**International Journal of Distributed Systems and Technologies**
Volume 15 • Issue 1

# A Hotspot-Aware Personalized News Recommendation Mechanism Based on DistilBERT-TC-MA

Qian He, Sichuan University of Media and Communications, China*
Ke Wang, Sichuan University of Media and Communications, China

## ABSTRACT

Aiming at the problems of existing news recommendation methods, such as inadequate exploration of the semantic information of news, neglecting potential hotspot features of news, and challenging the balance between user preferences and hotspot features, a hotspot-aware personalized news recommendation model (DistilBERT-TC-MA) is suggested, which integrates the distilled version of BERT (DistilBERT), text convolutional neural network (TextCNN), and multilayer attention (MA). First, it takes full advantage of DistilBERT, TextCNN, and self-attention mechanism to achieve news encoding. Following this, representations of trending news are dynamically aggregated using the attention mechanism, while user preferences are mined utilizing user click history. Finally, in order to successfully accomplish the click prediction of candidate news, the hotspot features, user preferences, and candidate news are ultimately combined using a click predictor. The experimental results of the suggested DistilBERT-TC-MA model on MIND dataset are better than several other advanced methods.

## KEYWORDS

Distilbert, Hot Feature Extraction, Multilayer Attentional Interaction, News Recommendation, Textcnn, User Preference Mining

The current trend shows that users prefer browsing news through connected devices. However, as the quantity of news continues to grow, users often feel overwhelmed in the ocean of information, making it challenging to swiftly and accurately discover news content that aligns with their preferences and needs. Consequently, achieving personalized news recommendations is a crucial solution to enhance the user's news reading experience. (Qin, J., & Lu, P., 2020; Tian, X., et al., 2021; Jiang, S., Zhao, H., & Guo, J., 2021).

The task of news recommendation is to deliver news resources to users that they might find interesting from a plethora of news information, effectively filtering out irrelevant news, and meeting diverse user demands for news information to enhance the quality of their reading experience. Many current research efforts (Talha et al., 2023) analyze user historical behavior, click records, and other personalized information to establish user interest models, uncover user interests, and more precisely

DOI: 10.4018/IJDST.339565 *Corresponding Author

predict user interests in different topics and keywords. Initially, many works adopt machine learning-based methods, such as (Liu et al., 2010), who design a model based on user click behavior. It utilizes a Bayesian method to learn user's interest representation based on the distribution of user clicks on articles of different news topics. However, such methods often require manual annotation of features and struggle to deeply explore the semantic information. Therefore, CNN, LSTM, and attention networks gradually become employed. For instance, Zhu et al. (2019) utilize two parallel max-pooling CNN to explore implicit feature representations. However, traditional CNNs face challenges in handling long-distance text interactions in news modeling. (Wu et al., 2019) propose a method based on a multi-head self-attention mechanism, enhancing the representation capability of news features. However, large pre-trained language models demonstrate stronger modeling capabilities in capturing complex contextual information in news text, and they are widely applied in personalized news recommendation. For example, (Zhang et al., 2021) incorporate title into the BERT, simultaneously capturing word-level and news-level representation to enhance text representation. It can be observed that the introduction of pre-trained models makes news recommendation more accurate and effective.

However, this entirely user-interest-based recommendation approach may lead to recommendations that are overly similar or even repetitive to the content users have clicked on, resulting in the creation of information bubbles. In fact, trending news significantly influences users. For instance, users interested in sports and military topics might also find daily trending news appealing. Therefore, user click intentions are not only related to individual interests but may also be associated with certain trending news. The current strategy for recommending trending news is to recommend the same trending news to all users, but this approach struggles to address the differentiation in user interests. Hence, some studies propose effective solutions. For example, (Jonnalagedda & Gauch et al., 2013) propose a hybrid method that combines news popularity and user profiles to recommend personalized popular news. They use cosine similarity (Harbouche et al., 2023) to calculate the match between a piece of news and popularity or individual interest. However, these studies often employ traditional machine learning methods, making it challenging to deeply explore potential hot point features and flexibly balance user interests and hot point features. In recent years, some research uses deep learning techniques to explore deep semantic features and learns user's interest representations through user's click history, effectively enhancing the expressive capabilities of news representation and user interest representation. However, there is currently limited work considering the use of deep learning to extract trending features and balance user interests and trending features.

Based on this, this paper designs a Hotspot-Aware News Recommendation Model using DistilBERT-TC-MA. To alleviate the model's size and parameter count for improved efficiency, and considering the challenges of effectively mining deep semantic information in news modeling with current deep learning methods, the model utilizes DistilBERT, TextCNN, and SelfAttention neural network models to effectively process news headline features and deeply explore the rich semantic information in news. In order to effectively balance personalized recommendations and trending news, enhancing both the personalization and diversity of recommendations, the paper proposes the use of attention networks in the Hotspot Information Mining and User Preference Mining modules. These modules interactively incorporate candidate news, hotspots, and user information. In summary, the paper makes the following contributions:

1) We design a Hotspot-Aware News Recommendation Model based on DistilBERT-TC-MA, aiming to explore the rich semantic features. The model effectively balances personalized recommendations and trending news, thereby enhancing the personalization and diversity of recommendations.
2) By adopting DistilBERT, TextCNN, and SelfAttention neural network models, we maintain model efficiency while deeply mining rich semantic information.
3) This paper introduces the Hotspot Information Mining and User Preference Mining modules. Through the incorporation of attention networks, these modules effectively facilitate interaction

between candidate news, hotspot information, and user preferences. This enhances the level of personalization in recommendations while maintaining appropriate attention to trending news.

The following content is organized as follows: Chapter 2 introduces related work, Chapter 3 presents the proposed methods, Chapter 4 covers experiments and analysis, and Chapter 5 concludes the work.

## RELATED WORK

### News Ranking

We rank potential news based on the modeling of user and news interactions to achieve personalized displays according to the user's individual interests. Relevance-based news ranking methods (Wu et al., 2023) typically involve sorting candidate news based on the user's personalized relevance. Numerous approaches straightforwardly evaluate the relevance between users and news by comparing their ultimate representations. For instance, (Goossen et al., 2011) gauge relevance by computing the cosine similarity of the CF-IDF feature vectors for users and news. Meanwhile, (Okura et al., 2017) predict relevance scores by utilizing the inner product between the representations of news and users. However, these methods have a potential issue—they tend to suggest news articles that are similar to those previously clicked by users. Therefore, some methods attempt to address this issue by striving to recommend content different from news previously clicked by the user. (Li et al., 2011) starts by ranking news articles according to their relevance to user interests. Subsequently, they refine the ranking list by incorporating factors such as news popularity and recency to generate the ultimate recommended list. Hence, the paper modifies the ranking approach by incorporating elements like news novelty, popularity, and timeliness. This aims to boost recommendation diversity and mitigate the identified issue. This is done without compromising the user experience during the process of exploring potential user interests.

