Diversifying Sentiments in News Recommendation

Mete Sertkan, Sophia Althammer, Sebastian Hofstätter and Julia Neidhardt

Christian Doppler Laboratory for Recommender Systems, TU Wien, Vienna, Austria

Abstract

Personalized news recommender systems are widely deployed to filter the information overload caused by the sheer amount of news produced daily. Recommended news articles usually have a sentiment similar to the sentiment orientation of the previously consumed news, creating a self-reinforcing cycle of sentiment chambers around people. Wu et al. introduced SentiRec – a sentiment diversity-aware neural news recommendation model to counter this lack of diversity.

In this work, we reproduce SentiRec without access to the original source code and data sample. We re-implement SentiRec from scratch and use the Microsoft MIND dataset (same source but different subset as in the original work) for our experiments. We evaluate and discuss our reproduction from different perspectives. While the original paper mainly has a user-centric perspective on sentiment diversity by comparing the recommendation list to the user's interaction history, we also analyze the intra-list sentiment diversity of the recommendation list. Additionally, we study the effect of sentiment diversification on topical diversity. Our results suggest that SentiRec does not generalize well to other data since the compared baselines already perform well, opposing the original work's findings. While the original SentiRec utilizes a rule-based sentiment analyzer, we also study a pre-trained neural sentiment analyzer. However, we observe no improvements in effectiveness nor in sentiment diversity. To foster reproducibility, we make our source code publicly available.

1. Introduction

Content-based recommenders usually recommend items to users similar to items they have liked in the past [1]. Also, recent well-performing neural news recommendation methods follow this principle. They model users based on their previously browsed news articles and, in turn, rank candidate news articles based on a relevance score considering the user model [2]. However, such approaches are prone to a lack of diversity. Especially since news with negative sentiment is more often clicked than positive ones, diversifying the sentiment is essential in news recommendations [3].

Taking all this into account, Wu et al. [3] introduced SentiRec, a sentiment diversity-aware neural news recommendation method. They learn sentiment-aware news representations by considering the content of the news and jointly training the recommendation model together with an auxiliary sentiment prediction task. Users are modeled by their previous clicked and non-clicked (i.e., seen but not clicked) news articles. The SentiRec approach regularizes and thus increases sentiment diversity by penalizing candidate news with similar sentiment compared to the users' overall sentiment orientation. In both sentiment regularization and sentiment

prediction tasks, VADER [4], a rule-based sentiment-analyzer, is utilized to determine the sentiment polarity score as the label.

In this work, we reproduce SentiRec without having access to the original source code or dataset. Our request for access to the original source code and data set has not been answered yet. Thus, we re-implement SentiRec from scratch and use the Microsoft MIND [2] dataset (same data source but different subset as in the original work) for our experiments. We evaluate our reproduction from different perspectives, namely i) effectiveness, ii) user-centric sentiment diversity, iii) intra-list sentiment diversity, and iiii) topical diversity. In our first evaluation perspective we aim to compare effectiveness trends from the original paper with our implementation and study:

RQ1 *How does our reproduced SentiRec implementation compare to the MIND [2] baselines concerning effectiveness?*

In contrast to the original work, our reproduction does not significantly outperform the baselines, which might be due to the dataset differences, highlighting the shortcomings of SentiRec regarding generalizability. We also employed a pre-trained neural sentiment analyzer (BERT-SA[1]) in addition to the rule-based one (VADER-SA [4]). When using BERT-SA, we observe no gains in recommendation performance and sentiment diversity compared to the VADER-SA setting. Our next evaluation perspective is user-centric sentiment diversity, as defined in the original paper; thus, we investigate:

RQ2 *How does our reproduced SentiRec implementation compare to the MIND [2] baselines concerning user-centric sentiment diversity?*

Opposing the original paper, we could not achieve the best user-centric sentiment diversity results by outperforming the random model while maintaining the best effectiveness. Moreover, we demonstrate that some baselines already reach sufficient user-centric sentiment diversity and significantly outperform SentiRec, (again) highlighting the lack of generalizability. While the original paper focuses on user-centric sentiment diversity by comparing the recommended list of news to the user's interaction history, our third perspective focuses on sentiment diversity between news articles within a recommendation list, i.e., intra-list sentiment diversity. Thus we investigate:

RQ3 *How does our reproduced SentiRec implementation compare to the MIND [2] baselines concerning intra-list sentiment diversity?*

In contrast to the user-centric evaluation and although been penalized for user-centric sentiment similarity, our reproduction significantly outperforms most baselines if intra-list sentiment diversity is considered. This calls for a discussion on whether to employ a user-centric or an intra-list diversification and further investigations. While the original paper only considers sentiment diversity, we also analyze topical diversity, and thus in our final evaluation perspective, we study:

RQ4 *How does our reproduced SentiRec implementation compare to the MIND [2] baselines concerning user-centric and intra-list topical diversity?*

The user-centric topical diversity compares the user's interaction history to the recommendation list. We demonstrate that the baselines already reach significantly better user-centric topical diversity than our Sentirec reproduction - highlighting the tradeoff between different objectives.

---

[1]https://huggingface.co/distilbert-base-uncased-finetuned-sst-2-english

In intra-list topical diversity, our reproduction reaches comparable results to the baselines (taking aside the random model).

