MERCY UNIVERSITY

"NLP PROJECT"

PROJECT PROPOSAL

Team Members:

Ann Rose Joju Edakkalathur
Dong Gyu Noh
Ravisankar Chengannagari

Submitted to

Professor Agnideven PalanisamySundar

PERSONALIZED NEWS RECOMMENDATION SYSTEM

ABSTRACT

In today's digital landscape, users are exposed to an overwhelming volume of information. News platforms, in particular, face the challenge of maintaining user engagement by delivering relevant and timely content. This project addresses that challenge by developing a personalized news recommendation system using the Microsoft News Dataset (MIND), a large-scale collection of anonymized user interactions. We will build and evaluate several recommendation models, ranging from traditional content-based and collaborative filtering to advanced deep learning approaches. The goal is to accurately predict which articles users are likely to engage with, enhancing personalization and overall user experience. Model performance will be compared using standard metrics such as NDCG and MAP to identify the most effective recommendation strategy.

PROBLEM STATEMENT

With the vast and fast-changing flow of online news, users often struggle to find articles that match their individual interests. Traditional platforms lack personalization, leading to missed relevant content and lower engagement. The challenge lies in designing a recommendation system that can understand user preferences, adapt to dynamic behavior, and provide meaningful suggestions in real time.
This project tackles that problem using the MIND-small dataset to explore and compare multiple recommendation strategies from simple popularity-based methods to deep learning models to determine which approach best predicts user interests and maximizes engagement.

MOTIVATION

Recommendation systems have become integral to modern digital experiences, powering platforms from e-commerce to streaming services. News recommendation, however, presents unique challenges due to rapidly changing content and user interests. Through this project, our team aims to gain hands-on experience in applying machine learning and NLP methods to a real-world dataset while contributing to more intelligent content delivery.
Beyond technical learning, we are motivated by the broader impact of personalization helping users discover information that truly matters. This project provides an opportunity to combine technical skills with meaningful problem-solving, preparing us for future work in AI, data science, and product innovation.

DESCRIPTION

This project focuses on building a personalized news recommendation system using the MIND-small dataset, which includes anonymized user behaviors and detailed article information. We

will implement and compare multiple approaches, starting from traditional content-based and collaborative filtering models to hybrid deep learning methods that integrate user behavior with textual features.

In this project, our team will divide responsibilities across data preprocessing, feature engineering, model development, and evaluation. Using Python along with libraries such as Pandas, Scikit-learn, and TensorFlow, we aim to construct a practical and interpretable system. By the end of the project, we will analyze the performance of different models and summarize key insights into effective strategies for personalized news recommendation.

BACKGROUND AND RELATED WORK

This project focuses on recommender systems, which aim to predict what users are most likely to engage with next. Traditional approaches include content-based filtering, which recommends items similar to a user's past preferences, and collaborative filtering, which learns from patterns shared among users. Earlier techniques like matrix factorization, such as Singular Value Decomposition (SVD), were effective but struggled with new users or items and couldn't fully capture changes in user interests over time.

Recent progress in deep learning has made recommender systems more adaptive. Models using Recurrent Neural Networks (RNNs) and Transformers can now learn from the sequence of user interactions to predict future interests. Hybrid approaches that combine textual features with collaborative signals have also improved personalization and accuracy. In this project, we'll explore this evolution from classical filtering methods to modern deep learning models using the MIND dataset, a large-scale news recommendation benchmark introduced by Microsoft Research (Wu, Wu, Qi, Huang, & Xie, 2020), to predict user click behavior in news recommendations.

DATA SOURCE AND PREPROCESSING PLAN

A. DATASET

For this project, we will use the MICROSOFT NEWS DATASET (MIND).

Link**:** https://www.kaggle.com/datasets/arashnic/mind-news-dataset

Each subset contains three major files:

- **news.tsv** - includes metadata for each news article (news ID, category, subcategory, title, abstract, URL, entities…)
- **behaviors.tsv** - contains user interaction logs (user ID, timestamps, clicked and non-clicked news…)

news.tsv

Variable	Description
news_id	Unique identifier for each news article; used to match with user interactions in *behaviors.tsv*.
category	Main category of the article (e.g., *Sports*, *Politics*, *Entertainment*).
subcategory	Specific subcategory under the main category (e.g., *Soccer*, *Elections*).
title	Headline or title of the news article; primary textual feature for content representation.
abstract	Short summary of the article; provides additional semantic context for embedding.
url	Original source URL of the article (not used for modeling).
title_entities	Named entities (people, organizations, places) automatically extracted from the title.
abstract_entities	Named entities automatically extracted from the abstract text.

behaviors.tsv

Variable	Description
impression_id	Unique identifier for each impression event (news recommendation instance).
user_id	Anonymized ID representing each user; used to group multiple sessions.
time	Timestamp of the impression event; indicates session timing for chronological modeling.
history	Space-separated list of news IDs previously clicked by the user; represents reading history.
impressions	List of candidate news articles displayed in that session with click labels (e.g., “N12345-1 N12346-0”); where 1 = clicked, 0 = skipped.

This rich dataset is ideal for building and comparing a wide range of recommendation algorithms.