### Preference-Based News Recommendation

Personalized recommendation systems have found extensive applications in the realm of the internet, aiming to provide users with the most relevant and valuable content based on their personal preferences and needs. In the domain of news recommendation, preference-based recommendation algorithms analyze users' historical behaviors, browsing records, and preferences to establish a model of user preferences. Consequently, these algorithms recommend news content that aligns with the user's preferences.

Preference-based news recommendation commonly employs collaborative filtering (CF). CF is a common approach that analyzes the similarity of behaviors among users. It identifies other users with similar preferences to the target user by leveraging their behaviors for recommendations (Dong et al., 2016). The CF method is employed for user rating prediction. In the calculation of user similarity, parameters related to news hotspots are incorporated to enhance the correlation coefficient formula, addressing the sparsity issue in the user rating matrix data. However, early collaborative filtering algorithms often only use descriptive features (such as ID) to construct user and news embeddings, neglecting the rich semantic information between user and news interactions.

Therefore, (Sun, C et al., 2021) propose a method based on SVD and K-means. It utilizes CF method to explore users' latent preferences. Liu & Liu (2022) employ technologies such as word vectors and neural topic models to extract semantic representation from news, acquire vectors representing news features and amalgamate all semantic features to form the user's preference vector. By generating a candidate set of news that users might find interesting, the effectiveness of news recommendations has been significantly enhanced. (Okura et al., 2017) use the similarity between news to learn news embedding representations and introduce topic information to enrich news modeling. While achieving

certain effectiveness, it struggles to learn deep semantic information. Therefore, deep learning techniques are gradually introduced into news recommendation systems. For instance, (Li, J et al., 2022) employ a multi-head attention approach to learn user preferences. (Meng, L., & Shi, C., 2020) propose a CalDN model, which through a bidirectional attention recurrent network, it captures various aspects of user reading preferences and provides personalized reading suggestions. (Tran et al., 2023) introduce the CupMar method, utilizing multiple attribute features in the news encoder to obtain rich news representations through neural network layers. Within the user configuration encoder, an analysis of the user's browsed news is conducted to deduce both long-term and short-term interests. This approach effectively captures the dynamics of user interests over time. (Wu et al., 2019) learn user short and long-term preferences from user click history and user ID embedding vectors. They use two methods to integrate and recommend news based on user preferences.

However, most deep learning methods require training from scratch and may struggle to achieve good generalization in scenarios with limited samples. Therefore, many research works are gradually adopting pre-training model approaches. (Huang et al., 2022) introduce an adaptive transformer model, aiming to capture profound interactions between users and candidate news by effectively combining historical clicked news and candidate news to unveil their inherent correlations. However, the majority of existing Pre-trained Language Models (PLMs) boast large sizes, housing hundreds of millions of parameters. Given the need for low-latency services catering to millions of users in many online news applications, (Wu et al., 2021) propose a teacher-student joint learning and distillation framework. This involves integrating the gradients of the teacher model into the updates of the student model, facilitating the more efficient transfer of valuable knowledge acquired by the teacher model. Hence, this paper considers the advantages of deep learning and knowledge distillation in news modeling. It utilizes DistilBERT to obtain embedding vectors for news titles, followed by deep learning methods to extract deep semantic information, effectively improving the recommendation performance.

## Hotspot-Based News Recommendation

While preference-based news recommendation algorithms can provide users with relatively accurate personalized recommendations, they often overlook the factors of news trends and real-time relevance. Due to the rapid update pace of news content, users may miss out on some important trending news, limiting the effectiveness of recommendation algorithms. Therefore, introducing information about news trends is necessary and beneficial for personalized news recommendations.

(Jonnalagedda, N., et al., 2016) collect numerous popular news from the Twitter website. The similarity between candidate news and all popular news is computed, with the cumulative addition of these similarity values serving as the popularity weight for the candidate news, ultimately predicting its level of popularity. However, news popularity may be related to other factors. (Tiwari. S et al., 2018) combine user preferences with popular news at their current location. They obtain user mobile location information and directly retrieve nearby hot news topics through the Twitter API (Application Program Interface). (Naterajans et al., 2016) consider article popularity, trending patterns, user profiles, and location preferences as necessary influencing factors. Therefore, they calculate the cosine similarity between news articles and popular tweets to obtain popular news features. Additionally, they calculate the TF-IDF values between news articles and tweets with trending patterns to obtain trending news. Finally, they design a manually adjustable parameter to help users choose the proportion of recommended popular news and trending news. Although this method can effectively obtain features of popular news, acquiring deep semantic information about hot news is crucial. (Yi. B et al. 2022) propose CDBE, which utilizes CNN, TextCNN, LSTM, and Bi-LSTM to predict hot news. However, there is a lack of interaction between candidate news, hot information, and user information, making it challenging to balance personalized recommendations while maintaining appropriate attention to hot news, resulting in a reduction in the diversity of the recommendation system. (F. Xu et al., 2022) introduce based on Multiple Perspectives (BTEC). This method uses the BERT model to vectorize

the body, title, events, and hotspots. It integrates these four perspectives to enhance the effectiveness of news recommendations.

Based on this, a hotspot-aware news recommendation model is proposed, named DistilBERT-TC-MA. It comprises four components: news encoder, hotspot information mining, user preference mining, and click predictor. Utilizing DistilBERT, TextCNN, and SelfAttention neural network models, it effectively processes news title features and deeply mines rich semantic information from the news. To balance personalized recommendations and hotspot news, and to enhance both personalization and diversity of recommendations, modules for hotspot information mining and user preference mining are designed. These modules use attention networks to interactively engage candidate news and hotspots with user information, effectively improving prediction performance.

## METHOD

### Problem Definition

The news title is a decisive factor influencing the user's clicking choice; therefore, this paper employs the news title as the input to the model. For a user $u$, their news click history is represented as $\{t_1^u, t_2^u, ...t_{N_u}^u\}$, where $t_i^u (i = 1,...N_u)$ is the title of the *i*-th news clicked by user $u$, and $N_u$ is the total number of news clicked by user $u$. Hotspot news within the past period V is denoted as $\{p_1, p_2, \ldots p_M\}$, where $p_j (j = 1,...M)$ represents the title of the *j*-th hotspot news during that period, and $M$ is the total number of hotspot news during that time. Each click history news title $t$ or hotspot news title $p$ consists of a series of words $[w_1, w_2, ...w_n]$, where $n$ is the number of words in the title. The model's prediction objective is to predict whether user $u$ will click on a candidate news $x$, previously unseen, based on their click history and hotspot news from the past period. This involves predicting the recommendation score for the news.

