

Received October 17, 2019, accepted November 11, 2019, date of publication November 21, 2019, date of current version December 5, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2954957

# Contextual Hybrid Session-based News Recommendation with Recurrent Neural Networks

GABRIEL DE SOUZA P. MOREIRA<sup>1,2</sup>, DIETMAR JANNACH<sup>3</sup>, ADILSON MARQUES DA CUNHA<sup>1</sup>,

<sup>1</sup>Department of Electrical Engineering and Computing, Instituto Tecnológico de Aeronáutica (ITA) - São José dos Campos, São Paulo, Brazil

<sup>2</sup>CI&T - Campinas, São Paulo, Brazil

<sup>3</sup>Department of Applied Informatics, University of Klagenfurt, Austria

Corresponding author: Gabriel de Souza P. Moreira (e-mail: gspmoreira@gmail.com).

arXiv:1904.10367v2 [cs.IR] 8 Dec 2019

**ABSTRACT** Recommender systems help users deal with information overload by providing tailored item suggestions to them. The recommendation of news is often considered to be challenging, since the relevance of an article for a user can depend on a variety of factors, including the user's short-term reading interests, the reader's context, or the recency or popularity of an article. Previous work has shown that the use of Recurrent Neural Networks is promising for the next-in-session prediction task, but has certain limitations when only recorded item click sequences are used as input. In this work, we present a contextual hybrid, deep learning based approach for session-based news recommendation that is able to leverage a variety of information types. We evaluated our approach on two public datasets, using a temporal evaluation protocol that simulates the dynamics of a news portal in a realistic way. Our results confirm the benefits of considering additional types of information, including article popularity and recency, in the proposed way, resulting in significantly higher recommendation accuracy and catalog coverage than other session-based algorithms. Additional experiments show that the proposed parameterizable loss function used in our method also allows us to balance two usually conflicting quality factors, accuracy and novelty.

**INDEX TERMS** Artificial Neural Networks, Context-Aware Recommender Systems, Hybrid Recommender Systems, News Recommender Systems, Session-based Recommendation

## I. INTRODUCTION

**R**ECOMMENDER Systems (RS) are nowadays widely used on modern online services, where they help users finding relevant content. Today, the application fields of recommenders range from the suggestion of items on e-commerce sites, over music recommendations on streaming platforms, to friend recommendations on social networks, where they can generate substantial business value [1], [2].

One of the earliest application domains is the recommendation of online *news* [3]. News recommendation is sometimes considered as being particularly difficult, as it has a number of distinctive characteristics [4]. Among other challenges, news recommenders have to deal with a constant stream of news articles being published, which at the same time can become outdated very quickly. Another challenge is that the system often cannot rely on long-term user preference profiles. Typically, most users are not logged in and

their short-term reading interests must be estimated from only a few logged interactions, leading to a *session-based recommendation problem* [5]. Finally, like in certain other application domains, a news RS has to find the right balance between recommending only items with the highest assumed relevance and the diversity and novelty of the recommendations as a whole [6]–[10].

In recent years, we observed an increased interest in the problem of session-based recommendation, where the task is to recommend relevant items given an ongoing user session. Recurrent Neural Networks (RNN) represent a natural choice for sequence prediction tasks, as they can learn models from sequential data. *GRU4Rec* [11] was one of the first neural session-based recommendation techniques, and a number of other approaches were proposed in recent years that rely on deep learning architectures, as in [12], [13].

However, as shown in [14]–[16], neural approaches that

only rely on logged item interactions have certain limitations and they can, depending on the experimental setting, be outperformed by much simpler approaches based, e.g., on nearest-neighbor techniques.

One typical way of improving the quality of the recommendations in sparse-data situations is adopt a hybrid approach and consider additional information to assess the relevance of an item [17]–[19]. Previous approaches in the context of session-based recommendation for example used content [20] or context information [21] for improved recommendations. In our work, we adopt a similar approach.

Differently from existing works, however, we consider multiple types of side information in parallel and rely on a corresponding system architecture that allows us to combine different information types. Specifically, we adopt the general conceptual model for news recommendation that we initially proposed in [22], and base our implementation on the corresponding meta-architecture for news recommender systems called *CHAMELEON* [23]. This meta-architecture was designed to address specific challenges of the news domain, like the fast decay of item relevance and extreme user- and item-cold start problems.

Going far beyond the initial analyses presented in these previous papers, we investigate, in this current work, the effects of using various information sources on different quality factors for recommendations, namely accuracy, coverage, novelty, and diversity. Furthermore, we propose a novel approach that allows us to balance potential trade-offs—e.g., accuracy vs. novelty—depending on the specific needs of a given application.

The Research Questions (RQ) of this work are as follows:

- *RQ1* - How does our technical approach perform compared to existing approaches for session-based recommendation?
- *RQ2* - What is the effect of leveraging different types of information on the quality of the recommendations?
- *RQ3* - How can we balance competing quality factors in our neural-based recommender system?

We answer these questions through a series of experiments based on two public datasets from the news domain. One of these datasets is made publicly available in the context of this research. Our experiments will show that (a) considering a multitude of information sources is indeed helpful to improve the recommendations along all of the considered quality dimensions and (b) that the proposed balancing approach is effective. To ensure the repeatability of our research, we publicly share the code that was used in our experiments, which not only includes the code for the proposed approach and the baselines, but also the code for data pre-processing, parameter optimization, and evaluation.

The rest of this paper is organized as follows. Next, in Section II, we review existing works and previous technical approaches. In Section III, we summarize the *CHAMELEON* meta-architecture and present details of our proposed method. In Section IV, the experimental design is described and in Section V we present and discuss our results.

The paper ends with a summary and outlook on future works in Section VI.

## II. BACKGROUND AND RELATED WORK

In this section, we will first review challenges of news recommendation in more detail and summarize the conceptual model for news recommendation presented in [22]. We will then discuss previous approaches of applying deep learning for certain recommendation tasks. Finally, we will briefly survey existing works on different quality factors for recommender systems.

### A. NEWS RECOMMENDER SYSTEMS

The problem of filtering and recommending news items has been investigated for more than 20 years now, see [24] for an early work in this area. Technically, a variety of approaches have been put forward in these years, from collaborative filtering approaches [25], [26], to content-based methods [27]–[32], or hybrid systems [27], [33]–[39], see also [3] and [40] for recent surveys.

#### 1) Challenges of News Recommendation

The main goal of personalized news recommendation is to help readers finding interesting stories that maximally match their reading interests [36]. The news domain has, however, a number of characteristics that makes the recommendation task particularly difficult, among them the following [3], [40]:

- *Extreme user cold-start* - On many news sites, the users are anonymous or not logged in. News portals have often very little or no information about an individual user's past behavior [26], [27], [36];
- *Accelerated decay of item relevance* - The relevance of an article can decrease very quickly after publication and can also be immediately outdated when new information about an ongoing development is available. Considering the recency of items is therefore very important to achieve high recommendation quality, as each item is expected to have a short shelf life [25], [40];
- *Fast growing number of items* - Hundreds of new stories are added daily in news portals [41]. This intensifies the item cold-start problem. However, fresh items have to be considered for recommendation, even if not too many interactions are recorded for them [26]. Scalability problems may arise as well, in particular for news aggregators, due to the high volume of new articles being published [3], [31], [40];
- *Users preferences shift* - The preferences of individual users are often not as stable as in other domains like entertainment [26]. Moreover, short-term interests of users can also be highly determined by their contextual situation [26], [42]–[44] or by exceptional situations like breaking news [39].

The technical approach chosen in our work takes many of these challenges into account. In particular, it supports the

consideration of short-term interests through the utilization of a neural session-based recommendation technique based on RNNs. Furthermore, factors like article recency [3], [45], [46] and general popularity [19] are taken into account along with the users' context. Finally, our next-article prediction approach supports online learning in a streaming scenario [47], and is able, due to its hybrid nature, to recommend items that were not seen in training data.

## 2) Factors Influencing the Relevance of News Items

Fig. 1 shows the conceptual background of our proposed solution. In this model, a number of factors can influence the relevance of a news article for an individual user, including article-related ones, user-related ones, and what we call global factors.

With respect to *article-related* factors, we distinguish between static and dynamic properties. Static properties refer to the article's content (text), its title, topic, mentioned entities (e.g., places and people) or other metadata [27], [48]. The reputation of the publisher can also add trust to an article [49], [50]. Some news-related aspects can also dynamically change, in particular its popularity [33], [51] and recency [38], [50]. On landing pages of news portals, those two properties are typically the most important ranking criteria and in comparative evaluations, recommending recently popular items often shows to be a comparably well-performing strategy [3].

When considering *user-related* factors, we distinguish between the users' (short-term and long-term) interests and contextual factors. Regarding the context, their location [52]–[54], their device [55], and the current time [31], [53] can influence the users' short term interests, and thus the relevance of a news article [31], [48]. In addition, the *referrer* URL can contain helpful information about a user's navigation and reading context [38].

Considering the user's long-term interests can also be helpful, as some user preferences might be stable over extended periods of time [26]. Such interests may be specific personal preferences (e.g., chess playing) or influenced by popular global topics (e.g., on technology). In this work, we address only short-term user preferences, since we focus on scenarios where most users are anonymous. In general, however, as shown in [56], it is possible to merge long-term and short-term interests by combining different RNNs when modeling user preferences.

Finally, there are global factors that can affect the general popularity of an item, and thus, its relevance for a larger user community. Such global factors include, for example, breaking news regarding natural disasters or celebrity news. Some topics are generally popular for many users (e.g., sports events like Olympic Games); and some follow some seasonality (e.g., political elections), which also influences the relevance of individual articles at a given point in time [33], [50], [51].

## B. DEEP LEARNING FOR RECOMMENDER SYSTEMS

Within the last few years, deep learning methods have begun to dominate the landscape of algorithmic research in RS, see [57] for a recent overview. In this specific instantiation of the *CHAMELEON* meta-architecture [23], we implement two major tasks using deep learning techniques: (a) learning article representations and (b) computing session-based recommendations.

### 1) Deep Feature Extraction from Textual Data for Recommendation

Traditional recommendation approaches to leverage textual either use bag-of-words or TF-IDF encodings to represent item content or meta-data descriptions [58]–[60] or they rely on topic modeling [61], [62]. A potential drawback of these approaches is that they do not take word orders and the surrounding words of a keyword into account [63].

Newer approaches therefore aim to extract more useful features directly from the text and use them for recommendation. Today's techniques in particular include words embeddings, paragraph vectors, Convolutional Neural Networks (CNNs), and RNNs [64]. Kim *et al.* [63], for example, proposed Convolutional Matrix Factorization (*ConvMF*), which combines a CNN with Probabilistic Matrix Factorization to leverage information from user reviews for rating prediction.

Similarly, Seo *et al.* [65] aim to jointly model user preferences and item properties using a CNN, using a local and global attention mechanism.

Using a quite different approach, Bansal *et al.* [64] used an RNN to learn representations from the textual content of scientific papers. Besides predicting ratings for a given article, they used multi-task learning to predict also item metadata such as genres or item tags from text.

Our work shares similarities with these previous works in that we extract features using deep learning, in our case with a CNN, based on pre-trained word embeddings. However, instead of predicting ratings, our approach learns a representation of an article's content by training a separate neural network for a side task—predicting article metadata attributes based on its text.

Differently from [64], we also do not rely on an end-to-end model to extract features and to recommend items. Instead, we rely on two different modules in order to ensure scalability, given the often huge amount of recorded user interactions and news articles published every day [3], [40]. The details of our approach will be discussed in Section III.

### 2) Deep Learning for Session-based Recommendation

RNNs are a natural choice for session-based recommendation scenarios as they are able to model sequences in datasets [66]. *GRU4Rec*, proposed by Hidasi *et al.* [11], represents one of the earliest approaches in that context. In their approach, the authors specifically use Gated Recurrent Units (GRU) to be better able to deal with longer sessions and the vanishing gradient problem of RNNs. Later on, a number

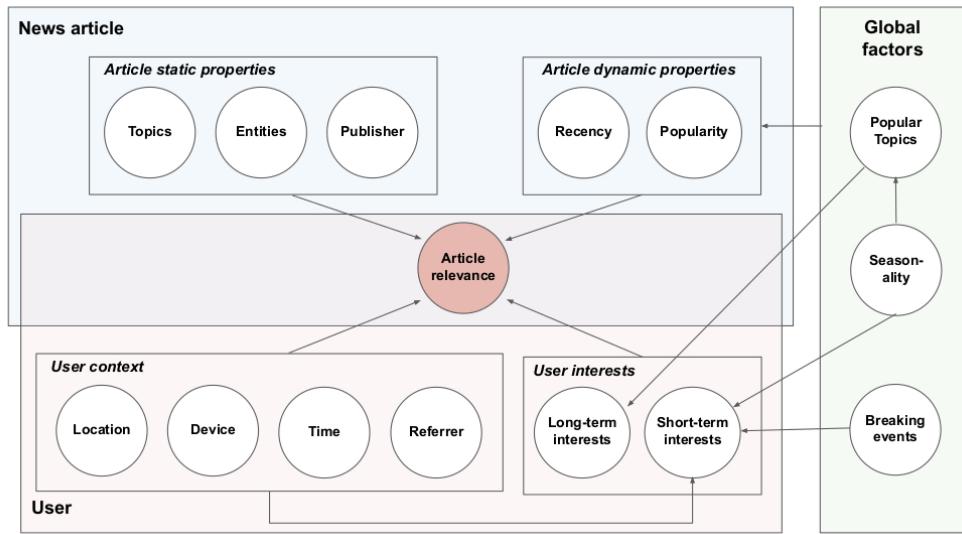


FIGURE 1: Conceptual model of news relevance factors

of improvements were published by the authors in terms of more effective loss functions [67].