The contributions of this work are as follow:

- We reproduce SentiRec [3] without having access to the original source code and dataset. Instead, we re-implement SentiRec from scratch and use the MIND [2] dataset. Although our implementation shows similar trends, we fail to reproduce the original findings, which might be caused due to dataset differences. In particular, the baselines in our experiments already show decent recommendation and sentiment diversity performance.
- We propose extending the experiment by using a pre-trained neural sentiment analyzer instead of a rule-based sentiment analyzer. However, we observe no gains in effectiveness nor sentiment diversity.
- We propose extending the experiment by considering user-centric topical diversity and intra-list topical and sentiment diversity. While the baselines outperform our reproduction if user-centric and intra-list topical diversity is considered, it significantly outperforms the baselines in intra-list sentiment diversity.
- We publish the first open implementation of SentiRec for the community at: https://github.com/MeteSertkan/newsrec

2. Background

The way how items are presented often influences the decision behavior of users [5]. Thus, when interacting with news articles, also their textual style plays an essential role [3, 6, 7] besides semantic or syntactic properties. However, these features are hard to engineer by hand. Recently, deep learning architectures have been increasingly used in recommendation scenarios [8]. These architectures have proven highly beneficial when capturing various patterns (e.g., user sessions, structure in pictures or language) or dealing with high complexity (e.g., multi-modal data, very dynamic settings, etc.). They usually follow an *end-to-end* feature extraction paradigm, where the recommendation model and the representation model (i.e., item and user encoder) are trained simultaneously. Thus handcrafted heuristics are avoided [9]. The trend has also reached the new recommendation domain. For example, NAML [10] uses attention networks to incorporate different views of a news article, e.g., title, abstract, category, etc., into the news; LSTUR [11] captures the short-term interest of users by applying GRU on recently clicked items and long-term interest by considering a user's whole history track; and NRMS [12] uses multi-head self-attention in combination with additive-attention to model news articles, and in turn, users. However, by only considering the content of the users' previous interactions, they are prone to recommend in a *"more of the same"* way, and consequently, they might lack diversity. Therefore, we study news diversification and, in particular, sentiment diversification. In this work, we re-implement, extend, and analyze SentiRec [3]. SentiRec learns sentiment-aware news representations using an auxiliary sentiment prediction task and introduces a sentiment regularization method to obtain sentiment-diverse recommendations. While sentiment-aware recommendations have been studied in the tourism domain [13, 14], movie domain [15, 16], and e-commerce [17, 18], to name a few, less attention has been paid to sentiment-aware recommendations in the news domain and nor to sentiment diversification.
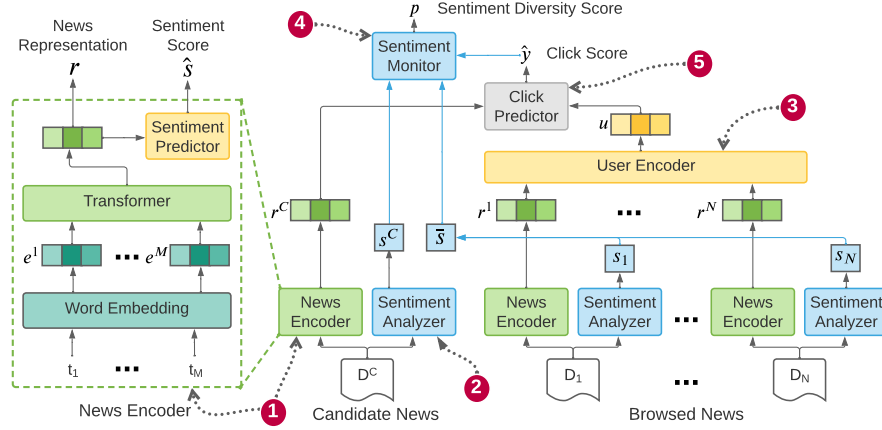
Figure 1: Overview of SentiRec [3] comprising following major components: **1** *News Encoder*, which learns to encode news by their content and simultaneously to predict a sentiment score based on the learned encoding; **2** *Sentiment Analyzer*, which assigns a sentiment score to each news article based on its content; **3** *User Encoder*, which models users based on their previous news interactions; **4** *Click Predictor*, which determines a score for a given user and candidate news pair; and **5** *Sentiment Monitor*, which monitors and regularizes the sentiment diversity.

3. Methods

3.1. SentiRec

SentiRec aims to optimize recommendation accuracy and sentiment diversity, which naturally leads to a trade-off between accuracy and diversity. The overall task is to rank candidate items based on a user's history of previous items. Given for a user u a history set H of n previously browsed news articles $[D_1, ..., D_n]$ with sentiment polarity scores $[s_1, ..., s_n]$, the aim is to rank a set C of p candidate news articles $[D_1^c, ..., D_p^c]$ (with sentiment polarity scores $[s_1^c, ..., s_p^c]$) by assigning each article a score i.e., $[\hat{y}_1, ..., \hat{y}_p]$. In particular, SentiRec seeks for sentiment diversity in the recommendation list. Higher diversity is achieved if top-ranked news articles have different sentiment polarity scores than the overall sentiment orientation $\bar{s} = mean([s_1, ..., s_N])$ of the user's previously browsed news. In the following we describe the different SentiRec components as shown in Figure 1.