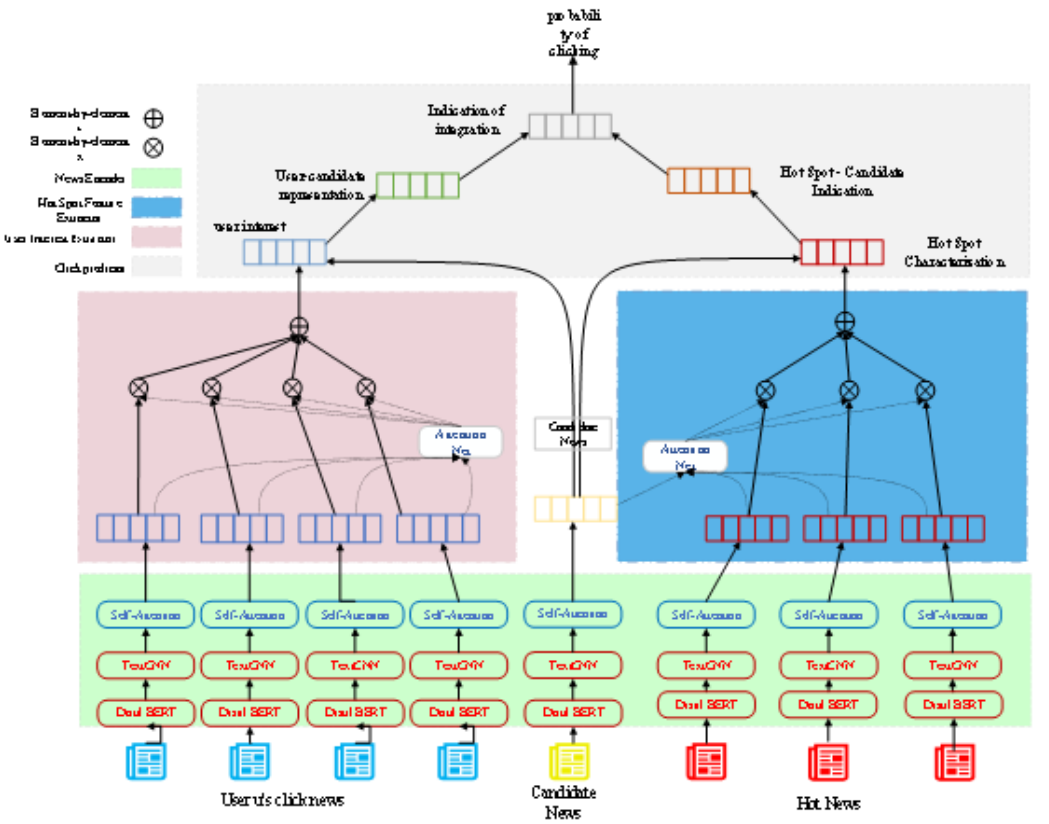
### DistilBERT-TC-MA

As shown in Figure 1, DistilBERT-TC-MA has four main parts: news encoder, hotspot information mining, user preference mining, and click predictor. The model takes as input the recent hot news, the user's history of clicked, and the candidate news. First is the news encoder, which utilizes DistilBERT to obtain word embeddings for the news, introducing TextCNN for a more flexible treatment of semantic features at different levels in the news title. This helps capture local information and patterns in the news title. SelfAttention networks are introduced to weight the features extracted by TextCNN, enhancing sensitivity to information. Next is the hotspot information mining module and user preference extraction module. By introducing attention networks, they effectively facilitate interaction between candidate news, hotspot information, and user information, balancing user preferences and hotspot features. Finally, the click prediction module predicts the probability of the user clicking on the candidate news.

### News Encoder

The user's decision to view news is highly dependent on the news headline. In the DistilBERT-TC-MA framework, the news encoder gains an understanding of news representation by analyzing news headlines. The news encoder employs a three-layer architecture, where DistilBERT is utilized in the first layer, TextCNN in the second, and the third layer incorporates the self-attention mechanism. DistilBERT is applied for its proficiency in capturing global contextual information, while TextCNN excels at extracting locally continuous text features. Despite TextCNN's effectiveness in capturing locally continuous text features, it may struggle to acquire discontinuous text features. To overcome this limitation, the model introduces a self-attention module in the third layer. This module enhances the accuracy of news representation by effectively capturing both c

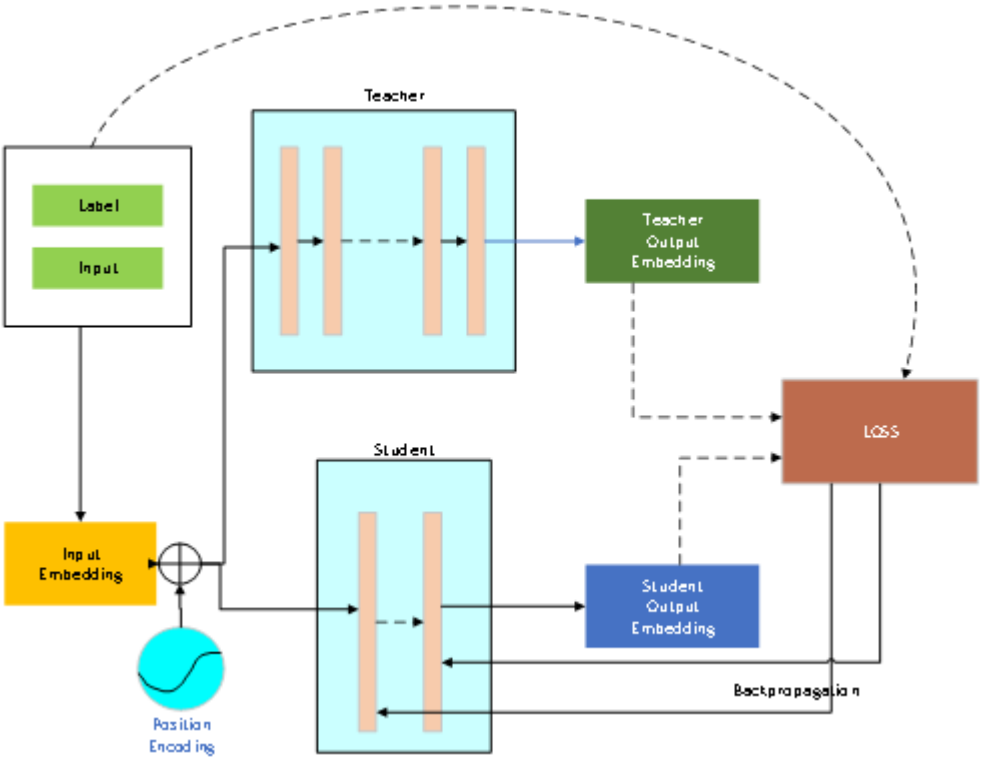
**Figure 1. Architecture of the DistilBERT-TC-MA**




The first layer consists of the DistilBERT word vector model. Utilizing the lightweight DistilBERT model to output character-level dynamic word vectors ensures both the model's training speed and the quality of the word vectors produced. The DistilBERT word vector model converts word sequences in news headlines into dynamic character-level word vectors with a length of n. The word sequences of clicked history news and hot news headlines are denoted as $t = w_{1:n} = [w_1, w_2, ...w_n]$ and $p = w_{1:n} = [w_1, w_2, ...w_n]$, respectively. These sequences are converted into a vector sequence, $w_{1:n} = [w_1, w_2, ...w_n] \in R^{n \times d}$, through pre-training embeddings in a large corpus, where d=768 represents the dimension of the word embedding vector. The intricacies of the DistilBERT model architecture are illustrated in Figure 2.