One limitation of *GRU4Rec* in the news domain is that the method can only recommend items that appeared in the training set, because it is trained to predict scores for a fixed number of items. Another potential limitation is that RNN-based approaches that only use item IDs for learning with no side information might not be much better or even worse in terms of prediction accuracy than simpler approaches. Detailed analyses of this phenomenon can be found in [14], [15], [47].

A number of works, however, exist that propose RNN-based approaches that use additional side information about the user's context or the items. In [68], for example, the authors extended *GRU4Rec* to additionally use image and textual descriptions of the items. Like in our work, they did not apply an integrated end-to-end approach, but extracted image features independently by using transfer learning from a pre-trained network [69] and used simple *TF-IDF* vectors for textual representations.

Contextual information was used in combination with RNNs, for example, in [70] or [71]. In [70], the authors consider not only the sequence of events when making predictions but also the type of the event, the time gaps between events, or the time of the day of an event, leading to what they call Contextual Recurrent Neural Networks for Recommendation (*CRNN*). Similarly, Twardowski [71] considers time as a contextual factor that is combined with item information within a hybrid approach.

A work that has certain similarities with ours in terms of the recommendation approach is the Recurrent Attention DSSM (*RA-DSSM*) model by Kumar [72].

The *RA-DSSM* is an adaptation for the news domain of the *Multi-View Deep Neural Network (MV-DNN)*, which extended the Deep Structured Semantic Model (*DSSM*) [73]

information retrieval architecture to recommender systems. The (*MV-DNN*) maps users and items to a shared semantic space and recommend items that have the highest similarity with the users in the mapped space.

Technically, the authors use a bidirectional *LSTM* layer with an attention mechanism [74]. Similarly to our instantiation of the *CHAMELEON* framework, they rely on RNNs as a base building block, use embeddings to represent textual content and implement a similarity-based loss function derived from *MV-DNN*. The *CHAMELEON* meta-architecture however, as will be discussed in Section III-A, lives at a higher level of abstraction than the specific *RA-DSSM* model.

Our solution also differs from *RA-DSSM* in a number of other dimensions. *RA-DSSM* for example uses *doc2vec* embeddings [75] to represent content, while we propose a specific neural architecture to learn textual representations based on pre-trained word embeddings for improved accuracy.

Furthermore, the *RA-DSSM* does not use any contextual information about users or articles, which may limit its accuracy in cold-start scenarios that are common in news recommendation. Article recency and popularity were not considered in their model as well. Additionally, we use a temporal evaluation protocol to emulate a more realistic scenario, described in Section IV-C, while their experiments do not mimic the dynamics of a news portal.

3) Deep Reinforcement Learning for News Recommendation  
 Reinforcement learning is an alternative technical approach for recommending online news, and often multi-arm (contextual) bandit models were applied for the task [76]. In [4], the authors propose a novel deep reinforcement learning technique for news recommendation. Differently from our problem setting, the authors focus on session-aware recommendations, where longer-term information about individual

users is available. Similarly to our work, however, the approach proposed in [4] relies on a number of features that we also used in our models, e.g., article metadata, recent click counts, and context features. In their problem setting with longer-term models, the authors in addition included a number of user-related pieces of information, which are typically not available in session-based recommendation task, e.g., preferences regarding different content categories over longer periods of time.

### C. BALANCING ACCURACY AND NOVELTY IN RECOMMENDER SYSTEMS

It is known for many years that prediction accuracy is not the only factor that determines the success of a recommender. Other quality factors discussed in the literature are, e.g., novelty, catalog coverage, diversity [77], or reliability [78]. In the context of news recommendation, the aspect of novelty is particularly relevant to avoid a “rich-get-richer” phenomenon where a small set of already popular articles get further promoted through recommendations and less popular or more recent items rarely make it into a recommendation list.

The novelty of a recommended item can be defined in different ways, e.g., as the non-obviousness of the item suggestions [79], or in terms of how different an item is with respect to what has already been experienced by a user or the community [80]. Recommending solely novel or unpopular items can, however, be of limited value when they do not match the users’ interests well. Therefore, the goal of a recommender is often to balance these competing factors, i.e., make somewhat more novel and thus risky recommendations, while at the same time ensuring high accuracy.

In the literature, a number of ways have been proposed to quantify the degree of novelty, including alternative ways of considering popularity information [81] or the distance of a candidate item to the user’s profile [3], [82], [83]. In [80], the authors propose to measure novelty as the opposite of popularity of an item, under the assumption that less popular (long-tail) items are more likely to be unknown to users and their recommendation will, hopefully, lead to higher novelty levels. In our work, we will also consider the novelty of the recommendations and adopt existing novelty metrics from the literature.

Regarding the treatment of trade-off situations, different technical approaches are possible. One can, for example, try to re-rank an accuracy-optimized recommendation list, either to meet globally defined quality levels [84] or to achieve recommendation lists that match the preferences of individual users [85]. Another approach is to vary the weights of the different factors to find a configuration that leads to both high accuracy and good novelty [86].

Finally, one can try to embed the consideration of trade-offs within the learning phase, e.g., by using a corresponding regularization term. In [87], the authors propose a method called Novelty-aware Matrix Factorization (*NMF*), which tries to simultaneously recommend accurate and novel items. Their proposed regularization approach is pointwise, mean-

ing that the novelty of each candidate item is considered individually.

In our recommendation approach, we consider trade-offs in the regularization term as well. Differently, from [87], however, our approach is not focused on matrix factorization, but rather on neural models that are derived from the *DSSM*. Furthermore, the objective function in our work uses a list-wise ranking approach to learn how to enhance the novelty level of the top-n recommendations.

## III. TECHNICAL APPROACH

The work presented in this paper is based on an instantiation of the *CHAMELEON* meta-architecture, which we presented in an initial version in [23]. The meta-architecture is designed for building session-based news recommendation systems, which are context-aware and can leverage additional content information.

We will discuss this meta-architecture next in Section III-A. Afterwards, in Section III-B, we provide information about the specific instantiation used for our experiments. Finally, in Section III-C propose a novel technical approach to balance accuracy and novelty based on a parameterizable loss function.

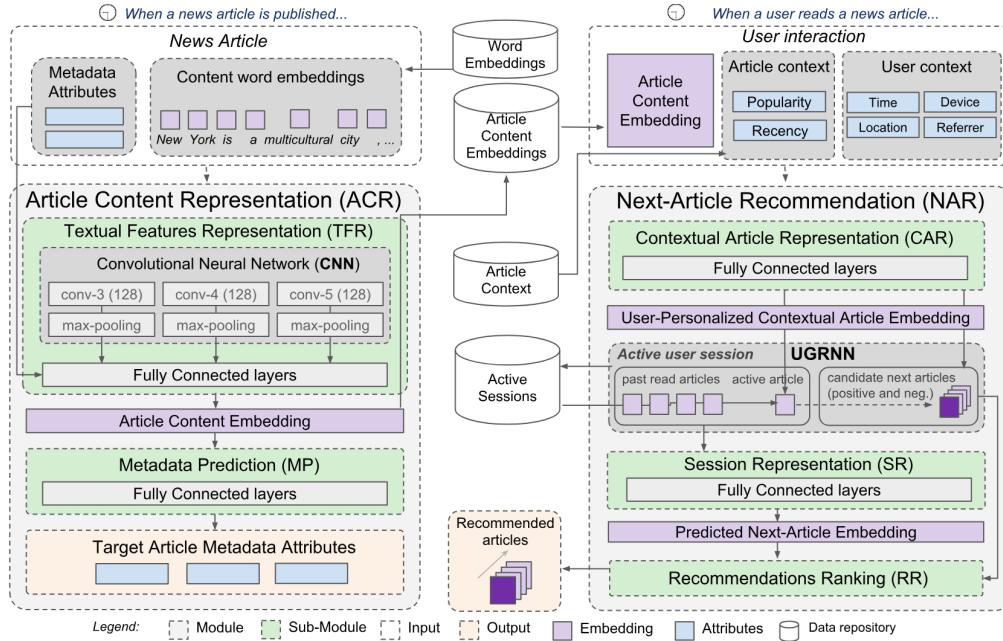
### A. THE CHAMELEON META-ARCHITECTURE

The *CHAMELEON* meta-architecture was designed to deal with some of the specific requirements of news recommendation, as outlined in Section II-A. Generally, when building a news recommender system, one has several design choices regarding the types of data that are used, the chosen algorithms, and the specific network architecture when relying on deep learning approaches. With *CHAMELEON*, we provide an architectural abstraction (a “meta-architecture”), which contains a number of general building blocks for news recommenders and which can be instantiated in various ways, depending on the particularities of the given problem setting.

Fig. 2 shows the *main building blocks* of the meta-architecture and also sketches how it was instantiated for the purpose of this research. At its core, *CHAMELEON* consists of two complementary modules, with independent life cycles for training and inference:

- The *Article Content Representation (ACR)* module used to learn a distributed representation (an embedding) of the articles’ content; and
- The *Next-Article Recommendation (NAR)* module responsible to generate next-article recommendations for ongoing user sessions.

In a *CHAMELEON*-based architecture, the *ACR* module learns an *Article Content Embedding* for each article independently from the recorded user sessions. This is done for scalability reasons, because training user interactions and articles in a joint process would be computationally very expensive, given the typically large amount of recorded user interactions. Instead, the internal model is trained for a side classification task—predicting target metadata attributes (e.g. news category, topic, tags) of an article.

FIGURE 2: An architecture instantiation of *CHAMELEON*

After training, the learned *Article Content Embeddings* are stored in a repository for further usage by the *Next Article Recommendation* module.

The *NAR* module, which provides recommendations for active sessions, is designed as a hybrid recommender system, considering both the recorded user interactions and the content of the news articles. It is also context-aware in that it leverages information about the usage context, e.g., location, device, previous clicks in the session, and the article's context — popularity and recency — which quickly decay over time. All these inputs are combined by feed-forward layers to produce what we call a *User-Personalized Contextual Article Embedding*. As a result, we obtain individualized article embeddings, whose representations depend on the user's context and other factors such as the article's current popularity and recency.

Generally, considering these additional factors can be crucial for the effectiveness of the recommendations, in particular as previous work has shown that RNNs without side information are often not much better than relatively simple algorithms [14], [15]. Additional details about the *CHAMELEON* meta-architecture can be found in [23].

## B. SPECIFIC INSTANTIATION

For the experiments conducted in this work, we used an instantiation of the *ACR* module that is similar to the one from [23]. Specifically, we extract features from textual content with a CNN. The *Article Content Embeddings* were trained to predict target article metadata attributes. In order to support multiple target attributes, a new loss function was designed to compute a weighted sum of classification losses for single-label (*softmax cross-entropy*) and multi-label at-

tributes (*sigmoid cross-entropy*), e.g., tags and keywords. The architecture of the *ACR* module and the training protocol is described in more detail in [23]. The input and output features for each dataset used in the experiments will be presented in Section IV-A.

Furthermore, the *NAR* module was instantiated with some improvements compared to [23]. Generally, the *NAR* module uses RNNs to model the sequence of user interactions. We empirically tested different RNN cells, like variations of *LSTM* [88] and *GRU* [89], whose results were very similar. At the end, we selected the *Update Gate RNN (UGRNN)* cell [90], as it led to slightly higher accuracy. The *UGRNN* architecture is a compromise between *LSTM/GRU* and a vanilla RNN. In the *UGRNN* architecture, there is only one additional gate, which determines whether the hidden state should be updated or carried over [90]. Adding a new (non bi-directional) RNN layer on top of the previous one also led to some accuracy improvement.

In a first step, the *NAR* module derives what we call a *User-Personalized Contextual Article Embedding* as described above. Specifically, in our instantiation, we consider the recent popularity of an article (e.g., by considering the clicks within the last hour) and its recency in terms of hours since its publication. As the user's context, we consider the time, location, device, and referrer type in case this information is available. The overall training phase of the *NAR* module then consists in learning a model that relates these *User-Personalized Contextual Article Embeddings* of the recommendable articles with the *Predicted Next Article Embeddings*, based on representations learned by the RNN from past session information.

Specifically, the optimization goal is to *maximize the sim-*

ilarity between the *Predicted Next-Article Embedding* and the *User-Personalized Contextual Article Embedding* corresponding to the next article actually read by the user in his or her session (positive sample), whilst minimizing its similarity with negative samples (articles not read by the user during the session)<sup>1</sup>. Using this strategy, a newly published article can be immediately recommended, as soon as its *Article Content Embedding* is added to the repository. Details regarding the optimization problem are described next.

### C. A PARAMETERIZABLE LOSS FUNCTION TO BALANCE ACCURACY AND NOVELTY

In this section, we describe the loss function of the *NAR* module, designed to optimize for accuracy (Section III-C1) and a newly proposed extension to balance accuracy and novelty (Section III-C2).