1 *News Encoder.* The task of the news encoder is to find a representation r^c of candidate news D^c as well as representations $[r_1, ..., r_N]$ of browsed news $[D_1, ..., D_N]$ by taking their title as input. It consists of an embedding layer followed by a transformer layer to obtain a representation r out of a sequence of terms. Since no details about the transformer layer were given, we follow the architecture of the closely related NRMS [12] model. Thus, we use multi-head self-attention for contextualization and additive-attention to obtain a unified embedding out of the contextualized word embeddings. The news encoder is jointly trained with an auxiliary sentiment prediction task in order to infuse sentiment awareness to the news representation. The sentiment score $\hat{s}$ is predicted using a linear layer, i.e, $\hat{s} = V_s \times r + v_s$, where V_s and v_s are learnable parameters and r is the news representation. As loss function the mean absolute error between predicted

sentiment scores $\hat{s}_i$ and the sentiment scores determined by the sentiment analyzer s_i is used as follows :

$$\mathcal{L}_{senti} = \frac{1}{S} \sum_{i=1}^{S} |\hat{s}_i - s_i| \tag{1}$$

❷ *Sentiment-Analyzer.* Given the title of a news article, the sentiment analyzer determines the sentiment polarity score ranging in [-1, 1], which is considered as the sentiment label of the respective news article. The original paper uses VADER [4] (a rule-based method) as sentiment analyzer (VADER-SA). In addition, we also study a pre-trained neural sentiment analyzer[2] (BERT-SA).

❸ *User Encoder.* The user encoder gets the sentiment-aware representations of the previously browsed news, i.e., $[r_1, ..., r_N]$, as input and uses a transformer layer (i.e., multi-head self-attention followed by additive attention according to NRMS [12]) to obtain a representation u of the user.

❹ *Click Predictor.* The click predictor uses the dot-product between user and candidate embedding, i.e, ur^c, to determine a click score $\hat{y}$.

❺ *Sentiment Monitor.* The sentiment monitor observes to what extent the sentiment polarity score (obtained by the sentiment analyzer) s^c of a candidate news article diverges from the users' overall sentiment orientation $\bar{s} = mean([s_1, ..., s_N])$ (i.e., the mean sentiment polarity score of the users browsing history). This diversity in sentiment is measured by $p = max(0, \bar{s}s^c\hat{y})$, where larger values of p indicate less sentiment diversity. The sentiment diversity score p is further used to regularize and steer the model into a more sentiment diverse direction. Following loss function is used for this purpose:

$$\mathcal{L}_{div} = \frac{1}{|S|} \sum_{i \in S} p_i \tag{2}$$

where S is the training set and p_i the sentiment diversity score of the i-th sample.

Negative sampling is used in order to create a labeled dataset for the recommendation task. For each clicked news in a user impression, K non-clicked samples from the same impression are randomly selected. The recommendation loss is the negative log-likelihood of the clicked samples and is defined as follows:

$$\mathcal{L}_{rec} = - \sum_{i \in S} log(\frac{exp(\hat{y}_i^+)}{exp(\hat{y}_i^+) + \sum_{j=1}^{K} exp(\hat{y}_{i,j}^-)}) \tag{3}$$

where $\hat{y}_i^+$ is the click score of i-th clicked news and $\hat{y}_{i,j}^-$ the click score of the j-th sample of the corresponding K negative samples, and S is the training set. The final loss function brings all three losses, i.e., recommendation loss, sentiment prediction loss, and sentiment diversity loss, together as follows:

$$\mathcal{L} = \mathcal{L}_{rec} + \lambda\mathcal{L}_{senti} + \mu\mathcal{L}_{div} \tag{4}$$

where λ and μ are hyperparameters controlling the influence of the sentimen prediction loss and sentiment diversity loss respectively.

---

[2]https://huggingface.co/distilbert-base-uncased-finetuned-sst-2-english

3.2. Evaluation Perspectives

We evaluate our reproduction from five different perspectives: effectiveness, user-centric sentiment diversity, intra-list sentiment diversity, user-centric topical diversity, and intra-list topical diversity. Note, in contrast to the intra-list diversity measures, the user-centric measures assess diversity in relation to the user's previous news consumption. We compare the results of our reproduction against all baselines and our extensions, using paired t-test with Bonferroni correction [19, 20].

Effectiveness. We evaluate effectiveness using AUC, MRR, $nDCG@5$, and $nDCG@10$.

User-Centric Sentiment Diversity. We evaluate user-centric sentiment diversity using the sentiment alignment metrics S_{MRR} and $S@K$, introduced by WU et al. [3], which is defined as follows:

$$S_{MRR} = max(0, \bar{s} \sum_{i=1}^{C} \frac{s_i^c}{i}), \quad S@K = max(0, \bar{s} \sum_{i=1}^{K} s_i^c) \tag{5}$$

where C is the length of the recommendation list (i.e, number of candidate items) and s_i^c is the sentiment polarity score of the news article ranked at position i in this list; and $\bar{s}$ is the overall sentiment orientation of the corresponding user. Hence, the closer top-ranked candidates' sentiment to the users' overall sentiment orientation, the higher the sentiment alignment metrics. Ergo, lower sentiment alignment scores indicate more sentiment-diverse recommendations.