The second layer is TextCNN. CNN (T. Kim, Y. Na, & S. Park., 2023) includes a convolutional layer and a pooling layer, which achieves the text classification task by extracting local features. TextCNN is a variant of CNN, and its main goal is to derive distinct local features through the specification of distinct filter kernel sizes. By defining various filter kernel sizes, TextCNN attempts to derive a variety of local features in order to obtain more realistic and representative features. The model is utilized predominantly in the classification of text. The middle portion of Figure 1 depicts the schematic structure of the model. and the corresponding filter convolution kernel sizes are 2, 3, and 4, respectively, as follows:

**Figure 2. Architecture of DistilBERT**




(1) A convolutional layer comprising three filters with convolutional kernel sizes of 2, 3, and 4 forms the basis of the TextCNN model. These filters are capable of extracting features at varying degrees. The formula for its calculation is:

$$h_i = f\left(\sum_{x=1}^{3}\sum_{y=1}^{3} w_{i(x,y)} \times c_{(x,y)} + b_i\right) \tag{1}$$

where, the activation function is denoted by $f$, $W_{i(x,y)}$ represents the weight of the $i$-th node in the output matrix corresponding to the filter input node $(x,y)$, $c_{(x,y)}$ represents the value of node $(x,y)$ in the filter, and $b_i$ represents the bias term for the $i$-th node. The calculation of the convolutional layer result $h_i$ is achieved by employing three filters with convolutional kernels set to 2, 3, and 4 for localized feature extraction.

(2) To accomplish the goal of dimensionality reduction, the pooling layer simultaneously reduces the size of feature vectors and network parameters while enabling the model to focus more on the precise location of non-features. This is achieved by pooling the output $h_i$ of the convolution operation and using the maximum pooling method.
(3) The fusion layer will process the output of the pooling layer in order to combine the features obtained from the three pooling layers into a single, representative feature vector, denoted as

$C = [C_1, C_2, ...C_m]$, where m represents the number of sliding windows, which depends on the sliding step size s.

The third layer is the self-attention network. To make up for the limitation of TextCNN in dealing with long sequence text data, it is utilized to acquire the discontinuous features of news text data. Meanwhile, by assigning different weights to news text features, important features are highlighted, and irrelevant features are discarded to save computational resources. In addition, some long headline sentences are composed of multiple components, and in order to better represent the overall semantics of the headline, it is imperative to focus on the $r$ different parts of the headline. $C$ is the self-attention network's input, and the output is its weight matrix $A$, which is calculated as shown in Equation (2):

$$A = soft\max(W_{s_2} \tanh(W_{s_1} C^T)) \tag{2}$$

where $W_{s_1} \in R^{N_f \times d_a}$ and $W_{s_2} \in R^{d_a \times r}$ denote the parameters $d_a$ and $r$ in the self-attention network, respectively, are hyperparameters that can be set. Then the matrix $A$ and the overall word representation $C$ are multiplied to obtain the weighted sum of the $r$ part, which is the final representation vector $e$ for each news headline. The calculation method is as follows:

$$E = AC \tag{3}$$

$$e = flatten(E) \tag{4}$$

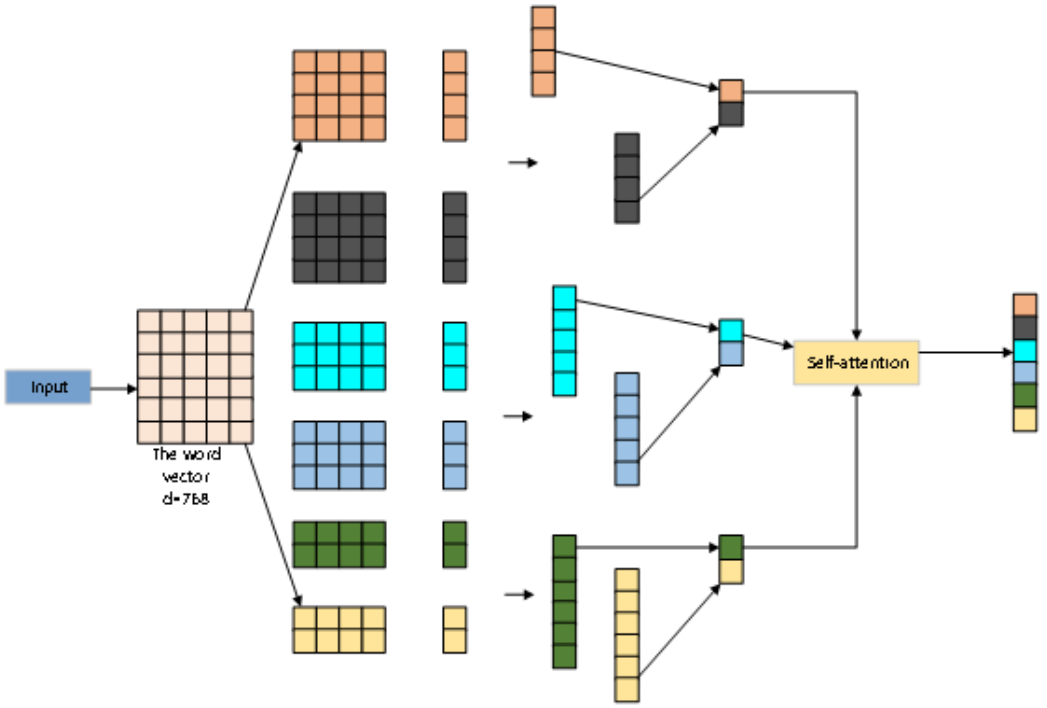
where $flatten(.)$ denotes the flattening operation that transforms the matrix $E \in R^{r \times N_f}$ into the vector $e \in R^{\delta}$, $\delta = rN_f$. The final news representation vector $e$ will be used as the input for the following hotspot information mining and user preference mining.

## Hotspot Information Mining

To assess the potential hotness of candidate news, the hotspot factor is introduced. By extracting features related to hotspots and calculating the degree of hotness of a candidate news, it is possible to judge its potential hotspot potential. To achieve this goal, a hotspot information miner is designed, which can perform hotspot information extraction by learning the feature representations of hotspot news over a period of time.