B. DATA PREPROCESSING AND CLEANING

To make the raw data suitable for modeling, we will perform several preprocessing steps:

- **Parsing User Behaviors:**
  Extract user–item interactions from *behaviors.tsv*, expanding impression lists into labeled pairs (clicked = 1, not clicked = 0).
- **Text Cleaning:**
  Preprocess titles and abstracts in *news.tsv* by converting to lowercase, removing punctuation and stop words, and applying lemmatization to normalize text.
- **Handling Missing and Invalid Data:**
  Detect and handle missing values or corrupted entries across files. Remove duplicate or empty records.
- **Merging Datasets:**
  Integrate user interactions with article information using *news_id* as a key to form a unified dataset.
- **Time-Based Splitting:**
  Split the dataset chronologically - earlier interactions for training and later for testing - to simulate realistic user behavior prediction.

PLANNED MODELS AND EVALUATION STRATEGY

We will transform textual and behavioral data into numerical representations suitable for both traditional and deep learning models:

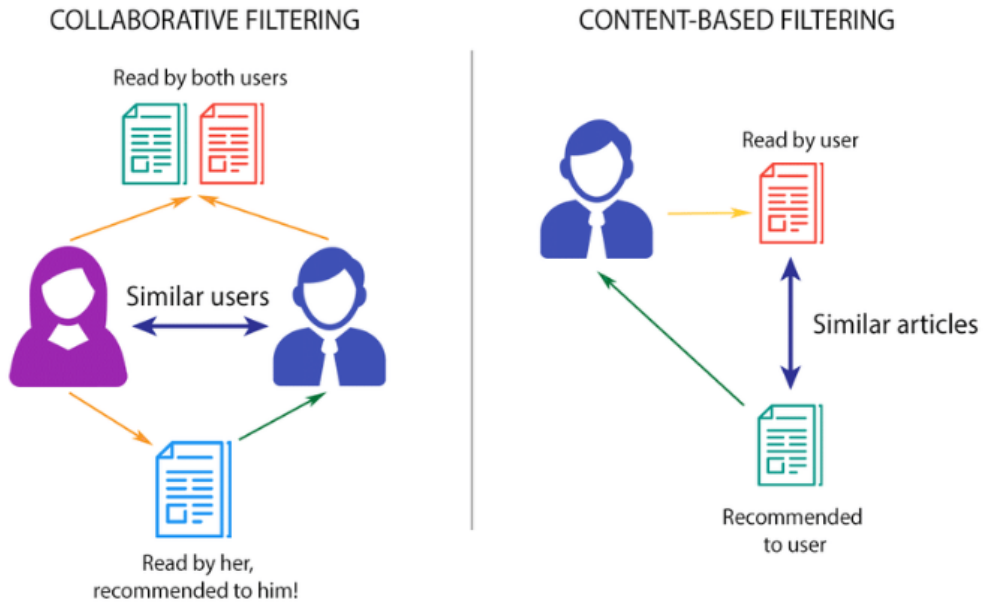
- **TF-IDF Vectors:** Represent article text features for content-based filtering.
- **User–Item Interaction Matrix:** Construct a sparse matrix of user clicks for collaborative filtering (SVD).
- **Embeddings:** Utilize pretrained entity embeddings and optionally train Word2Vec embeddings to capture semantic similarity between articles.
- **Categorical Encoding:** Encode category and subcategory as numerical or embedding features.

A. MODELING APPROACH

We will compare three types of models:

1. **Baseline Model:**
   Popularity-Based Recommender ranks the most-clicked articles overall as a simple benchmark.
2. **Classical Machine Learning Models:**

Content-Based Filtering using TF-IDF and cosine similarity.
Collaborative Filtering (SVD) via matrix factorization.



3. **Deep Learning Model:**
Hybrid Neural Recommender combining content embeddings (text/entities) and user embeddings to predict click probabilities.

B. EVALUATION STRATEGY

To evaluate recommendation quality, we will apply ranking-based metrics commonly used in

recommender systems:

1. **Precision@K / Recall@K:** Accuracy of top K predictions.
2. **MAP (Mean Average Precision):** Measures overall ranking quality.
3. **NDCG (Normalized Discounted Cumulative Gain):** Rewards higher-ranked relevant items.
4. **AUC:** For binary click prediction assessment.

All models will be evaluated using a time-based split to ensure that predictions are made on unseen future interactions, mirroring real-world deployment.

REFERENCES

1. Kaggle Dataset: https://www.kaggle.com/datasets/arashnic/mind-news-dataset
2. MIND: A large-scale dataset for news recommendation, Fangzhao Wu, Ying Qiao, Jiun-Hung Chen, Chuhan Wu, Tao Qi, JianXum Lian, Danyang Liu, Xing Xie, Jianfeng Gao, Winnie Wu, Ming Zhou: https://aclanthology.org/2020.acl-main.331.pdf
3. User Recommendation System based on MIND Dataset, Niran A. Abdulhussein, Ahmed J. Obaid: https://arxiv.org/pdf/2209.06131
4. IMAGES are taken from
    a. https://medium.com/swlh/news-recommendation-system-a8efde3cb233
    b. https://rishika-ravindran.medium.com/what-are-recommendation-systems-and-how-are-companies-using-them-a5b08ff4df42