Intra-List Sentiment Diversity *(not included in the original paper).* As the sentiment polarity score s_i of a news article is only one scalar, we compute the intra-list sentiment diversity by averaging the absolute difference of sentiment polarity scores s_i and s_j between each news pair in the Top-K list of recommended candidate articles:

$$ILS_S@K = \frac{2}{K(K-1)} \sum_{s_i, s_j \in C@K} |s_i - s_j| \tag{6}$$

The intra-list sentiment diversity score lies between 0 and 1, with 0 being maximal divers.

User-Centric Topical Diversity *(not included in the original paper).* We consider the news articles' categories (e.g., sports) and subcategories (e.g., soccer) to compute topical diversity. We represent a (sub)category of a news article with a 1-hot-encoding. We compute the user's category representation c_u by summing up all browsed news category representations. Similarly, we compute the recommendations list's category representation $c_{C@K}$ by summing up the category representations of the recommended top-K candidate news articles. We then measure diversity $T@K$ by taking cosine similarity between c_u and $c_{C@K}$. This leads to a measure between 0 and 1, with 0 being maximal divers. Similarly, we measure T_{MRR} with the difference being computing a weighted average of all candidates' category representations to obtain a representation c_{MRR} of the recommendation list, where the weight is the rank of corresponding news articles.

$$T_{MRR} = cos_{sim}(c_{MRR}, c_u), \quad T@K = cos_{sim}(c_{C@K}, c_u) \tag{7}$$

Intra-List Topical Diversity *(not included in the original paper).* We again represent a (sub)category of a news article with a 1-hot-encoding. We measure the intra-list topical diversity of the recommendation list by computing the average pairwise cosine similarity between the 1-hot-encoded category representations *c* of the recommended top-k news articles. This leads to a measure between 0 and 1, with 0 being maximal divers.

$$ILS_T@K = \frac{2}{K(K-1)} \sum_{c_i,c_j \in C@K} cos_{sim}(c_i, c_j) \tag{8}$$

4. Experimental Setting

Dataset. The dataset of the original paper is constructed from MSN News[3] logs collected from 0ctober 31, 2018, to January 29, 2019, but has not been open-sourced, and our access request has not been answered yet. Thus, we use the MIND [2] dataset - specifically the MIND-small[4] version - in our experiments, as it stems from the same source. It was randomly sampled from 50K users (with at least five clicks) during six weeks, from October 12 to November 22, 2019, where the first five weeks are for training and the last week for testing. One sample is composed of a timestamp, the user-id, a list of chronologically ordered news-ids representing the user's click history, and a list of shuffled candidate news-ids with corresponding labels (i.e., 1 for *clicked* and 0 for *seen but not clicked*). Detailed statistics of the datasets are summarized in Table 1. Mind-small has five times more users with about two times fewer impressions and on average about seven times fewer positive interactions per user (seven clicks vs. 49) than the SentiRec dataset.

Table 1
SentiRec dataset (as reported) and MIND-small dataset statistics.

Dataset	#Users	#News	#Impression	#Clicks	#Non-Clicks
SentiRec	10,000	42,255	445,230	489,644	6,651,940
MIND-small	50,000	65,238	230,117	347,727	8,236,715

Training. All models are trained on 90% of the training data. The remaining 10% is used to tune the hyperparameters by optimizing AUC. We use early-stopping with a minimum delta of 0.0001 AUC and patience of 5. Note that we use 300-dimensional Glove embeddings [21] in all models to initialize the word embedding layer and NLTK [22] word tokenizer for tokenization. Further, we limit the number of browsed news in each impression to 50 and the title length to 20 terms (smaller sequences are zero-padded).

Parameter Settings. We set the negative sampling ratio K to 4. We apply 20% dropout to the word embeddings. We use multi-head self-attention with 15 attention heads followed by an additive-attention layer with a 200-dimensional query vector. We use the ADAM [23] optimizer with a learning rate of 0.0001 and a batch size of 128. For the VADER-SA based model (*SentiRec_V*)

---

[3]https://www.msn.com/en-us/news
[4]https://msnews.github.io/index.html

we set $\lambda = 0.4$ and $\mu = 10$ and for the BERT-SA based model (*SentiRec$_B$*) we set $\lambda = 0.4$ and $\mu = 1$.

Baselines. We compare the reproduced and adapted models against following baselines suggested by the dataset providers [2]:

LSTUR [11] (not included in the original paper) - Neural news recommender capturing users' long- and short-term interests. We initialize the GRU network with the user embedding. We set the masking probability of the users' long-term interests to 50%. We apply 20% dropout to the word embeddings. The negative sampling ratio K is set to 4. For the CNN, we set the number of filters to 300 and the window size to 3. We use a 200-dimensional query vector for the additive-attention layer. We use the ADAM [23] optimizer with a learning rate of 0.0001 and a batch size of 256.

NAML [10] (not included in the original paper) & $NAML_T$ (adaptation of NAML as in the original paper) - Neural news recommender incorporating multiple views (i.e., title, category, and abstract) into the news representation. We limit the abstract length to 50 terms. We apply 20% dropout to the word embeddings. We set the category embeddings dimension to 100. The number of CNN filters is set to 400 and the window size to 3. We use 200-dimensional query vectors in the additive-attention layers. The negative sampling ratio K is set to 4. We use the ADAM [23] optimizer with a learning rate of 0.0001 and a batch size of 256. We also trained $NAML_T$ - a *"title only"* version - as used in the original paper [3]. We obtained the same parameters as NAML without the need for category dimensions.

