A Comparative Analysis of Text-Based Explainable Recommender Systems

Alejandro Ariza-Casabona
University of Barcelona, CLiC-UBICS
Barcelona, Spain
alejandro.ariza14@ub.edu

Ludovico Boratto
University of Cagliari
Cagliari, Italy
ludovico.boratto@acm.org

Maria Salamó
University of Barcelona, CLiC-UBICS
Barcelona, Spain
maria.salamo@ub.edu

ABSTRACT

One way to increase trust among users towards recommender systems is to provide the recommendation along with a textual explanation. In the literature, extraction-based, generation-based, and, more recently, hybrid solutions based on retrieval-augmented generation have been proposed to tackle the problem of text-based explainable recommendation. However, the use of different datasets, preprocessing steps, target explanations, baselines, and evaluation metrics complicates the reproducibility and state-of-the-art assessment of previous work among different model categories for successful advancements in the field. Our aim is to provide a comprehensive analysis of text-based explainable recommender systems by setting up a well-defined benchmark that accommodates generation-based, extraction-based, and hybrid approaches. Also, we enrich the existing evaluation of explainability and text quality of the explanations with a novel definition of feature hallucination. Our experiments on three real-world datasets unveil hidden behaviors and confirm several claims about model patterns. Our source code and preprocessed datasets are available at https://github.com/alarca94/text-exp-recsys24.

CCS CONCEPTS

• **Information systems** → **Recommender systems**; • **Computing methodologies** → *Natural language generation*; Information extraction.

KEYWORDS

Explainable Recommendation, Natural Language Explanations, Reproducibility, Feature Hallucination

ACM Reference Format:
Alejandro Ariza-Casabona, Ludovico Boratto, and Maria Salamó. 2024. A Comparative Analysis of Text-Based Explainable Recommender Systems. In *18th ACM Conference on Recommender Systems (RecSys '24), October 14–18, 2024, Bari, Italy*. ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3640457.3688069

RecSys '24, October 14–18, 2024, Bari, Italy

ACM ISBN 979-8-4007-0505-2/24/10
https://doi.org/10.1145/3640457.3688069

1 INTRODUCTION

Explanations are crucial in artificial intelligence, especially when models act as "black boxes", as it heavily affects users' trust and reliability of the system [2, 11, 32]. One important field in which explanations can directly affect the system's utility and subsequent revenue is recommendation. Explanations in recommender systems can be presented in different modalities and the most common ones are textual [1, 17, 18], visual [8], attribute [47], and graph-based [9, 13, 39] explanations. The focus of this paper is on the textual modality, hence models providing explanations in natural language. Text-based explainable recommenders often exploit historical user-item reviews to generate explanations, as reviews most likely contain the user's buying reasoning and/or hands-on experience with the product. Depending on how textual explanations are obtained from these existing reviews, text-based explainable recommenders can be further categorized as: *generation-based* [1, 7, 10, 17, 19], *extraction-based* [3, 14, 25, 35], and *hybrid* [4, 41, 44] models. Unfortunately, the lack of source code availability in the recent research and the fact that each category uses its own definition of explanation, set of evaluation metrics, and preprocessing techniques, among others, make reproducibility and inter-category comparisons challenging. The novelty of our work lies in the creation of a unified reproducible evaluation benchmark for text-based explainable recommendation and a comparative analysis of the current state-of-the-art, emphasizing potential improvement directions.

According to the literature, several highlights can be deduced from each model category. On the one hand, *generation-based* models form the explanations word by word in an autoregressive fashion by leveraging past user-item representations. Therefore, although these models have the potential to achieve a superior and more personalized explanation performance, they suffer from the sparsity problem present in recommendation scenarios and lack controllability in their generation process [15], leading to repetitive and shallow explanations [17, 31]. Furthermore, the lack of grounded facts about items when generating explanations leads them to the so-called *hallucination effect* [26, 30]. On the other hand, *extraction-based* models form an explanation by selecting the top-k explanation sentences from a candidate pool made of past user-item review sentences. Therefore, this type of models is limited to the variety and completeness of the selected candidate pool and may suffer in (semi-) cold-start scenarios where multiple relevant item attributes are missing. Fortunately, the hallucination effect can be easily overcome in extraction-based models by filtering the candidate pool of explanation sentences.

More recently, *hybrid* architectures have been proposed combining extraction-based and generation-based solutions in order to exploit the benefits of both, in a retrieval-augmented generation fashion. In particular, they often extract a set of candidate explanations that serve as grounded context to a generative model that is responsible for creating or overwriting them into a more personalized explanation.

Despite all three types of models being considered as text-based explainable recommenders, they are hardly ever compared to each other besides the simple baselines from each category. Therefore, as previously mentioned, it hinders a proper state-of-the-art assessment and the evolution of explainable recommender systems research. After carefully analyzing the most recent literature and resources, the lack of comparisons mostly lies on the difficulty to reproduce previous work due to missing code and data, the mismatch among evaluation criteria employed in each work and the use of different pre-processing techniques leading to skewed results. Henceforth, the contributions of our work can be described as:

- Set up a unified benchmark to evaluate generation-based, extraction-based, and hybrid models under the same conditions, including problem formulation, data preprocessing, and evaluation metrics.
- Propose a novel feature hallucination metric that allows to measure the hallucination effect on the explainable recommendation task.
- Provide a reproducible and detailed comparison under the proposed benchmark, including the new feature hallucination perspective. To the best of our knowledge, our work is the first one to compare state-of-the-art models from different text-based explainable categories, and its impact on future research can be highly beneficial.

2 RESEARCH PROCEDURE

In this section, a description of the model selection process, model replication and benchmark configuration (data preprocessing and evaluation criteria) for each research line is provided.

2.1 Paper Selection

We selected a mixture of baseline models often used in every text-based explanation category and recent proposals found by filtering papers published in top-tier information retrieval conferences[1] and journal publishers[2]. Snowball search seeded by baseline references and term-based search were primarily used to find potential candidates. The combination of several search terms i.e. "extraction-based", "generation-based", "recommender system", "recommendation", "explainable", were used to perform such publication term-based filtering. The search was further refined by excluding publications that did not explicitly mention a reference to their source code implementation, and we were unable to find their respective code repository in public version control websites. Out of 17 state-of-the-art potential candidates from last year (2023), 9 works did not provide the source code and were not considered reproducible for our comparison study [5, 6, 12, 27, 31, 33, 38, 45, 48]. Thus, our reproducible candidate pool was reduced to only 8 works: PRAG [40], EMER [50], P5 [10], AdaReX [43], SEQUER [1], POD [18], ERRA [4], and ExBERT [44]. PRAG [40] was also discarded due to

[1]CIKM, ECIR, FAccT, KDD, RecSys, SIGIR, WSDM, WWW, UMAP
[2]ACM, Elsevier, IEEE, Springer

Table 1: Relevant models for our study.

Model	Year	Status[1]	Split[2]	Datasets[3]	Tasks[4]
Generation-based Models					
NRT [19]	2017	R	Rand	AZ, Y	RP, EG
PETER [17]	2021	R	Rand	AZ, Y, TA	RP, EG,
SEQUER [1]	2023	R	Seq	AZ, Y	RP, EG, NIR
POD [18]	2023	R	Rand, Seq	AZ	EG, NIR, TNR
Extraction-based Models					
ESCOFILT [25]	2021	M	Rand	AZ	RP, EE
GREENer [35]	2022	M	Rand	RB, TA	FE, EE
Hybrid Models					
ERRA [4]	2023	M	Rand	AZ, Y, TA	RP, EE, EG
ExBERT [44]	2023	R	Rand	AZ, Y	RP, EE, EG

[1]Status - R: Reproducible, M: Partial Code Missing
[2]Split - Rand: Random, Seq: Sequential
[3]Datasets - AZ: Amazon [23], Y: Yelp[10], TA: TripAdvisor [34], RB: RateBeer [22]
[4]Tasks - EG/EE: Explanation Generation/Extraction, RP: Rating Prediction, FE: Feature Extraction, NIR/TNR: Next-Item/Top-N Recommendation

insufficient computational resources on our end. EMER [50] and P5 [10] were discarded due to the use of auxiliary information about the items. AdaReX [43] was discarded due to their focus on cross-domain recommendation, thus requiring significant changes to their implementation.

Table 1 presents the final selection for this study, providing an overview of our baselines and state-of-the-art model choices, along with their reproducibility status, train-test splitting criteria, dataset choices, and tasks for which each model is optimized.