#### 1) Optimizing for Recommendation Accuracy

Formally, we can describe the method for optimizing prediction accuracy as follows. The inputs for the *NAR* module, described later in Table 3, are represented by “ $i$ ” as the article ID, “ $uc$ ” as the user context, “ $ax$ ” as the article context, and “ $ac$ ” as the article textual content. Based on those inputs, we define “ $cae = \Psi(i, ac, ax, uc)$ ” as the *User-Personalized Contextual Article Embedding*, where  $\Psi(\cdot)$  represents a sequence of fully-connected layers with non-linear activation functions to combine the inputs for the RNN.

The symbol  $s$  stands for the user session (sequence of articles previously read, represented by their *cae* vectors), and “ $nae = \Gamma(s)$ ” denotes the *Predicted Next-Article Embedding*, where  $\Gamma(\cdot)$  is the output embedding predicted by the RNN as the next article.

In (1), the function  $R$  describes the relevance of an item  $i$  for a given user session  $s$  as the similarity between the *nae* vector predicted as the next-article for the session and the *cae* vectors from the recommendable articles.

$$R(s, i) = \text{sim}(nae, cae) \quad (1)$$

In the *NAR* module instantiation presented in [23], the  $\text{sim}(\cdot)$  function was simply the cosine similarity. For this study, it was instantiated as the element-wise product of the embeddings, followed by a number of feed-forward layers. This setting allows the network to flexibly learn an arbitrary matching function:

$$\text{sim}(nae, cae) = \phi(nae \odot cae), \quad (2)$$

where  $\phi(\cdot)$  represents a sequence of fully-connected layers with non-linear activation functions, and where the last layer outputs a single scalar representing the relevance of an article as the predicted next article. In our study,  $\phi(\cdot)$  consisted of a sequence of 4 feed-forward layers with a *Leaky ReLU* activation function [93], with 128, 64, 32, and 1 output units.

<sup>1</sup>The approach is inspired by the *DSSM* [73] and by later works that applied the idea for recommender systems [72], [91], [92] and which use a ranking loss function based on the similarity of embeddings.

The ultimate task of the *NAR* module is to produce a ranked list of items (top-n recommendation) that we assume the user will read next<sup>2</sup>. Using  $i \in D$  to denote the set of all items that can be recommended, we can define a ranking-based loss function for a problem setting as follows. The goal of the learning task is to maximize the similarity between the predicted next article embedding (*nae*) for the session and the *cae* vector of the next-read article (positive sample, denoted as  $i^+$ ), while minimizing the pairwise similarity between the *nae* and the *cae* vectors of the negative samples  $i^- \in D^-$ . i.e., those that were not read by the user in this session. Since  $D$  can be large in the news domain, we approximate it through a set  $D'$ , which is the union of the unit set of the read articles (positive sample)  $\{i^+\}$  and a set with random negative samples from  $D^-$ .

As proposed in [73], we compute the posterior probability of an article being the next one given an active user session with a *softmax* function over the relevance scores:

$$P(i | s, D') = \frac{\exp(\gamma R(i, s))}{\sum_{\forall i' \in D'} \exp(\gamma R(i', s))} \quad (3)$$

where  $\gamma$  is a smoothing factor (usually referred to as *temperature*) for the softmax function, which can be trained on a held-out dataset or which can be empirically set.

Using these definitions, the model parameters  $\theta$  in the *NAR* module are estimated to maximize the accuracy of the recommendations, i.e, the likelihood of correctly predicting the next article given a user session. The corresponding loss function to be minimized, as proposed in [73]:

$$\text{accuracy\_loss}(\theta) = \frac{1}{|C|} \sum_{(s, i^+, D') \in C} -\log(P(i^+ | s, D')), \quad (4)$$

where  $C$  is the set of user clicks available for training, whose elements are triples of the form  $(s, i^+, D')$ .

Since  $\text{accuracy\_loss}(\theta)$  is differentiable w.r.t. to  $\theta$  (the model parameters to be learned), we can use back-propagation on gradient-based numerical optimization algorithms in the *NAR* module.

#### 2) Balancing Recommendations Accuracy and Novelty

In order to incorporate the aspect of novelty of the recommendations directly in the learning process, we propose to include a novelty regularization term in the loss function of the *NAR* module. This regularization term has a hyper-parameter which can be tuned to achieve a balance between novelty and accuracy, according to the desired effect for the given application. Note that this approach is not limited to particular instantiations of the *CHAMELEON* meta-architecture, but can be applied to any other neural architecture which takes the article’s recent popularity as one of the inputs and uses a *softmax* loss function for training [73].

<sup>2</sup>This corresponds to a typical next-click prediction problem.

In our approach, we adopt the novelty definition proposed in [80], [94], which is based on the inverse popularity of an item. The underlying assumption of this definition is that less popular (long-tail) items are more likely to be unknown to users and their recommendation will lead to higher novelty levels [3].

The proposed novelty component therefore aims to bias the recommendations of the neural network toward more novel items. The corresponding regularization term is based on listwise ranking, optimizing the novelty of a recommendation list in a single step. The positive items (actually clicked by the user) are not penalized based on their popularity, only the negative samples. The novelty of the negative items is weighted by their probabilities to be the next item in the sequence (computed according to (3) in order to push those items to the top of the recommendation lists that are both novel and relevant).

Formally, we define the novelty loss component as:

$$\text{nov\_loss}(\theta) = \frac{1}{|C|} \sum_{(s, i^+, D'^-) \in C} \frac{\sum_{i \in D'^-} P(i | s, D'^-) * \text{novelty}(i)}{\sum_{i \in D'^-} P(i | s, D'^-)}, \quad (5)$$

where  $C$  is the set of recorded click events for training,  $D'^-$  is a random sample of the negative samples, not including the positive sample as in the accuracy loss function (4). The novelty values of the items are weighted by their predicted relevance  $P(i | s, D'^-)$  in order to push both novel and relevant items towards the top of the recommendations list.

The novelty metric in (6) is defined based on the *recent normalized popularity* of the items. The negative logarithm in (6) increases the value of the novelty metric for long-tail items. The computation of the *normalized popularity* sums up to 1.0 for all recommendable items (set  $I$ ), as shown in (7). Since we are interested in the recent popularity, we only consider the clicks an article has received within a time frame (e.g., in the last hour), as returned by the function `recent_clicks`( $\cdot$ ):

$$\text{novelty}(i) = -\log_2(\text{rec\_norm\_pop}(i) + 1), \quad (6)$$

$$\text{rec\_norm\_pop}(i) = \frac{\text{recent\_clicks}(i)}{\sum_{j \in I} \text{recent\_clicks}(j)} \quad (7)$$

a: Complete Loss Function

The complete loss function proposed in this work combines the objectives of accuracy and novelty:

$$L(\theta) = \text{accuracy\_loss}(\theta) - \beta * \text{nov\_loss}(\theta), \quad (8)$$

where  $\beta$  is the tunable hyper-parameter for novelty. Note that the novelty loss term is subtracted from the accuracy loss, as this term is higher when more novel items are

recommended. The values for  $\beta$  can either be set based on domain expertise or be tuned to achieve the desired effects.

#### IV. EXPERIMENTAL EVALUATION

We conducted a series of experiments to answer the research questions described above. In the context of *RQ1*, our goal was to compare our method (*CHAMELEON*) with existing session-based recommenders in the news domain. For *RQ2*, we try to understand the effects of leveraging different types of information on the quality of the recommendations. Finally, *RQ3* addresses the effectiveness of our approach on balancing the accuracy and novelty trade-off.

In this section, we first discuss our experimental design, including the used datasets and the evaluation approach. The results of the evaluation will be discussed later in Section V.

#### A. DATASETS

We use two public news portals datasets for our evaluation. The datasets contain recorded user interactions and information about the published articles:

- *Globo.com (G1)* dataset - Globo.com is the most popular media company in Brazil. This dataset was originally shared by us in [23]. With this work, we publish a second version<sup>3</sup>, which also includes contextual information. The dataset was collected from the G1 news portal, which has more than 80 million unique users and publishes over 100,000 new articles per month;
- *SmartMedia Adressa dataset* - This dataset contains approximately 20 million page visits from a Norwegian news portal [95]. In our experiments we used the full dataset, which is available upon request<sup>4</sup>, and includes article text and click events of about 2 million users and 13,000 articles.

Both datasets include the textual content of the news articles, article metadata (such as publishing date, category, and author), and logged user interactions (page views) with contextual information. Since we are focusing on session-based news recommendations and short-term users preferences, it is not necessary to train algorithms for long periods. Therefore, and because articles become outdated very quickly, we have selected for the experiments all available user sessions from the first 16 days for both datasets.

In a pre-processing step, like in [15], [39], [71], we organized the data into sessions using a 30 minute threshold of inactivity as an indicator of a new session. Sessions were then sorted by timestamp of their first click. From each session, we removed repeated clicks on the same article, as we are not focusing on the capability of algorithms to act as reminders as in [96]. Sessions with only one interaction are not suitable for next-click prediction and were discarded. Sessions with more than 20 interactions (stemming from *outlier* users with an unusual behavior or from bots) were truncated.

<sup>3</sup><https://www.kaggle.com/gspmoreira/news-portal-user-interactions-by-globocom>

<sup>4</sup><http://reclab.idi.ntnu.no/dataset>

The characteristics of the resulting pre-processed datasets are shown in Table 1. Coincidentally, the datasets are similar in many statistics, except for the number of articles. For the *G1* dataset, the number of recommendable articles (clicked by at least one user) is much higher than for the *Adressa* dataset. The higher Gini index of the articles' popularity distribution also indicates that the clicks in the *Adressa* dataset are more biased to popular articles, leading to a higher inequality in clicks distribution than for the *G1* dataset.

TABLE 1: Statistics of the datasets used for the experiments.

	<i>Globo.com (G1)</i>	<i>Adressa</i>
Language	Portuguese	Norwegian
Period (days)	16	16
# users	322,897	314,661
# sessions	1,048,594	982,210
# clicks	2,988,181	2,648,999
# articles	46,033	13,820
Avg. sessions length (# clicks / # sessions)	2.84	2.70
Gini index (of the article pop. distribution)	0.952	0.969

## B. COMPARED RECOMMENDATION APPROACHES

This section describes the implementation of a specific instantiation of *CHAMELEON* and of a number of baseline techniques.

### 1) CHAMELEON—Implementation Specifics

This instantiation of the *CHAMELEON* meta-architecture, presented in Fig. 2, was implemented using TensorFlow [97], a popular Deep Learning framework. We publish the source code for our neural architecture and for the baseline methods to make our experiments reproducible<sup>5</sup>.

The *Article Content Embeddings* were trained by the *ACR* module, whose input and target features for the classifier are described in Table 2. Within the *Next Article Recommendation (NAR)* module, rich features were extracted from the user interactions logs, as detailed in Table 3. The features were prepared to be used as input for both the *ACR* and *NAR* modules as follows.

Categorical features with low cardinality (i.e., with less than 10 distinct values) were one-hot encoded and features with high cardinality were represented as trainable embeddings. Numerical features were standardized with z-normalization. The dynamic features *Novelty* and *Recency* were normalized based on a sliding window of the recent clicks (within the last hour), so that they can accommodate both repeating changes in their distributions over time, e.g., within different periods of the day, and abrupt changes in global interest, e.g., due to breaking news.

### 2) Baseline Methods

In our experiments, we consider (a) different variants of our instantiation of the *CHAMELEON* meta-architecture to

<sup>5</sup>[https://github.com/gabrielspmoreira/chameleon\\_recsys](https://github.com/gabrielspmoreira/chameleon_recsys)

TABLE 2: Features used by the *Article Content Representation (ACR)* module.

Features	Type	Description
<b>Input features</b>		
Textual Content	Emb.	Article text represented as a sequence of word embeddings, pre-trained for the language of the dataset. <sup>1</sup>
Concepts, Entities, Locations, Persons	Categ.	Lists of categorical values extracted with NLP-techniques by <i>Adressa</i> [95]. Available only for the <i>Adressa</i> dataset.
<b>Target features</b>		
Category	Categ.	The category of the article, defined by editors.
Keywords*	Categ.	Human-labeled keywords for the <i>Adressa</i> dataset.

<sup>1</sup> Portuguese: Pre-trained Word2Vec *skip-gram* model (300 dimensions) available at <http://nilc.icmc.usp.br/embeddings>; Norwegian: a *skip-gram* model (100 dimensions) available at <http://vectors.nlpl.eu/repository> (model #100).

assess the value of considering additional types of information and (b) a number of session-based recommender algorithms, described in Table 4. While some of the chosen baselines appear conceptually simple, recent work has shown that some of them are able to outperform very recent neural approaches for session-based recommendation tasks [14], [15], [47]. Furthermore, the simple methods, unlike neural-based approaches, can be continuously updated over time and take newly published articles into account.

## C. EVALUATION METHODOLOGY

One main goal of our experimental analyses is to make our evaluations as realistic as possible. We therefore did not use the common evaluation approach of random train-test splits and cross-validation. Instead, we use the temporal offline evaluation method that we proposed in [23], which simulates a streaming flow of user interactions (clicks) and new articles being published, whose value quickly decays over time. Since in practical environments it is highly important to very quickly react to incoming events [99], [100], the baseline recommender methods were constantly updated over time.

