self.dropout(uservec)                    # added drop out
        itemvec_click = self.dropout(itemvec_click)

        # For each positive interaction,sample a random negative
        neg_sample = torch.randint_like(batch["click"],1,self.num_items)
        itemvec_noclick = self.itememb(neg_sample)
        itemvec_noclick = self.dropout(itemvec_noclick)  # Apply dropout to negative samples

        score_click = torch.sigmoid((uservec*itemvec_click).sum(-1).unsqueeze(-1))
        score_noclick =  torch.sigmoid((uservec*itemvec_noclick).sum(-1).unsqueeze(-1))

        # Compute loss as binary cross entropy (categorical distribution between the clicked and the no clicked item)
        scores_all = torch.concat((score_click, score_noclick), dim=1)
        target_all = torch.concat((torch.ones_like(score_click), torch.zeros_like(score_noclick)),dim=1)
        # loss = F.binary_cross_entropy(scores_all, target_all)
        # return loss
        loss = F.binary_cross_entropy(scores_all, target_all)
        reg_loss = self.reg * (self.useremb.weight.norm(2) + self.itememb.weight.norm(2)) # add regularization
        return loss + reg_loss
```

MF Method result

For example, we can randomly choose a News (ID: Nxxx) and using the item embedding learned from MF, to generate 5 most similar news article.

We can see how well the model works by also looking at the category or subcategory here.

E.g. N3259 47186 both in sports - football_nfl; N43083 46582 is lifestyle and foodanddrink

It means that the model can represent similar characteristics of news in terms of users preferences, so that is very likely prefered by similar users as well.

```python
ind = item2ind.get("N3259")
# This calculates the cosine similarity and outputs the 5 most similar articles w.r.t to ind in descending order
similarity = torch.nn.functional.cosine_similarity(itememb[ind], itememb, dim=1)
most_sim = news[~news.ind.isna()].iloc[(similarity.argsort(descending=True).numpy()-1)]
most_sim.head(5)
```

	itemId	category	subcategory	title	abstract	url	title_entities	abstract_entities	ind	n_click_t
1317	N3259	sports	football_nfl	Mayfield's postgame outfit gets all the memes	Social media had a field day with Baker Mayfie...	https://assets.msn.com/labs/mind/AAJNKsO.html	[{"Label": "Baker Mayfield", "Type": "P", "Wik...	[{"Label": "Baker Mayfield", "Type": "P", "Wik...	1318.0	
169	N43083	lifestyle	lifestylehomeandgarden	What It Was Like Inside the Homes of the Pilgrims	There's a lot of folklore surrounding the firs...	https://assets.msn.com/labs/mind/AAJOJle.html	[{"Label": "Pilgrims (Plymouth Colony)", "Type...	[{"Label": "Pilgrims (Plymouth Colony)", "Type...	170.0	
1912	N47186	sports	football_nfl	Report: Jaguars QB Gardner Minshew could be pl...	Minshew has taken Jacksonville by storm in 201...	https://assets.msn.com/labs/mind/AAJKOv1.html	[{"Label": "Gardner Minshew", "Type": "N", "Wi...	[{"Label": "Gardner Minshew", "Type": "N", "Wi...	1913.0	
3400	N22836	autos	autossema	Here's the Real 2020 Ford Bronco in Off-Road-R...	The production Bronco takes shape and if the r...	https://assets.msn.com/labs/mind/AAJQB8U.html	[{"Label": "Ford Bronco", "Type": "V", "Wikida...	[{"Label": "Ford Bronco", "Type": "V", "Wikida...	3401.0	
2974	N46582	foodanddrink	restaurantsandnews	Take a look at IKEA food courts around the world	Scroll through these different IKEA food court...	https://assets.msn.com/labs/mind/AAD7y76.html	[{"Label": "IKEA", "Type": "O", "WikidataId": ...	[{"Label": "IKEA", "Type": "O", "WikidataId": ...	2975.0	

MF Method result

```
# Example: Recommend top 5 news articles for a user
user_id = test['userIdx'].iloc[3616]  # Replace with the desired user index
recommended_news = recommend_news(user_id, mf_model, top_k=5)
print(f"Recommended news articles for user {user_id}: {recommended_news}")