For a given hot news as $\{p_1, p_2, ...p_M\}$, its vector representation is represented as $\{e(p_1), e(p_2), ...e(p_M)\}$. However, the hot news topics are multifaceted, and in order to characterize different aspects of the hot information, an attention mechanism is used to select the hot news that is relevant to the current candidate news. Specifically, for a hot news $p_j$ and a candidate news $x$, the candidate news representation $e(x)$ is taken as a query vector, and the hot news representation is represented as $e(p_j)$ The hot news miner learns the attention weight $\alpha_j$ of the jth hot news by computing the similarity across the hot news query vector and the hot news representation $e(p_j)$. The calculation method is shown in equation (5):

**Figure 3. The analytical framework based on TextCNN and Attention**




$$\alpha_j = soft\max(e\left(p_j\right)^T e(x)) = \frac{\exp(e(p_j)^T e(x))}{\sum_{j=1}^{M} \exp(e(p_j)^T e(x))} \tag{5}$$

For candidate news $x$, the final two-point feature representation $h$ is the attention-weighted sum of all hot news representations, computed as shown in (6):

$$h = \sum_{j=1}^{M} \alpha_j e(p_j) \tag{6}$$

## User Preference Mining

User preference mining is a methodology employed to acquire knowledge of user preference representations by analyzing the user's past click activity. Users' preferences are frequently complex, so in order to effectively represent different aspects of users' preferences, an attention mechanism is introduced, which is capable of selecting the click history associated with the current candidate news.

In this way, the suggested model enhances its capability to encompass a wide range of user preferences, thereby augmenting the accuracy and effectiveness of user preference extraction. For the clicked news $t_i^m$ and the current candidate news $x$ of user $u$, the preference query vector for the candidate news is denoted by $e(x)$, and the clicked news is denoted by. The user preference mining

learns the attentional weight $\beta_i$ of the ith clicked news by computing the similarity within the preference query vector and the clicked news, which is calculated as follows:

$$\beta_j = soft\max(e(t_i^u)^T e(x)) = \frac{\exp(e(t_i^u)^T e(x))}{\sum_{i=1}^{N_u} \exp(e(t_i^u)^T e(x))} \tag{7}$$

For candidate news $x$, the attention-weighted sum of all clicked news is denoted by $i_u$, which is calculated as follows:

$$i_u = \sum_{i=1}^{N_u} \beta_i e(t_i^u) \tag{8}$$

## Click Predictor

Using a click predictor, one can estimate the likelihood that a user will click on each prospective news item. The click predictor must take into account three vectors: the representation of candidate news, the representation of user preferences, and the representation of hotspot features. To generate the hotspot-candidate vector, the candidate news representation is initially spliced with the hotspot feature representation and the user preference representation correspondingly to obtain the hotspot-candidate vector $o_{h,x}$ and the user-candidate vector $o_{i_u,x}$. The calculation is as follows:

$$o_{h,x} = \tanh(w_1(h \oplus e(x)) + b_1) \tag{9}$$

$$o_{i_u,x} = \tanh(w_2(i_u \oplus e(x)) + b_2) \tag{10}$$

Then, the method of calculating the likelihood score of a user clicking on that candidate news is as follows:

$$o_{i_u,h,x} = fuse(o_{i_u}, o_{h,x}) \tag{11}$$

$$\hat{y}_{u,x} = sigmoid(wo_{i_u,h,x} + b) \tag{12}$$

The function $fuse(.,.)$ represents the fusion of the two vectors by the fusion method, and finally, the sigmoid nonlinear variation is employed as the activation function to predict the probability of user $u$ clicking on the current news candidate $x$, $\hat{y}_{u,x}$, where $w$ and $b$ represent learnable parameters. For function $fuse(.,.)$, a splicing method is used to splice the hotspot-candidate vector $o_{h,x}$ and the user-candidate vector $o_{i_u,x}$. After that, the result is fed into the linear layer and the tanh nonlinear layer in turn, where $w_6$ and $b_6$ are the learnable parameters, which are calculated as follows:

$$o_{i_u,h,x} = \tanh(w_6(o_{i_u,x} \oplus o_{h,x}) + b_6) \tag{13}$$

### Loss Function and Penalty Term

During training, the samples are divided into positive and negative samples, where the user's most recent selected news is taken as a positive sample, and the news that appeared but not clicked in the same time period is taken as a negative sample. Then, a training sample is denoted as $s = (\{t_1^u, t_2^u, ...t_{N_u}^u\}, \{p_1, p_2, ...p_M\}, x, y_{u,x})$, where $\{t_1^u, t_2^u, ...t_{N_u}^u\}$ represents the set of click history of user $u$, $\{p_1, p_2, ..., p_M\}$ represents the set of news in the same time period, and $x$ is the current candidate news. Each positive sample has a label of 1, which implies that $y_{u,x} = 1$; $y_{u,x} = 0$ represents each negative sample has a label of 0. Finally, every other sample is given a probability on its candidate news $x$, i.e. $\hat{y}_{u,x}$. The method of minimizing the negative log-likelihood function is utilized for training the model with the following formula:

$$L = -(\sum_{s \in S^+} y_{u,x} \log \hat{y}_{u,x} + \sum_{s \in S^-} (1 - y_{u,x}) \log(1 - \hat{y}_{u,x})) \tag{14}$$

where $S^+$ and $S^-$ denote the positive and negative sample sets, correspondingly.

$L_2$ regularization is used to decrease the complexity of the model while avoiding overfitting situations and improving the model's generalization ability l. In order to avoid the self-attention network in the news encoder which often gives the same weight to $r$ different parts of the news headline, citing the Reference (Lin, Z., et al., 2017), the penalty term $P$, is used, which is calculated as follows:

$$P = \left( \left\| AA^T - I \right\|_F \right)^2 \tag{15}$$

where $A$ is the weight matrix in the self-attention method, $I$ denotes the unit matrix, and $\|.\|_F$ represents the Frobenius Norm for a given proof. The coefficients of the $L_2$ regular term and the coefficients of the $P$ penalty term are set to $l_c$, $p_c$ and they are minimized together with the original loss function $L$.

## EXPERIMENTS

We make several experiments to investigate the effectiveness of DistilBERT-TC-MA. By comparing it with other baselines to assess prediction errors, we aim to address the following three questions:

1) How does the performance of the DistilBERT-TC-MA method compare to state-of-the-art approaches?
2) What is the impact of the crucial model designs in DistilBERT-TC-MA on the experiments?
3) How do different methods used in the designed modules affect the experiments?

### Details of the Experimental Environment and Dataset

The MIND dataset is collected from the anonymous behavioral logs of the Microsoft News website, comprising a large-scale news recommendation dataset (Wu, F et al., 2020). To construct the MIND dataset, one million users who had at least five news clicks within a six-week period from October 12 to November 22, 2019, were randomly sampled. MIND consists of 161,031 news articles, 1,000,000 users,

**Table 1. Configuration Settings**

| Environment | Configuration Details |
|---|---|
| IDE Parameters | Anaconda3-Windows-x86_64 |
| GPU | NVIDIA GeForce RTX 3090 Ti 24GB |
| Hard disk | 1T |
| CPU | Intel CoreI i7-8750H@2.20GHz |
| Programming language | Python 3.10 |
| Development Framework | TensorFlow 1.14.0 |

and 24,155,470 behavior logs. Each news article includes attributes such as news ID, title, abstract, body, category, and entities. Each behavior log contains information about click events, non-click events, timestamps, user identifiers, and the user's historical clicked news prior to that behavior log.