*CHAMELEON*'s *NAR* module supports online learning, as it is trained on mini-batches. In our training protocol, we decided to emulate a streaming scenario, in which each user session is used for training only once. Such a scalable approach is different from many model-based recommender systems, like *GRU4Rec* and *SR-GNN*, which require training for some epochs on a large set of recent user interactions to reach competitive accuracy results.

### 1) Evaluation Protocol

The evaluation process works as follows:

- The recommenders are continuously trained on the users' sessions ordered by time and grouped by hours. Each five hours, the recommenders are evaluated on sessions from the next hour, as exemplified in Fig. 3. With this interval of five hours (not a divisor of 24 hours),

TABLE 3: Features used by the *Next-Article Recommendation (NAR)* module

Group	Features	Type	Description
<b>Dynamic article features</b>			
Article Context	Novelty	Num.	The novelty of an article, computed based on its normalized recent popularity, as described in (6).
	Recency	Num.	Computed as the logarithm of the elapsed days (with hours represented as the decimal part) since an article was published: $\log_2((\text{current\_date} - \text{published\_date}) + 1)$ .
<b>Static article features</b>			
Id	Id	Emb.	Trainable embeddings for article IDs.
Content	ACE	Emb.	The <i>Article Content Embedding</i> representation learned by the ACR module.
Metadata	Category	Cat.	Article category
	Author *	Cat.	Article author
<b>User context features</b>			
Location	Country, Region, City*	Categ.	Estimated location of the user
Device	Device type	Categ.	Desktop, Mobile, Tablet, TV**
	OS	Categ.	Device operating system
	Platform**	Categ.	Web, mobile app
Time	Hour of the day	Num.	Hour encoded as cyclic continuous feature (using sine and cosine)
	Day of the week	Num.	Day of the week
Referrer	Referrer type	Categ.	Type of referrer: e.g., direct access, internal traffic, search engines, social platforms, news aggregators

\* Only available for the Adressa dataset.

\*\* Only available for the G1 dataset.

it was possible to sample different hours of the day across the dataset for evaluation. After the evaluation of the next hour was done, this hour is also considered for training, until the entire dataset is covered.<sup>6</sup> It is important to note that, while the most of the baseline methods were continuously updated during the evaluation hour, the neural methods—*CHAMELEON*, *SR-GNN*, and *GRU4Rec*—were not trained as evaluation progressed.<sup>7</sup> This allows us to emulate a realistic scenario in production where the neural network is trained and deployed once an hour to serve recommendations for the next hour;

- For each session in the evaluation set, we incrementally “revealed” one click after the other to the recommender, as done, e.g., in [11] and [56];
- For each click to be predicted, we created a set containing 50 randomly sampled recommendable articles *not* viewed by the user in the session (negative samples), plus the true next article (positive sample), as done in [101] and [102]. The sampling strategy was popularity-biased (i.e., the item sampling probability is proportional to its support), so that strong (popular) negative samples are always present. We then evaluate the algorithms in the task of ranking those 51 items;
- Given these rankings, standard information retrieval metrics can be computed.

For a realistic evaluation, it is important that the chosen negative samples consist of articles which would be of some interest to readers and which were also available

<sup>6</sup>Our datasets comprises 16 days. We used the first two days to learn an initial model for the session-based algorithms and report the averaged measures after that warm-up period.

<sup>7</sup>Additionally, as the original implementations of *SR-GNN* and *GRU4Rec* do not support fine tuning of previously trained models with more data, those models were trained (for some epochs) considering only sessions from the last 5 hours before each evaluation. On the other hand, *CHAMELEON*’s network was incrementally trained over time (except during evaluation).

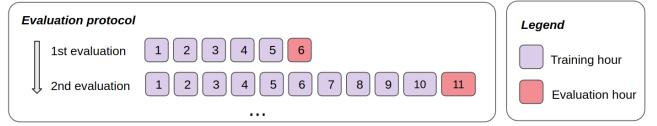


FIGURE 3: Illustration of the evaluation protocol. After training for 5 hours, we evaluate using the sessions of the next hour.

for recommendation in the news portal at a given point of time. For the purpose of this study, we therefore selected as recommendable articles the ones that received at least one click by any user in the preceding hour. To finally select the negative samples, we implemented a popularity-based sampling strategy similar to the one from [11].

## 2) Metrics

To measure quality factors such as accuracy, item coverage, novelty, and diversity, we have selected a set of top-N metrics from the literature. We chose the cut-off threshold at N=10, representing about 20% of the list containing the 51 sampled articles (1 positive sample and 50 negative samples).

The accuracy metrics used in our study were the *Hit Rate* (*HR@n*), which checks whether or not the true next item appears in the top-N ranked items, and the *Mean Reciprocal Rank* (*MRR@n*), a ranking metric that is sensitive to the position of the true next item in the list. Both metrics are common when evaluating session-based recommendation algorithms [11], [15], [47].

As an additional metric, we considered *Item Coverage* (*COV@n*), which is sometimes also called “aggregate diversity” [84]. The idea here is to measure to what extent an algorithm is able to diversify the recommendations and to make a larger fraction of the item catalog visible to the users. We compute coverage as the number of distinct articles

TABLE 4: Baseline session-based recommender algorithms used in the experiments.

Neural Methods	
<i>GRU4Rec</i>	A landmark neural architecture using RNNs for session-based recommendation [11]. For this experiment, we used the <i>GRU4Rec</i> v2 implementation, which includes the improvements reported in [67]. <sup>1</sup> We furthermore improved the algorithm's negative sampling strategy for the scenario of news recommendation. <sup>2</sup>
<i>SR-GNN</i>	A recently published state-of-the-art architecture for session-based recommendation based on Graph Neural Networks. In [98], the authors reported superior performance over other neural architectures such as <i>GRU4Rec</i> [11], <i>NARM</i> [13] and <i>STAMP</i> [12].
Association Rules-based Methods	
<i>Co-Occurrence (CO)</i>	Recommends articles commonly viewed together with the last read article in previous user sessions. This algorithm is a simplified version of the association rules technique, having two as the maximum rule size (pairwise item co-occurrences) ([15], [47]).
<i>Sequential Rules (SR)</i>	The method also uses association rules of size two. It however considers the sequence of the items within a session. A rule is created when an item $q$ appeared after an item $p$ in a session, even when other items were viewed between $p$ and $q$ . The rules are weighted by the distance $x$ (number of steps) between $p$ and $q$ in the session with a linear weighting function $w_{SR} = 1/x$ [15];
Neighborhood-based Methods	
<i>Item-kNN</i>	Returns the most similar items to the last read article using the cosine similarity between their vectors of co-occurrence with other items within sessions. This method has been commonly used as a baseline when neural approaches for session-based recommendation were proposed, e.g., in [11].
<i>Vector Multiplication Session-Based kNN (V-SkNN)</i>	This method compares the entire active session with past (neighboring) sessions to determine items to be recommended. The similarity function emphasizes items that appear later within the session. The method proved to be highly competitive in the evaluations in [14], [15], [47].
Other Methods	
<i>Recently Popular (RP)</i>	This method recommends the most viewed articles within a defined set of recently observed user interactions on the news portal (e.g., clicks during the last hour). Such a strategy proved to be very effective in the 2017 CLEF NewsREEL Challenge [99].
<i>Content-Based (CB)</i>	For each article read by the user, this method suggests recommendable articles with similar content to the last clicked article, based on the cosine similarity of their <i>Article Content Embeddings</i> .

<sup>1</sup> *GRU4Rec* v2 [67] was released on Jun 12, 2017 and is available at <https://github.com/hidasib/GRU4Rec>

<sup>2</sup> We exchanged the original negative sampling approaches used for training *GRU4Rec* by the sampling strategy described in Section IV-C1 (i.e., popularity-biased from recent clicks), and observed accuracy improvements for *GRU4Rec* in these experiments.

that appeared in any top-N list divided by the number of recommendable articles [103], i.e., those that were clicked at least once in the last hour.

To measure novelty and diversity, we adapted the evaluation metrics that were proposed in [8], [80], [94]. We provide details of their implementation in Appendix A. The novelty metrics *ESI-R@n* and *ESI-RR@n* are based on item popularity, returning higher values for long-tail items. The *ESI-R@n* (Expected Self-Information with Rank-sensitivity) metric includes a rank discount, so that items in the top positions of the recommendation list have a higher effect on the metric. The *ESI-RR@n* (Expected Self-Information with Rank- and Relevance-sensitivity) metric not only considers a rank discount, but also combines novelty with accuracy, as the relevant (clicked) item will have a higher impact on the metric if it is among the top-n recommended items. Our diversity metrics are based on the *Expected Intra-List Diversity (EILD)* metric. Analogously to the novelty metrics, there are variations to account for rank-sensitivity (*EILD-R@n*) and for both rank- and relevance-sensitivity (*EILD-RR@n*).

For our experiments, all recommender algorithms were tuned towards higher accuracy (*MRR@10*) for each dataset using random search on a hold-out validation set. The resulting best hyper-parameters are reported in Appendix B.

## V. RESULTS AND DISCUSSION

In this section, we present the main results and discuss our findings under the perspective of our research questions. For all tables presented in this section, best results for a metric are printed in bold face. If the best results are significantly different<sup>8</sup> from measures of all other algorithms, they are marked with \*\*\* when  $p < 0.001$ , with \*\* when  $p < 0.01$ , and with \* symbol when  $p < 0.05$ .

### A. EVALUATION OF RECOMMENDATION QUALITY (RQ1)

In this section, we first analyze the obtained accuracy results and then discuss the other quality factors.

#### 1) Accuracy Analysis

Table 5 shows the accuracy results obtained by the different algorithms in terms of the *HR@10* and *MRR@10* metrics. The reported values correspond to the average of the measures obtained for each evaluation hour, according to the evaluation protocol (Section IV-C).

In this comparison, our *CHAMELEON* instantiation outperforms the other baseline algorithms on both datasets and on both accuracy metrics by a large margin. The *SR* method performs second-best.

Generally, the observed difference between *CHAMELEON* and *SR* is higher for the *G1* dataset. This can be explained by

<sup>8</sup> As errors around the reported averages were normally distributed, we used paired Student's t-tests with Bonferroni correction for significance tests.

TABLE 5: Accuracy results for *G1* and *Adressa*

Algorithm	G1 dataset		Adressa dataset	
	HR@10	MRR@10	HR@10	MRR@10
<i>CHAMELEON</i>	<b>0.6738***</b>	<b>0.3458***</b>	<b>0.7018***</b>	<b>0.3421***</b>
<i>SR</i>	0.5900	0.2889	0.6288	0.3022
<i>Item-kNN</i>	0.5707	0.2801	0.6179	0.2819
<i>CO</i>	0.5689	0.2626	0.6131	0.2768
<i>V-SkNN</i>	0.5467	0.2494	0.6140	0.2723
<i>SR-GNN</i>	0.5144	0.2467	0.6122	0.2991
<i>GRU4Rec</i>	0.4669	0.2092	0.4958	0.2200
<i>RP</i>	0.4577	0.1993	0.5648	0.2481
<i>CB</i>	0.3643	0.1676	0.3307	0.1253

the facts that (a) the number of articles in the *G1* dataset is more than 3 times higher than in the other dataset and (b) the *G1* dataset has a lower popularity bias, see the *Gini index* in Table 1. As a result, algorithms that have a higher tendency to recommend popular items are less effective for datasets with a more balanced click distribution. Looking, for example, at the algorithm that simply recommends recently-popular articles (*RP*), we see that its performance is much higher for the *Adressa* dataset, even though the best obtained measures are almost similar for both datasets.

We can furthermore observe that other neural approaches (i.e., *SR-GNN* and *GRU4Rec*) were not able to provide better accuracy than non-neural baselines for session-based news recommendation. One of the reasons is that in a real-world scenario—as emulated in our evaluation protocol—those models cannot be updated as often as the baseline methods, due to challenges of asynchronous model training and frequent deployment. Furthermore, *CHAMELEON*'s architecture was designed to be able to recommend fresh articles not seen during training. *SR-GNN* and *GRU4Rec* in contrast, cannot make recommendations for items that were not encountered during training, which limits their accuracy in a realistic evaluation. In our datasets, for example, we found that about 3% (*Adressa*) to 4% (*G1*) of the item clicks in each evaluation hour were on fresh articles, i.e., on articles that were not seen in the preceding training hours.

From the two neural methods, the newer graph-based *SR-GNN* method was performing much better than *GRU4Rec* in our problem setting. However, as our detailed analysis in Section V-B will show, *SR-GNN* does not achieve the performance levels of *CHAMELEON*, even when *CHAMELEON* is not leveraging any additional side information other than the article ID (configuration *IC1* in Table 8).

In Fig. 4 and 5, we plot the obtained accuracy values (*MRR@10*) of the different algorithms along the 16 days, with an evaluation after every 5 hours. We can note that, after some training hours, *CHAMELEON* clearly recommends with higher accuracy than all other algorithms.

