Topic-Centric Explanations for News Recommendation

DAIRUI LIU, School of Computer Science, University College Dublin, Ireland
DEREK GREENE, School of Computer Science, University College Dublin, Ireland
IRENE LI, Information Technology Center, University of Tokyo, Japan
XUEFEI JIANG, School of Computer Science, University College Dublin, Ireland
RUIHAI DONG, School of Computer Science, University College Dublin, Ireland

News recommender systems (NRS) have been widely applied for online news websites to help users find relevant articles based on their interests. Recent methods have demonstrated considerable success in terms of recommendation performance. However, the lack of explanation for these recommendations can lead to mistrust among users and lack of acceptance of recommendations. To address this issue, we propose a new explainable news model to construct a topic-aware explainable recommendation approach that can both accurately identify relevant articles and explain why they have been recommended, using information from associated topics. Additionally, our model incorporates two coherence metrics applied to assess topic quality, providing a measure of the interpretability of these explanations. The results of our experiments on the MIND dataset indicate that the proposed explainable NRS outperforms several other baseline systems, while it is also capable of producing interpretable topics measured by coherence metrics. Furthermore, we present a case study through real-world examples showcasing the usefulness of our NRS for generating explanations.

CCS Concepts: • **Information systems** → **Recommender systems**; **Document topic models**.

Additional Key Words and Phrases: News Recommender Systems, Topic-Centric Explanations

ACM Reference Format:
Dairui Liu, Derek Greene, Irene Li, Xuefei Jiang, and Ruihai Dong. 2024. Topic-Centric Explanations for News Recommendation. *ACM Trans. Recomm. Syst.* 1, 1, Article 1 (January 2024), 26 pages. https://doi.org/10.1145/3680295

1 Introduction

With the development of online news services, such as Google News, millions of users can acquire news information from convenient platforms, rather than directly from traditional media sources. However, it can be difficult for users to browse all available sources in order to find relevant articles which match their specific interests. This has motivated the development of personalized *news recommender systems* (NRS), which aim to identify relevant news articles based on the personal interests of a given user. These recommendations can improve users' experience and save time when finding interesting news. This has led to considerable work [2, 34, 45, 47–49, 58] focused specifically on improving the recommendation performance of these systems. However, the provision of *explanations* for news recommendations has rarely been considered by researchers. This deficiency can lead to many problems [56]: 1) users may not trust results provided by the

Authors' Contact Information: Dairui Liu, dairui.liu@ucdconnect.ie, School of Computer Science, University College Dublin, Dublin, Ireland; Derek Greene, derek.greene@ucd.ie, School of Computer Science, University College Dublin, Dublin, Ireland; Irene Li, ireneli@ds.itc.u-tokyo.ac.jp, Information Technology Center, University of Tokyo, Tokyo, Japan; Xuefei Jiang, xuefei.jiang@ucdconnect.ie, School of Computer Science, University College Dublin, Dublin, Ireland; Ruihai Dong, ruihai.dong@ucd.ie, School of Computer Science, University College Dublin, Dublin, Ireland.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
© 2024 Copyright held by the owner/author(s).
ACM 2770-6699/2024/1-ART1
https://doi.org/10.1145/3680295

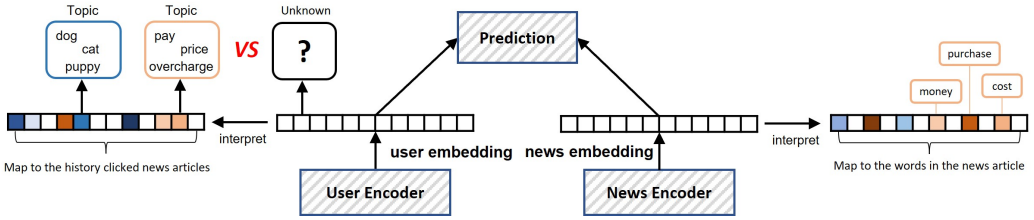


Fig. 1. A simplified illustration of a standard recommender making a prediction based on the user embedding and the candidate news embedding from the respective encoders. Typically an NRS generates only latent embeddings, which are difficult to understand. This work aims to interpret these embeddings using latent topics. We map the latent feature of the user embedding to the corresponding existing history article and identify the topic descriptors of this article based on the corresponding news embedding.

system for poor recommendations, since they are unaware of recommendation reasons; 2) the system may be less effective in persuading users to accept results; 3) this may further decrease the system's trustworthiness. Thus, providing explanation is critical in helping users to understand why corresponding news items have been presented to them by an NRS. Providing explanation can also be helpful for the system designer, allowing them to understand situations where the news recommendation process fails. The faithfulness of the generated explanations is particularly important in this context [18, 42]. In summary, an effectively explainable NRS, which is generally based on a standard NRS, should accurately recommend news and explain those recommendations simultaneously.

The workflow of a standard personalized NRS involves several key steps [50]. The first step is to recall a small set of candidate articles from a large-scale news pool when a user visits the news platform. The recommender will then rank these candidate articles according to the user's interests, as encoded in their individual profile. Subsequently, the system will display the top-k ranked articles to the user and record their subsequent behavior, which can be used as the basis for future recommendations. The NRS usually has no explicit user interest data (e.g. rating scores), so only implicit feedback (e.g. clicks) will be available to the system. During this process, the recommender is the core component of a standard NRS (see Fig. 1). This involves encoding the textual content of news and the corresponding user profile (e.g. history of clicked news) separately. News content modeling is important when attempting to build a high-performance NRS because clicked news articles usually reflect a user's interests. The predictions are generated through the collaborative contribution of user embedding and news embedding. However, the operation of the prediction process with a standard recommender is difficult to understand because the factors affecting the prediction are unclear. Thus, an intuitive approach to solve the problem is to identify the most critical factors determining whether the user will click on an article, which can lead to several benefits: 1) it can help the system designer to understand the underlying reasons behind the user's behavior, thereby allowing them to improve the recommendation process; 2) by presenting a transparent recommendation result to users, it can increase their trust and encourage them to accept the recommended news articles; 3) a more explainable system can lead to a better user experience, which helps build a trustworthy system. However, providing explanations alongside recommendations does not necessarily guarantee more accurate news recommendation results. Therefore, there is a crucial need to balance the accuracy and explainability of the recommendations. This consideration leads us to the research questions of our study:

- **RQ1**: Can the explainable news recommender system (NRS) accurately identify and recommend news articles that align with users' interests, thus outperforming other baseline systems in terms of recommendation performance?
- **RQ2**: Can the proposed method provide explanations with coherent topic representations that could help understand specific news recommendations?

It is worth noting that news articles usually contain rich textual metadata, such as title, abstract, and body content, which can be leveraged as a valuable resource for providing explanations. In this work, we not only process news articles for recommendations, but also extract topics from them to interpret generated news embedding as shown in Fig. 1 to build an explainable NRS. In Table 1, we see an example of the clicked news history from a random user from the real-world MIND dataset [51], which demonstrates the latent topics of the user's interests. The category is marked by the row color, while the topic indicators are highlighted by the font color. The selected user appears to have a broad range of interests, but is perhaps most interested in travel news. Thus, the user has a higher probability of clicking on an article from the "travel" category. However, explaining recommendations using fixed category information might lack nuance and flexibility since a high-level news category will generally consist of many sub-topics which might evolve over time. Thus, discovering topics in the news corpus and employing them as explanations is a core objective of our explainable NRS.

Table 1. An example of the history sequence of news article clicks by a given user $U91836$. We see that the user has broad interests involving topics around travel, food, finance, and health. We highlight some topic indicators, such as "*seafood*" and "*restaurant*", for the food-related topic.

News ID	*Category*	*Title*	*Body*
N51163	Lifestyle	He grew a 910-pound pumpkin and then used it as a boat.	Instead of making a giant jack lantern or a massive pie that could feed the whole town ...
N35656	Travel	Motorcyclist killed in crash on New Cut Road identified.	The name of a man killed in a motorcycle crash on New Cut Road close to Iroquois Park ...
N31402	Food and Drink	The Best Seafood Restaurant in All 50 States.	No matter where you are in the United States even in the most remote regions tasty ...
N29802	Finance	Barneys Is Sold for Scrap, Ending an Era.	For decades, Barneys New York epitomized a certain kind of aspirational Manhattan ...
N48390	Travel	First round of auctions begin for Joe Ley Antiques.	Joe Ley Antiques, which has been in business for 56 years, will be open and operating ...
N3142	Health	Teen wins science competition with liquid bandage invention.	A 14-year-old from California is America's new top young scientist. Eighth-grader Kara ...
N21773	Travel	Road built by biblical villain uncovered in Jerusalem.	Pontius Pilate is a man many Jews and Christians love to hate. For Christians, the Roman ...

News content can be accurately encoded with the help of existing methods, such as the multi-head self-attention mechanism [49] or convolutional neural networks [2]. However, these methods are not explainable. Therefore, we propose embedding an explainable news model [26] that can extract explainable topics when encoding news content, thereby potentially aiding in understanding the rationale behind specific news recommendations. Specifically, we aim to provide explanations generated in a *topic-centric* manner, which differs from general explanations of recommendations [56]. Compared to previous studies [27, 42, 55] on NRS explainability research, we generate explanations from a large news corpus with specific topic indicators, instead of simply providing a high-level category name. The key aspects of our work are as follows:

- Most existing news recommendation system studies [2, 47–49] focus on improving recommendation performance, while ignoring the interpretability of models and explanations about recommendations. Therefore, we propose to use an interpretable news encoding

model—Bi-level Attention-based Topical Model (BATM) [26] to learn an explainable news representation. Thus, we extract attention weights from multiple attention networks during the news modeling and user modeling procedure to generate topic-aware explanations. We achieve state-of-the-art performance on recommendation when compared with selected baselines, while also discovering interpretable topics from the news corpus.
- Although some researchers have presented visual explanations for recommendations to support model transparency, such work did not measure the interpretability of the model [47, 48]. In other words, some case studies were provided to show interpretability, but missing quantitative evaluation metrics were not considered by the authors. Thus, we propose to use quantitative topic coherence metrics [25, 33] to evaluate the quality of topics extracted by our model, which reflects the interpretability of these explanations. Although our primary goal is around explanation rather than topic modeling, our experimental results show that we can often generate high-quality topics and improve coherence scores by applying an entropy regularization strategy.

The remainder of this paper is organized into three parts: related work, methodology, and experimental results. Due to the lack of previous work around explainable NRS, we mainly concentrate on reviewing news recommendation methods in general in Section 2.1, with a short review of explainable methods in Section 2.2. We compare the recommendation performance of our explainable NRS with several baselines in Section 4.3. In addition to evaluating recommendation performance, we use two topic coherence metrics to assess the extracted topics and compare them with a standard topic model in Section 4.5. Finally, in Section 4.6 we present a case study showing the kinds of explanations generated by our approach for real recommendations. The source code associated with this work will be made available online[1].

2 Related Work

News recommendation, based on the personal interests of each user, is an active research area. Most researchers focus on the improvement of the performance of news recommendation systems, while few works consider recommendation explanations. In this section, we first introduce several popular approaches for news recommendation in Section 2.1. We then discuss explainable methods for the recommendation task from attention-based and topic-based perspectives in Section 2.2.