# Evaluate on the test set
test['recommended'] = test['userIdx'].apply(lambda x: recommend_news(x, mf_model, top_k=5))

print(test.head(5))
```

```
Recommended news articles for user 3725: ['N20567', 'N685', 'N33976', 'N18004', 'N25677']
   epochhrs  userId  click                                         noclicks  \
1  437106.0  U91836      0                  [0. 0. 0. 0. 0. 0. 0. 0. 0. 0]
2  437143.0  U73700      0  [0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. ...
6  437110.0  U46596      0                                     [0. 0. 0. 0]
7  437122.0  U79199      0                                        [0. 0. 0]
9  437145.0  U89744      0  [0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 189. 0...

   userIdx                                   recommended      pred
1      733  [N48325, N828, N56742, N15317, N35305]  0.037991
2     5117  [N57886, N16965, N57998, N30561, N61267]  0.049031
6     5122  [N27922, N17449, N22397, N40772, N58045]  0.436584
7     5123  [N23887, N45610, N46827, N4434, N57537]  0.400623
9     5124  [N13429, N5905, N16317, N51840, N43347]  0.426119
```

The NRS give recommendations by the MF, for an arbitrary user (test with U3725) and we get the 5 example news article to this user.

The NRS can adjust the TopN that show to the user after the ranking of scoring (which is (0,1) decimal numbers (pred)).

The values are from the product of two component matrix: user embedding and word embedding calculated before.

NRS Problem Pros&Cons

- **Advantages**:
- To the demand side: Enhance the user experience.