## 2) Analysis of Additional Quality Factors

The results obtained for the other recommendation quality factors investigated in our research—item coverage, novelty, and diversity—are shown in Table 6. The observations can be

summarized as follows:

TABLE 6: Evaluation of other quality factors for the *G1* and *Adressa* datasets

Recommender	<i>Item Coverage</i> <i>COV@10</i>	<i>Novelty</i>		<i>Diversity</i>	
		<i>ESI-R@10</i>	<i>ESI-RR@10</i>	<i>EILD-R@10</i>	<i>EILD-RR@10</i>
<b>G1 dataset</b>					
<i>CHAMELEON</i>	0.6373	6.4177	<b>0.7302***</b>	0.3620	<b>0.0419***</b>
<i>SR</i>	0.2763	5.9747	0.5747	0.3526	0.0374
<i>Item-kNN</i>	0.3913	6.5909	0.6301	0.3552	0.0361
<i>CO</i>	0.2499	5.5728	0.5126	0.3570	0.0352
<i>V-SkNN</i>	0.1355	5.1760	0.4411	0.3558	0.0339
<i>SR-GNN</i>	0.3196	5.4280	0.5093	0.3668	0.0350
<i>GRU4Rec</i>	0.6333	5.2332	0.3925	0.3662	0.0310
<i>RP</i>	0.0218	4.4904	0.3259	<b>0.3750***</b>	0.0296
<i>CB</i>	<b>0.6774</b>	<b>8.1531***</b>	0.5488	0.2789	0.0193
<b>Adressa dataset</b>					
<i>CHAMELEON</i>	0.7926	5.3410	<b>0.6083***</b>	0.2123	<b>0.0250***</b>
<i>SR</i>	0.4604	5.4443	0.5277	0.2188	0.0235
<i>CO</i>	0.4220	5.0789	0.4748	0.2138	0.0222
<i>Item-kNN</i>	0.5314	5.4675	0.5091	<b>0.2246</b>	0.0228
<i>V-SkNN</i>	0.1997	4.6018	0.4112	0.2112	0.0217
<i>SR-GNN</i>	0.5197	5.1013	0.5125	0.2214	0.0241
<i>GRU4Rec</i>	0.5143	5.0571	0.3782	0.2131	0.0184
<i>RP</i>	0.0542	4.1465	0.3486	0.2139	0.0200
<i>CB</i>	<b>0.8875***</b>	<b>7.6715***</b>	0.4104	0.0960	0.0060

- In terms of *item coverage* (*COV*), *CHAMELEON* has a much richer spectrum of articles that are included in its top-10 recommendations compared to other algorithms, suggesting a higher level of personalization. The only method with a higher coverage was the *CB* method, which however is not very accurate. This is expected for a method that is agnostic of an article's popularity.
- Looking at *novelty*, the *CB* method also recommends the least popular, and thus more novel articles, according to the *ESI-R* metric. This effect has been observed in other works such as [8], [104], which is expected as this is the only method that does not take item popularity into account in any form. *CHAMELEON* ranks third on this metric for the *G1* dataset and is comparable to the other algorithms for *Adressa*<sup>9</sup>. Looking at novelty in isolation is, however, not sufficient, which is why we include the relevance-weighted *ESI-RR* metric as well. When novelty and relevance are combined in one metric, it turns out that *CHAMELEON* leads to the best values on both datasets-
- Considering *diversity*, we can observe that most algorithms are quite similar in terms of the *EILD-R@10* metric. The *CB* method has the lowest diversity by design, as it always recommends articles with similar content. When article relevance is taken into account along with diversity with the *EILD-RR@10* metric, we again see that *CHAMELEON* is more successful than others in balancing diversity and accuracy.

<sup>9</sup>We will show later, in Section V-C, how the novelty of *CHAMELEON* can be increased based on the novelty regularization method proposed in Section III-C2.

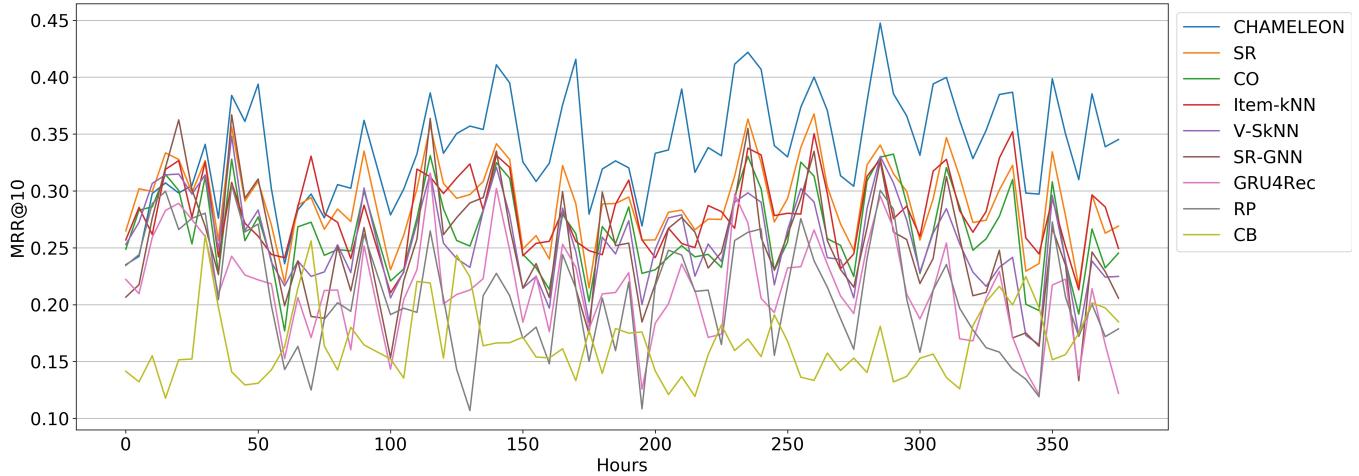


FIGURE 4: G1 (16 days) - Detailed results after every 5 hours (MRR@10)

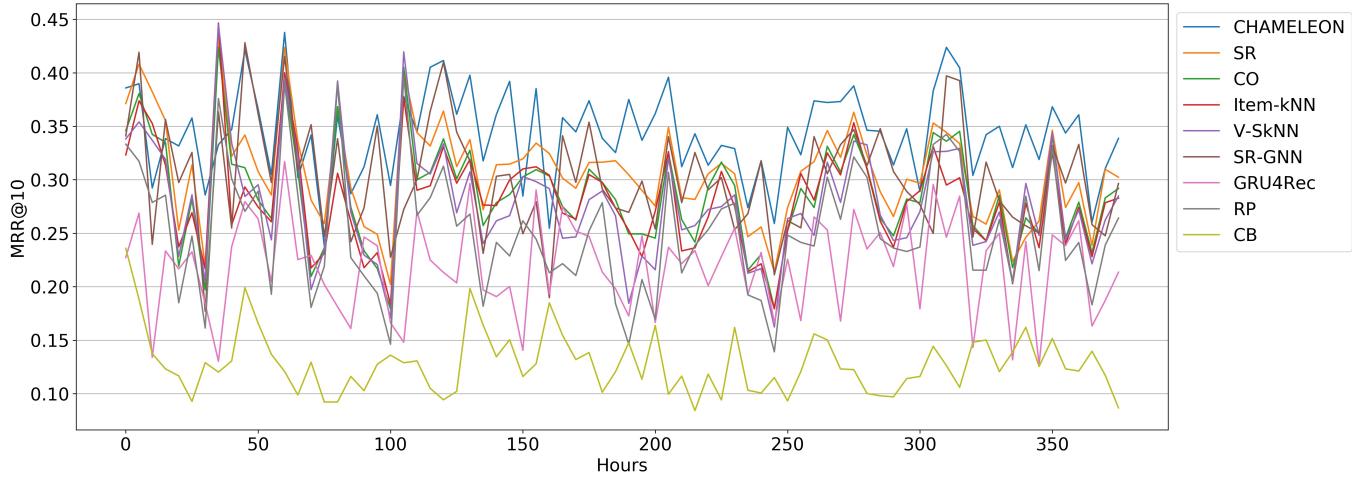


FIGURE 5: Adressa (16 days) - Detailed results after every 5 hours (MRR@10)

### B. ANALYZING THE IMPORTANCE OF INPUT FEATURES FOR THE NAR MODULE (RQ2)

*CHAMELEON* leverages a number of input features to provide more accurate recommendations, as shown in Table 3. In order to understand the effects of including those features in our model, we performed a number of additional experiments with features combined in different Input Configurations (IC)<sup>10</sup>. Table 7 shows five different configurations where we start only with the article IDs (*IC1*) and incrementally add more features until we have the model with all input features (*IC5*).

Note that we have included two variations of *IC3*: (a) using the *Article Content Embeddings (ACE)* learned with the *ACR* module, trained to predict article metadata attributes from text (supervised learning), and (b) using article embeddings trained with *doc2vec* [75] (unsupervised learning).

Table 8 shows the results of this study. We can generally see that both accuracy (*HR@10* and *MRR@10*) and item

TABLE 7: *Input Configurations (IC)* for the *NAR* module

Input config.	Feature Sets
<i>IC1</i>	Article Id
<i>IC2</i>	<i>IC1</i> + Article Context (Novelty and Recency)
<i>IC3 (ACE)</i>	<i>IC2</i> + Article Content represented as the <i>Article Content Embeddings</i> learned by the <i>ACR</i> module
<i>IC3 (doc2vec)</i>	<i>IC2</i> + Article Content represented as <i>doc2vec</i> embeddings
<i>IC4</i>	<i>IC3</i> + Article Metadata
<i>IC5</i>	<i>IC4</i> + User Context

coverage (*COV@10*) improve when more input features are considered in the *NAR* module. The largest improvements in terms of accuracy for both datasets can be observed when the feature set *Article Metadata (IC3)* are included. The feature sets of *User Context (IC5)* and *Article Context (IC2)* also played an important role when generating the recommendations.

We can also observe cases where measures become lower with the addition of new features. For both datasets, for

<sup>10</sup>This process is sometimes referred to as *ablation study*.

example, the diversity of *CHAMELEON*'s recommendations in terms of the *EILD-R* metric decreases with additional features, in particular when the *Article Content* features are included at *IC3*. This is expected, as recommendations become generally more similar when content features are used in a hybrid RS.

Looking at the two variations of configuration *IC3*, we can observe that for the *G1* dataset the textual content representation of *ACE* leads to a much higher accuracy than *doc2vec* embeddings. This confirms the usefulness of our specific way of encoding the textual content with the *ACR* module, based on word embeddings pre-trained in a larger corpus (e.g. Wikipedia).

For the *Adressa* dataset, however, the results with *ACE* and *doc2vec* are very similar<sup>11</sup>. A possible explanation for the difference between the datasets can lie in the nature of the available metadata of the articles, which are used as target attributes during training. In the *G1* dataset, for example, we have 461 article categories, which is much more than for the *Adressa* dataset, with 41 categories. Furthermore, the distribution of articles by category is more unbalanced for *Adressa* (Gini index = 0.883) than for *G1* (Gini index = 0.820). In theory, fine-grained metadata can lead to content embeddings clustered around distinctive topics, which may be useful to recommend related content.

TABLE 8: Effects of different input feature configurations on recommendation quality.

Recommender	HR@10	MRR@10	COV@10	ESI-R@10	EILD-R@10
<b>G1 dataset</b>					
<i>IC1</i>	0.5708	0.2674	0.6084	6.2597	<b>0.4515</b>
<i>IC2</i>	0.6073	0.2941	0.6095	6.1841	0.3736
<i>IC3 (doc2vec)</i>	0.6169	0.3003	0.6211	6.2115	(0.4504)
<i>IC3 (ACE)</i>	0.6472	0.3366	0.6296	6.1507	0.3625
<i>IC4</i>	0.6483	0.3397	0.6316	6.1573	0.3621
<i>IC5</i>	<b>0.6738***</b>	<b>0.3458*</b>	<b>0.6373</b>	<b>6.4177**</b>	0.3620
<b>Adressa dataset</b>					
<i>IC1</i>	0.6779	0.3260	0.7716	5.3296	<b>0.2190</b>
<i>IC2</i>	0.6799	0.3273	<b>0.8034</b>	5.2636	0.2187
<i>IC3 (doc2vec)</i>	0.6907	0.3339	0.7951	5.2856	(0.4565)
<i>IC3 (ACE)</i>	0.6906	0.3348	0.7820	5.2771	0.2103
<i>IC4</i>	0.6906	0.3362	0.7882	5.2900	0.2123
<i>IC5</i>	<b>0.7018***</b>	<b>0.3421**</b>	0.7926	<b>5.3410</b>	0.2123

### C. BALANCING ACCURACY AND NOVELTY WITH CHAMELEON (RQ3)

In this section, we analyze the effectiveness of our novel technical approach to balance accuracy and novelty within *CHAMELEON*, as described in Section III-C2. Specifically, we conducted a sensitivity analysis for the novelty regularization factor ( $\beta$ ) in the proposed loss function.

Table 9 shows the detailed outcomes of this analysis. As expected, increasing the value of  $\beta$  increases the novelty of the recommendations and also leads to higher item coverage.

<sup>11</sup>Except for the *EILD-R@10* metric, which cannot be compared because this metric uses different content embeddings (*ACE* or *doc2vec*) to compute similarities in this case.