NRMS [12] - Neural news recommender which utilizes multi-head self-attention within both the news encoder and the user encoder. We use multi-head self-attention with 15 attention heads followed by an additive-attention layer with a 200-dimensional query vector. We apply 20% dropout to the word embeddings. We set the negative sampling ratio K to 4. We use the ADAM [23] optimizer with a learning rate of 0.0001 and a batch size of 128.

5. Results and Analysis

In this section, we present and analyze our results and answer our previously stated research questions. We investigate whether the reproduced models perform as described in the original paper and study:

RQ1 *How does our reproduced SentiRec implementation compare to the MIND [2] baselines concerning effectiveness?*
We compare the recommendation performance (i.e., *AUC, MRR, nDCG@5,* and *nDCG@10*) of the reproduced model (i.e., *SentiRec$_V$*) against the baselines (i.e., *LSTUR* [11], *NAML & $NAML_T$* [10], *NRMS* [12], and *Random*), which is summarized in rows 1-6 of Table 2. Opposing the original work, our sentiment reproduction does not significantly outperform all baselines concerning recommendation effectiveness. Moreover, it performs similarly to the closely related *NRMS* baseline. Furthermore, utilizing a pre-trained neural sentiment analyzer instead of the rule-based one does not yield performance gains (compare rows 6 to 7 in Table 2).

RQ2 *How does our reproduced SentiRec implementation compare to the MIND [2] baselines concerning user-centric sentiment diversity?*
We investigate sentiment diversity by comparing the sentiment alignment scores (i.e., S_{MRR},

Table 2

Comparing effectiveness (i.e., AUC, MRR, nDCG@5, and nDCG@10). Higher effectiveness scores indicate better performance. Subscripts V (VADER-SA) and B (BERT-SA) indicate the used sentiment analyzer. Note, † indicates a statistically significant difference to $SentiRec_V$ at alpha 0.05.

	Model	AUC	MRR	nDCG @5	nDCG @10
1	Random	$.4994^\dagger$	$.2190^\dagger$	$.2236^\dagger$	$.2863^\dagger$
2	$NAML_T$	.6194	**.2982**	.3190	.3804
3	$NAML$	.6206	$.2913^\dagger$	.3185	$.3782^\dagger$
4	$LSTUR$	$.6210^\dagger$	$.2840^\dagger$	$.3101^\dagger$	$.3721^\dagger$
5	$NRMS$	**.6228**	.2946	.3191	.3817
6	$SentiRec_V$	<u>.6224</u>	<u>.2952</u>	**.3211**	<u>.3818</u>
7	$SentiRec_B$	.6219	.2942	<u>.3203</u>	**.3820**

Table 3

Comparing user-centric sentiment and topic alignment (i.e., S_{MRR}, S@5, S@10, T_{MRR}, T@5, T@10). Lower alignment scores indicate better diversity. Subscripts V (VADER-SA) and B (BERT-SA) indicate the used sentiment analyzer. Note, † indicates a statistically significant difference to $SentiRec_V$ at alpha 0.05.

	Model	VADER-SA Labels S_{MRR}	S@5	S@10	BERT-SA Labels S_{MRR}	S@5	S@10	T_{MRR}	T@5	T@10
1	$Random$	$\mathbf{.0086}^\dagger$	$\mathbf{.0150}^\dagger$	$\mathbf{.0188}^\dagger$	$\mathbf{.1095}^\dagger$	$\mathbf{.1748}^\dagger$	$\mathbf{.2638}^\dagger$	$\mathbf{.4315}^\dagger$	$\mathbf{.3680}^\dagger$	$\underline{.4428}^\dagger$
2	$NAML_T$	$.0157^\dagger$	$.0276^\dagger$	$.0382$	$.1741^\dagger$	$.2623^\dagger$	$.3933^\dagger$	$.5091^\dagger$	$.4570^\dagger$	$.5047^\dagger$
3	$NAML$	$\underline{.0131}^\dagger$	$\underline{.0210}^\dagger$	$\underline{.0248}^\dagger$	$\underline{.1132}^\dagger$	$\underline{.1749}^\dagger$	$\underline{.2936}^\dagger$	$\underline{.4504}^\dagger$	$\underline{.3744}^\dagger$	$\mathbf{.4270}^\dagger$
4	$LSTUR$	$.0158^\dagger$	$.0281^\dagger$	$.0412^\dagger$	$.1655^\dagger$	$.2637^\dagger$	$.4297^\dagger$	$.4735^\dagger$	$.4220^\dagger$	$.4867^\dagger$
5	$NRMS$	$.0149^\dagger$	$.0282$	$.0390$	$.1317^\dagger$	$.2317^\dagger$	$.3869^\dagger$	$.4883$	$.4353$	$.4926^\dagger$
6	$SentiRec_V$	$.0161$	$.0284$	$.0386$	$.1300$	$.2153$	$.3651$	$.4872$	$.4328$	$.4891$
7	$SentiRec_B$	$.0174^\dagger$	$.0325^\dagger$	$.0449^\dagger$	$.1560^\dagger$	$.2675^\dagger$	$.4330^\dagger$	$.4905^\dagger$	$.4414^\dagger$	$.4942^\dagger$