The experiments are organized into weekly time windows, and user click histories, hot news, and candidate news are selected based on the following rules: Users and candidate news come from users and their click records during an even-numbered week. Users click histories are derived from the user's click history in the preceding odd-numbered week. Hot news sets are also from news released in the preceding odd-numbered week. The hotness of news is determined by the total click count, and a hot news list is formed by sorting news in descending order of total click counts.

A sample contains n clicked news ($n \geq 0$) by the user, the top m hot news ($m \geq 0$) in the same time period of the clicked news, where the hot news is ranked in ascending order by the total number of clicks, a candidate news and a label of 1 or 0, as shown in Table 2. When a user truly clicks on a candidate news, the label is represented by 1, i.e., this sample is a positive sample, and conversely, when a user does not truly click on the candidate news, the label is indicated by 0, i.e., this sample is a negative sample. In this dataset, the training set and test set are divided with a ratio of 5:3.

## Baselines

**CalDN** (Meng, L., & Shi, C., 2020) captures user reading preferences from various aspects, including the environment, breaking news, and news content information, using a bidirectional attention recurrent network, providing personalized reading recommendations.

**MINER** (Li, J., et al., 2022) employs a multi-attention approach to learn and express user preferences. Additionally, it designs a category-aware attention mechanism, incorporating news category information as an explicit interest signal into the attention mechanism.

**CDBE** (Yi. B et al. 2022) predicts hot news using CNN, TextCNN, LSTM, and Bi-LSTM. However, there is a lack of interaction between candidate news, hot news information, and user information. This makes it challenging to maintain a personalized level while appropriately focusing on hot news, leading to a reduction in the diversity of the recommendation system.

**Table 2. Composition of a sample**

| Category | Number |
|---|---|
| Clicked News | n |
| Hot News | m |
| Candidate News | 1 |
| Label | 1/0 |

**BTEC** (F. Xu et al., 2022) utilizes the BERT model to vectorize news content, including the body, title, events, and hot topics. It integrates these vectors from four perspectives to enhance the effectiveness of news recommendations.

## Evaluation

To verify the efficacy of the suggested recommendation method, the following common evaluation metrics for recommender systems are used to measure the effectiveness of the method.

AUC, representing the area under the ROC curve, is primarily employed for assessing the accuracy of binary classification models and indicating the model's effectiveness in recommending items. Assuming M and N are the counts of positive and negative samples, respectively, and "$rank_i$" signifies the ranking of the predicted probability value for the *i*-th positive sample, the AUC calculation is as follows:

$$AUC = \frac{\sum_{i=1}^{M} rank_i - \frac{M(1+M)}{2}}{M \times N} \tag{16}$$

MRR, this indicator represents the inverse average of the rank of items in multiple recommendation lists, which reflects the position of the first relevant item in the recommendation list:

$$MRR = \frac{1}{Q}\sum_{i=1}^{Q}\frac{1}{rank_i} \tag{17}$$

NDCG@K stands for Normalized Discounted Cumulative Gain, which reflects the quality of the result in terms of the accuracy of a set of the entire sorted list. NDCG@K is calculated in the following:

$$NDCG@K = \frac{DCG@K}{IDCG@K} \tag{18}$$

$$IDCG@K = \sum_{i=1}^{K_{rel}} \frac{2^{r(i)} - 1}{\log_2(i+1)} \tag{19}$$

$$DCG@K = \sum_{i=1}^{K} \frac{2^{r(i)} - 1}{\log_2(i+1)} \tag{20}$$

where DCG@K indicates the discounted cumulative gain and IDCG@K represents the normalized DCG@K. $K$ indicates the discounted cumulative gain. Assuming that $r(i)$ denotes whether the ith item in the list is relevant or not, and has only two values, 0 and 1, then DCG@K is calculated as in (20). $K_{rel}$ represents the $K_{rel}$ most relevant items in the actual sorted list, then IDCG@K is calculated as in (18).

## Comparing the Results With Mainstream Methods RQ1

To answer the question 1, the proposed DistilBERT-TC-MA model is compared with several mainstream news recommendation methods serving as baselines. The test results are presented in Table 3.

It can be observed that MINER, utilizing a multi-head attention mechanism, outperforms CalDN, which employs bidirectional attention recurrent networks. CDBE, incorporating various neural networks such as CNN and LSTM, demonstrates an effective improvement in recommendation performance. BTEC, leveraging the pre-trained BERT model, offers a more comprehensive understanding of semantic information in text, including news titles and summaries, beyond simple keyword matching. However, considering the substantial parameter size of the BERT model and the effectiveness of various neural network fusions, we propose the DistilBERT-TC-MA model, employing DistilBERT, TextCNN, and self-attention networks. In comparison to MINER and CalDN, the self-attention network in our model captures discontinuous features in news text data. Compared to the BTEC model based on hot topics, our model uses the more concise and efficient DistilBERT model, outputting character-level dynamic vectors. In contrast to the CDBE model, our approach integrates user preference mining, combining attention networks with hot topic information related to candidate news.

## Ablation RQ2

To validate the effectiveness of each part in the DistilBERT-TC-MA, we conduct ablation experiments with individual modules. The testing time per sample, represented by "Time(s)," is measured.

As shown in Table 4, the absence of pre-trained models like DistilBERT may result in insufficient exploration of deep semantic information, affecting the model's comprehension of news content. Additionally, using this component increases the testing time per sample, attributed to the relatively higher algorithmic complexity of DistilBERT compared to other methods, which remains a significant factor affecting training speed. The absence of TextCNN might limit the efficient processing of news title features, and the lack of attention mechanisms like SelfAttention could lead to a performance

**Table 3. News recommendation results of different methods**

| Method | AUC (%) | MRR (%) | NDCG@5(%) | NDCG@10(%) |
|---|---|---|---|---|
| MINER | 56.13 | 25.54 | 26.67 | 34.73 |
| BTEC | 58.14 | 27.02 | 28.36 | 36.79 |
| CalDN | 60.67 | 28.68 | 31.34 | 38.37 |
| CDBE | 73.89 | 34.38 | 38.45 | 44.77 |
| **DistilBERT-TC-MA (ours)** | **75.14** | **36.28** | **40.17** | **46.59** |

**Table 4. Results of ablation experiments**

| Method | AUC (%) | MRR (%) | NDCG@5(%) | NDCG@10(%) | Time(s) |
|---|---|---|---|---|---|
| w/o hotspot information mining | 68.36 | 29.46 | 35.19 | 36.79 | 0.022 |
| w/o DistilBERT | 69.16 | 30.24 | 36.42 | 42.36 | 0.019 |
| w/o TextCNN | 72.46 | 34.54 | 37.85 | 41.03 | 0.020 |
| w/o self-attention network | 73.89 | 35.12 | 38.16 | 43.76 | 0.023 |
| **DistilBERT-TC-MA (ours)** | **75.14** | **36.28** | **40.17** | **46.59** | **0.025** |

drop when handling long sequences. The absence of these components may reduce the performance of the recommendation system, making it challenging to balance the demands of personalized recommendations and hot news. Therefore, the utilization of components such as DistilBERT, TextCNN, and SelfAttention enhances the news recommendation system's semantic understanding, feature extraction capability, and long sequence modeling capabilities, effectively improving the news recommendation performance of the proposed DistilBERT-TC-MA model.