2.2 Problem Formulation and Discrepancies

The aim of a text-based explainable recommender system is to generate a natural language explanation $\hat{e}_{u,i}$ to explain the recommendation of item i to user u. In order to create the training corpus, the user-item interaction history between a set of users $u \in U$ and a set of items $i \in I$ along with the corresponding user reviews $t \in T$ and additional information such as the ratings $r \in R$ is collected. Therefore, each interaction can originally be described as a set $\{u, i, t_{u,i}, r_{u,i}\}$. Next, the relevant item features for the user $f \in \mathcal{F}_{u,i}$ and recommendation explanations $e \in \mathcal{E}_{u,i}$ are extracted from those reviews, forming richer sets $\{u, i, \mathcal{E}_{u,i}, \mathcal{F}_{u,i}, r_{u,i}\}$ that are used to train and evaluate the models.

The common choice for feature extraction is the Sentires toolkit[3] [47, 49] that extracts multiple tuples of (feature, adjective, opinion, sentence) from each review. However, there is a significant difference on how generation-based and extraction-based models treat this information to build the final instance sets, particularly $\mathcal{E}_{u,i}$ and $\mathcal{F}_{u,i}$. On the one hand, *generation-based* models select a random extracted tuple per review as explanation for the given interaction, thus, reducing the review r, originally containing multiple sentences $\mathcal{E}$ and item features $\mathcal{F}$, to a single sentence e and feature word f. Therefore, the explainable models are only required to generate a single sentence as explanation. On the other hand, *extraction-based* models use the output of the feature extraction toolkit to form a set of item features for the whole corpus and

[3]https://github.com/evison/Sentires

manually filter invalid ones using their own domain knowledge. Then, reviews are tokenized into sentences and filtered by feature inclusion rules. Unlike generation-based proposals, there may be more than a single sentence as a valid explanation per interaction, and each explanation sentence may contain multiple item features. During evaluation of extraction-based models, the top-k most relevant sentences from the candidate pool are joined together and compared to the concatenation of groundtruth explanations for the corresponding instance.

Apart from the corpus creation, mismatches in terms of data preprocessing (lowercasing, tokenizers, etc.), as well as manual filtering of features and selection of evaluation metrics were overcome to enable the comparison of models from different categories and the construction of a unified benchmark. In terms of data preparation and sample creation, we closely followed the extraction-based approach. Despite multiple explanation sentences being available as a target, our benchmark is designed so that models are only required to predict a single explanation sentence, resembling the generation-based approach in that regard. The difference lies in the fact that the prediction does not necessarily need to match a randomly selected sentence out of all possible explanations. Consequently, several adjustments to the generation-based training (see Section 2.3.1) and the unified evaluation criteria (see Section 2.5) are necessary.

2.3 Model and Implementation Details

An overview of each selected model per category and the modifications to their implementations (if any) are presented below:

2.3.1 Generation-based Models. Although generation-based models are still expected to generate a single explanation sentence, the explanation target may contain more than one sentence. Thus, a sampler is required to decide which of those target sentences will be generated by the model at each training iteration. By randomly selecting an explanation sentence each time a user-item interaction is sampled, the model gains access to a more complete representation than simply fixing the explanation choice in the beginning.

NRT [4] *[19]:* Strong generation-based model that utilizes an RNN architecture to generate explanations. It is trained to provide a rating prediction and the explanation generation simultaneously.

PETER [4] *[17]:* First generation-based explainable model that uses a Transformer-like architecture to perform multiple tasks i.e. rating prediction, context prediction and explanation generation.

SEQUER [5] *[1]:* Generation-based model that extended PETER by including the user interaction sequence as the language model context and added next-item prediction to the multitask setup. It also explores different masking alternatives, differentiating prefix and decoding inputs for improved representation capabilities.

POD [4] *[18]:* Transformer-like model inspired on P5 [10] that uses a pre-trained multitask Large Language Model [28] as a backbone and finetunes it to perform Top-N recommendation, next-item recommendation and explanation generation. Its novelty lies in a prompt distillation technique.

[4]https://github.com/lileipisces/NLG4RS
[5]https://github.com/alarca94/sequer-recsys23

2.3.2 Extraction-based Models.

ESCOFILT [6] *[25]:* Extraction-based baseline that applies unsupervised text clustering (k-Means) on review sentences encoded with Sentence-BERT [29] to create user and item profiles. It is able to use these profiles to perform rating prediction. Although this model is able to provide sentences as explanations for its predictions, the original source code does not provide this functionality. Therefore, similar to [35], we use the k sentences closest to the item cluster centroids to provide the explanation. As our evaluation consists of generating a single explanation sentence per interaction, if k is greater than 1, a random sampling is performed on those candidate sentences.

GREENer [7] *[35]:* Extraction-based model that constructs a heterogeneous graph of user-item-attribute-sentence nodes from the historical reviews of both users and items, and proposes a graph neural network solution followed by a Deep Cross Network [36] in order to rank the candidate sentences and attributes present in the graph. In their original implementation, they further re-rank the selected sentences using Integer Linear Programming (ILP) to balance relevance and content redundancy. In our experiments (see Section 3.3), we provide results with and without the re-ranking post-processing step. We thank the authors for providing us the code/instructions to run their model. However, since the code to generate the heterogeneous graph from the dataset was incomplete, we created the graphs using the most recent history of users and items. This ensured that the shape distribution of the graphs closely matched that of the original dataset.

2.3.3 Hybrid Models.

ERRA [8] *[4]:* Hybrid architecture composed of a retrieval mechanism to extract relevant attributes and explanation sentences from the corpus, followed by a Transformer-like architecture inspired in PETER. The retrieval mechanism uses Sentence-BERT to semantically encode sentences and item attributes. Despite the source code being available, several discrepancies were found when compared to the information stated in their published article. The most noticeable one was the lack of an aspect discriminator in the loss function and model architecture. The code for the attributes and sentence extraction was also missing. Our consideration on the matter was to follow the publication details as guideline and implement the missing parts accordingly.

ExBERT [9] *[44]:* Inspired by PETER, ExBERT uses a Transformer-like architecture with multitask learning. However, they replace causal masking with bidirectional masking for text generation and add a new loss term called "matched explanation prediction". Furthermore, they include an extraction mechanism using a Sentence-BERT encoder to create pseudo user and item profiles that are then fed into the model. Note that authors used an EMA (Exponential Moving Average) training for stabilization. In addition, the user and item profiles were built by computing semantic similarity between historical explanations and the target explanation and extracting

[6]https://github.com/reinaldncku/ESCOFILT
[7]https://github.com/HCDM/XRec
[8]https://github.com/Complex-data/ERRA
[9]https://github.com/zhanhuijing/ExBERT

Table 2: Dataset Statistics and filtering thresholds.

	Yelp	TripAdvisor	RateBeer
k-core	5	5	20
Unique Reviews	4	4	20
# Users	18,173	22,985	4,731
# Items	12,680	7,192	5,292
# Reviews	186,549	270,922	693,620
# Features	498	503	570
Avg. # Reviews / User	10.27	11.79	146.61
Avg. # Exp. / Review	2.47	4.34	3.79
Avg. # Words / Exp.	15.53	19.90	14.12
Avg. # Feats. / Exp.	2	3.32	3.96
95% Exp. Length	32	42	23
Sparsity	0.08	0.16	2.77

Table 3: Description of the evaluation metrics.

Type	Metric	Description
Explainability	DIV	Feature Diversity among explanations.
	FCR	Feature Coverage Ratio at corpus level.
	FP	Average Feature Precision.
	FR	Average Feature Recall.
	FHR	Feature Hallucination at explanation level.
	D-FHR	Feature Hallucination at document level.
Text Quality	USR	Unique Sentence Ratio.
	AL	Average Length of the explanations.
	IDF/W	Inverse Document Frequency per Word in the explanations.
	Rep-L	Content Repetition at unigram level.
	Seq-Rep-2	Content Repetition at bigram level.
	BLEU-n	n-gram similarity metric used in machine translation.
	ROUGE-n	n-gram similarity metric used in text summarization.
	BERT-S	Semantic similarity score based on BERT.

the top-k most similar sentences. This approach is supposed to yield the upper bound of the model performance assuming a perfect extraction mechanism, but it is an unrealistic setup. Therefore, we also present the results for this model with a more naive profile creation in Section 3.3.