- To the supply side: It can attract higher volume of traffic to the news websites. News organisations can help users navigate so fostering stronger and deeper connections with audiences over time (Vrijenhoek et al., Citation2021).
- **Disadvantages**:
- To the demand side: There may be the formation of an informed democratic citizenry or citizens’ information behaviour (Moeller et al., Citation2016).
- To the supply side: The existed NRS may primarily driven by algorithms based on user preferences and popularity metrics rather than on human judgement, this can ultimately also affect journalistic selection and creation practices (Carlson, Citation2018; Møller, Citation2022b; Napoli, Citation2014).

(E.g. the fixed pattern or tone in the news articals)

News recommendation characteristics

News recommendation has unique challenges:

- Dynamic Content: News articles are constantly being updated.
- Cold-Start Problem: New articles and users frequently appear with no prior interaction history.
- Implicit Feedback: User interactions are typically implicit (clicks) rather than explicit ratings.

Thus, we will use the Neural Network as one of the main model implementation, as they can learn the features from word embeddings and do not depend much on the historical/interactive data. We will fine-tune the parameters of the Network to minimize loss using cross-entropy.

Competitor analysis

- The multi dimensions that we focus on, in terms of "user experience":

Usability, Usefulness, Effectiveness or Satisfactory interaction with the system. (Konstan and Riedl 2012; Knijnenburg et al. 2012).

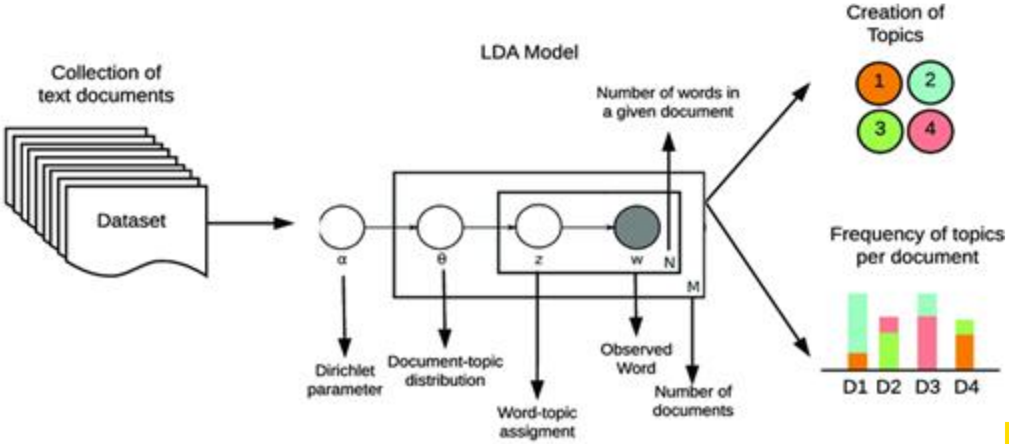
- News sources and agencies