2.1 News Recommendation Methods

Recommendation methods can be categorized into three broad branches: *collaborative filtering* (CF), *content-based filtering* (CBF), and hybrid methods [1]. The CF approach makes predictions primarily based on the interaction between users and news, without knowing the news features in advance. However, this method often suffers from a severe cold start problem [24]. CBF methods can address this problem by introducing the content associated with users and news, which is our primary research focus. Since it is often easier for researchers to solve a problem in news recommendations using CBF methods, they have become particularly popular in recent years. Many researchers have demonstrated that CBF is usually more effective than a pure CF method [38], though CBF cannot handle the large number of temporary and anonymous users that are common in a real-world NRS. Deep learning (DL) models have gradually become predominant in this area because of their performance when dealing with content-based news recommendation [38]. Thus, with the success of DL-based CBF, we have looked at research works that share a similar technical paradigm [50], including news modeling, user modeling, and news ranking procedures. Multiple studies [2, 45, 47–49, 54] have proven the effectiveness of the recommendation paradigm for modeling news and

[1]https://github.com/Ruixinhua/ExplainedNRS

user representation separately. Next, we discuss successful methods for modeling news and user representation, some of which will provide baselines for our experiments.

2.1.1 News Modeling. The main goal of news modeling is to comprehend the characteristics and content of news, which is the core problem of the news recommender system. Due to the short lifespan of news items, the performance of CF methods [11] which represent news articles only by their IDs, is usually sub-optimal compared to CBF methods. Thus, we focus only on CBF news modeling methods, incorporating content features to represent news. These methods have traditionally extracted content-based features from the text of articles to construct a vector space model (VSM) representation [39]. Some strategies rely on hand-crafted features, such as concept frequency-inverse document frequency (CF-IDF) [15] and its enhancements [12]. These manually curated approaches necessitate significant human effort and domain knowledge to provide a basis for understanding news articles' semantics. Thus, these techniques are often impractical for comprehending semantic interrelations within news texts, primarily due to the incurred costs of manual annotation.

To better encode the semantic meaning of news article content, dense embedding-based representations have been proposed as an alternative to sparse VSM models [32]. Such modern models in natural language processing (NLP) can be helpful when encoding news content for recommendations. For instance, researchers have proposed an embedding-based news recommendation (EBNR) method [34], based on a variant of a de-noising autoencoder to learn representations from article texts. Other embedding-based methods, such as the deep structured semantic model (DSSM) [17], have applied a deep neural network (DNN) on existing embeddings to learn hidden news representations. However, some studies use the popular word2vec embedding method [32] for constructing dense vectors to represent the words appearing in news articles. These embedding-based models use simple DNNs to model news text, but can sometimes fail to capture contextual information accurately. Thus, some researchers have employed more complex neural networks, like a 3-D convolutional neural network (CNN) [23] or a knowledge-aware CNN (KCNN) [45], to mine deep semantic relationship. For instance, DKN [45] attempts to discover latent knowledge-level connections from news article titles by using a word-entity-aligned KCNN to learn a knowledge-enhanced news representation. This method also incorporates a knowledge graph (KG) to encode entities using KG embedding algorithms such as TransD [20]. Similarly, other methods also enhance news modeling by involving more CNNs, such as DAN [58], which adopts a combination of two parallel CNNs, built from news titles and named entities, respectively.

While these CNN-based methods can effectively learn contextual representations for news items, they are often not good at capturing and highlighting informative words because of the difference in news content informativeness. Thus, some authors have introduced attention mechanisms [3, 44] to address the problem. For example, NPA [48] uses a personalized word-level attention-based CNN to learn attentive news representations. Similar to NPA, LSTUR [2] and NAML [47] both learn a combined representation of multiple news-related metadata, including titles, categories, subcategories, and news body content through CNNs and attention networks. Here category and subcategory embeddings provide multi-view information to provide a better understanding of news content. Further studies employ advanced attention models like NRMS [49] and FedRec [37] because of the effectiveness of attention mechanisms. NRMS [49] uses a multi-head self-attention (MHSA) network to capture word-level relations and applies another additive attention network to learn informative news representations. FedRec [37] utilizes a news encoder with a combination of CNN, MHSA, and additive attention network to form a comprehensive representation of article titles.

2.1.2 User Modeling. During the recommendation task, it is also essential to understand a user's interests from their profile, which usually consists of the history of their click behavior. Thus, user modeling is typically determined by modeling interactions, such as by inferring a user's interests from the sequence of their clicks. For example, EBNR [34] considers a variant of the recurrent neural network (RNN)–gated recurrent unit (GRU) network, to generate user representations with history sequences. Similarly, RA-DSSM [22] adopts an attention-based bi-directional long short-term memory (Bi-LSTM, another variant of RNN) network to capture changing and diverse user interests. Moreover, LSTUR [2] applies a GRU network and ID embeddings for short-term and long-term user interest modeling respectively.

In contrast, other methods like NAML [47] and KRED [28] only deal with the click sequences using a news-level attention network, which does not significantly affect the performance of recommendations. Similarly, DKN [45] uses a candidate-aware attention network to form user representation by the relevance of clicked news and candidate news. And NPA [48] considers using a personalized attention network with ID embeddings involved for user representation learning. Also, DAN [58] employs the combination of attentive LSTM and candidate-aware attention network for user interest modeling. Moreover, NRMS [49] uses a more complex attention model, which is composed of a multi-head self-attention network and an additive attention network to learn contextual user representations.

2.2 Explainable Methods

Explainable recommenders aim to provide explanations for the suggestions that they produce to indicate why a particular news item is being presented to a user [43, 56]. In recent years, with the success of deep learning, some models (i.e., LSTUR) have achieved impressive performance on news recommendations. However, these models are normally considered as a black box [2, 34, 45, 47–49, 58], making their outputs difficult to understand. Therefore, our focus shifts towards explainability through attention-based explanation and topic modeling, making use of text data. Attention scores help identify key text elements in recommendations, while topic modeling assists in understanding user preferences, making recommendations more transparent.

2.2.1 Attention-based Explanation. One intuitive way to provide explanations is to look at attention scores. For example, the work by Seo et al. [40] applies a CNN to model review texts, and this method can show which part of the given review is more important for the output, based on attention scores. Similarly, some studies [8, 29, 53] also consider the attention score over review words to explain recommendations. Chen et al. [8] propose an approach to select relevant user reviews as explanations in their rating prediction model. They use an attention mechanism to analyze both user and item reviews, allowing them to identify high-quality reviews that can be used as explanations. Specifically, the attention module is applied to determine the usefulness of reviews, ensuring that only the most relevant and informative reviews are selected as explanations. Lu et al. [29] integrate matrix factorization and an attentional GRU network on user-item rating data and customer reviews. The resulting user attention network is able to provide explanations by highlighting keywords and phrases based on attention scores. Xie et al. [53] introduce an attention-based personalized recommendation framework to probe the significance of specific sentences and highlight the most influential terms in reviews according to attention scores for different users.

While the attention mechanism appears to be a straightforward method to help explainability, Bastings and Filippova [4], Jain and Wallace [19] argue that such a method may not be able to provide 'meaningful' explanations. They suggest that saliency methods, like gradient-based techniques, outperform the use of attention weights as interpretations when identifying the most significant features of the input sequence that lead to the predicted outcome. However, Wiegreffe and Pinter

[46] claim that, despite the fact that explanations provided by attention mechanisms are not always faithful, in practice, this does not invalidate the plausibility of using attention as an explanation. Bibal et al. [5] furnish a comprehensive review of the ongoing discussion regarding attention-based explainability within NLP domains, endorsing the value of attention-based explanations. They consider that, except for performance consideration, if attention can be used for explanation rather than additional saliency methods, it would be beneficial. We concur that incorporating the attention mechanism within the explanation process in recommendation tasks could be beneficial, circumventing the need for saliency methods.

2.2.2 Explanation with Topics. In addition to the methods described above, researchers have also leveraged ideas from topic modeling to help improve explainable recommendations. The explanations here take the form of relevant *topic descriptors* (i.e., ranked lists of top words associated with topics) [33]. The Latent Dirichlet Allocation (LDA) topic modeling [6] has been widely applied to investigate user preferences, which are often visualized using topic word clouds [31]. Furthermore, Luostarinen and Kohonen [30] propose the first content-based news recommender system directly using the LDA model. Other probabilistic graphical models are also studied for explainable recommendations. Wu and Ester [52] propose the Factorized Latent Aspect ModEl model (FLAME) that learns personalized preferences using item reviews. A word cloud is generated on the hotel description for hotel recommendations on the TripAdvisor corpus. Similarly, Zhao et al. [57] utilize a probabilistic graphical model that leverages sentiment, aspect, and region information for point-of-interest (POI) recommendation, making it possible to provide personalized topical-aspect explanations.

More recently, neural topic modeling [13] has been demonstrated to generate more useful topics when working with large, heavy-tailed vocabularies. Panwar et al. [35] propose the Topic Attention Networks for Neural Topic Modeling (TAN-NTM) framework. It first encodes a document using an LSTM module, and then a topic-aware module is applied to produce the outputs. A novel attention mechanism is used to learn topic-word distribution as well as the correlation of relevant words and the corresponding topic. This model shows promising results in both document classification and topic-guided keyphrase generation. To better align user preference and content information, Guo et al. [16] propose the Topic-aware Disentangled Variational AutoEncoder (TopicVAE) model. This method first extracts topic-level item representations using an attention-based module and then adopts a variational autoencoder to model topic-level disentangled user representation. Experiments show that this model outperforms other selected baselines on recommendations, as well providing interpretability on disentangled representations. In this paper, we also focus on generating explanations using a similar attention mechanism to extract topics for the news recommendation task.

2.2.3 Evaluation of Explanation with Topics. Quantitatively evaluating explanations poses a challenge due to the absence of a ground truth for comparison. Thus, most studies resort to case studies for assessing the effectiveness of the generated explanation. Notably, Gedikli et al. [14] investigate different explanation types and unveil that explanations enriched with pertinent textual tags can elevate the user-perceived transparency and augment user satisfaction. Zhao et al. [57] demonstrate the most representative words unearthed by the model to validate the effectiveness of topical-aspect explanations. Similarly, Guo et al. [16] present the most relevant topic indicators from reviews as an evaluation of the explanation of user preference. As a viable alternative to case study analysis, O'Callaghan et al. [33] and Panwar et al. [35] also use automated topic coherence evaluation metrics like $NPMI$ [7, 25] and $Word2Vec$ similarity [33] as a quantitative validation to evaluate topic quality. In addition to the standard case study approach, later in Section 4.5 we also

adopt a coherence strategy to assess the topics extracted by our model to demonstrate the quality of the generated explanation.

3 Methodology

This section describes our proposed news modeling method [26] and its application in the context of a general personalized news recommendations framework. We first formulate the recommendation problem and describe the process to generate explanations for the news recommendation scenario.

