Deep Learning Model for News Quality Evaluation Based on Explicit and Implicit Information

Guohui Song[1,2], Yongbin Wang[1,*], Jianfei Li[1] and Hongbin Hu[1]

[1]State Key Laboratory of Media Convergence and Communication, Communication University of China, Beijing, 100024, China

[2]School of Computer and Cyber Sciences, Communication University of China, Beijing, 100024, China

*Corresponding Author: Yongbin Wang. Email: ybwang@cuc.edu.cn

ABSTRACT

Recommending high-quality news to users is vital in improving user stickiness and news platforms' reputation. However, existing news quality evaluation methods, such as clickbait detection and popularity prediction, are challenging to reflect news quality comprehensively and concisely. This paper defines news quality as the ability of news articles to elicit clicks and comments from users, which represents whether the news article can attract widespread attention and discussion. Based on the above definition, this paper first presents a straightforward method to measure news quality based on the comments and clicks of news and defines four news quality indicators. Then, the dataset can be labeled automatically by the method. Next, this paper proposes a deep learning model that integrates explicit and implicit news information for news quality evaluation (EINQ). The explicit information includes the headline, source, and publishing time of the news, which attracts users to click. The implicit information refers to the news article's content which attracts users to comment. The implicit and explicit information affect users' click and comment behavior differently. For modeling explicit information, the typical convolution neural network (CNN) is used to get news headline semantic representation. For modeling implicit information, a hierarchical attention network (HAN) is exploited to extract news content semantic representation while using the latent Dirichlet allocation (LDA) model to get the subject distribution of news as a semantic supplement. Considering the different roles of explicit and implicit information for quality evaluation, the EINQ exploits an attention layer to fuse them dynamically. The proposed model yields the Accuracy of 82.31% and the F-Score of 80.51% on the real-world dataset from Toutiao, which shows the effectiveness of explicit and implicit information dynamic fusion and demonstrates performance improvements over a variety of baseline models in news quality evaluation. This work provides empirical evidence for explicit and implicit factors in news quality evaluation and a new idea for news quality evaluation.

KEYWORDS

Deep learning; news quality; communication studies; classification

1 Introduction

With the rapid development of internet technology, news platforms based on personalized recommendation algorithms, such as Google News and Toutiao, have gradually emerged. These

platforms aggregate the news content provided by online content service providers, rapidly change the way of information dissemination, and progressively become the primary way for users to obtain information. Given the growing amount of Internet traffic, news platforms often rely on personalized recommendation methods and popularity prediction methods to allocate resources better to meet the evolving information needs of readers. Therefore, how to identify high-quality news that attracts users has become a hot spot in current research.

At present, deep learning technology has achieved great success in information recommendation and has been widely used in various fields, such as medicine [1], banking [2], law [3], education [4], biology [5], and insurance [6]. In news communication, news platforms often use deep learning techniques to recommend potentially widespread information to users. These methods are mainly based on the user's historical behavior to cater to the preferences of the user [7–9] or push potentially high-click news articles to the user [10–12]. However, these studies focus on optimizing news clicks or catering to users' preferences while ignoring the quality of the articles they recommend, which can lead to under-recommended content news. Human editors or automated quality inspection systems may be unable to detect some low-quality information on news platforms. These low-quality news articles may not gain users' attention and discussion, and regularly recommending them seriously damage the user experience and the competitiveness of online news platforms. Therefore, news platforms need a quality evaluation mechanism that can evaluate the quality of news before it is published, which helps the news platform push information that can attract users' attention and increase the competitiveness and engagement of the news platform.

News quality evaluation methods can divide into news popularity prediction [10–12] and clickbait detection [13–16]. Clickbait detection is often considered a classification problem to determine whether a headline is misleading. News articles with misleading headlines are considered low-quality because their contents do not satisfy users' expectations. However, clickbait detection methods cannot detect high-quality news because not all non-clickbait news articles are of high quality since their content may not catch users' attention. The researches focus on news popularity prediction mainly to predict the popularity of news based on the number of news readers have viewed. Most of these studies considered news popularity prediction a classification or regression problem. Unpopular news articles are regarded as low quality since their inability to attract users to click. However, high-popularity news may be due to catchy headlines whose content does not attract attention. Therefore, neither clickbait detection nor popularity prediction can effectively assess the quality of news. In addition, some studies consider the dwell time of news reading as an indicator of the quality of the news [17,18]. But the dwell time of news reading may be closely related to the length of the news content. At the same time, neither news popularity prediction nor clickbait detection can reflect whether the news has received widespread public attention and discussion. Therefore, defining the quality of news is an open issue that needs to be combined with practical application scenarios. In addition, there are many difficulties in assessing the quality of news. First, there are currently no datasets with news-quality labels. Secondly, given the labeled data, how to use the various features of news to design models to predict the quality of news is another vital issue accurately. Thirdly, the research object of news quality evaluation is mainly aimed at news headlines, measuring the relationship between news headlines and contents. The role of news headlines, content, and other features in evaluating news quality has not been well studied.