TABLE 9: Evaluation of *CHAMELEON*'s loss regularization factor for novelty ( $\beta$ )

Reg. factors	ESI-R@10	MRR@10	COV@10
<b>G1 dataset</b>			
$\beta = 0.0$	6.4177	<b>0.3458</b>	0.6373
$\beta = 0.1$	6.9499	0.3401	0.6785
$\beta = 0.2$	7.7012	0.3222	0.6962
$\beta = 0.3$	8.5763	0.2933	0.7083
$\beta = 0.4$	9.3054	0.2507	0.7105
$\beta = 0.5$	<b>9.8012</b>	0.2170	<b>0.7123*</b>
<b>Adressa dataset</b>			
$\beta = 0.0$	5.3410	<b>0.3421</b>	0.7926
$\beta = 0.1$	5.8279	0.3350	0.8635
$\beta = 0.2$	7.5561	0.2948	0.9237
$\beta = 0.3$	9.4709	0.2082	0.9353
$\beta = 0.4$	10.2500	0.1560	<b>0.9376</b>
$\beta = 0.5$	<b>10.5184</b>	0.1348	0.9365

Correspondingly, the accuracy values decrease with higher levels of novelty. Fig. 6 shows a scatter plot that illustrates some effects and contrasts of the obtained results in our evaluation. The trade-off between accuracy (*MRR@10*) and novelty (*ESI-R@10*) for *CHAMELEON* can be clearly identified. We also plot the results for the baseline methods here for reference. This comparison reveals that tuning  $\beta$  helps us to end up with recommendations that are both more accurate and more novel than the ones by the baselines. Fig. 6 also illustrates the differences between the two datasets. Due to the uneven distribution of the *Adressa* dataset, the performance improvements over the *RP* baseline, which recommends recently popular items, are smaller than for the *G1* dataset.

## VI. SUMMARY AND FUTURE WORKS

In this final section, we first summarize the major findings of our work and then give an outlook on future research directions in this area.

### A. SUMMARY

We have proposed a novel approach for session-based news recommendation, which in particular addresses domain-specific problems such as a) the short lifetime of the recommendable items and b) the lack of longer-term preference profiles of the users. The main technical contribution of our work lies in the combination of content and context features and a sequence modeling technique based on Recurrent Neural Networks. Furthermore, we propose a novel way to balance potentially conflicting optimization goals like accuracy and novelty through a parameterizable loss function.

The individual technical components that were developed in our work were integrated into a configurable open-source news recommendation framework for session-based recommendations. Experimental evaluations on two public news datasets revealed that a) the proposed hybrid approach leads to higher prediction accuracy and b) that our approach to balance conflicting optimization goals is effective.

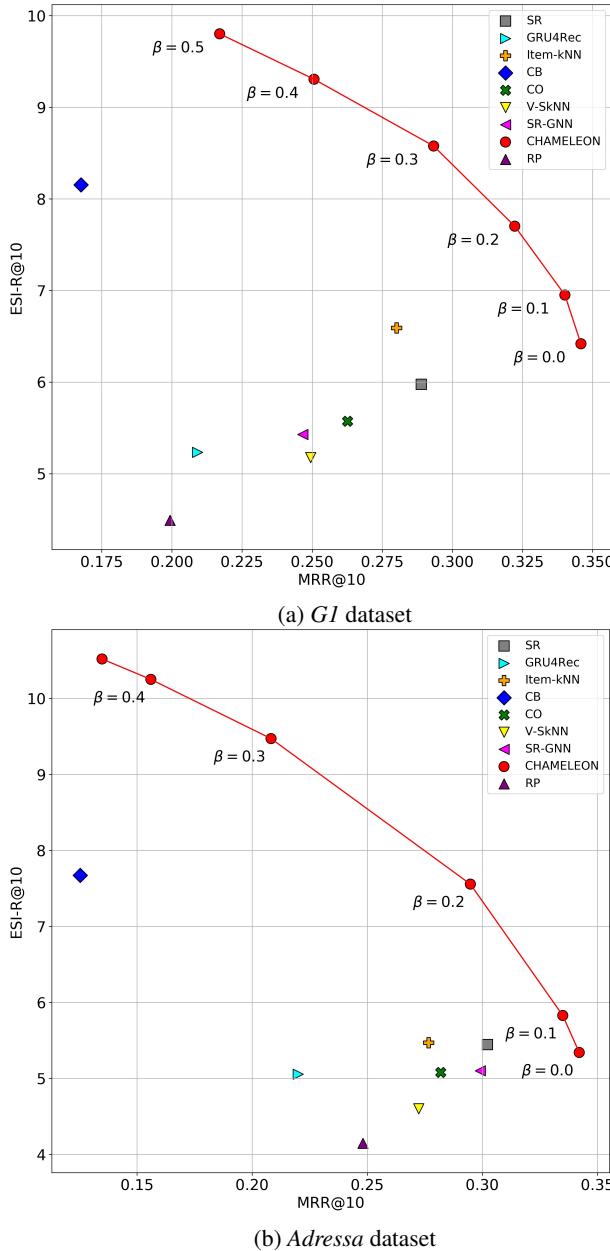


FIGURE 6: Trade-off between Accuracy (MRR@10) and Novelty (ESI-R) for different values of  $\beta$ .

## B. FUTURE WORKS

With respect to future works, our plan is to further investigate differences between existing algorithms in terms of their capability of dealing with the constant item cold-start problem, which is omnipresent in news portals.

Another specific challenge that we have not addressed so far and which was not investigated to a large extent in the literature as well is that of “outliers” in the user profiles. Specifically, there might be a certain level of noise in the user profiles. In the case of news recommendation, this could be random clicks by the user or user actions that result from a click-bait rather than from genuine user interest. As proposed

in previous works [105]–[107], we plan to identify such outliers and noise in the context of session-based recommendation to end up with a better estimate of the true user intent within a session.

Furthermore, we will investigate the role of emotions as a further contextual factor, see, e.g. [108], [109], both in the form of trying to consider the sentiment of a given news article and the current emotional state of the user.

Finally, our next immediate goals include the exploration of mechanisms within CHAMELEON that allow us to balance more than two quality factors, with a particular look at enhancing the diversity of the recommendations while preserving accuracy.

## ACKNOWLEDGMENTS

G. Moreira thanks CI&T for supporting this research in its R&D departments (D1 / Lab23), Globo.com for sharing a dataset and their technical challenges, and also Ecossistema Negocios Digitais Ltda for their support for this article.

## APPENDIX A NOVELTY AND DIVERSITY METRICS

In our studies, we use novelty and diversity metrics adapted from [80] and [94], which we tailored to fit our specific problem of session-based news recommendation. Generally, for the purpose of this investigation, novelty is evaluated in terms of *Long-Tail Novelty*. Items with high novelty correspond to long-tail items, i.e., items that were clicked on by few users, whilst low novelty items correspond to more popular items.

### A. ESI-R@N

The *Expected Self-Information with Rank-sensitivity* metric, presented in (9), was adapted from the *MSI* metric proposed by [8] with the addition of a rank discount. The term  $-\log_2 p(i)$  represents the core of this metric, which comes from the *self-information* (also known as *surprisal*) metric of Information Theory, which quantifies the amount of information conveyed by the observation of an event [8]. Applying the  $\log(\cdot)$  function emphasizes the effect of highly novel items. We define  $L = i_1, \dots, i_N$  as a recommendation list of size  $N = |L|$ .

$$\text{ESI-R}(L) = \frac{1}{\sum_{j=1}^N \text{disc}(j)} \sum_{k=1}^N -\log_2 p(i_k) \times \text{disc}(k) \quad (9)$$

In this setting, the probability  $p(i)$  of an item being part of a random user interaction under free discovery is the normalized recent popularity, i.e.,  $p(i) = \text{rec\_norm\_pop}(i)$ , previously presented in (7). In (9),  $\text{disc}(\cdot)$  is a logarithmic rank discount, defined in (10), that maximizes the impact of novelty for top ranked items, under the assumption that their characteristics will be more visible to users compared to the rest of the top-n recommendation list:

$$\text{disc}(k) = \frac{1}{\log_2(k+1)} \quad (10)$$

### B. ESI-RR@N

Analyzing quality factors like accuracy, novelty, and diversity in isolation can be misleading. Some Information Retrieval (IR) metrics, such as  $\alpha$ -nDCG, therefore consider novelty contributions only for relevant items for a given query [8]. As proposed by [80], a relevance-sensitive novelty metric should likewise assess the novelty level based on the recommended items that are actually relevant to the user.

Thus, we used a variation of a novelty metric to account for relevance—*Expected Self-Information with Rank- and Relevance-sensitivity (ESI-RR@n)*. It weights the novelty contribution by the relevance of an item for a user  $p(\text{rel}|i, u)$  [8]. We adapt the proposal from [94]:

$$p(\text{rel}|i, u) = \text{relevance}(i, u) = \begin{cases} 1.0, & \text{if } i \in I_u \\ b, & \text{otherwise} \end{cases}, \quad (11)$$

where  $I_u$  is the set of items the user interacted within the ongoing session, and  $b$  is a background probability of an unobserved interaction (negative sample) being also somewhat relevant for a user. The lower the value of  $b$  (e.g.,  $b = 0$ ) the higher the influence of relevant items (accuracy) in this metric. The author of [94] used an empirically determined value of  $b = 0.2$ , based on his experiments on balancing diversity and novelty. In our study, we arbitrarily set  $b = 0.02$ , so that all the 50 negative samples would sum up to the same relevance (1.0) of a positive (clicked) item.

Equation (12) shows how we compute the *ESI-RR@n* metric.

$$\text{ESI-RR}(L) = C_k \sum_{k=1}^N -\log_2 p(i_k) \times \text{disc}(k) \times \text{relevance}(i_k, u), \quad (12)$$

Equation (13) defines the term  $C_k$ , which computes the weighted average based on ranking discount.

$$C_k = \frac{1}{\sum_{k'=1}^N \text{disc}(k')} \quad (13)$$

Like in [94], the relevance is not normalized, so that more relevant items among the top-n recommendations lead to a global higher novelty.

### C. EILD-R@N

Diversity was measured based on the *Expected Intra-List Diversity* metric proposed by [80], with variations to account for rank-sensitivity (*EILD-R@n*) and for both rank- and relevance-sensitivity (*EILD-RR@n*).

Intra-List Diversity measures the dissimilarity of the recommended items with respect to the other items in the recommended list. In our case, the distance metric  $d(\cdot)$  defined in (14) is the cosine distance.

$$d(a, b) = (1 - \text{sim}(a, b))/2, \quad (14)$$

Here,  $a$  and  $b$  are the *Article Content Embeddings* of two articles and  $\text{sim}(a, b)$  is their cosine similarity. As the cosine similarity ranges from -1 to +1, the cosine distance is scaled to the range [0,1].

The *Expected Intra-List Diversity with Rank-sensitivity (EILD-R@n)* metric, defined in (15), is the average intra-distance between items pairs weighted by a logarithmic rank discount  $\text{disc}(\cdot)$ , defined in (10). Given a recommendation list  $L = i_1, \dots, i_N$  of size  $N = |L|$ , we compute the *EILD-R@n* metric as follows.

$$\begin{aligned} \text{EILD-R}(L) = & \frac{1}{\sum_{k'=1}^N \text{disc}(k')} \\ & \sum_{k=1}^N \text{disc}(k) \frac{1}{\sum_{l'=1: l' \neq k}^N \text{rdisc}(l', k)} \\ & \sum_{l=1: l \neq k}^N d(i_k, i_l) \times \text{rdisc}(l, k) \end{aligned} \quad (15)$$

The term  $\text{rdisc}(l, k)$ , defined in (16), represents a relative ranking discount, considering that an item  $l$  that is ranked before the target item  $k$  has already been discovered. In this case, items ranked after  $k$  are assumed to lead to a decreased diversity perception as the relative rank between  $k$  and  $l$  increases.

$$\text{rdisc}(l, k) = \text{disc}(\max(0, l - k)) \quad (16)$$

### D. EILD-RR@N

The *Expected Intra-List Diversity with Rank- and Relevance-sensitivity* finally measures the average diversity between item pairs, weighting items by rank discount and relevance, analogously to the *ESI-RR@n* metric:

$$\begin{aligned} \text{EILD-RR}(L) = & C_k \sum_{k=1}^N \text{disc}(k) \times \text{relevance}(i_k, u) C_l \\ & \sum_{l=1: l \neq k}^N d(i_k, i_l) \text{rdisc}(k, l) \times \text{relevance}(i_l, u) \end{aligned} \quad (17)$$

Here,  $C_k$  (13) and  $C_l$  (18) are normalization terms representing a weighted average based on rank discounts.

$$C_l = \frac{1}{\sum_{l'=1: l' \neq k}^N \text{rdisc}(k, l')} \quad (18)$$

## APPENDIX B FINAL ALGORITHMS

### HYPER-PARAMETERS

In Table 10, we present the best hyper-parameters found for each algorithm and dataset. They were tuned for accuracy (*MRR@10*) on a hold-out validation set, by running random search within defined ranges for each hyper-parameter. The methods *CO*, *RP*, and *CB* do not have hyper-parameters. More information about the hyper-parameters can be found in

the shared code and in the papers where the baseline methods were proposed.