3.1 Problem Formulation

Given a user u, let the browsing history of the user (an ordered set of news article documents) be denoted by $\mathcal{H}$, and the candidate news document set by C. The history set consists of all their historical clicked news documents $\mathcal{H} = \{N_1, N_2, \ldots, N_i, \ldots, N_H\}$, where H is the maximum amount of the historical news (i.e., we keep only the last H news articles). The candidate set contains several news documents and their corresponding binary labels $C = \{N_1\text{-}l_1, N_2\text{-}l_2, \ldots, N_i\text{-}l_i, \ldots, N_C\text{-}l_C\}$, where C is the number of the candidate documents for the current impression (C may vary for different impressions) and l_i reveals whether the user u will click this news. We then calculate relevance scores $\mathcal{S} = \{s_1, s_2, \ldots, s_i, \ldots, s_C\}$ for u and recommend the top-ranked news based on these scores. In this setting, the first problem (RQ1) is how to acquire an accurate ranked list of candidate documents to match the click preference of u. The second associated problem (RQ2) is how best to explain the ranking based on the information available.

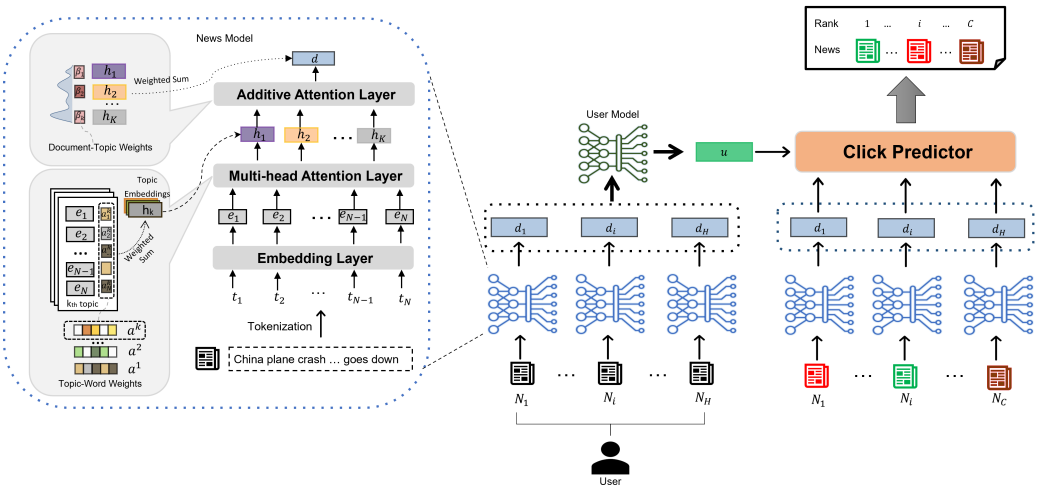


Fig. 2. The architecture of the proposed news recommendation framework, which consists of a model that encodes both news articles and users.

3.2 News Recommendations

We now introduce the proposed news recommendation workflow and its key components, as illustrated in Fig. 2. We first describe the encoder used for news modeling and then take the output of the news encoder to model the user representation. At the same time, the news encoder also

generates candidate news representations, which provide the input for the ranking module, along with the user representation. Finally, we outline the training strategy and loss function used to train the recommendation system.

3.2.1 News Modeling. This task aims to learn semantic news representations from documents using a shared news encoder module. Our proposed encoder, the Bi-level Attention-based Topical Model (BATM) [26], contains an embedding layer and two attention layers, which can generate news embeddings for recommendations and extract meaningful topics for explanations. For a given input document N_i, we first tokenize the document into a set of tokens $T_i = \{t_1, t_2, \ldots, t_n, \ldots, t_N\}$, where N is the maximum length of the tokenized set. Then we convert tokens T_i to embedding vectors $E_i = \{e_1, e_2, \ldots, e_n, \ldots, e_N\}$ in the embedding layer, where the embedding e_n represents word vector of the token t_n. We use pre-trained word embedding vectors to enhance the semantic meanings accordingly, improving recommendations' performance and topics' quality.

After the embedding layer, we pass word vectors E_i into two attention layers. The first one utilizes a multiple-topic attention mechanism to allow the model to focus on different positions in the document from different representation subspaces. We employ K attention heads to capture K topics among news corpus by topic-term weights $\mathcal{A}$, and here are the procedures of this layer: 1) we compute attention values g^k through feed-forward neural networks as shown in Eq. 1; 2) then use the *softmax* function to get the distribution of the normalized weights; 3) finally, we calculate the weighted sum of word embedding vectors using the attention weights to acquire the topic vector h_k:

$$\mathrm{g}_j^k = v_k^T \tanh\left(\mathrm{W}_K e_j + b_k\right) \tag{1}$$

$$\alpha_j^k = \frac{\exp(\mathrm{g}_j^k)}{\sum_n^N \exp(\mathrm{g}_n^k)} \tag{2}$$

$$h_k = \sum_j^N \alpha_j^k e_j \tag{3}$$

In the above the learnable parameters are $v_k \in \mathbb{R}^{D_K}$, $\mathrm{W}_K \in \mathbb{R}^{D_E \times D_K}$, and $b_k \in \mathbb{R}^{D_K}$, where D_K is the projected dimension of each attention in the middle and D_E is the embedding dimension of word vectors. We extract normalized attention weights $\mathcal{A} = \{\alpha^1, \alpha^2, \ldots, \alpha^k, \ldots, \alpha^K\}$ as topic-term weight distribution, where $\alpha^k \in \mathbb{R}^N$. The output topic vectors are $\{h_1, h_2, \ldots, h_k, \ldots, h_K\}$, where $h_k \in \mathbb{R}^{D_H}$ and D_H is the dimension of the topic vectors and document representations.

Finally, we feed the topic vectors into the additive attention layer, which generates the document-topic distribution $\mathcal{B}$ and the document representation d_i. This is achieved as follows: 1) we compute document topic attention values μ_k by Eq. 4; 2) we normalize attention weights by *softmax* function to get document-topic weights; 3) we acquire the document vector d_i through weighted sum up topic vectors according to normalized weights:

$$\mu_k = V_I^T \tanh\left(\mathrm{W}_I h_k + b_I\right) \tag{4}$$

$$\beta_k = \frac{\exp(\mu_k)}{\sum_k^K \exp(\mu_k)} \tag{5}$$

$$d_i = \sum_k^K \beta_k h_k \tag{6}$$

The trained parameters here are $V_I \in \mathbb{R}^{D_I}, \mathrm{W}_I \in \mathbb{R}^{D_E \times D_I}$, and $b_I \in \mathbb{R}^{D_I}$, where D_I is projected dimension of additive attention. Document-topic weights $\mathcal{B}^i = \{\beta_1, \beta_2, \ldots, \beta_k, \ldots, \beta_K\}$ reflects importance of each topic to the specific news N_i. The news representation $d_i \in \mathbb{R}^{D_H}$ is the final

output of the news encoder, where the users' history news representations are used to form the user representation.

3.2.2 User Modeling. This aspect is another core component of learning a user representation from the browsing history news $\mathcal{H} = \{d_1, d_2, \ldots, d_i, \ldots, d_H\}$. Inspired by previous studies [47–49], we use an additive attention network to encode news history. The procedure for acquiring a user representation is similar to Eqs.4 to 6. Specifically, we calculate user-news attention weights γ by normalizing attention values θ_i with a *softmax* function. Then the user vector u is calculated as the weighted sum of the user's history news representations using user-news attention weights.

$$\theta_i = V_U^T \tanh\left(\mathrm{W}_U d_i + b_U\right) \tag{7}$$

$$\gamma_i = \frac{\exp(\theta_i)}{\sum_j^H \exp(\theta_j)} \tag{8}$$

$$u = \sum_i^H \gamma_i d_i \tag{9}$$

Here $V_U^T \in \mathbb{R}^{D_U}$, $W_U \in \mathbb{R}^{D_H \times D_U}$, and $b_U \in \mathbb{R}^{D_U}$ are trainable parameters, and D_U is the dimension of the projection layer. We determine the significance of the news item N_i via the user-news weight γ_i, which is also used to select the most relevant news articles for the recommendations task.

In addition to attentive methods, sequential methods, such as Gated Recurrent Unit (GRU) networks [9], can also be effective in modeling a user's interest [2, 34]. Thus, we further consider the impact of using a GRU network as the user model, where the user vector is computed as follow [10]:

$$r_t = \sigma(W_{ir} d_t + W_{hr} o_{t-1} + b_r) \tag{10}$$

$$z_t = \sigma(W_{iz} d_t + W_{hz} o_{t-1} + b_z) \tag{11}$$

$$n_t = \tanh\left(W_{in} d_t + b_{in} + r_t * (W_{hn} o_{t-1} + b_{hn})\right) \tag{12}$$

$$o_t = (1 - z_t) * n_t + z_t * o_{t-1} \tag{13}$$

Accordingly, $W_{ir}, W_{hr}, W_{iz}, W_{hz}, W_{in}, W_{hn} \in \mathbb{R}^{D_H \times D_H}$ and $b_r, b_z, b_{in}, b_{hn} \in \mathbb{R}^{D_H}$ are learnable hidden weights and biases. We represent the input news representation at time t with d_t, the hidden states at time $t-1$ and t with o_{t-1} and o_t, respectively, and the reset, update, and new gates at time t with r_t, z_t, and n_t, respectively. We denote the sigmoid function with σ, and the Hadamard product with $*$.

We denote the first strategy that uses additive methods as the *BATM-ATT* model, and the second strategy incorporating the GRU network as the *BATM-GRU* model. The output user representation $u \in \mathbb{R}^{D_H}$ reflects a user's interest, which will be employed to determine the click probabilities of corresponding groups of candidate news articles.

3.2.3 Click Predictor and Training Strategy. By obtaining the user representation u and the set of candidate news representations $D = \{d_1, d_2, \ldots, d_i, \ldots, d_C\}$, we can employ the simple inner product to calculate the news click probability score, which is inspired by recent research [2, 47–49]. The probability score s_i of the news N_i is computed as $s_i = u^T d_i$, which determines the recommendation rank among the candidate news items. For the model training procedure, we use the negative sampling strategy [17, 49] to train our ranking model using the NCE loss. We treat each clicked news item as a positive sample of the candidate news set, and randomly select M non-clicked news items from the same impression as negative samples. Then we jointly calculate scores of positive

and negative to acquire the NCE loss. Finally, the loss $\mathcal{L}_{NCE}$ is given by the negative log-likelihood of all positive samples of this impression:

$$\mathcal{L}_{NCE} = -\sum_{p}^{P} \log \frac{\exp(s_p^+)}{\exp(s_p^+) + \sum_m^M \exp(s_m^-)} \tag{14}$$

where P represents the number of positive training samples in the impression and s_m^- denotes the m^{th} negative sample in the same session linked to the p^{th} positive sample.

3.3 Recommendation Explanations

3.3.1 Topic Extraction. After training the proposed recommendation system, we analyze the attention weights extracted from the trained *BATM-ATT* only to generate explanations because the *BATM-GRU* model is not fully explainable. An explanation is provided to clarify why the model recommends such news items, and we usually only care about the several foremost recommended news items. The core aspect of our explanations are the extracted topics used to reveal the relatedness between the user's browsing history news and the ranked candidate news. Thus, we propose to use the quantitative topic coherence metrics [25, 33] to globally evaluate the quality of topics extracted by the *BATM-ATT* model in Section 4.5, which can reflect the trustworthiness of these explanations. Later in Section 4.6 we present a case study to further motivate the idea of using topics to validate the relevance of related news articles in a real-world example.