To address the above limitations of existing research, this study proposes a deep learning model for online news quality evaluation, named EINQ, based on explicit and implicit news information. This study proposes a straightforward method to measure news quality based on the comments and clicks of each piece of news. The number of clicks represents the breadth of news dissemination, while

the number of comments is more reflective of the publicity effect of the news. Therefore, the quality of the news in this study mainly considers the ability of the news headlines, content, and other features to attract users to click and comment. According to this method, this paper defines four quality indicators to evaluate news quality. The explicit information includes the headline, source, and publishing time of the news, which users can see before they click. The implicit information refers to the news article's content and is the main factor that triggers user comments. The implicit and explicit information affect users' click and comment behavior differently. After obtaining the embedded representation of these two types of information, we dynamically fuse them through the attention mechanism for evaluating news quality.

In summary, the main contributions of this paper are as follows:

(1) This work proposes the concepts and indicators of news quality. To the best of our knowledge, this is the first attempt to incorporate the comments and clicks of news in news quality evaluation to evaluate the ability of news articles to elicit clicks and comments from users.

(2) This work proposes a deep learning model, EINQ, that integrates explicit and implicit news information for news quality evaluation. More importantly, EINQ incorporates an attention unit based on the implicit and explicit information of the news article, which can adaptively learn their combinations for classification tasks.

(3) This study verifies the effectiveness of EINQ as a whole, provides empirical evidence for explicit and implicit factors in news quality evaluation, and provides a new idea for news quality evaluation.

The rest of this paper is organized as follows. The next section summarizes the current work related to this study. Section 3 defines four quality indicators to evaluate news quality and describes ways to automate data labeling. Section 4 formally introduces the EINQ model for evaluating news quality. Section 5 presents the experimental results and detailed analysis. Section 6 finally concludes this paper.

2 Related Work

News quality evaluation is widely used in clickbait detection, and its original research intention is to detect some online news outlets trying to attract users to click through gimmicky headlines. Clickbait detection is generally defined as a binary classification problem that aims to determine whether a news headline is misleading. There is a large amount of research related to clickbait detection in academia. Kaur et al. [13] proposed a two-phase hybrid convolutional neural network (CNN) and long short-time memory network (LSTM) models for modeling short topic content. The experiment showed that clickbait such as Shocking/Unbelievable, Hypothesis/Guess, and Reaction are the highest in numbers among the other clickbait headlines published online. Indurthi et al. [14] modeled clickbait strength prediction as a regression problem. While previous methods have relied on traditional machine learning or recurrent neural networks (RNN), this study rigorously investigates using a transformer regression model for clickbait strength prediction. The experiment results with a benchmark dataset result in a new state-of-the-art for the clickbait intensity prediction task. Since previous researches usually focus on the semantic information of the English clickbait corpus, some studies have begun to focus on the Chinese clickbait problem. Liu et al. [15] constructed a Chinese WeChat clickbait dataset. They proposed an effective deep method by integrating semantic, syntactic, and auxiliary information, which respectively use Bidirectional Encoder Representation from Transformers (BERT) and Bidirectional Long Short-Term Memory (Bi-LSTM) network with an attention mechanism to encode title semantics. In addition, they propose an improved Graph Attention Network (GAT) to aggregate local syntactic structures of titles and use an attention mechanism to capture valuable

structures. The experiment results prove that model performance is better than compared baseline methods. Pujahari et al. [16] proposed a hybrid categorization technique for separating clickbait and non-clickbait articles by combining features, sentence structure, and clustering. The experimental results indicate that the proposed hybrid model is more robust, reliable, and efficient than any particular categorization technique. Some studies suggested that the use of clickbait can enhance the effectiveness of official propaganda. Meanwhile, some works focused on detecting fake news, considered low-quality due to the spread of false information. Zhang et al. [19] proposed a deep learning-based fast fake news detection model for cyber-physical social services, which adopted a convolution-based neural computing framework to extract feature representation for Chinese news texts. Such a design can ensure both processing speed and detection ability in scenes of Chinese short texts. The experiment results show that the proposal has a lower training time cost and higher classification accuracy than baseline methods. Some studies use the credibility of news to rate the quality of news. Romanou et al. [20] proposed a methodology for creating scientific news article representations by modeling the directed graph between the scientific news articles and the cited scientific publications. The network used for the experiments comprises scientific news articles, their topic, the cited research literature, and corresponding authors. The experiment results show promising applications of graph neural network approaches in knowledge tracing and scientific news credibility assessment. Mosallanezhad et al. [21] incorporated auxiliary information (e.g., user comments and user-news interactions) into a reinforcement learning-based model to address the diverse nature of news domains and expensive annotation costs. Extensive experiments on real-world datasets illustrate the effectiveness of the proposed model, especially when limited labeled data is available in the target domain.