2.4 Data Preparation

The evaluation of the models was conducted on three datasets commonly used in the review-based recommendation literature: TripAdvisor [34], Yelp[10], and RateBeer [22]. The reasoning behind our dataset selection was to favor under-explored yet popular datasets. Results for the commonly explored Amazon dataset partitions will be included in an extended analysis left as future work. Our preprocessing phase consisted of three steps. Following [42], we first extracted the valid item features and explanation sentences for each dataset from the user reviews. Then, we filtered each of these datasets via iterative k-core filtering to ensure a minimum item and user coverage. Given the fact that multiple users wrote the same review for several interactions, we additionally removed users with a minimum unique review threshold. Last, the RateBeer dataset was further cleansed by deduplicating user-item interactions, enforcing a minimum number of 6 tokens per explanation sentence and removing potential "aggregated" users by a maximum user occurrence threshold of 700.

Table 2 shows a summary of the preprocessed dataset statistics and the threshold values used per dataset. In terms of dataset split, due to the fact that some models make use of historical data, we performed a user-level temporal split similar to sequential recommendation with ratios 8:1:1 for train, validation and test sets respectively. Items appearing in the validation and test sets are required to appear in the training set at least once. Additional data preparation such as minimum context size or negative sampling was performed on a per model basis.

2.5 Evaluation Criteria

Although the groundtruth explanation $\mathcal{E}_{u,i}$ may be multiple sentences with each sentence containing one or several item features, the explainable model is only asked to predict one of those $\hat{e}_{u,i}$. Therefore, the evaluation metrics are computed in a one-to-many fashion and the maximum score is kept.

Unlike extraction-based models, generation-based approaches require a maximum text length to generate the explanation token by token. Previous studies often generated the same number of tokens for all datasets, which corresponded to roughly the average length of explanations. However, datasets statistics (see Table 2) differ from one another and some domains may require richer explanations than others. Despite the fact that shorter explanation sentences are preferred over long ones, we propose to cover at least 95% of the explanations in terms of number of words during the generation process. Concretely, we allow generation-based models to generate 35/45/25 tokens for Yelp/TripAdvisor/RateBeer datasets.

In our study, a combination of 14 metrics (i.e., 12 coming from previous work in extraction-based, generation-based and hybrid approaches and 2 novel ones) are used to evaluate the models. The metrics can be classified into two types (see Table 3 for a brief description of each metric). One type that accounts for item attributes in the produced explanations and measures the *explainability* capabilities of the model and another type that measures the *quality of the text*. Following [17, 35] for *explainability* metrics, DIV, FCR [16], FP and FR are computed. Following [4, 17, 41] for text quality metrics, USR [16], Avg. Len. (AL), IDF per word (IDF/W), Rep/L, Seq-Rep-2 [37], BLEU [24] (machine translation), ROUGE [21] (text summarization), and BERTScore (BERT-S) [46] are the selected metrics. Additionally, due to the importance of good quality explanations to improve the trustworthiness of the recommender system and the recent limitations of Large Language Models generating non-reliable or fake content, we propose two novel metrics to measure the degree of hallucination in terms of item features being mentioned in the generated explanations. The novel sentence-level Feature Hallucination Ratio (FHR) and the Document-level FHR (D-FHR) are defined in Equations (1), (2), respectively. Both metrics are normalized in the range $[0, 1]$, where 0 indicates the absence of hallucinations in the provided explanations.

$$FHR = \frac{1}{|\mathcal{D}|} \sum_{u,i}^{\mathcal{D}} \frac{|\hat{\mathcal{F}}_{u,i} - \mathcal{F}_i|}{|\hat{\mathcal{F}}_{u,i}|} \tag{1}$$

[10]https://www.yelp.com/dataset

Where $\mathcal{D}$ is the test dataset, $\hat{\mathcal{F}}_{u,i}$ is the set of features included by the model in the explanation sentence for the recommendation of item i to user u and $\mathcal{F}_i$ is the set of features mentioned in the dataset about item i.

$$D\text{-}FHR = \frac{\sum_{u,i}^{\mathcal{D}} (\hat{\mathcal{F}}_{u,i} - \mathcal{F}_i) \neq \varnothing}{|\mathcal{D}|} \quad (2)$$

Last, in terms of model hyperparameters, we use the ones reported by their respective authors. In order to facilitate hyperparameter search to future studies, our source code provides a python script to optimize hyperparameters using HyperOpt's Tree Parzen Estimators from the Ray library [20]. Additionally, we report average results of four independent runs per experiment.

3 EXPERIMENTAL RESULTS

Our study aims to evaluate different categories text-based explainable recommender systems and establish a common benchmark to assess the current state-of-the-art. Therefore, three research questions will be primarily addressed:

RQ1 What is the current state-of-the-art in text-based explainable recommender systems in terms of text quality and explainability performance?

RQ2 To what extent are text-based explainable models hallucinating item attributes?

RQ3 Which are the inner mechanisms of text-based explainable model that play a bigger role in their performance?

3.1 Explainability and Text Quality Comparison (RQ1)

Table 4 presents the results discussed in this section.

Evaluation perspectives and metrics complementarity: In the problem of text-based explainability, there are many factors that need to be considered. For example, shorter explanations are often preferred to longer ones as long as they are readable and provide user-targeted content[11]. The relevance of explanations for a particular user is another important aspect, and can be achieved by retrieving or generating explanations that contain the item features that strongly impact that user's behavior towards the recommended item. This user personalization can be measured by the feature diversity (DIV) as well as the feature precision (FP) and recall (FR). However, no single metric can account for all of these factors and complementary metrics are required to extract meaningful conclusions. Otherwise, single metric scores may lead to partial conclusions and an erroneous analysis. For instance, as shown in Table 4, despite the fact that ERRA achieves high text similarity scores with the groundtruth explanations according to BLEU and ROUGE metrics in the TripAdvisor dataset, its USR score is really low, indicating a poor explanation personalization. One possible reason for this output could be if the model opts to learn a very generic explanation including words and features really popular within the corpus documents, ignoring some other situational features. Also, the fact that a model is able to cover most of the features learned from the training dataset in their explanations as indicated by the FCR metric does not mean that the model is actually retrieving relevant features for the user at a given interaction measured by FP and FR, indicating poor user and item representation capabilities of the model. This shortcoming noticeably happens to extraction-based solutions over all three datasets.

Effect of data sparsity on explanations: As a general remark, the sparsity of the dataset plays an important role on all model categories, specially for generation-based approaches. Yelp is the dataset with the biggest data sparsity and the representation capabilities of generation-based models seem heavily affected, leading to reduced FCR, FP and FR scores. They tend to generate explanations using more popular item features, thus decreasing the feature diversity and leading to more repetitive explanations overall, as indicated by the USR metric. The generation-based model that suffers the most is POD, as shown in Table 4. Despite using a pretrained LLM that supposedly can already generate good quality text, the finetuning step results to be a challenging task in this scenario, as it requires the model to learn useful user/item representations that properly map those users and items to the already learned text semantic space. Note that SEQUER, PETER and NRT learn from scratch both representation spaces (text and recommendation) simultaneously. Nevertheless, the fact that generation-based models are the most affected ones does not mean that extraction-based models are indifferent to sparsity. For instance, GREENer uses embedding vectors to propagate user and item information through the heterogeneous graphs. Therefore, it experiences a drop of ∼50% in terms of FP among other metrics (DIV, BERTScore) when we compare its results in RateBeer and Yelp. This is not the case for ESCOFILT as it mainly relies on the textual content of the reviews to extract the explanations, and the user-item representations are only used for the rating prediction task, so interaction sparsity plays a less important role on its explainable effectiveness. With regard to hybrid approaches, their extraction mechanism from user and item profiles seem to mitigate the effect of data sparsity in their performance to certain extent, specially in the case of ExBERT, obtaining higher feature precision and recall as well as overall text quality in all three datasets.

Explanation length and content information: Regarding explanation length, extraction-based approaches favor the extraction of longer sentences as opposed to generation-based models that tend to generate shorter sentences. This is important to consider because precision-based metrics often benefit short-length text generation. This is the case for BLEU which, despite applying a brevity penalty to sentences shorter, may still struggle at estimating the brevity threshold when explanation sentences from user reviews vary a lot in terms of length as stated in [41]. In any case, short and concise explanations with informative item attributes are preferred by the users. Therefore, apart from the average explanation length, the system needs to maximize the IDF/W to maximize the information quantity and minimize the token repetition i.e. Rep/L and SR2. Extraction-based models excel at these aspects due to their candidate pool being formed by human-written sentences containing the reasoning behind their interaction behavior. Interestingly, the use of a pre-trained LLM (POD) that already knows how to structure sentences is able to achieve a really competitive performance regarding content repetition, information quantity per word and sufficient length brevity in all three datasets. The

[11]We consider explanation lengths to be optimal if they resemble the average explanation length of groundtruth explanations in user reviews.