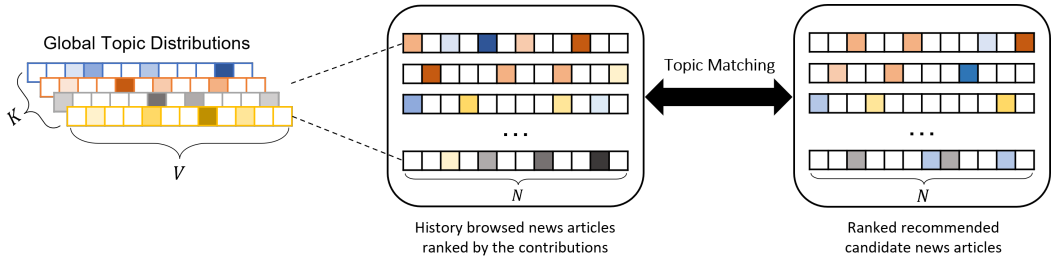


Fig. 3. An illustration of the process of using topics as recommendation explanations, where different colors represent different topics. The more saturated color indicates a higher topic weight for a given word.

The generation of explanations consists of two steps, as shown in Fig. 3. Here we revisit the recommendation example previously discussed in Section 3.1. The first phase involves the global topic distribution from the multiple-topic attention layer, as calculated by Eq. 1 and Eq. 2 using embedded word vectors as inputs. Assume that there are K global topics in total and V words in the corpus, so the resulting topic distribution $\mathcal{T}$ is a $K \times V$ weight matrix. Moreover, we extract the *top-M* most important words from the global topic distribution $\mathcal{T}$ for each topic and view them as being the topic's *descriptors*, which are used for the topic's quality evaluation. The next phase involves identifying the news contribution among the browsing history $\mathcal{H}$ for the candidate set C. Here we can recognize the news contribution by using user-news attention weights γ as calculated by Eq. 7 and Eq. 8. A news article from $\mathcal{H}$ should correlate highly with the clicked candidate news articles when they focus on similar topics. We select the most relevant topics according to the document-topic distributions $\mathcal{B}$ for each news article. To verify the topic relevance among the most contributed history clicked news articles and recommended news articles, we highlight the highest-scoring words from a subset topic distribution $\mathcal{T}$, which takes the form of a $K \times N$ matrix where N is the length of the corresponding news article. Thus, we can validate the relevance of the history $\mathcal{H}$ and candidate C from the semantic meaning of those highlighted words.

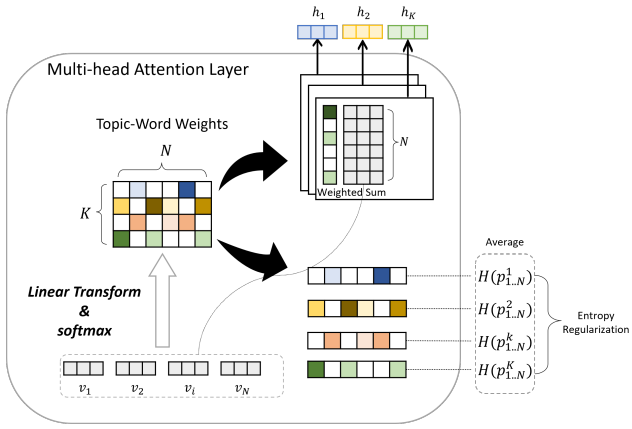


Fig. 4. Calculation procedure of entropy regularization procedure in the multi-head attention layer.

3.3.2 Enhancing Topic Extraction with Entropy Regularization. In the multi-head attention-based news recommendation model described earlier, each attention head corresponds to a topic within the news document. However, a significant challenge arises when most topics exhibit uniform focus across various words, thereby reducing the interpretability of the recommendations. We desire each attention head to encapsulate a unique and unambiguous semantic. For instance, one attention head could be focused on the election, assigning greater weights to keywords such as *republican*, *democratic*, and *party*. In contrast, another attention head could concentrate on the entertainment industry, especially the film industry, giving higher weights to terms like *director*, *actor*, and *studio*. To quantify the concentration degree of each attention head, we leverage information entropy. Each attention head yields a weight probability distribution over a news document, symbolized by $\boldsymbol{p}$, and the information entropy of this distribution is computed as demonstrated in Eq 15.

$$H(\boldsymbol{p}) = -\sum_i p_i \log_2 p_i \tag{15}$$

The value of $H(\boldsymbol{p})$ is larger when weights are evenly spread across words, whereas it is smaller when the weights are chiefly concentrated on a select few words and most other words are assigned lower values. The goal for this model is to have a lower entropy, enabling each attention head to focus on fewer, yet more pertinent keywords. To achieve this, we introduce a regularization term. We calculate the entropy for each head for every news item and use the average of these calculations as the regularization term, which is illustrated in Eq 16.

$$\mathcal{L} = \mathcal{L}_{NCE} + \lambda \cdot \frac{1}{MK} \sum_i^M \sum_j^K H(\boldsymbol{p}) \tag{16}$$

where $\mathcal{L}_{NCE}$ is negative log-likelihood loss, M represents the number of news involved in the training process, K is the number of topics, and λ represents the entropy regularization coefficient.

This entropy regularization term influences the overall weight distributions across all attention heads, penalizing news items that demonstrate uniformly distributed weights and promoting a more focused emphasis on key terms.

4 Experiments

We now provide details of experimental evaluations, including a description of the news dataset, baseline models, and experimental settings for the news recommendation task. We also provide an evaluation of the global topics generated by our approach using topic coherence metrics, together with a case study in Section 4.6.

4.1 Data

We evaluated our proposed model on a news recommendation task with a real-world news recommendation dataset, MIcrosoft News Dataset (MIND) [51]. MIND is a large-scale English news recommendation collection that consists of 1 million anonymized users and more than 160k English news articles, retrieved during 6 weeks from October 12 to November 22, 2019. In addition to the articles themselves, over 15 million impression logs were collected during this time period, involving more than 24 million user clicks. Each impression log represents a one-time recommendation that includes the IDs of news shown to a user when the user browsed the news platform during a specific time slot, along with the click behaviors on these news articles. The content related to each news article includes its title, abstract, category, and URL. We add further information based on the original news dataset by including news body content for our experiments.

Table 2. Summary information of the versions of MIND considered in our experiments.

MIND Version	#Users	#Impressions	#Clicks	#News	#Avg. Len
MIND-OFFICIAL	1,000,000	15,777,377	24,155,470	161,013	639.57
MIND-LARGE	750,434	2,609,219	3,958,501	130,379	593.56
MIND-SMALL	94,057	230,117	347,727	65,238	638.43

#Avg. Len denotes the average token count in news content (title, abstract, and body), with tokens split by spaces. MIND-Official statistic data is from the MIND [51]. MIND-LARGE and MIND-SMALL are publicly available versions.

The final publicly-released MIND dataset contains only two weeks' impression logs from November 9 to November 22, 2019, where only the first week's logs are labeled. MIND released two versions of the dataset from the first week's data, named MIND-LARGE and MIND-SMALL respectively. The larger dataset contains all behaviors from the first week, while the smaller one includes only the impression logs for a subset of the users. We list details of the official MIND and publicly available versions in Table 2. Our evaluation is based on the MIND-SMALL and MIND-LARGE, which, while robust in its coverage of numerous user interactions across a diverse range of news categories, is collected over a relatively short and fixed period. This limitation may introduce temporal biases—such as increased interest in specific topics due to current events—that are not representative of longer-term news consumption patterns. Despite these limitations, the dataset's large user base and substantial interaction data provide a valuable foundation for evaluating the news recommendation system.

In our evaluations, we randomly divide the data into training/validation/test sets. As for the news recommendation (NR) task, we randomly split the first week's log into three sets for both MIND-SMALL and MIND-LARGE versions: training set (from November 9 to November 14), validation set (November 15), and test set (November 15). We select the best hyperparameters for different models based on the validation set, and compare their relative performance on the test set.

4.2 Baseline Models and Experimental Settings

For the purpose of assessing recommendations performance, we compare our model with two groups of baselines. The first group relies on popularity-based strategies:

- MostPop recommends news articles by measuring their popularity, quantified by the aggregate number of views within the dataset.
- RecencyPop assigns news popularity determined by temporal proximity that news articles are more recent to the users' engagement time are deemed more significant.
- TopicPop recommends popular news aligned with the categories of a user's historical clicks, prioritizing news that matches previously browsed topics.

The second group consists of several popular deep neural models designed for news recommendation, and comparisons are summarized in Table 3:

- DKN [45] applies a word-entity-aligned KCNN on news representations learning and a candidate-aware attention network for recommendations[2].
- NAML [47] leverages two separate CNN to encode news title and body text while using linear layers to encode category and subcategory. The news representations are learned from a multi-view of text, category, and subcategory representations. And modeling user behavior using another attention network.
- NRMS [49] uses multi-head self-attentions for both news and user modeling.
- LSTUR [2] involves an attention-based CNN on news representations learning and a GRU network to model short-term user interest along with user id for long-term user interest.
- NPA [48] uses a personalized word-level attention-based CNN to learn news representations, while another personalized attention network is employed to learn user representations.

Table 3. Comparison of Different Nenural News Recommendation Models

Model	News Modeling	User Modeling
DKN [45]	Word-entity-aligned KCNN	Candidate-aware attention network
NAML [47]	Separate CNNs for title and body	Attention network
NRMS [49]	Multi-head self-attention	Multi-head self-attention
LSTUR [2]	Attention-based CNN	GRU with user ID
NPA [48]	Word-level attention-based CNN	Personalized attention network
BATM-NR	Bi-level Attention Topic Model (BATM)	Attention or GRU

We divide experimental configurations into two categories: general system settings and model hyperparameter settings. Consistent with previous studies [2, 45, 47–49], we employ the same training procedure for all systems to ensure fairness. The general system settings are as follows:

(1) We set the maximum length of a news article to 100 tokens, with a maximum length of 30 for the news title and 70 for the news body, respectively.
(2) we sampled 50 most recent browsed articles from a user's history for user representation learning.
(3) The negative sampling ratio M was set to 4 during training to pair each positive sample with 4 negative samples during the training procedure.
(4) We employed Adam [21] as the model optimization technique during gradient descent for all deep neural models.

[2]We use dot product instead of a linear neural network to predict click probability for a fair comparison.

In terms of model hyperparameter settings, we initialized the embedding layer with pre-trained Glove embedding [36] and set the embedding dimension D_E to 300. To mitigate overfitting, we added a dropout layer [41] to each layer for all models in comparison. In our experiments, we conducted a thorough hyperparameter tuning for all deep neural models involved in the comparison. The candidate hyperparameters tested on the validation set included:

(1) Batch size: [16, 32, 64, 128]
(2) Dropout rate: [0, 0.2, 0.5, 0.8]
(3) Learning rate: [0.005, 0.001, 0.0005]
(4) Model-specific unique hyperparameters

Following previous work [51], we consider four ranking metrics to evaluate the performance of news recommendations: Group AUC, MRR, nDCG@5, and nDCG@10. After extensive evaluation experiments on the validation set, we identified the optimal hyperparameter settings for each model monitored by the Group AUC metric, which demonstrated the best performance. These settings are summarized in Table 4.