## REFERENCES

- [1] C. A. Gomez-Uribe and N. Hunt, "The Netflix recommender system: Algorithms, business value, and innovation," *Transactions on Management Information Systems*, vol. 6, no. 4, pp. 13:1–13:19, 2015.
- [2] D. Lee and K. Hosanagar, "Impact of recommender systems on sales volume and diversity," in *Proceedings of the Thirty Fifth International Conference on Information Systems (ICIS'14)*, 2014.
- [3] M. Karimi, D. Jannach, and M. Jugovac, "News recommender systems—survey and roads ahead," *Information Processing & Management*, vol. 54, no. 6, pp. 1203–1227, 2018.
- [4] G. Zheng, F. Zhang, Z. Zheng, Y. Xiang, N. J. Yuan, X. Xie, and Z. Li, "Drm: A deep reinforcement learning framework for news recommendation," in *Proceedings of the 2018 World Wide Web Conference*, ser. WWW '18, 2018, pp. 167–176.
- [5] M. Quadrana, P. Cremonesi, and D. Jannach, "Sequence-aware recommender systems," *ACM Computing Surveys (CSUR)*, vol. 51, no. 4, p. 66, 2018.
- [6] I. Szpektor, Y. Maarek, and D. Pelleg, "When relevance is not enough: promoting diversity and freshness in personalized question recommendation," in *Proceedings of the 22nd International Conference on World Wide Web (WWW'13)*, 2013, pp. 1249–1260.
- [7] S. Vargas, L. Baltrunas, A. Karatzoglou, and P. Castells, "Coverage, redundancy and size-awareness in genre diversity for recommender systems," in *Proceedings of the 8th ACM Conference on Recommender Systems (RecSys'14)*, 2014, pp. 209–216.
- [8] P. Castells, N. J. Hurley, and S. Vargas, "Novelty and diversity in recommender systems," in *Recommender Systems Handbook*, F. Ricci, L. Rokach, and B. Shapira, Eds. Springer US, 2015, pp. 881–918.
- [9] L. Wu, Q. Liu, E. Chen, N. J. Yuan, G. Guo, and X. Xie, "Relevance meets coverage: A unified framework to generate diversified recommendations," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 7, no. 3, p. 39, 2016.
- [10] P. Cheng, S. Wang, J. Ma, J. Sun, and H. Xiong, "Learning to recommend accurate and diverse items," in *Proceedings of the 26th International Conference on World Wide Web (WWW'17)*, 2017, pp. 183–192.
- [11] B. Hidasi, A. Karatzoglou, L. Baltrunas, and D. Tikk, "Session-based recommendations with recurrent neural networks," in *Proceedings of Fourth International Conference on Learning Representations (ICLR'16)*, 2016.
- [12] Q. Liu, Y. Zeng, R. Mokhosi, and H. Zhang, "STAMP: short-term attention/memory priority model for session-based recommendation," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD'18*, 2018, pp. 1831–1839.
- [13] J. Li, P. Ren, Z. Chen, Z. Ren, T. Lian, and J. Ma, "Neural attentive session-based recommendation," in *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, ser. CIKM '17, 2017, pp. 1419–1428.
- [14] D. Jannach and M. Ludewig, "When recurrent neural networks meet the neighborhood for session-based recommendation," in *Proceedings of the Eleventh ACM Conference on Recommender Systems (RecSys'17)*, 2017, pp. 306–310.
- [15] M. Ludewig and D. Jannach, "Evaluation of session-based recommendation algorithms," *User-Modeling and User-Adapted Interaction*, vol. 28, no. 4–5, pp. 331–390, 2018.
- [16] M. Ludewig, N. Mauro, S. Latifi, and D. Jannach, "Performance comparison of neural and non-neural approaches to session-based recommendation," in *Proceedings of the 2019 ACM Conference on Recommender Systems (RecSys 2019)*, 2019.
- [17] T. K. Paradarami, N. D. Bastian, and J. L. Wightman, "A hybrid recommender system using artificial neural networks," *Expert Systems with Applications*, vol. 83, pp. 300–313, 2017.
- [18] D. Kim, C. Park, J. Oh, and H. Yu, "Deep hybrid recommender systems via exploiting document context and statistics of items," *Information Sciences*, vol. 417, pp. 72–87, 2017.
- [19] N. Jonnalagedda, S. Gauch, K. Labille, and S. Alfarhood, "Incorporating popularity in a personalized news recommender system," *PeerJ Computer Science*, vol. 2, p. e63, 2016.
- [20] T. X. Tuan and T. M. Phuong, "3d convolutional networks for session-based recommendation with content features," in *Proceedings of the Eleventh ACM Conference on Recommender Systems*, ser. RecSys '17, 2017, pp. 138–146.
- [21] L. Zhang, P. Liu, and J. A. Gulla, "A deep joint network for session-based news recommendations with contextual augmentation," in *Proceedings of the 29th on Hypertext and Social Media*, ser. HT '18, 2018, pp. 201–209.

TABLE 10: Best hyper-parameters per algorithm and dataset

Method	Parameter	Description	G1	Addressa
<i>CHAMELEON</i>	batch_size	Number of sessions considered for each mini-batch	256	64
	learning_rate	Learning rate for each training step (mini-batch)	1e-4	3e-4
	reg_l2	$L_2$ regularization of the network's parameters	1e-5	1e-4
	softmax_temperature	Used to control the “temperature” of the softmax function	0.1	0.2
	CAR_embedding_size	Size of the User-Personalized Contextual Article Embedding	1024	1024
	rnn_units	Number of units in an RNN layer	255	255
<i>SR</i>	rnn_num_layers	Number of stacked RNN layers	2	2
	max_clicks_dist	Maximum number of clicks to walk back in the session from the currently viewed item.	10	10
	dist_between_clicks_decay	Decay function for the distance between two items clicks within a session (e.g., linear, same, div, log, quadratic)	div	div
<i>Item-kNN</i>	reg_lambda	Smoothing factor for the popularity distribution to normalize item vectors for co-occurrence similarity	20	20
	alpha	Balance between normalizing with the support counts of the two items. 0.5 gives cosine similarity, 1.0 gives confidence.	0.75	0.50
<i>V-SkNN</i>	sessions_buffer_size	Buffer size of last processed sessions	3000	3000
	candidate_sessions_sample_size	Number of candidates near the sessions to sample	1000	2000
	nearest_neighbor_session_for_scoring	Nearest neighbors to compute item scores	500	500
	similarity	Similarity function (e.g., Jaccard, cosine)	cosine	cosine
	sampling_strategy	Strategy for sampling (e.g., recent, random)	recent	recent
	first_session_clicks_decay	Decays the weight of first user clicks in active session when finding neighbor sessions (e.g. same, div, linear, log, quadratic)	div	div
<i>SR-GNN</i>	batch_size	Batch size	128	128
	n_epochs	Number of training epochs	10	10
	hidden_size	Number of units on hidden state	200	200
	l2_lambda	Coefficient of the $L_2$ regularization	1e-5	2e-5
	propagation_steps	GNN propagation steps	1	1
	learning_rate	Learning rate	1e-3	1e-3
	learning_rate_decay	Learning rate decay factor	0.15	0.1
	learning_rate_decay_steps	number of steps after which the learning rate decays	3	3
<i>GRU4Rec</i>	nonhybrid	Enables/disables the Hybrid mode	True	True
	batch_size	Batch size	128	128
	n_epochs	Number of training epochs	3	3
	optimizer	Training optimizer	Adam	Adam
	loss	The loss type	bpr-max-0.5	bpr-max-0.5
	layers	Number of GRU units in the layers	[300]	[300]
	dropout_p_hidden	Dropout rate	0.0	0.0
	learning_rate	Learning rate	1e-4	1e-4
	l2_lambda	Coefficient of the $L_2$ regularization	1e-5	2e-5
<i>GroupLens</i>	momentum	if not zero, Nesterov momentum will be applied during training with the given strength	0	0
	embedding	Size of the embedding used, 0 means not to use embedding	0	0

- [22] G. d. S. P. Moreira, “Chameleon: A deep learning meta-architecture for news recommender systems,” in Proceedings of the Doctoral Symposium at 12th ACM Conference on Recommender Systems (RecSys’18), 2018, pp. 578–583.
- [23] G. d. S. P. Moreira, F. Ferreira, and A. M. d. Cunha, “News session-based recommendations using deep neural networks,” in Proceedings of the 3rd Workshop on Deep Learning for Recommender Systems (DLRS) at ACM RecSys’18, 2018, pp. 15–23.
- [24] J. A. Konstan, B. N. Miller, D. Maltz, J. L. Herlocker, L. R. Gordon, and J. Riedl, “GroupLens: applying collaborative filtering to usenet news,” Communications of the ACM, vol. 40, no. 3, pp. 77–87, 1997.
- [25] A. S. Das, M. Datar, A. Garg, and S. Rajaram, “Google news personalization: scalable online collaborative filtering,” in Proceedings of the 16th International Conference on World Wide Web (WWW’07), 2007, pp. 271–280.
- [26] J. Díez Peláez, D. Martínez Rego, A. Alonso Betanzos, Ó. Luaces Rodríguez, and A. Bahamonde Rionda, “Metrical representation of readers and articles in a digital newspaper,” in Proceedings of the 10th ACM Conference on Recommender Systems (RecSys 2016), 2016.
- [27] L. Li, D. Wang, T. Li, D. Knox, and B. Padmanabhan, “Scene: a scalable two-stage personalized news recommendation system,” in Proceedings of the 34th International Conference on Research and Development in Information Retrieval (SIGIR’11), 2011, pp. 125–134.
- [28] M. Capelle, F. Frasincar, M. Moerland, and F. Hogenboom, “Semantics-based news recommendation,” in Proceedings of the 2nd International Conference on Web Intelligence, Mining and Semantics, 2012, p. 27.
- [29] H. Ren and W. Feng, “Concert: A concept-centric web news recommendation system,” in Proceedings of the International Conference on Web-Age Information Management, 2013, pp. 796–798.
- [30] I. Ilievski and S. Roy, “Personalized news recommendation based on implicit feedback,” in Proceedings of the 2013 International News Recommender Systems Workshop and Challenge, 2013, pp. 10–15.
- [31] I. Mohallick and Ö. Özgöbek, “Exploring privacy concerns in news recommender systems,” in Proceedings of the International Conference on Web Intelligence (WI’17), 2017, pp. 1054–1061.
- [32] T. Bogers and A. Van den Bosch, “Comparing and evaluating information retrieval algorithms for news recommendation,” in Proceedings of the 2007 ACM conference on Recommender systems. ACM, 2007, pp. 141–144.
- [33] W. Chu and S.-T. Park, “Personalized recommendation on dynamic content using predictive bilinear models,” in Proceedings of the 18th International Conference on World Wide Web (WWW’09), 2009, pp. 691–700.
- [34] J. Liu, P. Dolan, and E. R. Pedersen, “Personalized news recommendation based on click behavior,” in Proceedings of the 15th International Conference on Intelligent User Interfaces, 2010, pp. 31–40.
- [35] J. Rao, A. Jia, Y. Feng, and D. Zhao, “Personalized news recommendation using ontologies harvested from the web,” in International Conference on Web-Age Information Management, 2013, pp. 781–787.
- [36] C. Lin, R. Xie, X. Guan, L. Li, and T. Li, “Personalized news recom-