$S@5$, and $S@10$ − lower scores indicate higher sentiment diversity) of our reproduced model, i.e., $SentiRec_V$, and the baselines (see rows 1-6 in Table 3). In the original work [3], SentiRec outperforms all baselines in sentiment diversity - even the Random model - while maintaining the highest recommendation performance scores. We can not confirm these findings. Moreover, our results suggest that the baselines already perform well in all aspects, i.e., recommendation performance and sentiment diversity. In particular, we do not observe large margins in sentiment diversity as in the original paper While the original paper studies sentiment diversity with a user-centric focus, it is also essential to investigate sentiment diversity within a recommended list of news articles; thus, we ask:

RQ3 *How does our reproduced SentiRec implementation compare to the MIND [2] baselines concerning intra-list sentiment diversity?*

We compute the intra-list sentiment similarity at cutoff K, i.e., $ILS_S@K$, by considering the pairwise differences of news articles within a top K recommendation list. Table 4 (rows 1-7)

Table 4

Comparing sentiment- and topic-based intra-list similarity (i.e., $ILS_S@5$, $ILS_S@10$, $ILS_T@5$, $ILS_T@10$). Lower intra-list similarity scores indicate better diversity. Subscripts V (VADER-SA) and B (BERT-SA) indicate the used sentiment analyzer. Note, † indicates a statistically significant difference to $SentiRec_V$ at alpha 0.05.

	Model	VADER-SA Labels		BERT-SA Labels		$ILS_T@5$	$ILS_T@10$
		$ILS_S@5$	$ILS_S@10$	$ILS_S@5$	$ILS_S@10$		
1	*Random*	$.2393^\dagger$	$.2394^\dagger$	$.5047^\dagger$	$.5045^\dagger$	$\mathbf{.0774}^\dagger$	$\mathbf{.0775}^\dagger$
2	$NAML_T$	$.2336^\dagger$	$.2377^\dagger$	$.4770^\dagger$	$.4863^\dagger$	$.1396^\dagger$	$.1089^\dagger$
3	$NAML$	$.2600^\dagger$	$.2480^\dagger$	$.5221^\dagger$	$.5049^\dagger$	$.3377^\dagger$	$.1886^\dagger$
4	$LSTUR$	$\underline{.2313}$	$\underline{.2347}$	$.4826^\dagger$	$.4826$	$\underline{.1223}^\dagger$	$.1026$
5	$NRMS$	$.2376^\dagger$	$.2393^\dagger$	$.4700$	$.4819$	$.1290$	$.1016$
6	$SentiRec_V$	$\mathbf{.2310}$	$\mathbf{.2337}$	$\underline{.4682}$	$\underline{.4812}$	$.1289$	$\underline{.1013}$
7	$SentiRec_B$	$.2423^\dagger$	$.2404^\dagger$	$\mathbf{.4444}^\dagger$	$.4648^\dagger$	$.1429^\dagger$	$.1063^\dagger$

summarizes our outcomes. A lower intra-list similarity score indicates better diversity. In contrast to our user-centric diversity findings, where the baselines already exhibit decent performance, we observe that our reproduced model, i.e., *Sentirec$_V$*, significantly outperforms most baselines concerning intra-list sentiment diversity. In comparison, the $NAML$ baseline shows the worst performance. Suggesting that additional modalities might foster user-centric sentiment diversity (see Table 3) but hurt intra-list sentiment diversity by recommending top K news articles with a rather higher sentiment similarity. Effectiveness and sentiment diversity are the emergent perspectives to evaluate SentiRec; in addition to those, we also focus on topical diversity and investigate:

RQ4 *How does our reproduced SentiRec implementation compare to the MIND [2] baselines concerning user-centric and intra-list topical diversity?*

We adapt the user-centric sentiment alignment metrics and introduce user-centric topical alignment metrics, i.e., T_{MRR} and $T@K$, by considering the categorical membership of the news articles. Lower T_{MRR} / $T@K$ indicate higher diversity. The last three columns of Table 3 summarize our analysis. The *Random* model recommends the most topically diverse news articles to the users' previously browsed news articles, except if the top 10 recommendations are considered, where the $NAML$ model excels. The $NAML$ and the $LSTUR$ baselines significantly reach better user-centric topical diversity than our reproduced *SentiRec* models while maintaining reasonable recommendation performance – demonstrating the competitiveness of the baseline models. If we consider intra-list topical diversity $ILS_T@K$ (see Table 4 last two columns), which is defined by the pairwise categorical differences within the recommendation list, the *Random* Model recommends the most diverse news articles. Our reproduction, *SentiRec$_V$*, outperforms the $NAML$ models and is on par with the $LSTUR$ and $NRMS$ baselines.

6. Discussion

Overall, we cannot confirm the findings of the original work, where they outperformed all baselines in effectiveness and user-centric sentiment diversity. We argue that the effectiveness and diversity discrepancies between the original SentiRec and our reproduction are due to dataset differences highlighting the shortcomings of *SentiRec* concerning generalizability. Our dataset contains five times more users and about 23K more news than the original paper; however, it contains relatively few positive feedback (i.e., clicks) and spans only over six weeks (compared to nine weeks). Thus, inherently more diverse behavior is contained in the used dataset than in the original paper. One might argue that the sentiment diversity issue in our sample is not as prevalent as in the sample of the original work. However, we demonstrate that the $NAML$ baseline significantly outperforms our reproduction and gets close to the *Random* model's performance. This highlights that there is room for improvement, which is not utilized by the *SentiRec*'s diversification approach.