Table 4. Optimal Hyperparameter Settings for Compared Models

Models	Batch Size	Dropout	Learning Rate	Model-specific unique hyperparameters
LSTUR [2]	32	0.2	0.0005	User embeddings: user ids initialization (ini); Number of CNN filters: 300
NAML [47]	32	0.2	0.0005	Number of CNN filters: 300
NPA [48]	32	0.2	0.0005	Number of CNN filters: 300
NRMS [49]	32	0.2	0.0005	Head number: 20; Head dimension: 20
DKN [45]	32	0.2	0.0005	Number of CNN filters: 300
BATM-NR	32	0.2	0.0005	Topic number: 70; Entropy Regularization: 0.001

We repeat each experiment with the best model hyperparameter setting independently 5 times with a fixed set of random seeds. The performance scores on the test set are reported in Section 4.3.

4.3 News Recommendations Performance

The overall performances of all baselines and two variants of our model are summarized in Table 5. All the numbers in the table are percentages numbers with '%' omitted. The overall best are boldfaced, and the previous best results are underlined, respectively.

Table 5. Recommendation performance for different models, in terms of AUC, MRR, nDCG@5, and nDCG@10.

Models	MIND-SMALL				MIND-LARGE			
	AUC	MRR	nDCG@5	nDCG@10	AUC	MRR	nDCG@5	nDCG@10
MostPop	53.17±0.00	26.87±0.00	28.12±0.00	34.15±0.00	53.12±0.00	26.96±0.00	28.17±0.00	34.03±0.00
RecencyPop	54.01±0.00	26.94±0.00	28.05±0.00	34.16±0.00	53.97±0.00	27.07±0.00	28.10±0.00	33.99±0.00
TopicPop	57.45±0.00	27.64±0.00	29.30±0.00	35.43±0.00	56.76±0.00	27.70±0.00	29.27±0.00	35.34±0.00
LSTUR [2]	67.26±0.13	31.44±0.16	35.20±0.18	41.43±0.19	69.31±0.16	33.38±0.22	37.39±0.21	43.57±0.17
NAML [47]	67.14±0.20	31.58±0.28	35.20±0.29	41.52±0.28	69.24±0.17	33.96±0.27	37.85±0.21	44.02±0.18
NPA [48]	66.24±0.25	31.06±0.13	34.37±0.19	40.69±0.15	68.95±0.21	33.37±0.37	37.26±0.35	43.39±0.34
NRMS [49]	66.58±0.17	31.44±0.15	34.99±0.19	41.21±0.16	69.09±0.13	33.25±0.42	37.19±0.33	43.43±0.29
DKN [45]	66.95±0.25	31.12±0.28	34.94±0.29	41.13±0.29	68.89±0.11	33.24±0.15	37.26±0.11	43.43±0.11
BATM-NR	**68.13±0.15**	**32.62±0.08**	**36.36±0.07**	**42.46±0.09**	**69.73±0.16**	**34.17±0.22**	**38.11±0.16**	**44.25±0.23**

From the results, we can make a number of important observations. Firstly, we see that TopicPop achieves better performance than the other popularity-based methods, primarily because it utilizes category information from users' historical browsing news articles, enabling more accurate recommendations. However, all deep neural models outperform the popularity-based approaches,

benefiting from their advanced capabilities in processing and representing the textual content of news articles, thereby providing more sophisticated and contextually relevant recommendations.

Second, all deep neural models achieved similar performance levels based on the experimental settings described in Section 4.2. This is because we employed the same pre-trained embedding parameters (GloVe) to initialize the embedding layer, which means the number of trainable parameters of models is close[3]. Another reason is that we applied a similar architecture which contains a news encoder and a user encoder and makes predictions using the results of dot-product between news representations and user representations. Thus, the difference in results is due to news modeling and user modeling design.

Third, the performances of all models in the MIND-LARGE set are significantly better compared to the results in the context of the MIND-SMALL set. NAML model generally achieved the best results among the selected baselines, except in some cases (e.g., the AUC metric) where LSTUR model performs slightly better than the NAML model, as it can obtain information from news text, category, and subcategory to form informative representations. LSTUR model is also competitive, as it uses a sequence model (GRU) for user modeling, effectively capturing user interests, which is also the reason we tried a GRU variant of our model.

Finally, our models *BATM-NR* outperform all baselines on the MIND-SMALL set and the MIND-LARGE set for all metrics. We can observe that NAML model performs comparable to our models, though our models are slightly better overall. The encoding of category and subcategory in NAML model may fuse explicit topic information into the final news representations, which is very similar to our topic modeling module. However, our model can extract latent topics from the news texts and acquire topic representation from texts instead of categories and subcategories. Overall, these experiments demonstrate the effectiveness of our proposed models for news recommendation.

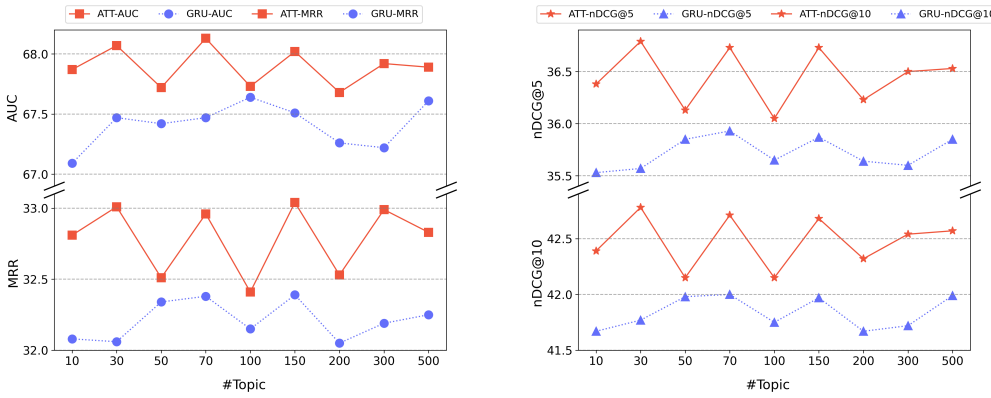
4.4 Recommendation Effectiveness Study

In this section, we explore various factors that influence the performance of our proposed *BATM-NR* system. Our focus is primarily on three key aspects: the number of topics, the selection of User Encoders, and the λ value of Entropy Regularization. Given the complexity and interplay of multiple variables within our models, we adopt a controlled approach to isolate the impact of each factor. This study allows us to find the optimal configurations for our recommendation systems.

4.4.1 Influence of Topic Numbers and User Encoders. We study the impact of different topic numbers on the recommendation performance of the *BATM-NR* system on the MIND-SMALL dataset. Our study focuses on the model without incorporating Entropy Regularization, examining two distinct user encoders within the *BATM-NR* framework. Firstly, the range of topic numbers considered in our experiment includes [10, 30, 50, 70, 100, 150, 200, 300, 500]. This allows us to examine the effect of topic granularity on recommendation performance comprehensively. The recommendation performance corresponding to each topic number are shown in Fig. 5. In this figure, the red straight lines represent the performance of the model utilizing a fully attention-based mechanism (*BATM-ATT*), whereas the blue dot lines denote the model employing GRU as the user encoder (*BATM-GRU*).

Our results reveal an oscillating performance trend as we change the number of topics in a relatively small range. Notably, the *BATM-ATT* system demonstrates comparable and superior performance when the topic number is set to either 30 or 70, indicating these as the optimal settings for topic granularity in this context. Furthermore, *BATM-ATT* consistently outperformed *BATM-GRU* across different topic settings, leading to its selection as our final model.

[3]The number of embedding layer's parameters usually make up 90% of the model in total



(a) AUC and MRR scores

(b) nDCG@5 and nDCG@10 scores

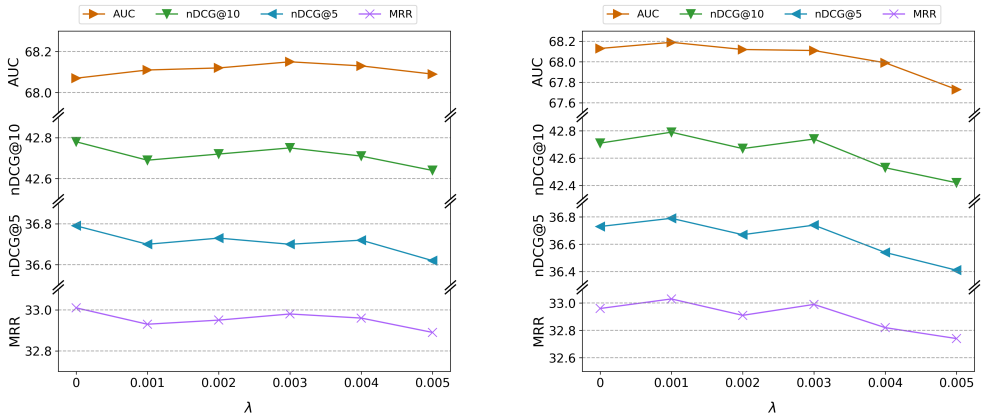
Fig. 5. The influence of topic numbers on the recommendation performance of two *BATM-NR* variants.

4.4.2 Influence of Entropy Regularization. We now explore the effect of Entropy Regularization on the recommendation performance of the chosen *BATM-ATT* model. Entropy Regularization, as defined in Eq. 16, incorporates a coefficient λ which plays a pivotal role in shaping the model's understanding of topic distribution. However, the direct impact of this regularization on the recommendation process is not straightforward. Our study examines the effect of varying λ within the range of [0, 0.001, 0.002, 0.003, 0.004, 0.005], under two distinct settings of topic numbers: 30 and 70. The results of this exploration are depicted in Fig. 6, from which several key observations can be drawn.

Firstly, the influence of λ on recommendation performance is relatively minor, as indicated by the small variations in metrics within the same topic number setting. This suggests that while Entropy Regularization contributes to a more refined topic distribution, its direct impact on recommendation quality is limited. Secondly, an initial improvement in recommendation performance is noted as λ increases, followed by a decline after surpassing a certain threshold. This pattern indicates a delicate balance between the beneficial and detrimental effects of Entropy Regularization on the model's recommendation capabilities. Finally, considering the overall recommendation performance, a λ value within 0 to 0.003 appears to offer an optimal balance. Beyond this range, the effectiveness of the recommendations diminishes, suggesting that excessive regularization may hinder the model's predictive accuracy.