Table 4: Explanation generation results. For each dataset, bold numbers indicate leading results, while underlining highlights second-best scores. Statistical significance between the leading and second-best scores is indicated by * for a *p-value* < 0.05.

	Methods	Explainability						Text Quality										
		DIV↓	FCR↑	FP↑	FR↑	FHR↓	D-FHR↓	USR↑	AL	IDF/W↑	Rep/L↓	SR2↓	B1↑	B4↑	R1↑	R2↑	RL↑	BERT-S↑
RateBeer	NRT	0.409	0.592	0.394	0.113	0.011	0.031	0.381	12.243	3.223	0.149	0.036	50.024	8.389	39.993	12.501	33.682	0.623
	PETER	0.581	0.325	0.351	0.102	0.009	0.023	0.085	12.039	3.079	0.142	0.037	49.026	6.924	38.704	11.172	32.435	0.617
	SEQUER	0.402	0.577	0.404	0.118	0.010	0.029	0.357	12.001	3.219	0.161	0.047	51.087	9.192	40.898	13.314	34.456	0.627
	POD	1.243	0.541	0.378	0.119	0.026	0.071	0.170	10.585	3.519	0.091	0.008	49.666	**12.216***	40.342	**16.213***	36.048	0.639
	ESCOFILT	**0.090***	1.000	0.162	0.042	**0.000***	**0.000***	**0.932***	**13.525**	**4.485***	**0.072***	**0.002**	28.617	1.004	23.688	2.291	19.163	0.550
	GREENer	0.827	**1.000**	0.348	0.113	0.031	0.125	0.568	12.353	3.962	0.083	0.002	45.668	10.582	39.498	15.484	35.165	0.638
	ERRA	0.451	0.344	0.388	0.103	0.008	0.019	0.192	12.008	3.120	0.183	0.060	48.551	7.060	39.499	11.859	33.098	0.615
	ExBERT	0.204	0.821	**0.476***	**0.139***	0.020	0.051	0.774	11.714	3.354	0.144	0.034	**54.412***	11.381	**43.723***	14.889	**36.601***	**0.643***
TripAdvisor	NRT	0.947	0.224	0.324	0.106	0.048	0.152	0.060	13.824	2.606	0.259	0.070	50.146	6.444	37.839	9.413	29.202	0.618
	PETER	0.960	0.141	0.314	0.105	0.068	0.218	0.021	14.235	2.624	0.269	0.074	49.357	6.225	37.504	9.370	28.933	0.607
	SEQUER	0.961	0.139	0.319	0.101	0.060	0.190	0.021	13.101	2.597	0.226	0.050	51.276	6.413	37.765	9.198	29.379	0.620
	POD	2.209	0.033	0.252	0.076	0.113	0.300	0.001	12.219	2.855	**0.083**	0.008	43.652	5.033	35.506	6.853	28.201	0.615
	ESCOFILT	**0.112***	**1.000**	0.180	0.046	**0.000***	**0.000***	**0.898***	**18.774**	**4.132***	0.115	0.005	31.897	1.299	24.853	2.631	18.782	0.561
	GREENer	0.209	0.999	0.283	0.078	0.030	0.094	0.732	16.956	3.621	0.118	**0.004**	42.927	14.582	33.899	10.606	27.582	0.613
	ERRA	1.577	0.050	0.364	0.112	0.020	0.061	0.002	12.592	2.437	0.263	0.108	53.760	7.385	39.167	10.369	30.325	0.611
	ExBERT	0.275	0.995	**0.551***	**0.175***	0.073	0.189	0.780	15.084	3.316	0.180	0.041	**59.192***	**30.327***	**50.286***	**28.555***	**43.345***	**0.687***
Yelp	NRT	1.059	0.104	0.302	0.162	0.096	0.137	0.052	10.353	2.555	0.227	0.047	38.106	4.020	33.693	6.751	26.608	0.621
	PETER	1.342	0.104	0.315	0.180	0.077	0.118	0.023	10.547	2.491	0.207	0.039	40.055	4.310	34.085	7.069	26.762	0.626
	SEQUER	1.569	0.075	0.301	0.172	0.105	0.164	0.016	**10.316**	2.412	0.201	0.023	39.605	4.114	34.021	6.857	26.941	**0.629***
	POD	1.337	0.005	0.188	0.072	0.272	0.340	0.000	12.052	3.409	0.074	**0.002**	27.091	1.735	22.228	2.984	17.417	0.581
	ESCOFILT	**0.054***	**0.999***	0.137	0.068	**0.000***	**0.000***	**0.839***	14.626	**4.289***	0.087	0.003	23.072	1.172	20.952	1.675	16.210	0.565
	GREENer	0.082	0.977	0.182	0.077	0.047	0.097	0.758	10.820	3.855	**0.067**	0.002	28.108	3.464	24.665	3.210	19.830	0.589
	ERRA	0.709	0.032	0.336	0.111	0.063	0.070	0.004	8.928	2.277	0.307	0.210	33.701	3.226	35.122	6.536	27.137	0.549
	ExBERT	0.301	0.257	**0.386***	**0.205***	0.160	0.233	0.248	10.935	2.905	0.243	0.084	**40.553**	**4.753***	**35.940**	**8.154***	**28.366***	0.612

latter may be conditioned on the beam search decoding strategy favoring shorter texts [19]. Unfortunately, hybrid approaches using Sentence-Transformers to encode sentences in their retrieval stage do not benefit from these LLM capabilities in the explanation generation step, since these embedding representations are simply used as input to a generative language model that is learned from scratch.

Analyzing category-specific model differences: Apart from analyzing aggregated aspects of each category as a whole, it is important to assess their respective state-of-the-art to allow future models to select better baselines. In the case of generation-based approaches, both SEQUER and POD outperform their respective baselines in terms of text quality metrics and explainable capabilities when a domain is not excessively sparse. However, they both suffer in sparse recommendation domains, especially POD that is not able to benefit from collaborative signals to align the recommendation and text representation spaces. SEQUER proves that sequential recommendation as context may be beneficial in multiple domains to the explanation generation task, enhancing their semantic text quality and feature relevance as per the BERTScore, BLEU-1, ROUGE-L and FP metrics. In extraction-based approaches, GREENer can be considered the state-of-the-art in terms of explainable capabilities. Despite the fact that GREENer is limited by past user reviews to extract the explanation sentences, the semantic similarity that it is able to achieve is on par with that of generation-based models. However, the message propagation through the graph of unrefined user and item representations in sparse domains penalizes greatly this model's choices. On the contrary, ESCOFILT builds the user and item profiles via unsupervised clustering of textual embeddings, avoiding the negative effect of sparsity but consistently leading to suboptimal explanation results. Finally, in terms of hybrid approaches, ExBERT completely outperforms the other approaches, including ERRA, in the most important metrics. Looking at the differences between the two, ERRA directly uses the extracted user and item profiles into the Transformer Decoder input while ExBERT uses a Transformer Encoder-Decoder architecture to obtain proper contextualized query representations that are fed to the bidirectional decoder. Also, the use of a matched explanation prediction task in ExBERT may help alleviate the problem of generic content generation. Surprisingly, this task and the retrieval mechanism do not aid the model into avoiding the hallucination of unseen item attributes, as the FHR is ~2-5x higher than their generation-based counterparts.

Observations RQ1. *Generation-based models often allow for more personalized explanations containing relevant feature attributes for the users and good semantic similarity with respect to the groundtruth explanations. Extraction-based models provide more informative explanations with lower content repetitiveness and a bigger feature coverage over all three datasets. However, both types of models suffer from the data sparsity and poor user-item representations. As such, generation-based models opt to explain recommendations with the most popular features, thus minimizing the feature coverage and unique sentence ratios in order to maintain the text quality and feature relevance. Extraction-based models, despite maintaining a good feature coverage, struggle to extract explanations with relevant item features to the user. Overall, ExBERT (hybrid approach) seems to limit the weaknesses of extraction-based and generation-based approaches and maintain state-of-the-art performance over all metrics in all three datasets.*

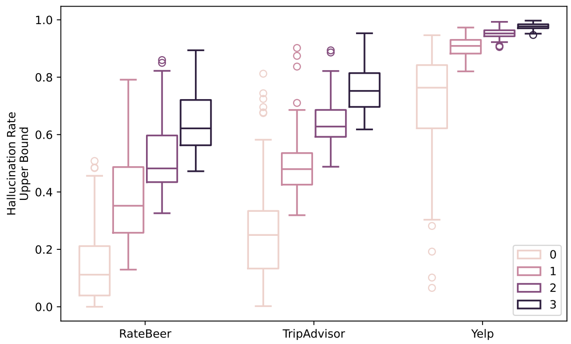

Figure 1: Hallucination rate upper bound in the test partition per dataset and features grouped by popularity in quartiles. Groups 0 to 3 are sorted in decreasing popularity order.