Such as CNN, BBC, New York Times, The Washington Post

Through their news webpage, as well as the mobile apps

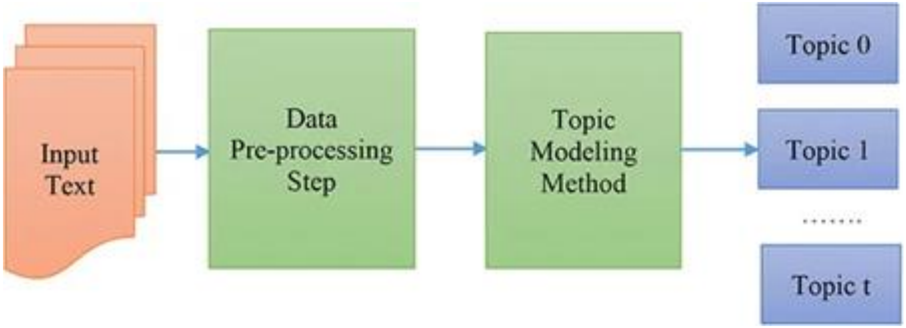
Competitor analysis

- **The existed methods:**
- Topic modeling and **Latent Dirichlet Allocation** (LDA) methods :
- LDA-based recommendation systems work by extracting latent topics from the textual content associated with items or user interactions. Instead of relying solely on explicit user-item interactions (such as ratings or clicks), these systems consider the semantic context of the items.



Competitor analysis

- **How LDA works and its effectiveness:**
- For instance, in a news recommendation system, LDA can discover topics like politics," "technology," or "sports," and then recommend articles to users based on their historical interests in these topics.
- These systems excel at understanding the underlying themes and content structures within news articles, enabling them to deliver personalized and relevant news feeds to users.



The word cloud for the topics:

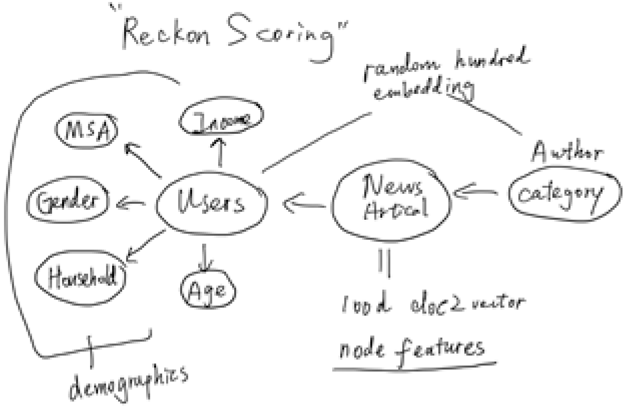
Attempt for GNN design

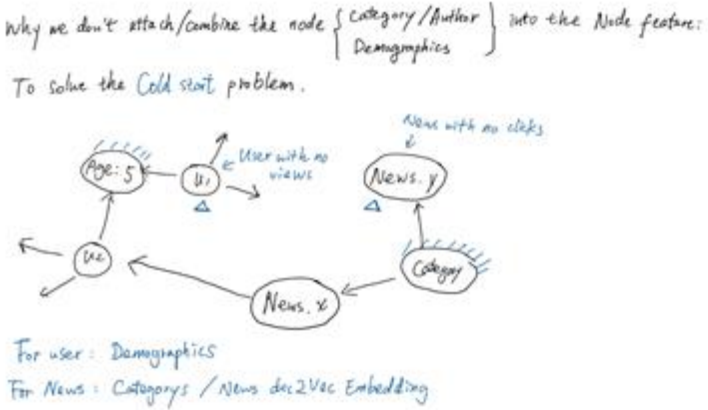
- Reason why GNN may work

Highly dynamic user behavior: News readers may have long-term or short-term preferences that evolve over time, either gradually or abruptly.

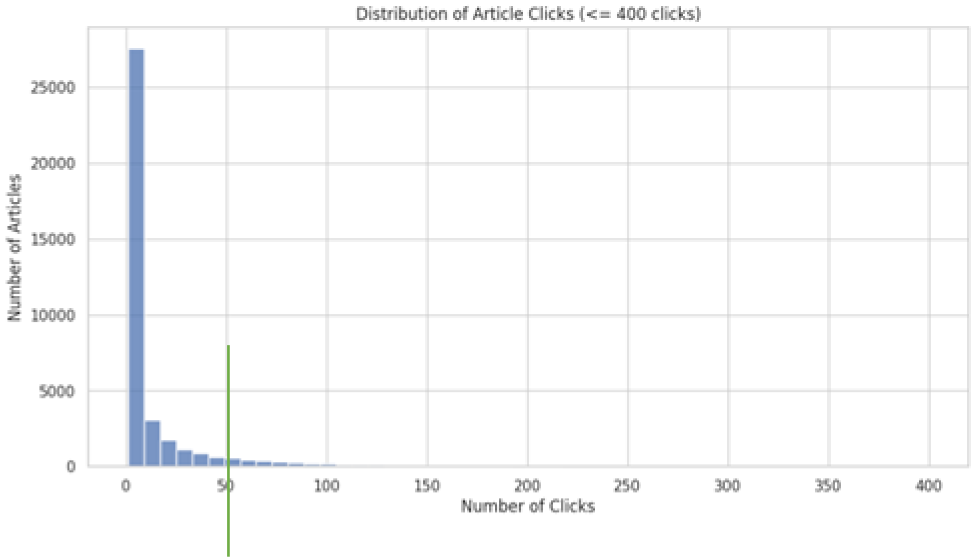
The GNN is useful when the entities have complex properties.

- We are also considering the problems with GNN, as discussed before, where collaborative filtering may depend on more sophisticated datasets having user interactions information.





Evaluation



1. Imbalanced density of Number of clicks distribution.