4.5 Evaluation of Explainability

Explainability in this context stems from understanding recommendation behaviors through a topic modeling perspective, enhancing transparency and trustworthiness by interpreting the meanings of hidden vectors via attention scores, as mentioned in Section 3.3. Quantitatively evaluating such explainability poses a challenge due to the absence of standardized benchmarks or ground truth, a dilemma we acknowledge in Section 2.2. Unlike other attention-based explanations detailed in Section 2.2.1, our approach employs topic modeling to understand and explain recommendation behaviors. As described in Section 3.3, both historical and candidate news articles can be modeled by the *BATM* model [26], which describes news articles through document-topic and topic-word distributions. The extracted topic descriptors can reflect the user's interest in certain topics. Thus,



(a) Topic number is set to 30

(b) Topic number is set to 70

Fig. 6. Impact of Entropy Regularization (λ) on the recommendation performance of *BATM-ATT* for different topic numbers.

recommendations can be interpreted as the alignment between the topics described in candidate news articles and those of interest to the user, illustrated by Fig. 3. The quality of these extracted topics can indicate the explainability of the recommendation process which heavily relies on the matching of generated topics. Therefore, we propose using topic coherence as a quantitative measure of explainability, which assesses the semantic interpretability of the top terms that are typically used to describe the topics extracted by the topic modeling algorithm.

Our assessment of explainability now pivots to using topic coherence to evaluate the ability of topic modeling algorithms in extracting coherent topics. Unlike the classical topic modeling algorithm, Latent Dirichlet Allocation (LDA), which primarily aims to extract topics, our models focus on achieving better recommendations instead. Consequently, the evaluation of proposed models is divided into two distinct aspects: 1) recommendation performance, 2) the capability to extract coherent topics. The comprehensive analysis of the model's recommendation effectiveness was previously detailed in Section 4.4. Thus, we focus on the second aspect of our evaluation by leveraging topic coherence as a quantifiable metric to assess this dimension.

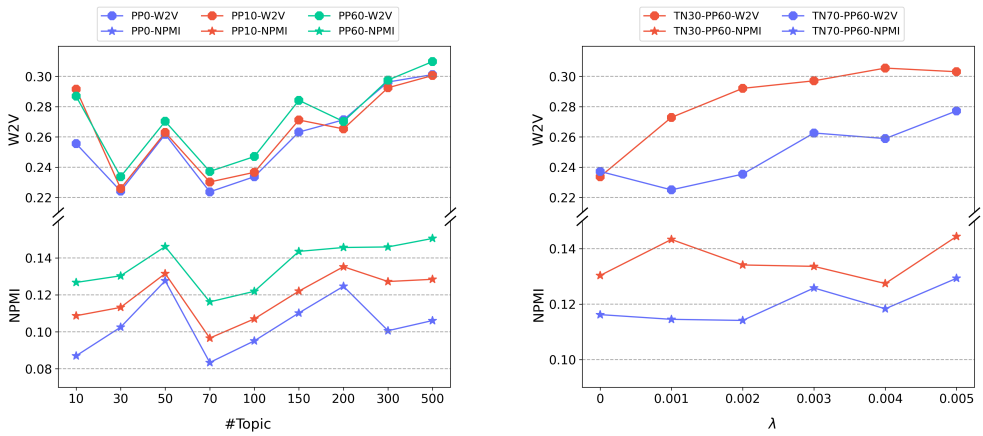
After training the model, we calculate the topic coherence scores using two different metrics from the topic modeling literature: *NPMI* [7, 25] and *Word2Vec*(*W2V*) similarity [33]. These metrics are used to evaluate the coherence of the extracted topic descriptors. The reason for using these two metrics is that the topic coherence metric *NPMI* is regarded as positively correlated with human intuition [25], while *W2V* similarity is designed for embedding-based methods. We select the *top-M* highest scoring words as topic descriptors and calculate *NPMI* scores and *W2V* similarity scores of them. For a given topic k, suppose we obtain a topic descriptor set $T_k = \{t_1^k, t_2^k, \ldots, t_n^k, \ldots, t_N^k\}$, so we compute the *NPMI* scores and *W2V* similarity scores as follow:

$$\mathrm{NPMI}(T_k) = \frac{1}{\binom{M}{2}} \sum_{j=2}^{N} \sum_{i=1}^{j-1} \frac{\log \frac{P\left(t_j^k, t_i^k\right)+\epsilon}{P\left(t_i^k\right)P\left(t_j^k\right)}}{-\log P\left(t_i^k, t_j^k\right)+\epsilon} \tag{17}$$

$$\mathrm{W2V}(T_k) = \frac{1}{\binom{M}{2}} \sum_{j=2}^{N} \sum_{i=1}^{j-1} \mathrm{similarity}(e_j^k, e_i^k) \tag{18}$$

Here e_j^k and e_i^k in Eq. 18 are the word vectors of tokens t_j^k and t_i^k from the trained model. We set $M = 10$ for each topic and take the average values of all topics. The word co-occurrence probabilities of t_j^k and t_i^k are counted from the reference corpus $\mathcal{S}_{\text{news}}$, which contain all news articles from the MIND dataset.

4.5.1 Effect of Post-Processing. Corpus preprocessing can influence the resulting downstream topic descriptors and can further affect topic evaluation metrics. The processing pipeline of a topic model mainly involves three steps: 1) filter out stopwords using the default spaCy English stopword list[4]; 2) remove tokens that appear in more than 90% of documents; 3) remove tokens that appear in fewer than N documents. We use the pipeline as the preprocessing procedure to train an LDA model and evaluate it with the same processing pipeline. However, the preprocessing pipeline of our deep models is quite different as we keep most tokens (e.g., stopwords are preserved) for training to retain semantic information. Thus, we apply the same pipeline as the LDA model after the training process. As it is performed after training, we name it post-processing (denoted as *PP-N*, such as *PP*10) for our *BATM-ATT* model. For the model evaluated on the original training dataset without post-processing, we denote it as *PP*0.



(a) Effects of *PP-N* on models with different numbers of topics.

(b) Effects of λ for models with 30 and 70 topics on coherence scores.

Fig. 7. Impact of post-processing and entropy regularization coefficient λ on topic coherence in *BATM-ATT* models.

As illustrated in Fig 7(a), we observe that employing post-processing strategies leads to a significant improvement in topic coherence scores. This improvement can be attributed to our model's adaptive nature in managing the recommendation task, which does not heavily rely on the processing pipeline. However, the model may exhibit sensitivity to non-informative words like stopwords, which are typically filtered out during post-processing. In contrast to the findings discussed in

[4]https://github.com/explosion/spaCy/blob/v3.0.5/spacy/lang/en/stop_words.py

Section 4.4, we observe that while the model achieves optimal recommendation results with topic numbers at 30 and 70, its proficiency in topic extraction is not equally impressive. This discrepancy suggests that an improvement in topic coherence scores does not always contribute to an improved recommendation performance. It indicates a complex interplay between the tasks of topic extraction and recommendation, where optimizing for one does not necessarily guarantee optimal outcomes for the other.

4.5.2 Effect of Entropy Regularization. We now investigate the influence of Entropy Regularization coefficient λ on topic coherence scores. As discussed in Section 4.5.1, we recognize that post-processing enhances coherence scores, while models with a topic count of 30 and 70 demonstrate less impressive outcomes. Thus, our exploration is anchored on the $PP60$ setting with topic numbers at 30 and 70, as illustrated in Fig. 7. A key observation from our analysis is that larger values of the regularization coefficient λ, coupled with more topics, generally result in higher $NPMI$ and $W2V$ similarity scores. This pattern indicates that stringent regularization together with a diverse topic range elevates the internal consistency of individual topics. Moreover, it implies that the coherence of identified topics is effectively increased by the entropy regularization process. This could be due to the fact that more topics result in a model with a broader spectrum to capture different semantic aspects of the news articles, and a stronger regularization term encourages the model to focus on a smaller subset of more relevant keywords, thereby increasing the coherence within each topic. When comparing these results with the recommendation performance as shown in Fig. 6, it is clear that finding a balance between high-quality recommendations and high topic coherence is crucial. For optimal recommendation efficacy, a topic number of 70 with λ set to 0.001 is advisable, whereas the peak topic coherence scores are attained with a topic number of 30 and λ at 0.005. Nevertheless, we advocate for a λ range of 0.001 to 0.003, coupled with a topic number setting of either 30 or 70, to strike a balanced chord between recommendation precision and topic coherence.

4.6 Case Study

Traditional attention-based methods do not yield outputs that can be quantitatively analyzed using topic coherence metrics. This limitation necessitated our reliance on qualitative case analyses to demonstrate our model's comparative advantages. Therefore, we present two case studies as presented in Fig. 8 and Fig. 9 to visually illustrate how our *BATM-ATT* model generates topics for real-world examples, following the method outlined in Section 3.3. Additionally, we compare our topic-based explanations with another attention-based approach, specifically the NPA model's attention weights, as shown in Table 6.

Table 6. An example from the attention weights of NPA [48] on the word-level attention module

	User 1	User 2
Interaction History	'no doubt' kyler murray loves football more than baseball warriors by far most hated nba team in state-by-state survey the most famous actor the same age as you winners and losers from nfl week 17	holiday movie guide 2018 : every movie you should see celebrate the holidays with your favorite tv shows spider-man : into the spider-verse' swings to top box office spot the most famous actor the same age as you
Candidate News	5 most desirable super bowl matchups year of the superheroes : the highest-grossing movies of 2018	5 most desirable super bowl matchups year of the superheroes : the highest-grossing movies of 2018

A case involves the user browsed articles $\mathcal{H}$ and the ranked candidate news articles C shown to the user. We observe that the NPA model generally identifies a singular, dominant topic for an individual user. This is exemplified in Table 6, where the model highlights a user's specific interest area. For instance, with User 1, the focus is clearly on sports-related terms, suggesting a strong preference towards sports news. In contrast, our proposed *BATM-ATT* model comprehensively considers a user's interests from a topic modeling perspective, rather than focusing solely on a