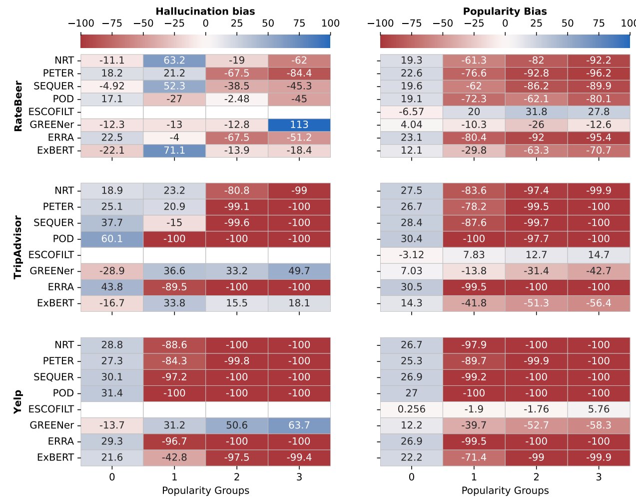

Figure 2: Heatmap of hallucination bias (left column) and popularity bias (right column) for each model in percentage. Note that the bias range is only bounded on the negative side (-100%), thus the positive range of the heatmap is set to a maximum of 100% for symmetric purposes. ESCOFILT is left blank in the left column as their pool of explanation sentences is only formed by past item explanations, thus its chances to hallucinate item features is zero.

3.2 The Problem of Feature Hallucination (RQ2)

In this section, we aim to analyze the feature hallucination effect in relation to the feature popularity and possible biases learned by the models. A feature is considered to be hallucinated in an explanation if there was no mention of this feature for the recommended item in any user review of the whole dataset. In other words, feature popularity has a direct effect on the feature hallucination, as it directly affects on its chances of being hallucinated. That is, the more popular a feature is, the less ratio of items will be left for it to be hallucinated. In order to visualize this phenomenon, we compute the maximum hallucination rate per feature in the test set and divide into four equal sized groups based on their feature popularity in the training set (popularity defined as the ratio of occurrence in user reviews). The resulting boxplot chart per popularity group and dataset is presented in Figure 1. It clearly depicts how, on average, the more popular a feature is (Group 0) the lower the chance of being hallucinated becomes in all three dataset. For instance, in the RateBeer dataset, the majority of the features belonging to the most popular group (Group 0) can be hallucinated in less than 20% of the test instances while the least popular group (Group 3) can be hallucinated in ∼65% of the instances.

However, despite this inverse "popularity-hallucination chance" relation, measuring to how extent one affects the other remains an unknown factor. In our attempt to shed light on the matter, we tackle the problem from a bias perspective in terms of both, popularity and hallucination (see Figure 2). Concretely, popularity and hallucination distributions will be computed as the expected ratio of features belonging to each popularity group. Popularity distributions will be calculated for all instances and hallucination distributions only for those instances where a model produces a hallucination.

On the one hand, the popularity bias for each feature group is based on relative distance between the distribution of predicted features in the models' explanations and the feature distribution in the training reviews. Overall, there is a model tendency to have a positive popularity bias towards the most popular group (Group 0), specially for generation-based and hybrid models. This effect appears to increase together with the data sparsity. In fact, several models (ERRA, SEQUER, POD) opt not to use the least popular groups at all, to the extreme point that POD only uses features from the most popular group to generate the explanations in Yelp. Extraction-based methods are by far the least affected in terms of popularity bias, although the state-of-the-art GREENer model still seems to suffer in Yelp with a relative decrease of 58.3% in popularity for the Group 3.

On the other hand, in terms of hallucination, not all features are eligible for each instance, deviating it from a uniform group distribution and making the hallucination bias definition not so trivial. Therefore, each model requires its own reference group-based hallucination distribution based on an unbiased version of itself and the instances where it produced a hallucination. We assume that an unbiased version of the model will hallucinate features from each group in the same proportion as their group popularity for the set of features that can actually be hallucinated. In this way, we correct the shift from the uniform distribution caused by the phenomenon depicted in Figure 1 at the instance level. Thus, hallucination bias is defined as the relative distance between the distributions of the biased and unbiased versions of the model. Figure 2 indicates that a positive popularity bias does not always incur in the same relative increase in terms of group hallucination. Data sparsity and model complexity play an important role in this regard as well. For instance, most models experiment an even bigger increase in hallucination bias for the group 0 with respect to popularity bias in the most sparse domain (Yelp). Thus, over-generalization of the popular features under poor and unreliable representation capabilities of the models lead to increased hallucinations using popular features, which negatively impact on user personalization and user's trust on the system's explanations. For instance, in NRT, if the popularity

Table 5: Ablation analysis on three text-based explainable recommenders (SEQUER, GREENer, ExBERT). Bold numbers indicate leading scores for each ablation. Statistical significance is indicated by * for a *p-value* < 0.05.

	Methods	Explainability						Text Quality										
		DIV↓	FCR↑	FP↑	FR↑	FHR↓	D-FHR↓	USR↑	AL	IDF/W↑	Rep/L↓	SR2↓	B1↑	B4↑	R1↑	R2↑	RL↑	BERT-S↑
Ratebeer	SEQUER	0.402	0.577	**0.404**	**0.118**	**0.010***	**0.029***	0.357	12.001	3.219	**0.161**	**0.047**	**51.087**	**9.192**	**40.898**	**13.314**	**34.456**	**0.627**
	SEQUER-FIXED	**0.387***	**0.620***	0.401	0.117	0.012	0.033	**0.396***	12.079	**3.243***	0.164	0.048	50.772	9.077	40.720	13.142	34.232	0.626
	GREENer	**0.827**	**1.000**	0.348	0.113	0.031	0.125	**0.568**	12.353	**3.962**	**0.083**	0.002	45.668	10.582	39.498	15.484	35.165	0.638
	GREENer-NOILP	0.897	1.000	**0.361***	**0.121***	**0.027**	**0.114**	0.528	12.320	3.882	0.084	**0.002**	**46.907**	**11.679***	**40.708***	**16.597***	**36.297***	**0.645***
	ExBERT	**0.204***	**0.821***	**0.476***	**0.139***	0.020	0.051	**0.774***	11.714	**3.354***	**0.144**	**0.034**	**54.412***	**11.381***	**43.723***	**14.889***	**36.601***	**0.643***
	ExBERT-NAIVE	0.397	0.680	0.389	0.116	**0.014***	**0.040***	0.450	12.095	3.239	0.147	0.035	50.486	8.871	40.224	12.760	33.989	0.627
Tripadvisor	SEQUER	0.961	0.139	**0.319**	**0.101***	0.060	0.190	0.021	13.101	2.597	**0.226**	**0.050***	**51.276***	**6.413***	**37.765***	**9.198***	**29.379***	**0.620**
	SEQUER-FIXED	**0.781***	**0.249***	0.310	0.097	**0.060**	**0.181**	**0.068***	13.299	**2.628**	0.237	0.059	49.704	6.066	37.185	8.829	28.880	0.616
	GREENer	**0.209**	**0.999**	**0.283**	0.078	0.030	0.094	0.732	16.956	3.621	**0.118**	**0.004**	**42.927**	**14.582**	**33.899**	**10.606**	**27.582**	**0.613**
	GREENer-NOILP	0.251	0.999	0.276	**0.082**	**0.027**	**0.089**	**0.751**	17.784	**3.650**	0.123	0.005	42.460	13.926	33.751	10.378	27.282	0.611
	ExBERT	**0.275***	**0.995***	**0.551***	**0.175***	**0.073***	**0.189***	**0.780***	15.084	**3.316**	**0.180***	**0.041**	**59.192***	**30.327***	**50.286***	**28.555***	**43.345***	**0.687***
	ExBERT-NAIVE	1.351	0.115	0.290	0.112	0.094	0.306	0.013	15.584	2.641	0.239	0.052	50.299	6.048	37.298	9.214	28.658	0.613
Yelp	SEQUER	1.569	0.075	**0.301***	**0.172***	**0.105***	**0.164***	0.016	10.316	2.412	**0.201**	**0.023**	**39.605**	**4.114***	**34.021***	**6.857***	**26.941***	**0.629***
	SEQUER-FIXED	**1.303***	**0.134***	0.290	0.163	0.122	0.181	**0.036***	10.383	**2.502***	0.213	0.039	38.751	3.931	33.676	6.579	26.582	0.622
	GREENer	0.082	**0.977**	**0.182***	**0.077***	0.047	0.097	0.758	10.820	3.855	**0.067**	**0.002**	28.108	3.464	**24.665**	3.210	**19.830**	**0.589**
	GREENer-NOILP	**0.044***	0.977	0.142	0.068	**0.038**	**0.080**	**0.909***	12.387	**4.116***	0.074	0.002	**28.574**	**5.044***	24.373	**3.557***	19.713	0.583
	ExBERT	**0.301***	**0.257***	**0.386***	**0.205***	0.160	0.233	**0.248***	10.935	**2.905***	0.243	0.084	**40.553**	**4.753***	**35.940***	**8.154***	**28.366***	0.612
	ExBERT-NAIVE	1.534	0.009	0.313	0.160	**0.089***	**0.131***	0.003	10.029	2.406	**0.172***	**0.012***	38.420	3.825	33.479	6.528	26.727	**0.627***