As mentioned, the $NAML$ [10] model outperforms all other models (except the *Random* model) regarding user-centric sentiment diversity while maintaining comparable recommendation performance to our *SentiRec* reproductions. Besides the title of a news article, it also considers category, subcategory, and abstract. Thus, we reason that considering different modalities supports the diversification task. Note, in the original paper $NAML$ is fed with only one modality (i.e., title) - in this work denoted as $NAML_T$.

Besides the user-centric view of sentiment diversity, we also analyze a more generic perspective, i.e., intra-list sentiment diversity. We demonstrate that our reproduction achieves an outstanding intra-list sentiment diversity, although optimized for user-centric sentiment diversity. Setting both perspectives alongside opens the room for the following question, which we will tackle in future work: Which view of sentiment diversity should we optimize while maintaining user satisfaction? Optimizing for the user-centric perspective is more conservative. This will rank news articles with an orthogonal sentiment to the overall sentiment of the user's news consumption higher. Such an approach has a strong nudging power but might drop user satisfaction by recommending more the "unusual". On the other hand, optimizing for the intra-list perspective is more relaxed by suggesting news articles with different sentiments. However, it bears the risk that users might still follow their previous behavior and consume, for example, only negative news articles.

Our final evaluation perspective, which the original work does not consider, is topical diversity. In particular, we consider categorical differences between recommended news articles and the users' browsed news, i.e., user-centric topical diversity and categorical differences within the news articles in the recommendation list, i.e., intra-list topical diversity. In both measures, the *Random* model achieves the most topically diverse recommendations. Setting aside the *Random* model, while in the user-centric perspective, our reproduction *Sentirec$_V$* is outperformed by most baselines, in the intra-list perspective, it is on par or better than the baselines. With different sentiment distributions within news categories, we plan to analyze whether topical diversification already yields sentiment diversification and higher user satisfaction in future work.

7. Conclusion

This work aims to reproduce SentiRec [3] - a sentiment diversity-aware neural news recommendation model - without having access to the original source code and dataset. We re-implement SentiRec from scratch and make it publicly available. We use the MIND [2] dataset, which has the same source as the original paper, albeit a different time period. Overall, we can not confirm the significant findings of the SentiRec paper. The reproduced model does not outperform the random model in (user-centric) sentiment diversity while maintaining the best recommendation performance compared to the baselines as in the original work. Moreover, our results suggest that the baselines already perform well. In particular, the NAML [10] model delivers the most sentiment-diverse recommendations (w.r.t. to the users' overall consumption behavior) apart from the random model while holding a comparable recommendation performance to all other baselines. We conclude that these discrepancies are due to dataset differences high-lighting the shortcomings of SentiRec concerning generalizability.

In addition to the original paper, we also consider the topical diversity of the recommended list compared to the users' previous user history. Similar to previously, we show that the baselines, particularly *NAML*, significantly yield better topical diversity than our reproduced *Sentirec* model.

In addition to a rule-based sentiment analyzer, as used by Wu et al. [3], we conducted our experiments with a pre-trained neural sentiment analyzer to study whether a neural model leads to better sentiment labels and thus to improved overall training performance. However, we do not observe improvements in recommendation performance or sentiment diversity.

While the original paper only focuses on sentiment diversity by comparing the users' overall user history with the recommendation list (i.e., user-centric diversity), we also investigate the sentiment and the topical diversity between news articles within the recommendation list (intra-list diversity). In contrast to the user-centric evaluation, the intra-list evaluation shows that our *SentiRec* reproduction significantly outperforms most baselines, while the strong *NAML* baseline performs poorly.

We discuss our different evaluation perspectives (i.e., user-centric/intra-list sentiment and topical diversity). We plan to conduct offline and online experiments to compare and combine them in future work. Furthermore, we plan to include other auxiliary information into the end-to-end recommendation model, such as emotion awarness and diversity. Ultimately, we want to create recommendation models that optimize for a broad range of goals and benefit society by more responsible recommendations.

Acknowledgments

References

[1] F. Ricci, L. Rokach, B. Shapira, Recommender Systems: Introduction and Challenges, Springer US, Boston, MA, 2015, pp. 1–34. URL: https://doi.org/10.1007/978-1-4899-7637-6_1. doi:10.1007/978-1-4899-7637-6_1.

[2] F. Wu, Y. Qiao, J.-H. Chen, C. Wu, T. Qi, J. Lian, D. Liu, X. Xie, J. Gao, W. Wu, M. Zhou, MIND: A large-scale dataset for news recommendation, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 3597–3606. URL: https://www.aclweb.org/anthology/2020.acl-main.331. doi:10.18653/v1/2020.acl-main.331.

[3] C. Wu, F. Wu, T. Qi, Y. Huang, SentiRec: Sentiment diversity-aware neural news recommendation, in: Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing, Association for Computational Linguistics, Suzhou, China, 2020, pp. 44–53. URL: https://www.aclweb.org/anthology/2020.aacl-main.6.

[4] C. Hutto, E. Gilbert, Vader: A parsimonious rule-based model for sentiment analysis of social media text, Proceedings of the International AAAI Conference on Web and Social Media 8 (2014). URL: https://ojs.aaai.org/index.php/ICWSM/article/view/14550.