## Component Analysis RQ3

To explore the effectiveness of using different approaches in different components, we conducted the following experiments.

### *Embedding Layer Model Comparison*

As shown in Table 5, DistilBERT exhibits significant advantages over traditional word embedding methods such as Word Embeddings, Word2Vec, and Glove. DistilBERT excels in understanding the contextual meaning of vocabulary, capturing deep semantic information, and adapting to different tasks through transfer learning. Its attention mechanism for global relationships enables it to focus more on the overall context when handling sequential data. Therefore, in this study, DistilBERT is employed to learn word embedding vectors.

### *Effectiveness of Hotspot Feature Extraction*

To assess the impact of the attention network in hotspot information mining on model performance, the model containing a self-attention network and the model using a direct averaging method without a self-attention network are experimentally compared, as depicted in Table 6. In contrast to the direct averaging method without an attention network, the experimental findings indicate that the hotspot information mining method containing a self-attention network improves by 5.36%, 3.72%, 4.78%, and 5.22% on AUC, MRR, NDCG@5, and NDCG@10, respectively. Incorporating a self-attention network enhances the ability to discern the significance levels of hotspot features, which helps enhance the model's characterization ability for hotspot news, thus effectively improving the news recommendation effect.

**Table 5. Comparison of the recommendation effect of DistilBERT and other three word embedding models**

| Model | AUC (%) | MRR (%) | NDCG@5(%) | NDCG@10(%) |
|---|---|---|---|---|
| Embedding | 69.19 | 32.27 | 35.07 | 41.03 |
| Word2vec | 71.86 | 33.46 | 36.79 | 41.54 |
| Glove | 73.67 | 24.28 | 38.27 | 44.25 |
| **DistilBERT (ours)** | **75.14** | **36.28** | **40.17** | **46.59** |

**Table 6. Effects of attention network in hotspot extractor**

| Method | AUC (%) | MRR (%) | NDCG@5(%) | NDCG@10(%) |
|---|---|---|---|---|
| Direct averaging without self-attentive networks | 69.78 | 32.56 | 35.39 | 41.37 |
| Contains self-attentive networks | 75.14 | 36.28 | 40.17 | 46.59 |

**Table 7. Comparison of 3 different fusion methods in the click predictor**

| Method | AUC (%) | MRR (%) | NDCG@5(%) | NDCG@10(%) |
|---|---|---|---|---|
| Summation | 71.35 | 32.16 | 36.85 | 43.16 |
| Multiplication | 73.46 | 34.37 | 37.89 | 43.97 |
| Concatenation (ours) | 75.14 | 36.28 | 40.17 | 46.59 |

### *Analysis of Fusion Methods for Click Predictor*

To enhance the fusion of user preferences and hotspot features, three different calculation methods were explored for fusing click prediction among them. These three calculation methods include summation, multiplication, and concatenation.

The summation operation involves element-wise addition of multiple feature maps to obtain the average value of the features, thereby reducing the impact of noise. The multiplication operation entails element-wise multiplication of multiple feature maps to enhance semantic information while preserving detailed information. Concatenation involves joining multiple feature maps along the depth dimension to achieve a more comprehensive feature representation. For instance, in the encoder-decoder architecture, concatenating lower-level features from the encoder with higher-level features from the decoder enhances the perception capabilities of the decoder.

Table 7 shows that compared with Summation and Multiplication, the Concatenation method improves at least 1.68%, 1.91%, 2.28% and 2.62% in AUC, MRR, NDCG@5 and NDCG@10. It can be seen that in the process of feature fusion, the Concatenation method can retain the useful information of useful various features more completely, which can help the neural network to dynamically fuse multiple features, so as to effectively improve the prediction effect.

## CONCLUSION

A news recommendation system model based on DistilBERT-TC-MA was designed to address the challenge faced by existing deep learning methods in effectively extracting deep semantic information in news modeling. By combining DistilBERT, TextCNN, and SelfAttention neural network models, this system can efficiently process news headline features, deeply mine rich semantic information, and enhance the understanding and modeling capabilities of news content. Additionally, the model addresses the issue of balancing personalized recommendations and hot news by introducing hot topic mining and user preference extraction modules, utilizing attention networks to interact with candidate news, hot topics, and user information. This design aims to improve the personalization and diversity of recommendations, allowing the recommendation system to better adapt to user interests and current trends, thereby enhancing user satisfaction and recommendation effectiveness. Overall, the innovation of this model lies in the comprehensive use of different neural network models and the improvement of the news recommendation system's performance through hot topic awareness and personalized mining, which has practical implications for enhancing user experience and information recommendation quality in real-world applications.

While this model has achieved certain effectiveness, the use of pre-trained models and self-attention networks increases the model's complexity, impacting recommendation efficiency. Additionally, there are other available news features that can be utilized for news modeling, such as news summaries and categories. Enriching news features can contribute to a more effective understanding of news representations. In our future work, we plan to use more efficient methods to enhance the overall performance of the model and consider incorporating additional available news features.

## REFERENCES

An, M., Wu, F., Wu, C., Zhang, K., Liu, Z., & Xie, X. (2019). Neural news recommendation with long- and short-term user representations. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics,* 336-345. doi:10.18653/v1/P19-1033

Bao, Y., Sun, Z., Qiang, Z., Lin, T., & Zheng, H. (2022). Hot News Prediction Method Based on Natural Language Processing Technology and Its Application. *Automatic Control and Computer Sciences*, *56*(1), 83–94. doi:10.3103/S0146411622010023

Chen, Y., & Zhao, C. G. (2022). Knowledge Graph and GNN-Based News Recommendation Algorithm With Edge Computing Support. *International Journal of Distributed Systems and Technologies*, 1–11.

Dogra, V., Singh, A., Verma, S., Kavita, Jhanjhi, N.Z., & Talib, M.A. (2021). Analyzing DistilBERT for Sentiment Classification of Banking Financial News. *Intelligent Computing and Innovation on Data Science*. 582.