bias is 26.7%, a reliable representation of the features in terms of hallucination would be expected to result in a hallucination bias lower than 26.7%, but it increases to 28.8%. On the contrary, multiple models are able to keep reliable explanations in terms of feature hallucination at the RateBeer dataset. For instance, despite the fact that the generation-based NRT baseline has a positive bias of 19.3% for the most popular group, the association capabilities of the model for this type of features and the recommended items is good enough to partially avoid hallucinations of this group. However, most of the models still have problems using features of the least popular groups to explain the recommendations.

Observations RQ2. *In terms of both, feature popularity and hallucination bias, extraction-based approaches are by far the least affected. Generation-based and hybrid approaches present a considerable popularity bias towards the most popular features that is exacerbated by the data sparsity. Furthermore, their over-generalization of popular features in such scenarios lead to increased hallucinations for the positive group in the generated explanations. Conversely, the least popular group of features is majorly avoided to explain the recommendations which also limits the user personalization capabilities of generation-based and hybrid models.*

3.3 Exploratory Ablation Studies (RQ3)

Table 5 presents the results of three small ablation studies including a benchmark design choice for generation-based models, the effect of re-ranking in the extraction-based state-of-the-art model and one concern about ExBERT retrieval mechanism.

Generation-based performance with fixed explanation sampling: In Section 2.3.1, we stated a training design choice for generation-based models. Our hypothesis was that randomly sampling an explanation sentence to be generated by the model at training time, could help the model learn a better representation of an item's features and favor a more coherent set of explanations during training than simply fixing the sampler at the beginning and, potentially, avoid confusions between subsequent explanation generations for the same user-item interactions. For this test, we selected the state-of-the-art generation-based SEQUER model and denoted the variation with the fixed sampler as SEQUER-FIXED. According to the results in Table 5, our alternate sampling allows the model to provide explanations using more relevant features on average (see FP and FR), and provide higher quality texts (see ROUGE, BLEU and BERTScore), at the expense of slightly lower user personalization (see USR and DIV). Results support our hypothesis, as alternate sampling allows the model to see a variety of explanations and features for each user-item interaction in the training set.

Effect of explanation diversity re-ranking on extraction relevance: Extraction-based approaches usually extract top-k sentences that are used as explanations. In cases where k is greater than one, it may be advantageous to choose sentences that are diverse to improve the relevant attribute coverage. In order to achieve accurate but diverse sentences in its top selection, GREENer applies a re-ranking of the sentences before retrieving the explanation. However, in our benchmark, the model is only asked to retrieve an explanation sentence. Therefore, diversity is no longer required among explanation sentences, but within the extracted sentence content. In such cases, we aim to determine how the re-ranking mechanism embedded in GREENer actually affects the top-1 extracted sentence of the model in terms of user explanation relevance. The version without the re-ranking is denoted GREENer-noILP. It is clear that the re-ranking mechanism penalizes the model in both, explainability and text quality, in the RateBeer dataset, but helps it in the Yelp dataset. Therefore, in a dataset where the model exhibits better performance in their original ranking, re-ranking incurs a greater diversity-relevance trade-off.

Importance of accurate retrieval in hybrid approaches: The last ablation study correspond to our concern about ExBERT retrieval mechanism mentioned in Section 2.3.3. The model uses the historical explanation sentences that better match the groundtruth explanation for a given recommendation interaction. However, the groundtruth explanation should be hidden from the retrieval mechanism and, despite the quality of semantic representation capabilities of the selected Sentence-Transformer, the results may serve as a performance upper bound of what a hybrid architecture can achieve, should the retrieval method be optimal. For this experiment, we replace the retrieval mechanism with a random sampling using the same historical window size and denote the variant as ExBERT-NAIVE. The performance of the model consistently decreases on all three datasets, leading the model to the generation of less personalized and relevant explanations.

Observations RQ3. *First, generation-based approaches benefit from a random sampling of explanations sentences to generate during training as opposed to a fixed sampler from the traditional setup. Second, diversity rerankers in extraction-based models heavily penalizes the relevance of the top-1 explanation with accurate rankings of the model but can actually help achieve better explanations when data sparsity leads to noisier and more unreliable explanation rankings. Third, a noisy retrieval method in hybrid approaches results in performances comparable to generation-based approaches without the retrieval mechanism, while an optimal retrieval step lead to state-of-the-art results in explainability and text quality metrics.*

3.4 Qualitative Summary

The quantitative analysis is supported in many instances by the qualitative results. Table 6 shows the explanation prediction of each model for a test instance that accurately serve to sum up their general behavior. In the table, the content repetition problem is present in ERRA's and PETER's explanations. In addition, the feature hallucination occurs in the generation-based NRT baseline indicating an erroneous aroma of the recommended item. Furthermore, extraction-based models provide a more diverse set of features but tend to extract longer explanations. Finally, we noticed the lack of user personalization of models such as POD resorting to the same explanation sentence for up to ∼95% of the samples at Yelp.

Table 6: Example of explanation on the RateBeer dataset for all models. Bold words represent predicted features. Underlined words indicate the feature is hallucinated. Italic words indicate the predicted feature is in the groundtruth explanation.

Model	Explanation
NRT	**strong** ***aroma*** of ***orange*** and <u>**grapefruit**</u> , very **sweet** and **malty** , a little too **sweet** for my liking
PETER	nice **amber color** , nice **head** , nice **lacing** , nice **lacing** , nice **malty** ***aroma*** , a little **sweet** , a bit
SEQUER	**sweet** ***aroma*** , **dark amber** with a thin **head** , **sweet malty flavor** , very **hoppy** , **bitter finish**
POD	**malty** ***aroma*** , **dark brown** with a thin **tan head** , **malty** with a
ESCOFILT	beer of the month club , **pours deep red** to **brown** with a decent **head** , ***aroma*** of marmalade , **strawberry** cake , ***malt*** and **vanilla** , **flavor** of british **hops** , **floral** notes , some biscuit ***flavors*** and long **bitter finish** with **tobacco** notes , medium **bodied** , nice .
GREENer	**strong** ***citrus hop aroma*** , ***orange*** with a tightly packed off **white head** , thin ***body*** , very **grassy** with ***citrus hop*** notes and building ***bitterness***
ERRA	nice **malty** ***aroma*** , nice **malty** ***aroma*** , nice **malty** ***aroma*** , nice **malty** ***aroma*** , nice ***hop aroma*** , nice ***hop bitterness***
ExBERT	***aroma*** is a bit of **caramel** , some **earthy hops** , and a bit of a **metallic** note , but not much else

4 CONCLUSIONS AND OPEN ISSUES

This paper presents a comparative analysis of text-based explainable recommender systems under a unified evaluation benchmark to assess the current state-of-the-art under different model categories: generation-based, extraction-based, and hybrid. In addition to traditional evaluation, we also present a novel definition of feature hallucination and measure the hallucination effect in relation to their popularity bias. Our analysis identifies extraction-based models as good candidates in terms of feature hallucination and popularity bias, specially under sparse domains where the learnt user and item representations may become unreliable. Generation-based models stand out in terms of feature relevance and text similarity with respect to the groundtruth explanations, indicating greater user personalization in rich domains. In addition, hybrid models reach a middle ground and seem to harness the advantages of both extraction-based and generation-based models while mitigating their limitations. Nonetheless, open issues are still present in state-of-the-art text-based explainable recommendation including aspects such as content repetitiveness, over-generalization of popular attributes, and sparsity impact on models' hallucinations. Finally, user-centric evaluations are left for future work.