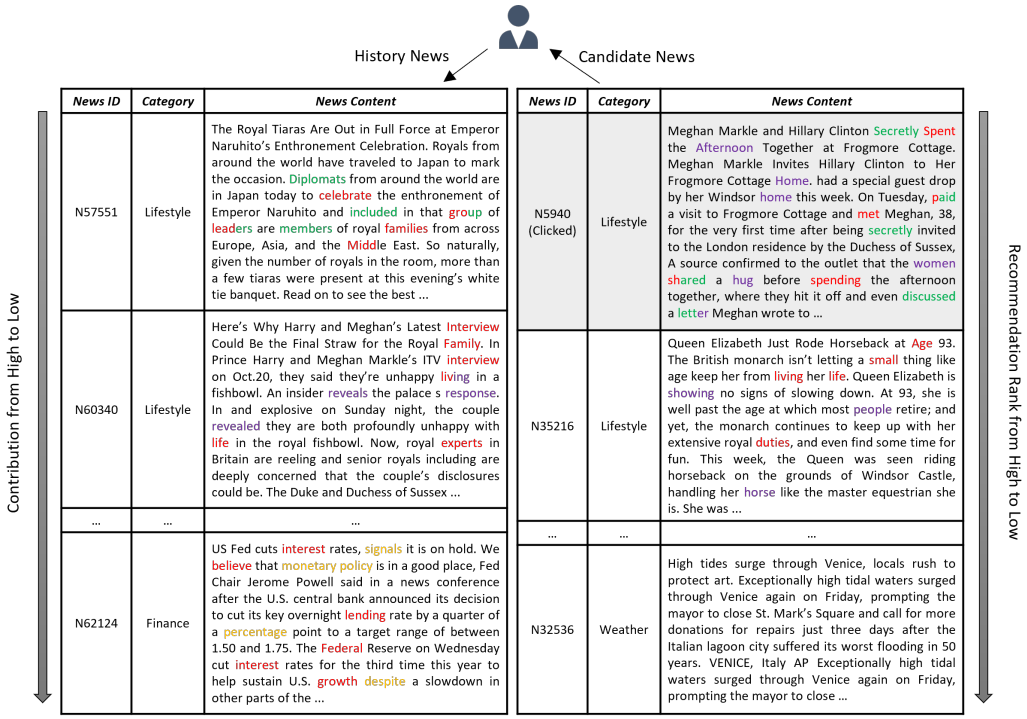
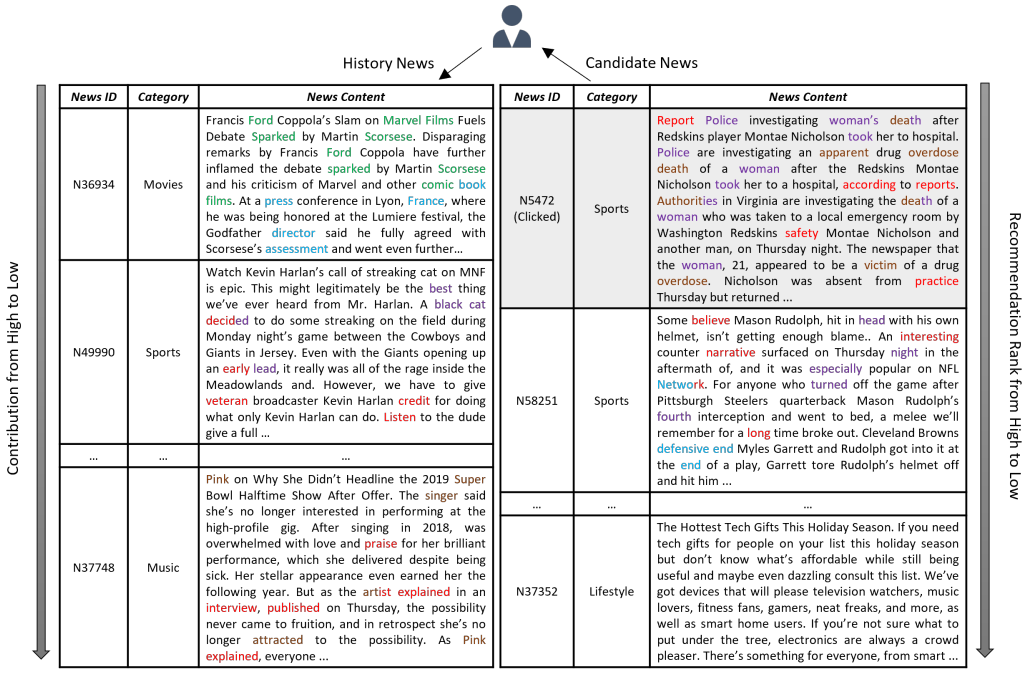


Fig. 8. A typical example of a recommendation case selected from the MIND test set. The topics for the news articles are highlighted with a consistent set of colors, each corresponding to a different topic. Since a word can potentially be associated with more than one topic, some words may have two or three colors. A color that appears more often signifies that the corresponding topic contributes more to an article than the other topics.

single interest. This enables it to capture the varied facets of a user's preferences, offering a more comprehensive and personalized news recommendation experience.

We highlight those descriptor words from three of the most important topics for each news article, where each topic is highlighted with a unique color. Firstly, we notice that the topic in red T_{red} is activated across nearly all articles. The exact meaning of the topic T_{red} is hard to summarize, as it is always highly related to the specific article itself. For example, it focuses on words such as "interview", "family", "living", and "life" for the article *N60340*, which are about daily life. However, for a financial news article *N62124*, the highlighted words now include "interest", "believe", "lending", and "federal", which are often used in the financial domain. We also observe that news articles under the same category usually discuss similar topics, and we observed that the example case also reflects this nature. Take the category "lifestyle" as an example, except for the general topic T_{red}, there are four more topics that concern words related to lifestyle for all four news articles. Among these four topics, the topic in green T_{green} and the topic in purple T_{purple} only appear in the articles on lifestyle. Thus, the last observation is that some topics are closely related to a specific subject, such as the topic in orange T_{orange}, which is highly related to the political economy topic, and attends to words like "signals", "monetary" and "policy". For further examples of topics generated by our approach, see the Appendix.



News ID	Category	News Content
N36934	Movies	Francis Ford Coppola's Slam on Marvel Films Fuels Debate Sparked by Martin Scorsese. Disparaging remarks by Francis Ford Coppola have further inflamed the debate sparked by Martin Scorsese and his criticism of Marvel and other comic book films. At a press conference in Lyon, France, where he was being honored at the Lumiere festival, the Godfather director said he fully agreed with Scorsese's assessment and went even further...
N49990	Sports	Watch Kevin Harlan's call of streaking cat on MNF is epic. This might legitimately be the best thing we've ever heard from Mr. Harlan. A black cat decided to do some streaking on the field during Monday night's game between the Cowboys and Giants in Jersey. Even with the Giants opening up an early lead, it really was all of the rage inside the Meadowlands and. However, we have to give veteran broadcaster Kevin Harlan credit for doing what only Kevin Harlan can do. Listen to the dude give a full ...
...	...	...
N37748	Music	Pink on Why She Didn't Headline the 2019 Super Bowl Halftime Show After Offer. The singer said she's no longer interested in performing at the high-profile gig. After singing in 2018, was overwhelmed with love and praise for her brilliant performance, which she delivered despite being sick. Her stellar appearance even earned her the following year. But as the artist explained in an interview, published on Thursday, the possibility never came to fruition, and in retrospect she's no longer attracted to the possibility. As Pink explained, everyone ...

News ID	Category	News Content
N5472 (Clicked)	Sports	Report Police investigating woman's death after Redskins player Montae Nicholson took her to hospital. Police are investigating an apparent drug overdose death of a woman after the Redskins Montae Nicholson took her to a hospital, according to reports. Authorities in Virginia are investigating the death of a woman who was taken to a local emergency room by Washington Redskins safety Montae Nicholson and another man, on Thursday night. The newspaper that the woman, 21, appeared to be a victim of a drug overdose. Nicholson was absent from practice Thursday but returned ...
N58251	Sports	Some believe Mason Rudolph, hit in head with his own helmet, isn't getting enough blame.. An interesting counter narrative surfaced on Thursday night in the aftermath of, and it was especially popular on NFL Network. For anyone who turned off the game after Pittsburgh Steelers quarterback Mason Rudolph's fourth interception and went to bed, a melee we'll remember for a long time broke out. Cleveland Browns defensive end Myles Garrett and Rudolph got into it at the end of a play, Garrett tore Rudolph's helmet off and hit him ...
...	...	...
N37352	Lifestyle	The Hottest Tech Gifts This Holiday Season. If you need tech gifts for people on your list this holiday season but don't know what's affordable while still being useful and maybe even dazzling consult this list. We've got devices that will please television watchers, music lovers, fitness fans, gamers, neat freaks, and more, as well as smart home users. If you're not sure what to put under the tree, electronics are always a crowd pleaser. There's something for everyone, from smart ...

Fig. 9. Another case with the same format as Fig. 8, sampled from the MIND test set.

5 Conclusion and Future Work

In this paper, we presented a novel recommender architecture that harnesses a bi-level attention framework to decouple the news recommendations process as topic capturing, topic importance recognition, and decision-making process to benefit explainability. We conducted experiments on a real-world news recommendation dataset where we compared our approach to several state-of-the-art alternatives. Results indicate that our model can achieve better performance while also successfully capturing intuitive meanings in the form of topical features, thus improving its explainability and transparency. Furthermore, we applied two topic coherence metrics to quantitatively evaluate the quality of the topics generated by our model, in the context of recommendation task, thereby validating the interpretability of our model. For future work, we suggest three distinct directions. First, we intend to explore the use of topic features to further improve news recommendation performance. Second, some strategies can be applied to extract more interpretable topics measured by coherence scores. Also, it is interesting to conduct a user study to determine how well our explainable recommender works in a real-world setting. Finally, it is important to investigate the impact of cognitive load due to explanations in recommender systems and explore ways to optimize the balance between providing helpful insights and maintaining user engagement without overwhelming them.

Acknowledgments

This research was conducted with the financial support of Science Foundation Ireland 12/RC/2289_P2 at Insight the SFI Research Centre for Data Analytics at University College Dublin.

References

[1] Gediminas Adomavicius and Alexander Tuzhilin. 2005. Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge and Data Engineering* 17 (2005), 734–749.

[2] Mingxiao An, Fangzhao Wu, Chuhan Wu, Kun Zhang, Zheng Liu, and Xing Xie. 2019. Neural News Recommendation with Long- and Short-term User Representations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 336–345.

[3] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural Machine Translation by Jointly Learning to Align and Translate. In *3rd International Conference on Learning Representations, ICLR*.

[4] Jasmijn Bastings and Katja Filippova. 2020. The elephant in the interpretability room: Why use attention as explanation when we have saliency methods? In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*. Association for Computational Linguistics, Online, 149–155. https://doi.org/10.18653/v1/2020.blackboxnlp-1.14

[5] Adrien Bibal, Rémi Cardon, David Alfter, Rodrigo Wilkens, Xiaoou Wang, Thomas François, and Patrick Watrin. 2022. Is Attention Explanation? An Introduction to the Debate. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Dublin, Ireland, 3889–3900. https://doi.org/10.18653/v1/2022.acl-long.269

[6] David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *The Journal of Machine Learning Research* 3 (2003), 993–1022.

[7] Gerlof Bouma. 2009. Normalized (pointwise) mutual information in collocation extraction. *Proceedings of GSCL* 30 (2009), 31–40.

[8] Chong Chen, Min Zhang, Yiqun Liu, and Shaoping Ma. 2018. Neural Attentional Rating Regression with Review-level Explanations. In *Proceedings of the 2018 World Wide Web Conference*. ACM, 1583–1592.

[9] Kyunghyun Cho, Bart van Merrienboer, Çaglar Gülçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. In *Conference on Empirical Methods in Natural Language Processing*.

[10] PyTorch Contributors. 2022. *PyTorch GRU Documentation*. https://pytorch.org/docs/stable/generated/torch.nn.GRU.html Access on 2023-01-18.

[11] Abhinandan S. Das, Mayur Datar, Ashutosh Garg, and Shyam Rajaram. 2007. Google news personalization: scalable online collaborative filtering. In *Proceedings of the 16th International Conference on World Wide Web*.

[12] Emma de Koning, Frederik Hogenboom, and Flavius Frasincar. 2018. News Recommendation with CF-IDF+. In *CAiSE*.

[13] Adji Bousso Dieng, Francisco J. R. Ruiz, and David M. Blei. 2020. Topic Modeling in Embedding Spaces. *Trans. Assoc. Comput. Linguistics* 8 (2020), 439–453.

[14] Fatih Gedikli, Dietmar Jannach, and Mouzhi Ge. 2014. How should I explain? A comparison of different explanation types for recommender systems. *International Journal of Human-Computer Studies* 72, 4 (2014), 367–382.

