

Aanisha Newaz

November 16, 2025

CP8339 Recommender System

Dr. Cherie (Chen) Ding

Assignment 2

1. Introduction

In this assignment, the experiment from the article *LLM-Generated Fake News Induces Truth Decay in News Ecosystem: A Case Study on Neural News Recommendation*, introduced by Wu et al. (2019), was replicated. The authors experimented on LSTUR and NRMS and how LLM-generated fake news affect the models. A new dataset was used that was a reconstructed version of the GossipCop dataset, to which user-news interactions were added. Additionally, GPT 4o-mini and Llama 3.1 were used to make LLM-generated real and fake news. Moreover, the experiment is conducted in 4 phases:

1. Phase 0: No LLM news
2. Phase 1: LLM news only in candidate lists
3. Phase 2: LLM news also appears in user history
4. Phase 3: LLM news also enters model training data.

The purpose of this study is to find how LLM-generated fake news could impact a neural news ecosystem once they are injected into the candidates list, user history, and the training data, and how news rankings are affected.

Due to limitations of not having the dataset used by the authors, this assignment will only do Phase 0 using the original GossipCop dataset. Since Phase 0 includes only human written articles, we will attempt to match the baseline statistics from figure 1 of the original paper.

Rec.	Type	MRR	nDCG@5	nDCG@10	Ratio@5	Ratio@10
LSTUR	HF	17.15	45.63	43.91	43.37	42.11
	HR	18.83	54.37	56.09	56.62	57.89
	(RRA)	9.80%	19.15%	27.74%	30.55%	37.47%
NRMS	HF	17.36	46.28	44.84	43.98	43.10
	HR	18.62	53.72	55.17	56.02	56.90
	(RRA)	7.26%	16.08%	23.04%	27.38%	32.02%

Figure 1: Performance comparison of human-written fake and real news. RRA represents the relative advantage of real news over fake news. A positive RRA indicates that real news ranks higher, while a negative one suggests the opposite. HF/HR: Human-Written Fake/Real. Rec.: Recommender.

2. Model Description

For this assignment, the NRMS (Neural News Recommendation with Multi-Head Self-Attention) model proposed by Wu et al. (2019) was used. A public implementation of NRMS, originally designed for the MIND news recommendation dataset, was cloned from GitHub and adapted. Because the original repository was optimized for the MIND dataset, once

Aanisha Newaz

November 16, 2025

CP8339 Recommender System

Dr. Cherie (Chen) Ding

the model was made to work using the tutorials provided, additional modifications were made to accommodate for the GossipCop dataset as well, including preprocessing and the evaluation pipeline.

Synthetic behavior logs were generated to mimic the MIND behavior structure so the NRMS framework could be reused without rewriting the entire data pipeline.

3. Dataset

Attempts were made to contact the authors to request for the exact study dataset, but no responses were received. Hence, as mentioned in the introduction, the original GossipCop dataset was used instead.

The GossipCop dataset was converted into a MIND-style dataset in which user-news interactions were synthetically created and formatted. This is because the behaviour log is more important to the experiment and the model, rather than the original classification style.

During training, the MIND dataset was then sampled to a size of 500. This is due to the full dataset exceeding runtime constraints.

4. Experimental Setup

- Base model: NRMS (Wu et al., 2019)
- Training dataset: sampled MIND behaviors (500 impressions)
- Evaluation dataset: converted GossipCop (2,214 impressions)
- Training epochs: 3
- Metrics computed separately for:
 - HF (fake news)
 - HR (real news)
- Metrics reported:
 - MRR
 - nDCG@5
 - nDCG@10
 - Ratio@5
 - Ratio@10
 - RRA (Relative Ranking Advantage)

This setup reproduces Phase 0 of the original study: only human-written fake and real news, no LLM content had been added.

5. Results

Based on our results in figure 2, real news (HR) and fake news (HF) performed similarly on MRR and nDCG@k, with fake news having a slight advantage. This is different from our baseline from the original paper in figure 4, and it is likely due to the limited amount of training data used. However, real news has done much better than fake news for the Ratio@k metrics, meaning that real news ranks much higher in the top-k. When our model was trained using 2000 behaviours in figure 3, the performance improved much more, with the RRA being positive

across all metrics. Hence, it is appropriate to say that our baseline model prefers real news to be recommended to its users.

=== NRMS (GossipCop Val, HF vs HR) ===					
Rec. Type	MRR	nDCG@5	nDCG@10	Ratio@5	Ratio@10
HF	40.79	48.86	55.03	63.69	63.53
HR	40.17	48.75	54.57	80.39	80.02
(RRA)	-1.51%	-0.24%	-0.83%	26.23%	25.95%

Figure 2: Assignment experiment results after training on 500 behaviours

=== NRMS (GossipCop Val, HF vs HR) ===					
Rec. Type	MRR	nDCG@5	nDCG@10	Ratio@5	Ratio@10
HF	39.19	47.63	53.79	64.18	63.53
HR	41.21	49.74	55.38	80.62	80.02
(RRA)	5.16%	4.43%	2.94%	25.60%	25.95%

Figure 3: Assignment experiment results after training on 200 behaviours

NRMS	HF	17.36	46.28	44.84	43.98	43.10
	HR	18.62	53.72	55.17	56.02	56.90
	(RRA)	7.26%	16.08%	23.04%	27.38%	32.02%

Figure 4: Original study baseline

Aanisha Newaz

November 16, 2025

CP8339 Recommender System

Dr. Cherie (Chen) Ding

6. Discussion

The reproduction of the study's Phase 0 experiment either shows better results in some statistics compared to the original study (Ratio@k), as well as worse (MRR and nDCG@k). Based on our result, the NRMS model by default prefers real news over fake news in populating the top rankings. However, fake news can still rank competitively on position-based metrics. Using synthetic user-news interactions were sufficient for this experiment, and the statistics gathered from this experiment replication can be used as a baseline for Phase 1, 2, and 3 once the original study's dataset is accessible.

7. Conclusion

This assignment partially replicates the study done by Hu et al (2025) by implementing the NRMS model and the original GossipCop dataset with synthetic additions. Although only Phase 0 was conducted due to dataset limitations, the results produced appropriate statistics that could serve as baselines for future studies on Phases 1, 2 and 3.

Aanisha Newaz

November 16, 2025

CP8339 Recommender System

Dr. Cherie (Chen) Ding

References

Article:

Beizhe Hu, Qiang Sheng, Juan Cao, Yang Li, and Danding Wang. 2025. LLM-Generated Fake News Induces Truth Decay in News Ecosystem: A Case Study on Neural News Recommendation. In Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '25). Association for Computing Machinery, New York, NY, USA, 435–445. <https://doi.org/10.1145/3726302.3730027>

NRMS Base Model:

<https://github.com/YadaYuki/news-recommendation-llm?tab=readme-ov-file>

Dataset:

<https://www.kaggle.com/datasets/akshaynarayananb/gossipcop?resource=download>