[5] D. Jannach, M. Zanker, A. Felfernig, G. Friedrich, Online consumer decision making, Cambridge University Press, 2010, p. 234–252. doi:10.1017/CBO9780511763113.012.

[6] R. El Baff, H. Wachsmuth, K. Al Khatib, B. Stein, Analyzing the Persuasive Effect of Style in News Editorial Argumentation, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 3154–3160. URL: https://www.aclweb.org/anthology/2020.acl-main.287. doi:10.18653/v1/2020.acl-main.287.

[7] M. Sertkan, J. Neidhardt, H. Werthner, Documents, topics, and authors: Text mining of online news, in: 2019 IEEE 21st Conference on Business Informatics (CBI), volume 01, 2019, pp. 405–413. doi:10.1109/CBI.2019.00053.

[8] S. Zhang, L. Yao, A. Sun, Y. Tay, Deep learning based recommender system: A survey and new perspectives, ACM Comput. Surv. 52 (2019). URL: https://doi.org/10.1145/3285029. doi:10.1145/3285029.

[9] Y. Deldjoo, M. Schedl, P. Cremonesi, G. Pasi, Recommender systems leveraging multimedia content, ACM Comput. Surv. 53 (2020). URL: https://doi.org/10.1145/3407190. doi:10.1145/3407190.

[10] C. Wu, F. Wu, M. An, J. Huang, Y. Huang, X. Xie, Neural news recommendation with attentive multi-view learning, arXiv preprint arXiv:1907.05576 (2019).

[11] M. An, F. Wu, C. Wu, K. Zhang, Z. Liu, X. Xie, Neural news recommendation with long- and short-term user representations, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, pp. 336–345. URL: https://www.aclweb.org/anthology/P19-1033. doi:10.18653/v1/P19-1033.

[12] C. Wu, F. Wu, S. Ge, T. Qi, Y. Huang, X. Xie, Neural news recommendation with multi-head self-attention, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural

Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 6389–6394. URL: https://www.aclweb.org/anthology/D19-1671. doi:10.18653/v1/D19-1671.

[13] H. Wang, Y. Fu, Q. Wang, H. Yin, C. Du, H. Xiong, A location-sentiment-aware recommender system for both home-town and out-of-town users, in: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '17, Association for Computing Machinery, New York, NY, USA, 2017, p. 1135–1143. URL: https://doi.org/10.1145/3097983.3098122. doi:10.1145/3097983.3098122.

[14] P. Padia, K. H. Lim, J. Cha, A. Harwood, Sentiment-aware and personalized tour recommendation, in: 2019 IEEE International Conference on Big Data (Big Data), 2019, pp. 900–909. doi:10.1109/BigData47090.2019.9006442.

[15] C. Orellana-Rodriguez, E. Diaz-Aviles, W. Nejdl, Mining affective context in short films for emotion-aware recommendation, in: Proceedings of the 26th ACM Conference on Hypertext & Social Media, HT '15, Association for Computing Machinery, New York, NY, USA, 2015, p. 185–194. URL: https://doi.org/10.1145/2700171.2791042. doi:10.1145/2700171.2791042.

[16] C. Musto, G. Rossiello, M. de Gemmis, P. Lops, G. Semeraro, Combining text summarization and aspect-based sentiment analysis of users' reviews to justify recommendations, in: Proceedings of the 13th ACM Conference on Recommender Systems, RecSys '19, Association for Computing Machinery, New York, NY, USA, 2019, p. 383–387. URL: https://doi.org/10.1145/3298689.3347024. doi:10.1145/3298689.3347024.

[17] D. Hyun, C. Park, M.-C. Yang, I. Song, J.-T. Lee, H. Yu, Review sentiment-guided scalable deep recommender system, in: The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR '18, Association for Computing Machinery, New York, NY, USA, 2018, p. 965–968. URL: https://doi.org/10.1145/3209978.3210111. doi:10.1145/3209978.3210111.

[18] A. Da'u, N. Salim, Sentiment-aware deep recommender system with neural attention networks, IEEE Access 7 (2019) 45472–45484. doi:10.1109/ACCESS.2019.2907729.

[19] J. Urbano, H. Lima, A. Hanjalic, Statistical significance testing in information retrieval: An empirical analysis of type i, type ii and type iii errors, in: Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR'19, Association for Computing Machinery, New York, NY, USA, 2019, p. 505–514. URL: https://doi.org/10.1145/3331184.3331259. doi:10.1145/3331184.3331259.

[20] M. D. Smucker, J. Allan, B. Carterette, A comparison of statistical significance tests for information retrieval evaluation, in: Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management, CIKM '07, Association for Computing Machinery, New York, NY, USA, 2007, p. 623–632. URL: https://doi.org/10.1145/1321440.1321528. doi:10.1145/1321440.1321528.

[21] J. Pennington, R. Socher, C. D. Manning, Glove: Global vectors for word representation, in: Empirical Methods in Natural Language Processing (EMNLP), 2014, pp. 1532–1543. URL: http://www.aclweb.org/anthology/D14-1162.

[22] S. Bird, E. Klein, E. Loper, Natural language processing with Python: analyzing text with the natural language toolkit, " O'Reilly Media, Inc.", 2009.

[23] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, arXiv preprint

arXiv:1412.6980 (2014).