[15] Frank Goossen, Wouter IJntema, Flavius Frasincar, Frederik Hogenboom, and Uzay Kaymak. 2011. News personalization using the CF-IDF semantic recommender. In *Proceedings of the International Conference on Web Intelligence, Mining and Semantics*. 1–12.

[16] Zhiqiang Guo, Guohui Li, Jianjun Li, and Huaicong Chen. 2022. TopicVAE: Topic-aware Disentanglement Representation Learning for Enhanced Recommendation. In *MM '22: The 30th ACM International Conference on Multimedia*. ACM, 511–520.

[17] Po-Sen Huang, Xiaodong He, Jianfeng Gao, Li Deng, Alex Acero, and Larry Heck. 2013. Learning deep structured semantic models for web search using clickthrough data. In *Conference on Information and Knowledge Management*.

[18] Alon Jacovi and Yoav Goldberg. 2020. Towards Faithfully Interpretable NLP Systems: How Should We Define and Evaluate Faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 4198–4205. https://doi.org/10.18653/v1/2020.acl-main.386

[19] Sarthak Jain and Byron C. Wallace. 2019. Attention is not Explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, 3543–3556.

[20] Guoliang Ji, Shizhu He, Liheng Xu, Kang Liu, and Jun Zhao. 2015. Knowledge Graph Embedding via Dynamic Mapping Matrix. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (volume 1: Long papers)*. 687–696.

[21] Diederik P. Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. *CoRR* abs/1412.6980 (2014).

[22] Vineet Kumar, Dhruv Khattar, Shashank Gupta, Manish Gupta, and Vasudeva Varma. 2017. Deep Neural Architecture for News Recommendation. In *CLEF (Working Notes)*.

[23] Vaibhav Kumar, Dhruv Khattar, Shashank Gupta, and Vasudeva Varma. 2017. Word Semantics Based 3-D Convolutional Neural Networks for News Recommendation. In *IEEE International Conference on Data Mining Workshops (ICDMW)*.

IEEE, 761–764.

[24] Xuan Nhat Lam, Thuc Vu, Trong Duc Le, and Anh Duc Duong. 2008. Addressing cold-start problem in recommendation systems. In *Proceedings of the 2nd International Conference on Ubiquitous Information Management and Communication*. 208–211.

[25] Jey Han Lau, David Newman, and Timothy Baldwin. 2014. Machine Reading Tea Leaves: Automatically Evaluating Topic Coherence and Topic Model Quality. In *Conference of the European Chapter of the Association for Computational Linguistics*.

[26] Dairui Liu, Derek Greene, and Ruihai Dong. 2022. A Novel Perspective to Look At Attention: Bi-level Attention-based Explainable Topic Modeling for News Classification. In *Findings of the Association for Computational Linguistics: ACL 2022*. ACL, Dublin, Ireland, 2280–2290.

[27] Danyang Liu, Jianxun Lian, Zheng Liu, Xiting Wang, Guangzhong Sun, and Xing Xie. 2021. Reinforced Anchor Knowledge Graph Generation for News Recommendation Reasoning. *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining* (2021).

[28] Danyang Liu, Jianxun Lian, Shiyin Wang, Ying Qiao, Jiun-Hung Chen, Guangzhong Sun, and Xing Xie. 2020. KRED: Knowledge-Aware Document Representation for News Recommendations. In *Proceedings of the 14th ACM Conference on Recommender Systems* (Virtual Event, Brazil) *(RecSys '20)*. Association for Computing Machinery, New York, NY, USA, 200–209. https://doi.org/10.1145/3383313.3412237

[29] Yichao Lu, Ruihai Dong, and Barry Smyth. 2018. Coevolutionary Recommendation Model: Mutual Learning between Ratings and Reviews. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web, WWW 2018*. ACM, 773–782.

[30] Tapio Luostarinen and Oskar Kohonen. 2013. Using Topic Models in Content-Based News Recommender Systems. In *Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA 2013)*. Linköping University Electronic Press, Sweden, Oslo, Norway, 239–251.

[31] Julian J. McAuley and Jure Leskovec. 2013. Hidden factors and hidden topics: understanding rating dimensions with review text. In *Proceedings of the 7th ACM Conference on Recommender Systems, RecSys '13*. ACM, 165–172.

[32] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systemss* 26 (2013).

[33] Derek O'Callaghan, Derek Greene, Joe Carthy, and Pádraig Cunningham. 2015. An analysis of the coherence of descriptors in topic modeling. *Expert Systems with Applications* 42 (2015), 5645–5657.

[34] Shumpei Okura, Yukihiro Tagami, Shingo Ono, and Akira Tajima. 2017. Embedding-based News Recommendation for Millions of Users. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

[35] Madhur Panwar, Shashank Shailabh, Milan Aggarwal, and Balaji Krishnamurthy. 2021. TAN-NTM: Topic Attention Networks for Neural Topic Modeling. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers)*. Association for Computational Linguistics, 3865–3880.

[36] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Doha, Qatar, 1532–1543. https://doi.org/10.3115/v1/D14-1162

[37] Tao Qi, Fangzhao Wu, Chuhan Wu, Yongfeng Huang, and Xing Xie. 2020. Privacy-Preserving News Recommendation Model Learning. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. Association for Computational Linguistics, Online, 1423–1432. https://doi.org/10.18653/v1/2020.findings-emnlp.128

[38] Shaina Raza and Chen Ding. 2021. News recommender system: a review of recent progress, challenges, and opportunities. *Artificial Intelligence Review* (2021), 1–52.

[39] Gerard Salton, Anita Wong, and Chung-Shu Yang. 1975. A vector space model for automatic indexing. *Commun. ACM* 18, 11 (1975), 613–620.

[40] Sungyong Seo, Jing Huang, Hao Yang, and Yan Liu. 2017. 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 2017*. ACM, 297–305. https://doi.org/10.1145/3109859.3109890

[41] Nitish Srivastava, Geoffrey E. Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* 15 (2014), 1929–1958.

[42] Maartje ter Hoeve, Anne Schuth, Daan Odijk, and M. de Rijke. 2018. Faithfully Explaining Rankings in a News Recommender System. *ArXiv* abs/1805.05447 (2018).

[43] Nava Tintarev and Judith Masthoff. 2015. *Explaining Recommendations: Design and Evaluation*. Springer US, Boston, MA, 353–382.

[44] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Neural Information Processing Systems*. 5998–6008.

[45] Hongwei Wang, Fuzheng Zhang, Xing Xie, and Minyi Guo. 2018. DKN: Deep Knowledge-Aware Network for News Recommendation. *Proceedings of the 2018 World Wide Web Conference* (2018).

[46] Sarah Wiegreffe and Yuval Pinter. 2019. Attention is not not Explanation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China, 11–20. https://doi.org/10.18653/v1/D19-1002

[47] Chuhan Wu, Fangzhao Wu, Mingxiao An, Jianqiang Huang, Yongfeng Huang, and Xing Xie. 2019. Neural News Recommendation with Attentive Multi-View Learning. In *International Joint Conference on Artificial Intelligence (IJCAI 2019)*.

[48] Chuhan Wu, Fangzhao Wu, Mingxiao An, Jianqiang Huang, Yongfeng Huang, and Xing Xie. 2019. NPA: Neural News Recommendation with Personalized Attention. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*.

[49] Chuhan Wu, Fangzhao Wu, Suyu Ge, Tao Qi, Yongfeng Huang, and Xing Xie. 2019. Neural News Recommendation with Multi-Head Self-Attention. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 6389–6394.

[50] Chuhan Wu, Fangzhao Wu, Yongfeng Huang, and Xing Xie. 2023. Personalized News Recommendation: Methods and Challenges. *ACM Transactions on Information Systems* 41, 1, Article 24 (jan 2023), 50 pages. https://doi.org/10.1145/3530257

[51] Fangzhao Wu, Ying Qiao, Jiun-Hung Chen, Chuhan Wu, Tao Qi, Jianxun Lian, Danyang Liu, Xing Xie, Jianfeng Gao, Winnie Wu, and Ming Zhou. 2020. MIND: A Large-scale Dataset for News Recommendation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 3597–3606.

[52] Yao Wu and Martin Ester. 2015. FLAME: A Probabilistic Model Combining Aspect Based Opinion Mining and Collaborative Filtering. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining, WSDM 2015*. ACM, 199–208.

[53] Jin Xie, Fuxi Zhu, Xuefei Li, Sheng Huang, and Shichao Liu. 2021. Attentive preference personalized recommendation with sentence-level explanations. *Neurocomputing* 426 (2021), 235–247. https://doi.org/10.1016/j.neucom.2020.10.041

[54] Boming Yang, Dairui Liu, Toyotaro Suzumura, Ruihai Dong, and Irene Li. 2023. Going Beyond Local: Global Graph-Enhanced Personalized News Recommendations. In *Proceedings of the 17th ACM Conference on Recommender Systems, RecSys 2023, Singapore, Singapore, September 18-22, 2023*, Jie Zhang, Li Chen, Shlomo Berkovsky, Min Zhang, Tommaso Di Noia, Justin Basilico, Luiz Pizzato, and Yang Song (Eds.). ACM, 24–34. https://doi.org/10.1145/3604915.3608801

[55] Mingwei Zhang, Guiping Wang, Lanlan Ren, Jianxin Li, Ke Deng, and Bin Zhang. 2022. METoNR: A meta explanation triplet oriented news recommendation model. *Knowl. Based Syst.* 238 (2022), 107922.

[56] Yongfeng Zhang and Xu Chen. 2020. Explainable Recommendation: A Survey and New Perspectives. *Found. Trends Inf. Retr.* 14 (2020), 1–101.

[57] Kaiqi Zhao, Gao Cong, Quan Yuan, and Kenny Q. Zhu. 2015. SAR: A sentiment-aspect-region model for user preference analysis in geo-tagged reviews. In *31st IEEE International Conference on Data Engineering, ICDE 2015, Seoul, South Korea, April 13-17, 2015*. IEEE Computer Society, 675–686.

[58] Qiannan Zhu, Xiaofei Zhou, Zeliang Song, Jianlong Tan, and Li Guo. 2019. DAN: Deep Attention Neural Network for News Recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

A Experimental Environment

Our experiments were conducted on a High-Performance Computing cluster running the Linux operating system. We used PyTorch 1.8.0 as the backend. The GPU type is NVIDIA Tesla V100 and A100 with 32GB and 40GB GPU memory respectively. We ran each experiment 5 times with fixed random seeds, each in a single thread.

B Global Topic Examples

Table 7. Examples of topics identified by our approach, in terms of extracted topic descriptors, with coherence scores calculated via C_{NPMI} and C_{W2V}.

Topic Descriptor	C_{NPMI}	C_{W2V}
dog cat terrier canine kennel pup retriever dogs pups canines	0.5131	0.6544
song songs album music guitar piano sound soundtrack audio nice	0.3526	0.5055
undergraduate admissions faculty universities graduate bachelor university students enrolled colleges	0.3546	0.6040
loan loans mortgages million mortgage river download borrowers lenders equity	0.3201	0.4386
pastas entrees salmon seafood appetizer mussels appetizers shrimp lobster oysters	0.5793	0.5952
oven ingredients bake recipes cooking baking protein crust butter recipe	0.5937	0.5214