ACKNOWLEDGMENTS

This work was partially supported by the FairTransNLP-Language Project (MCIN/AEI/10.13039/501100011033/FEDER,UE) and by the Generalitat de Catalunya (2021 SGR 00313). Maria belongs to WAI, UB, and Associated unit to CSIC by IIIA.

REFERENCES

[1] Alejandro Ariza-Casabona, Maria Salamó, Ludovico Boratto, and Gianni Fenu. 2023. Towards Self-Explaining Sequence-Aware Recommendation. In *Proceedings of the 17th ACM Conference on Recommender Systems, RecSys 2023, Singapore, Singapore, September 18-22, 2023*. ACM, 904–911. https://doi.org/10.1145/3604915.3608847

[2] Mustafa Bilgic and Raymond Mooney. 2005. Explaining Recommendations: Satisfaction vs. Promotion. In *Proceedings of Beyond Personalization, IUI*, Vol. 5. http://www.cs.iit.edu/~ml/pdfs/bilgic-iui05-wkshp.pdf

[3] Chong Chen, Min Zhang, Yiqun Liu, and Shaoping Ma. 2018. Neural Attentional Rating Regression with Review-level Explanations. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web, WWW 2018, Lyon, France, April 23-27, 2018*. ACM, 1583–1592. https://doi.org/10.1145/3178876.3186070

[4] Hao Cheng, Shuo Wang, Wensheng Lu, Wei Zhang, Mingyang Zhou, Kezhong Lu, and Hao Liao. 2023. Explainable Recommendation with Personalized Review Retrieval and Aspect Learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*. Association for Computational Linguistics, 51–64. https://doi.org/10.18653/V1/2023.ACL-LONG.4

[5] Zhixuan Chu, Hongyan Hao, Xin Ouyang, Simeng Wang, Yan Wang, Yue Shen, Jinjie Gu, Qing Cui, Longfei Li, Siqiao Xue, James Y. Zhang, and Sheng Li. 2023. Leveraging Large Language Models for Pre-trained Recommender Systems. *CoRR* abs/2308.10837 (2023). https://doi.org/10.48550/ARXIV.2308.10837 arXiv:2308.10837

[6] Anthony M. Colas, Jun Araki, Zhengyu Zhou, Bingqing Wang, and Zhe Feng. 2023. Knowledge-Grounded Natural Language Recommendation Explanation. In *Proceedings of the 6th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP, BlackboxNLP@EMNLP 2023, Singapore, December 7, 2023*. Association for Computational Linguistics, 1–15. https://doi.org/10.18653/V1/2023.BLACKBOXNLP-1.1

[7] Li Dong, Shaohan Huang, Furu Wei, Mirella Lapata, Ming Zhou, and Ke Xu. 2017. Learning to Generate Product Reviews from Attributes. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017, Valencia, Spain, April 3-7, 2017, Volume 1: Long Papers*.

Association for Computational Linguistics, 623–632. https://doi.org/10.18653/v1/e17-1059

[8] Shijie Geng, Zuohui Fu, Yingqiang Ge, Lei Li, Gerard de Melo, and Yongfeng Zhang. 2022. Improving Personalized Explanation Generation through Visualization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*. Association for Computational Linguistics, 244–255. https://doi.org/10.18653/v1/2022.acl-long.20

[9] Shijie Geng, Zuohui Fu, Juntao Tan, Yingqiang Ge, Gerard de Melo, and Yongfeng Zhang. 2022. Path Language Modeling over Knowledge Graphs for Explainable Recommendation. In *Proceedings of the ACM Web Conference 2022* (Virtual Event, Lyon, France) *(WWW '22)*. Association for Computing Machinery, New York, NY, USA, 946–955. https://doi.org/10.1145/3485447.3511937

[10] Shijie Geng, Shuchang Liu, Zuohui Fu, Yingqiang Ge, and Yongfeng Zhang. 2022. Recommendation as Language Processing (RLP): A Unified Pretrain, Personalized Prompt & Predict Paradigm (P5). In *Proceedings of the 16th ACM Conference on Recommender Systems* (Seattle, WA, USA) *(RecSys '22)*. Association for Computing Machinery, New York, NY, USA, 299–315. https://doi.org/10.1145/3523227.3546767

[11] Jonathan L. Herlocker, Joseph A. Konstan, and John Riedl. 2000. Explaining collaborative filtering recommendations. In *CSCW 2000, Proceeding on the ACM 2000 Conference on Computer Supported Cooperative Work, Philadelphia, PA, USA, December 2-6, 2000*. ACM, 241–250. https://doi.org/10.1145/358916.358995

[12] Yidan Hu, Yong Liu, Chunyan Miao, Gongqi Lin, and Yuan Miao. 2023. Aspect-Guided Syntax Graph Learning for Explainable Recommendation. *IEEE Trans. Knowl. Data Eng.* 35, 8 (2023), 7768–7781. https://doi.org/10.1109/TKDE.2022.3221847

[13] Xiaowen Huang, Quan Fang, Shengsheng Qian, Jitao Sang, Yan Li, and Changsheng Xu. 2019. Explainable Interaction-driven User Modeling over Knowledge Graph for Sequential Recommendation. In *Proceedings of the 27th ACM International Conference on Multimedia, MM 2019, Nice, France, October 21-25, 2019*. ACM, 548–556. https://doi.org/10.1145/3343031.3350893

[14] Trung-Hoang Le and Hady W. Lauw. 2020. Synthesizing Aspect-Driven Recommendation Explanations from Reviews. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*. ijcai.org, 2427–2434. https://doi.org/10.24963/IJCAI.2020/336

[15] Jiacheng Li, Zhankui He, Jingbo Shang, and Julian J. McAuley. 2023. UCEpic: Unifying Aspect Planning and Lexical Constraints for Generating Explanations in Recommendation. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD 2023, Long Beach, CA, USA, August 6-10, 2023*. ACM, 1248–1257. https://doi.org/10.1145/3580305.3599535

[16] Lei Li, Yongfeng Zhang, and Li Chen. 2020. Generate Neural Template Explanations for Recommendation. In *CIKM '20: The 29th ACM International Conference on Information and Knowledge Management, Virtual Event, Ireland, October 19-23, 2020*. ACM, 755–764. https://doi.org/10.1145/3340531.3411992

[17] Lei Li, Yongfeng Zhang, and Li Chen. 2021. Personalized Transformer for Explainable Recommendation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*. Association for Computational Linguistics, 4947–4957. https://doi.org/10.18653/v1/2021.acl-long.383

[18] Lei Li, Yongfeng Zhang, and Li Chen. 2023. Prompt Distillation for Efficient LLM-based Recommendation. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management, CIKM 2023, Birmingham, United Kingdom, October 21-25, 2023*. ACM, 1348–1357. https://doi.org/10.1145/3583780.3615017

[19] Piji Li, Zihao Wang, Zhaochun Ren, Lidong Bing, and Wai Lam. 2017. Neural Rating Regression with Abstractive Tips Generation for Recommendation. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, Shinjuku, Tokyo, Japan, August 7-11, 2017*. ACM, 345–354. https://doi.org/10.1145/3077136.3080822

[20] Richard Liaw, Eric Liang, Robert Nishihara, Philipp Moritz, Joseph E. Gonzalez, and Ion Stoica. 2018. Tune: A Research Platform for Distributed Model Selection and Training. *CoRR* abs/1807.05118 (2018). arXiv:1807.05118 http://arxiv.org/abs/1807.05118

[21] Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*. Association for Computational Linguistics, Barcelona, Spain, 74–81. https://aclanthology.org/W04-1013

[22] Julian J. McAuley, Jure Leskovec, and Dan Jurafsky. 2012. Learning Attitudes and Attributes from Multi-aspect Reviews. In *12th IEEE International Conference on Data Mining, ICDM 2012, Brussels, Belgium, December 10-13, 2012*. IEEE Computer Society, 1020–1025. https://doi.org/10.1109/ICDM.2012.110

[23] Jianmo Ni, Jiacheng Li, and Julian McAuley. 2019. Justifying Recommendations using Distantly-Labeled Reviews and Fine-Grained Aspects. 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, 188–197. https://doi.org/10.18653/v1/D19-1018

[24] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA*. ACL, 311–318. https://doi.org/10.3115/1073083.1073135

[25] Reinald Adrian Pugoy and Hung-Yu Kao. 2021. Unsupervised Extractive Summarization-Based Representations for Accurate and Explainable Collaborative Filtering. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*. Association for Computational Linguistics, 2981–2990. https://doi.org/10.18653/V1/2021.ACL-LONG.232