Dong, Y., Liu, S., & Chai, J. (2016). Research of hybrid collaborative filtering algorithm based on news recommendation. *2016 9th international congress on image and signal processing, biomedical engineering and informatics (CISP-BMEI).*

Fan, L., Xu, F., Sun, Y. Y., & Zhou, H. (2022). News Recommendation Algorithm Based on Multiple Perspectives, *2022 International Conference on Computer Network, Electronic and Automation (ICCNEA),* 21-25. doi:10.1109/ICCNEA57056.2022.00016

Goossen, F., IJntema, W., Frasincar, F., Hogenboom, F., & Kaymak, U. (2011). News personalization using the CF-IDF semantic recommender. *Proceedings of the International Conference on Web Intelligence, Mining and Semantics.* doi:10.1145/1988688.1988701

Harbouche, K., Khentout, C., Djoudi, M., & Alti, A. (2023). Measuring Similarity of Educational Items Using Data on Learners' Performance and Behavioral Parameters: Application of New Models SCNN-Cosine and Fuzzy-Kappa. *Ingenierie des Systemes d'Information*, *28*(1), 1–11. doi:10.18280/isi.280101

Huang, J., Han, Z., Xu, H., & Liu, H. (2022). Adapted transformer network for news recommendation. *Neurocomputing*, *469*, 119–129. doi:10.1016/j.neucom.2021.10.049

Jiang, S., Zhao, H., & Guo, J. (2021). A Review: Personalized News Recommendation fusion with Topic Model. *2021 3rd International Academic Exchange Conference on Science and Technology Innovation (IAECST),* 612-617. doi:10.1109/IAECST54258.2021.9695777

Jonnalagedda, N., & Gauch, S. (2013). Personalized news recommendation using twitter. *2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT).*

Jonnalagedda, N., Gauch, S., Labille, K., & Alfarhood, S. (2016). Incorporating popularity in a personalized news recommender system. *PeerJ. Computer Science*, *2*, e63. doi:10.7717/peerj-cs.63

Kim, T., Na, Y., & Park, S. (2023). Multi-Head Convolutional Neural Network Compression based on High-Order Principal Component Analysis. *2023 International Conference on Electronics, Information, and Communication (ICEIC),* 1-4. doi:10.1109/ICEIC57457.2023.10049909

Li, J., Zhu, J., Bi, Q., Cai, G., Shang, L., Dong, Z., Jiang, X., & Liu, Q. (2022). *MINER: Multi-interest Matching Network for News Recommendation. In Proc of the Findings of the Association for Computational Linguistics*. ACL.

Li, L., Wang, D., Li, T., Knox, D., & Padmanabhan, B. (2011). Scene: a scalable two-stage personalized news recommendation system. *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval.* doi:10.1145/2009916.2009937

Lin, Z., Feng, M., Santos, C.N., Yu, M., Xiang, B., Zhou, B., & Bengio, Y. (2017). A Structured Self-attentive Sentence Embedding. *ArXiv, abs/1703.03130.*

Liu, J., Dolan, P., & Pedersen, E. R. (2010). Personalized news recommendation based on click behavior. *Proceedings of the 15th international conference on Intelligent user interfaces*. doi:10.1145/1719970.1719976

Liu, R., & Liu, L. (2022). Design of personalized recommendation method for entertainment news based on collaborative filtering algorithm. *2022 International Symposium on Advances in Informatics, Electronics and Education (ISAIEE).* doi:10.1109/ISAIEE57420.2022.00108

Meng, L., & Shi, C. (2020). A Context-aware Interest Drift Network for Session-based News Recommendations, *2020 IEEE 6th International Conference on Computer and Communications (ICCC),* 1967-1971.

Natarajan, S., & Moh, M. (2016). Recommending news based on hybrid user profile, popularity, trends, and location. *2016 International Conference on Collaboration Technologies and Systems (CTS)*, 204-211. doi:10.1109/CTS.2016.0050

Okura, S., Tagami, Y., Ono, S., & Tajima, A. (2017). Embedding-based news recommendation for millions of users. *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*. doi:10.1145/3097983.3098108

Qin, J., & Lu, P. (2020). Application of News Features in News Recommendation Methods: A Survey. In P. Qin, H. Wang, G. Sun, & Z. Lu (Eds.), *Data Science. ICPCSEE 2020. Communications in Computer and Information Science* (p. 1258). Springer. doi:10.1007/978-981-15-7984-4_9

Sun, C., Sun, G., Ding, Z., Liu, Q., & Ma, Z. (2021). A News Recommendation Algorithm Based on SVD and Improved K-means, *2021 International Conference on Networking, Communications and Information Technology (NetCIT),* 130-134. doi:10.1109/NetCIT54147.2021.00033

Talha, M. M., Khan, H. U., Iqbal, S., Alghobiri, M., Iqbal, T., & Fayyaz, M. (2023). Deep Learning in News Recommender Systems: A Comprehensive Survey, Challenges and Future Trends. *Neurocomputing*, *562*, 126881. doi:10.1016/j.neucom.2023.126881

Tian, X., Ding, Q., & Liao, Z. H. (2021). Survey on deep learning based news recommendation algorithm. *Journal of Frontiers of Computer Science and Technology*, *15*(6), 971–998.

Tiwari, S., Pangtey, M. S., & Kumar, S. (2018). Location Aware Personalized News Recommender System Based on Twitter Popularity. *International Conference on Computational Science and Its Applications*, 650-658. doi:10.1007/978-3-319-95171-3_51

Tran, D. H., Sheng, Q. Z., Zhang, W. E., Tran, N. H., & Khoa, N. L. D. (2023). CupMar: A deep learning model for personalized news recommendation based on contextual user-profile and multi-aspect article representation. *World Wide Web (Bussum)*, *26*(2), 713–732. doi:10.1007/s11280-022-01059-6

Wu, C., Wu, F., An, M., Huang, J., Huang, Y., & Xie, X. (2019). NPA: neural news recommendation with personalized attention. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2576-2584. doi:10.1145/3292500.3330665

Wu, C., Wu, F., Ge, S., Qi, T., Huang, Y., & Xie, X. (2019). Neural news recommendation with multi-head self-attention. *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP).*

Wu, C., Wu, F., Huang, Y., & Xie, X. (2023). Personalized news recommendation: Methods and challenges. *ACM Transactions on Information Systems*, *41*(1), 1–50. doi:10.1145/3530257

Wu, F., Qiao, Y., Chen, J., Wu, C., Qi, T., Lian, J., Liu, D., Xie, X., Gao, J., Wu, W., & Zhou, M. (2020). MIND: A Large-scale Dataset for News Recommendation. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 3597-3606. doi:10.18653/v1/2020.acl-main.331

Wu, F., Qiao, Y., Chen, J. H., Wu, C., Qi, T., Lian, J., & Zhou, M. et al. (2020, July). Mind: A large-scale dataset for news recommendation. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 3597-3606. doi:10.18653/v1/2020.acl-main.331

Zhang, Q., Li, J., Jia, Q., Wang, C., Zhu, J., Wang, Z., & He, X. (2021). UNBERT: User-News Matching BERT for News Recommendation. *IJCAI (United States)*, 3356–3362. doi:10.24963/ijcai.2021/462

Zhao, N., & Xu, H. (2021). Distributed Recommendation Considering Aggregation Diversity. *International Journal of Distributed Systems and Technologies*, *12*(3), 83–97. doi:10.4018/IJDST.2021070105

Zhu, Q., Zhou, X., Song, Z., Tan, J., & Guo, L. (2019). DAN: deep attention neural network for news recommendation. *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*. doi:10.1609/aaai.v33i01.33015973

*Qian He graduated from Sichuan Normal University in 2002, Worked in Sichuan University of Media and Communications, associate professor. His research interests include Integrated Media Communication and Intelligent Media Technology.*