Too many news articles may lack of large enough click size, since the number of user is much larger than items

2. User-item interaction only has click (1) / noclick (0).

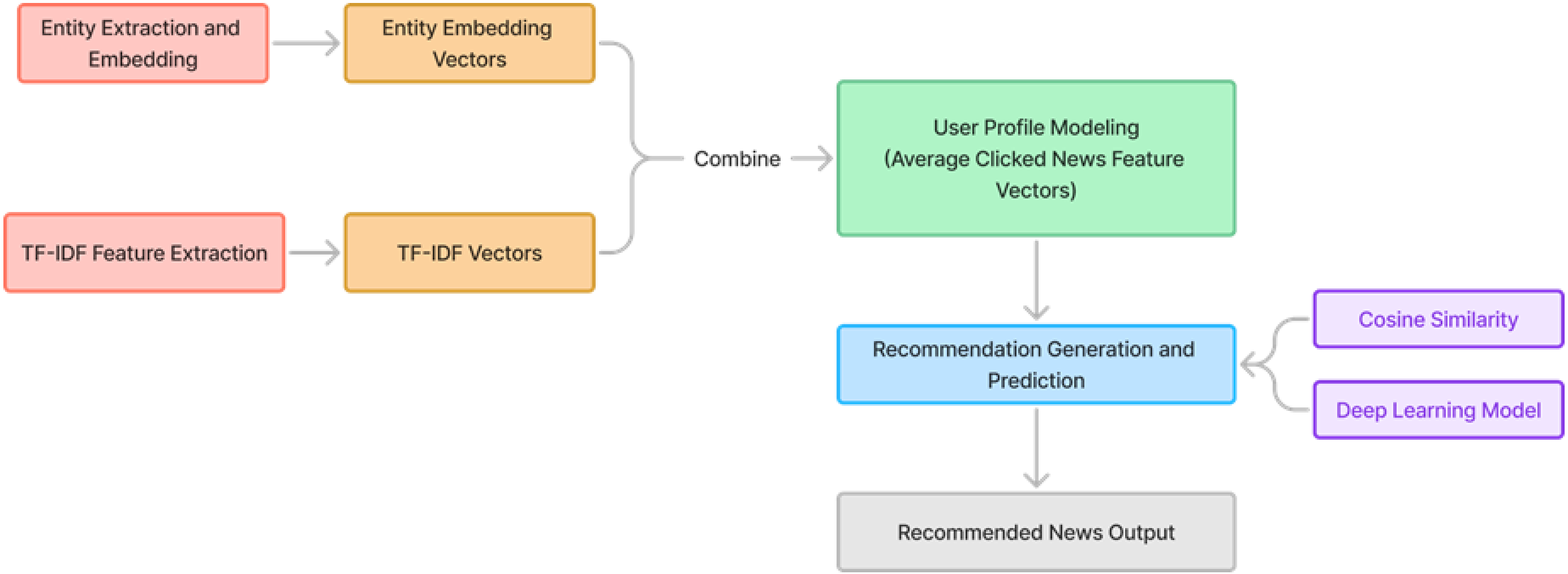
Thus is a binary classification problem for the NN in the training process.

Lack of demographics of users to utilize in content-based filtering

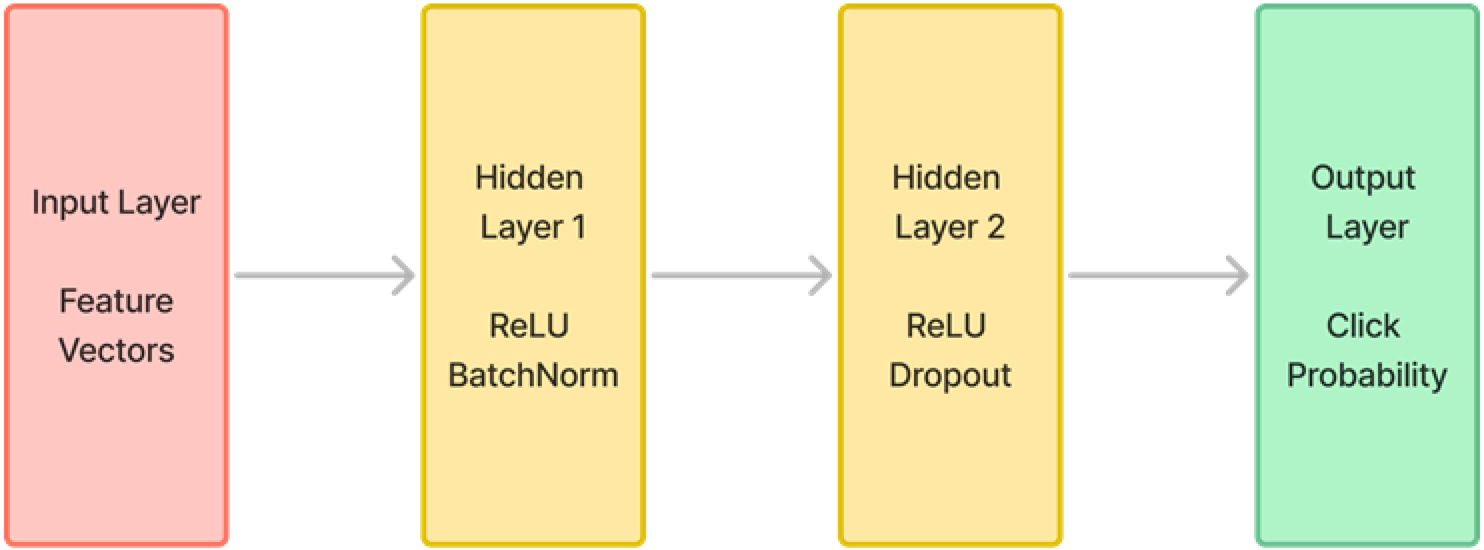
Model: Neural Collaborative Filtering (NCF)

- User and Item Embeddings: Captures latent features from interactions
- Content-Based Features: Incorporates TF-IDF vectors for user's history and current news
- Combined Features: Merges collaborative and content-based data
- Output: Predicts click probability for news articles

Deep Learning Model



Deep Learning Model



Conclusion

GNN		• Can capture complex high-order relationships between users and news • Modeling deeper interactions through message-passing mechanisms
MF	AUC 0.69	• Handle sparse and incomplete data, reduce the dimensionality and complexity of data • Suffer from overfitting and underfitting problem, may affect the accuracy and generalization of the recommendations
LibFM		• Performs well when handling high-dimensional sparse data • Can effectively capture interactions between features, offering more expressiveness than linear models
NCF	AUC 0.8	• Able to understand high-dimensional sparse data • The embeddings can effectively capture the interaction and extract the features
Deep Learning model	AUC 0.67	• Requires manual design and extraction of features • Unable to capture complex relationships

Thanks for your watching