[26] Yifu Qiu, Yftah Ziser, Anna Korhonen, Edoardo Maria Ponti, and Shay B. Cohen. 2023. Detecting and Mitigating Hallucinations in Multilingual Summarisation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*. Association for Computational Linguistics, 8914–8932. https://doi.org/10.18653/V1/2023.EMNLP-MAIN.551

[27] Yuanpeng Qu and Hajime Nobuhara. 2024. Generating Explanations for Explainable Recommendations Using Filter-Enhanced Time-Series Information. *IEEE Access* 12 (2024), 78480–78495. https://doi.org/10.1109/ACCESS.2024.3408252

[28] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *J. Mach. Learn. Res.* 21 (2020), 140:1–140:67. http://jmlr.org/papers/v21/20-074.html

[29] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. 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 2019, Hong Kong, China, November 3-7, 2019*. Association for Computational Linguistics, 3980–3990. https://doi.org/10.18653/V1/D19-1410

[30] Mobashir Sadat, Zhengyu Zhou, Lukas Lange, Jun Araki, Arsalan Gundroo, Bingqing Wang, Rakesh R. Menon, Md. Rizwan Parvez, and Zhe Feng. 2023. DelucionQA: Detecting Hallucinations in Domain-specific Question Answering. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*. Association for Computational Linguistics, 822–835. https://doi.org/10.18653/V1/2023.FINDINGS-EMNLP.59

[31] Jie Shuai, Le Wu, Kun Zhang, Peijie Sun, Richang Hong, and Meng Wang. 2023. Topic-enhanced Graph Neural Networks for Extraction-based Explainable Recommendation. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval* (<conf-loc>, <city>Taipei</city>, <country>Taiwan</country>, </conf-loc>) *(SIGIR '23)*. Association for Computing Machinery, New York, NY, USA, 1188–1197. https://doi.org/10.1145/3539618.3591776

[32] Rashmi R. Sinha and Kirsten Swearingen. 2002. The role of transparency in recommender systems. In *Extended abstracts of the 2002 Conference on Human Factors in Computing Systems, CHI 2002, Minneapolis, Minnesota, USA, April 20-25, 2002*. ACM, 830–831. https://doi.org/10.1145/506443.506619

[33] Ningning Sun, Yue Kou, Xiangmin Zhou, Derong Shen, Dong Li, and Tiezheng Nie. 2023. Learning Implicit Sentiment for Explainable Review-Based Recommendation. In *Databases Theory and Applications - 34th Australasian Database Conference, ADC 2023, Melbourne, VIC, Australia, November 1-3, 2023, Proceedings (Lecture Notes in Computer Science, Vol. 14386)*. Springer, 59–72. https://doi.org/10.1007/978-3-031-47843-7_5

[34] Hongning Wang, Yue Lu, and Chengxiang Zhai. 2010. Latent aspect rating analysis on review text data: a rating regression approach. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, July 25-28, 2010*. ACM, 783–792. https://doi.org/10.1145/1835804.1835903

[35] Peng Wang, Renqin Cai, and Hongning Wang. 2022. Graph-based Extractive Explainer for Recommendations. In *WWW '22: The ACM Web Conference 2022, Virtual Event, Lyon, France, April 25 - 29, 2022*. ACM, 2163–2171. https://doi.org/10.1145/3485447.3512168

[36] Ruoxi Wang, Bin Fu, Gang Fu, and Mingliang Wang. 2017. Deep & Cross Network for Ad Click Predictions. In *Proceedings of the ADKDD'17* (Halifax, NS, Canada) *(ADKDD'17)*. Association for Computing Machinery, New York, NY, USA, Article 12, 7 pages. https://doi.org/10.1145/3124749.3124754

[37] Sean Welleck, Ilia Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, and Jason Weston. 2020. Neural Text Generation With Unlikelihood Training. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net. https://openreview.net/forum?id=SJeYe0NtvH

[38] Huiqiong Wu, Guibing Guo, Enneng Yang, Yudong Luo, Yabo Chu, Linying Jiang, and Xingwei Wang. 2024. PESI: Personalized Explanation recommendation with Sentiment Inconsistency between ratings and reviews. *Knowl. Based Syst.* 283 (2024), 111133. https://doi.org/10.1016/J.KNOSYS.2023.111133

[39] Yikun Xian, Zuohui Fu, S. Muthukrishnan, Gerard de Melo, and Yongfeng Zhang. 2019. Reinforcement Knowledge Graph Reasoning for Explainable Recommendation. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2019, Paris, France, July 21-25, 2019*. ACM, 285–294. https://doi.org/10.1145/3331184.3331203

[40] Zhouhang Xie, Sameer Singh, Julian J. McAuley, and Bodhisattwa Prasad Majumder. 2023. Factual and Informative Review Generation for Explainable Recommendation. In *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*. AAAI Press, 13816–13824. https://doi.org/10.1609/AAAI.V37I11.26618

[41] Aobo Yang, Nan Wang, Renqin Cai, Hongbo Deng, and Hongning Wang. 2022. Comparative Explanations of Recommendations. In *WWW '22: The ACM Web Conference 2022, Virtual Event, Lyon, France, April 25 - 29, 2022*. ACM, 3113–3123. https://doi.org/10.1145/3485447.3512031

[42] Aobo Yang, Nan Wang, Hongbo Deng, and Hongning Wang. 2021. Explanation as a Defense of Recommendation. In *WSDM '21, The Fourteenth ACM International Conference on Web Search and Data Mining, Virtual Event, Israel, March 8-12, 2021*. ACM, 1029–1037. https://doi.org/10.1145/3437963.3441726

[43] Yi Yu, Kazunari Sugiyama, and Adam Jatowt. 2023. AdaReX: Cross-Domain, Adaptive, and Explainable Recommender System. In *Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region, SIGIR-AP 2023, Beijing, China, November 26-28, 2023*. ACM, 272–281. https://doi.org/10.1145/3624918.3625331

[44] Huijing Zhan, Ling Li, Shaohua Li, Weide Liu, Manas Gupta, and Alex C. Kot. 2023. Towards Explainable Recommendation Via Bert-Guided Explanation Generator. In *IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP 2023, Rhodes Island, Greece, June 4-10, 2023*. IEEE, 1–5. https://doi.org/10.1109/ICASSP49357.2023.10096389

[45] Jingsen Zhang, Xu Chen, Jiakai Tang, Weiqi Shao, Quanyu Dai, Zhenhua Dong, and Rui Zhang. 2023. Recommendation with Causality enhanced Natural Language Explanations. In *Proceedings of the ACM Web Conference 2023, WWW 2023, Austin, TX, USA, 30 April 2023 - 4 May 2023*. ACM, 876–886. https://doi.org/10.1145/3543507.3583260

[46] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net. https://openreview.net/forum?id=SkeHuCVFDr

[47] Yongfeng Zhang, Guokun Lai, Min Zhang, Yi Zhang, Yiqun Liu, and Shaoping Ma. 2014. Explicit Factor Models for Explainable Recommendation Based on Phrase-Level Sentiment Analysis. In *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval* (Gold Coast, Queensland, Australia) *(SIGIR '14)*. Association for Computing Machinery, New York, NY, USA, 83–92. https://doi.org/10.1145/2600428.2609579

[48] Yuting Zhang, Ying Sun, Fuzhen Zhuang, Yongchun Zhu, Zhulin An, and Yongjun Xu. 2024. Triple Dual Learning for Opinion-based Explainable Recommendation. *ACM Trans. Inf. Syst.* 42, 3 (2024), 70:1–70:27. https://doi.org/10.1145/3631521

[49] Yongfeng Zhang, Haochen Zhang, Min Zhang, Yiqun Liu, and Shaoping Ma. 2014. Do users rate or review?: boost phrase-level sentiment labeling with review-level sentiment classification. In *The 37th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '14, Gold Coast, QLD, Australia - July 06 - 11, 2014*, Shlomo Geva, Andrew Trotman, Peter Bruza, Charles L. A. Clarke, and Kalervo Järvelin (Eds.). ACM, 1027–1030. https://doi.org/10.1145/2600428.2609501

[50] Jihua Zhu, Yujiao He, Guoshuai Zhao, Xuxiao Bu, and Xueming Qian. 2023. Joint Reason Generation and Rating Prediction for Explainable Recommendation. *IEEE Trans. Knowl. Data Eng.* 35, 5 (2023), 4940–4953. https://doi.org/10.1109/TKDE.2022.3146178