- dation via implicit social experts," *Information Sciences*, vol. 254, pp. 1–18, 2014.
- [37] L. Li, L. Zheng, F. Yang, and T. Li, "Modeling and broadening temporal user interest in personalized news recommendation," *Expert Systems with Applications*, vol. 41, no. 7, pp. 3168–3177, 2014.
- [38] M. Trevisiol, L. M. Aiello, R. Schifanella, and A. Jaimes, "Cold-start news recommendation with domain-dependent browse graph," in *Proceedings of the 8th ACM Conference on Recommender systems (RecSys'14)*, 2014, pp. 81–88.
- [39] E. V. Epure, B. Kille, J. E. Ingvaldsen, R. Deneckere, C. Salinesi, and S. Albayrak, "Recommending personalized news in short user sessions," in *Proceedings of the Eleventh ACM Conference on Recommender Systems (RecSys'17)*, 2017, pp. 121–129.
- [40] Ö. Özgöbek, J. A. Gulla, and R. C. Erdur, "A survey on challenges and methods in news recommendation," in *Proceedings of the WEBIST'14*, 2014, pp. 278–285.
- [41] A. Spangher, "Building the next new york times recommendation engine," <https://open.blogs.nytimes.com/2015/08/11/building-the-next-new-york-times-recommendation-engine/>, Aug 2015.
- [42] P. G. Campos, F. Díez, and I. Cantador, "Time-aware recommender systems: a comprehensive survey and analysis of existing evaluation protocols," *User Modeling and User-Adapted Interaction*, vol. 24, no. 1–2, pp. 67–119, 2014.
- [43] B. Kille, F. Hopfgartner, T. Brodt, and T. Heintz, "The plista dataset," in *Proceedings of the 2013 International News Recommender Systems Workshop and Challenge at ACM RecSys'13*, 2013, pp. 16–23.
- [44] H. Ma, X. Liu, and Z. Shen, "User fatigue in online news recommendation," in *Proceedings of the 25th International Conference on World Wide Web (WWW'16)*, 2016, pp. 1363–1372.
- [45] K. F. Yeung and Y. Yang, "A proactive personalized mobile news recommendation system," in *Proceedings of the 2010 Developments in E-systems Engineering*, 2010, pp. 207–212.
- [46] M. Bieliková, M. Kompan, and D. Zelenák, "Effective hierarchical vector-based news representation for personalized recommendation," *Computer Science and Information Systems*, vol. 9, no. 1, pp. 303–322, 2012.
- [47] M. Jugovac, D. Jannach, and M. Karimi, "Streamingrec: A framework for benchmarking stream-based news recommenders," in *Proceedings of the Twelfth ACM Conference on Recommender Systems (RecSys '18)*, 2018, pp. 306–310.
- [48] A. Said, J. Lin, A. Bellogín, and A. de Vries, "A month in the life of a production news recommender system," in *Proceedings of the 2013 Workshop on Living labs for Information Retrieval Evaluation*, 2013, pp. 7–10.
- [49] A. Lommatzsch, "Real-time news recommendation using context-aware ensembles," in *Proceedings of the 36th European Conference on IR Research (ECIR'14)*, vol. 14, 2014, pp. 51–62.
- [50] J. A. Gulla, C. Marco, A. D. Fidjestøl, J. E. Ingvaldsen, and Ö. Özgöbek, "The intricacies of time in news recommendation," in *Extended Proceedings of the 24th Conference On User Modelling, Adaptation And Personalization (UMAP'16)*, 2016.
- [51] A. Lommatzsch, B. Kille, and S. Albayrak, "Incorporating context and trends in news recommender systems," in *Proceedings of the International Conference on Web Intelligence (WI'17)*, 2017, pp. 1062–1068.
- [52] B. Fortuna, C. Fortuna, and D. Mladenić, "Real-time news recommender system," in *Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 2010, pp. 583–586.
- [53] A. Montes-García, J. M. Álvarez-Rodríguez, J. E. Labra-Gayo, and M. Martínez-Merino, "Towards a journalist-based news recommendation system: The wesomender approach," *Expert Systems with Applications*, vol. 40, no. 17, pp. 6735–6741, 2013.
- [54] M. Tavakolifard, J. A. Gulla, K. C. Almeroth, J. E. Ingvaldesn, G. Nygreen, and E. Berg, "Tailored news in the palm of your hand: a multi-perspective transparent approach to news recommendation," in *Proceedings of the 22nd International Conference on World Wide Web (WWW'13)*, 2013, pp. 305–308.
- [55] H. J. Lee and S. J. Park, "Moners: A news recommender for the mobile web," *Expert Systems with Applications*, vol. 32, no. 1, pp. 143–150, 2007.
- [56] M. Quadrana, A. Karatzoglou, B. Hidasi, and P. Cremonesi, "Personalizing session-based recommendations with hierarchical recurrent neural networks," in *Proceedings of the 11th ACM Conference on Recommender Systems (RecSys'17)*, 2017.
- [57] S. Zhang, L. Yao, and A. Sun, "Deep learning based recommender system: A survey and new perspectives," *CoRR*, vol. abs/1707.07435, 2017. [Online]. Available: <http://arxiv.org/abs/1707.07435>
- [58] P. Melville, R. J. Mooney, and R. Nagarajan, "Content-boosted collaborative filtering for improved recommendations," in *Proceedings of the AAAI/IAAI'02*, 2002, pp. 187–192.
- [59] D. Agarwal and B.-C. Chen, "Regression-based latent factor models," in *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'09)*, 2009, pp. 19–28.
- [60] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Information Processing & Management*, vol. 24, no. 5, pp. 513–523, 1988.
- [61] C. Wang and D. M. Blei, "Collaborative topic modeling for recommending scientific articles," in *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'11)*, 2011, pp. 448–456.
- [62] P. K. Gopalan, L. Charlin, and D. Blei, "Content-based recommendations with poisson factorization," in *Proceedings of the Advances in Neural Information Processing Systems (NIPS'14)*, 2014, pp. 3176–3184.
- [63] D. Kim, C. Park, J. Oh, S. Lee, and H. Yu, "Convolutional matrix factorization for document context-aware recommendation," in *Proceedings of the 10th ACM Conference on Recommender Systems (RecSys'16)*, 2016, pp. 233–240.
- [64] T. Bansal, D. Belanger, and A. McCallum, "Ask the gru: Multi-task learning for deep text recommendations," in *Proceedings of the 10th ACM Conference on Recommender Systems (RecSys'16)*, 2016, pp. 107–114.
- [65] S. Seo, J. Huang, H. Yang, and Y. Liu, "Interpretable convolutional neural networks with dual local and global attention for review rating prediction," in *Proceedings of the Eleventh ACM Conference on Recommender Systems (RecSys'17)*, 2017, pp. 297–305.
- [66] T. Donkers, B. Loepf, and J. Ziegler, "Sequential user-based recurrent neural network recommendations," in *Proceedings of the Eleventh ACM Conference on Recommender Systems (RecSys'17)*, 2017, pp. 152–160.
- [67] B. Hidasi and A. Karatzoglou, "Recurrent neural networks with top-k gains for session-based recommendations," in *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*. ACM, 2018, pp. 843–852.
- [68] B. Hidasi, M. Quadrana, A. Karatzoglou, and D. Tikk, "Parallel recurrent neural network architectures for feature-rich session-based recommendations," in *Proceedings of the 10th ACM Conference on Recommender Systems (RecSys'16)*, 2016, pp. 241–248.
- [69] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'15)*, 2015, pp. 1–9.
- [70] E. Smirnova and F. Vasile, "Contextual sequence modeling for recommendation with recurrent neural networks," in *Proceedings of the 11th ACM Conference on Recommender Systems (RecSys'17)*, 2017, pp. 2–9.
- [71] B. Twardowski, "Modelling contextual information in session-aware recommender systems with neural networks," in *Proceedings of the 10th ACM Conference on Recommender Systems (RecSys'16)*, 2016, pp. 273–276.
- [72] V. Kumar, D. Khattar, S. Gupta, M. Gupta, and V. Varma, "Deep neural architecture for news recommendation," in *Working Notes of CEUR Workshop Proceedings at the 8th International Conference of the CLEF Initiative*, 2017.
- [73] P.-S. Huang, X. He, J. Gao, L. Deng, A. Acero, and L. Heck, "Learning deep structured semantic models for web search using clickthrough data," in *Proceedings of the 22nd ACM International Conference on Conference on Information & Knowledge Management*, 2013, pp. 2333–2338.
- [74] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.
- [75] Q. Le and T. Mikolov, "Distributed representations of sentences and documents," in *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, 2014, pp. 1188–1196.
- [76] L. Li, W. Chu, J. Langford, and R. E. Schapire, "A contextual-bandit approach to personalized news article recommendation," in *Proceedings of the 19th International Conference on World Wide Web (WWW 2010)*, 2010, pp. 661–670.

- [77] T. Di Noia, J. Rosati, P. Tomeo, and E. Di Sciascio, "Adaptive multi-attribute diversity for recommender systems," *Information Sciences*, vol. 382, pp. 234–253, 2017.
- [78] J. Bobadilla, A. Gutiérrez, F. Ortega, and B. Zhu, "Reliability quality measures for recommender systems," *Information Sciences*, vol. 442, pp. 145–157, 2018.
- [79] J. L. Herlocker, J. A. Konstan, L. G. Terveen, and J. T. Riedl, "Evaluating collaborative filtering recommender systems," *ACM Transactions on Information Systems (TOIS)*, vol. 22, no. 1, pp. 5–53, 2004.
- [80] S. Vargas and P. Castells, "Rank and relevance in novelty and diversity metrics for recommender systems," in *Proceedings of the fifth ACM Conference on Recommender Systems (RecSys'11)*, 2011, pp. 109–116.
- [81] T. Zhou, Z. Kuscsik, J.-G. Liu, M. Medo, J. R. Wakeling, and Y.-C. Zhang, "Solving the apparent diversity-accuracy dilemma of recommender systems," *Proceedings of the National Academy of Sciences*, vol. 107, no. 10, pp. 4511–4515, 2010.
- [82] M. Nakatsuji, Y. Fujiwara, A. Tanaka, T. Uchiyama, K. Fujimura, and T. Ishida, "Classical music for rock fans?: novel recommendations for expanding user interests," in *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, 2010, pp. 949–958.
- [83] J. Rao, A. Jia, Y. Feng, and D. Zhao, "Taxonomy based personalized news recommendation: Novelty and diversity," in *Proceedings of the International Conference on Web Information Systems Engineering*, 2013, pp. 209–218.
- [84] G. Adomavicius and Y. Kwon, "Improving aggregate recommendation diversity using ranking-based techniques," *IEEE Transactions on Knowledge and Data Engineering*, vol. 24, no. 5, pp. 896–911, 2011.
- [85] M. Jugovac, D. Jannach, and L. Lerche, "Efficient optimization of multiple recommendation quality factors according to individual user tendencies," *Expert Systems With Applications*, vol. 81, pp. 321–331, 2017.
- [86] F. Garcin, C. Dimitrakakis, and B. Faltings, "Personalized news recommendation with context trees," in *Proceedings of the 7th ACM Conference on Recommender Systems (RecSys'13)*, 2013, pp. 105–112.
- [87] L. Coba, P. Symeonidis, and M. Zanker, "Novelty-aware matrix factorization based on items' popularity," in *Proceedings of the International Conference of the Italian Association for Artificial Intelligence*, 2018, pp. 516–527.
- [88] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
- [89] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *arXiv preprint arXiv:1412.3555*, 2014.
- [90] J. Collins, J. Sohl-Dickstein, and D. Sussillo, "Capacity and trainability in recurrent neural networks," in *Proceedings of the 5th International Conference on Learning Representations (ICLR'17)*, 2017.
- [91] A. M. Elkahky, Y. Song, and X. He, "A multi-view deep learning approach for cross domain user modeling in recommendation systems," in *Proceedings of the 24th International Conference on World Wide Web (WWW'15)*, 2015, pp. 278–288.
- [92] Y. Song, A. M. Elkahky, and X. He, "Multi-rate deep learning for temporal recommendation," in *Proceedings of the 39th International Conference on Research and Development in Information Retrieval (SIGIR'16)*, 2016, pp. 909–912.
- [93] A. L. Maas, A. Y. Hannun, and A. Y. Ng, "Rectifier nonlinearities improve neural network acoustic models," in *Proceedings of the 30th International Conference on Machine Learning (ICML'13)*, vol. 30, 2013, pp. 3–9.
- [94] S. Vargas, "Novelty and diversity evaluation and enhancement in recommender systems," PhD thesis, Universidad Autónoma de Madrid, 2015.
- [95] J. A. Gulla, L. Zhang, P. Liu, Ö. Özgöbek, and X. Su, "The addressa dataset for news recommendation," in *Proceedings of the International Conference on Web Intelligence (WI'17)*, 2017, pp. 1042–1048.
- [96] L. Lerche, D. Jannach, and M. Ludewig, "On the value of reminders within e-commerce recommendations," in *Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization, (UMAP'16)*, 2016.
- [97] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin et al., "Tensorflow: Large-scale machine learning on heterogeneous distributed systems," *arXiv preprint arXiv:1603.04467*, 2016.
- [98] S. Wu, Y. Tang, Y. Zhu, L. Wang, X. Xie, and T. Tan, "Session-based recommendation with graph neural networks," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 346–353.
- [99] C. A. Ludmann, "Recommending news articles in the clef news recommendation evaluation lab with the data stream management system odysseus," in *Working notes of the Conference and Labs of the Evaluation Forum (CLEF'17)*, 2017.
- [100] B. Kille, A. Lommatzsch, F. Hopfgartner, M. Larson, and T. Brodt, "Clef 2017 newsreel overview: Offline and online evaluation of stream-based news recommender systems," *CEUR Workshop Proceedings*, 2017.
- [101] Y. Koren, R. Bell, and C. Volinsky, "Matrix factorization techniques for recommender systems," *IEEE Computer*, vol. 42, no. 8, 2009.
- [102] P. Cremonesi, Y. Koren, and R. Turrin, "Performance of recommender algorithms on top-n recommendation tasks," in *Proceedings of the fourth ACM conference on Recommender systems*, 2010, pp. 39–46.
- [103] D. Jannach, L. Lerche, I. Kamehkosh, and M. Jugovac, "What recommenders recommend: an analysis of recommendation biases and possible countermeasures," *User Modeling and User-Adapted Interaction*, vol. 25, no. 5, pp. 427–491, 2015.
- [104] Ó. Celma and P. Herrera, "A new approach to evaluating novel recommendations," in *Proceedings of the 2008 ACM Conference on Recommender Systems (RecSys'08)*, 2008, pp. 179–186.
- [105] R. Saia, L. Boratto, and S. Carta, "A semantic approach to remove incoherent items from a user profile and improve the accuracy of a recommender system," *Journal of Intelligent Information Systems*, vol. 47, pp. 111–134, 03 2016.
- [106] A. Said and A. Bellogín, "Coherence and inconsistencies in rating behavior: Estimating the magic barrier of recommender systems," *User Modeling and User-Adapted Interaction*, vol. 28, no. 2, pp. 97–125, 2018.
- [107] R. Saia, L. Boratto, and S. Carta, "Semantic coherence-based user profile modeling in the recommender systems context," in *Proceedings of the International Conference on Knowledge Discovery and Information Retrieval (KDIR 2014)*, 2014, pp. 154–161.
- [108] E. Poirson and C. D. Cunha, "A recommender approach based on customer emotions," *Expert Systems with Applications*, vol. 122, pp. 281 – 288, 2019.
- [109] J. Mizgajski and M. Morzy, "Affective recommender systems in online news industry: how emotions influence reading choices," *User Modeling and User-Adapted Interaction*, vol. 29, no. 2, pp. 345–379, 2019.

•••