On the other hand, some studies predict the quality of news before the news is released based on news headlines or content. The study mainly indicates the popularity of news based on the number of clicks, and news with low clicks is considered low-quality. Xiong et al. [10] believed that attractiveness and timeliness of news are two prominent drivers of news clicks and proposed a deep news click prediction method to integrate attractiveness, timeliness, and text of news for news click prediction. The experiment results show the effectiveness of attractiveness and timeliness in click-through prediction. Saeed et al. [11] predicted the popularity of a news item published on a specific website by exploiting the initial tweeting behavior of the news item on Twitter. The temporal characteristics of a news item are exploited as the news propagates via tweets. The experiment results show the effectiveness of temporal propagation patterns in predicting news popularity. Sun et al. [12] proposed an explicit Time embedding-based Cascade Attention Network (TCAN) as a novel popularity prediction architecture for large-scale information networks, which integrates temporal attributes into node features via an available time embedding approach to learn the representation of cascade graphs and cascade sequences. The experiment results show that TCAN outperforms other representative baselines while maintaining good interpretability. Wu et al. [17] proposed an effective news quality evaluation method based on the distributions of users' reading dwell time on the news, which incorporate news quality information into user interest modeling by designing a content-quality attention network to select clicked news based on news semantics and qualities. Extensive experiments on two real-world datasets show that the model can effectively improve the overall quality of recommended news and reduce the recommendation of low-quality news. Omidvar et al. [18] believed that the quality of headlines is determined by the number of clicks on the news and the time that users stay on the news content, and proposed a deep learning model for evaluating the quality of news headlines by four indicators. Stokowiec et al. [22] believed that the quality of news headlines is the primary factor that triggers user clicks, so they only used Bi-LSTM to predict the number of clicks on the news. The experiment results

show that using pre-trained word vectors in the embedding layer improves the results of LSTM models, especially when the training set is small. Weissburg et al. [23] proposed an interpretable attention-based model to assess the impact of a post's title on its popularity while controlling for the time of posting. The experiment results show that the title's emotion, narrative ability, and political power are essential in promoting the post's popularity.

In addition, some studies attempted to assess news quality from the feature extraction perspective. Rieis et al. [24] extracted features from the content of 69,907 news articles to find ways to improve the quality of headlines. The findings suggest that positive or negative headlines attract more clicks, while neutral headlines get fewer clicks. Kim et al. [25] proposed a novel generative model, the headline click-based topic model (HCTM), that extends latent Dirichlet allocation (LDA) to reveal the effect of topical context on the click-value of words in headlines. The experiment results show that by jointly taking topics and clicks into account, the model can detect changes in user interests within topics. Wu et al. [26] designed an encoder-decoder model using a large-scale training corpus to generate high-quality news headlines by reducing the importance of easy-to-classify sentences. In this part of the study, the characteristics affecting the quality of news headlines are identified by multiple feature extraction methods. Experimental results show that this method significantly improves the quality of headline editing compared to previous methods. Zhou et al. [27] proposed a joint predictor for both users' click and dwell time in session-based settings, which adopts a novel three-layered RNN structure that includes a fast-slow layer for very short sessions and an attention layer for noisy sessions. Experiments demonstrate that this model outperforms state-of-the-art user clicks and dwell time prediction methods. Romanou et al. [28] constructed a novel system for evaluating the quality of news articles which automatically collects contextual information about news articles in real time and provides quality indicators about their validity and trustworthiness. The news' quality indicators include social media discussions, news content, and source. Choi et al. [29] employed a traditional social science survey of over 7800 news audiences and implemented natural language processing for 1500 news articles concerning public affairs. Results suggest believability, depth, and diversity are more important in predicting news quality than readability, objectivity, factuality, and sensationalism. Alam et al. [30] proposed a score-based technique to quantify the quality of online news articles through news article metadata, content analysis, and epidemiological entity extraction. The experiment results show significant enhancement in the data quality by filtering irrelevant news.

From the above-related works, we can see different definitions of news quality in practical applications. Clickbait and fake news detection methods focus on identifying low-quality news headlines that are gimmicky or misleading. However, these methods fail to identify high-quality news that attracts users. News popularity prediction methods mainly indicate the popularity of news based on the number of clicks, and news with low clicks is considered low-quality. However, user click behavior is mainly triggered by the attractiveness of news headlines. The news popularity prediction methodology does not consider whether the news content attracts users and sparks widespread discussion among users. The methods based on feature extraction are mainly used to identify the features that affect the quality of news, such as words, sentences, emotions, etc. However, the interaction between news headlines, content, and other metadata features in evaluating news quality has not been thoroughly studied. To address the limitations of existing research, this paper defines news quality as the ability of news articles to elicit clicks and comments from users, which represents whether the article can attract widespread attention and discussion. Based on the above definition, this paper proposes a deep learning model that integrates explicit and implicit news information for news quality evaluation.

3 Materials and Methods

3.1 Materials

3.1.1 Data Collection

This work collected an online news articles dataset from a widely used Chinese news platform Toutiao (www.toutiao.com). Compared with other news platforms, Toutiao's users are more active, with more click and comment behavior. The news articles come from the five items at the top of the homepage, and the publish date is between November 05, 2020, and June 15, 2022. The collected data consists of multiple-dimensional information such as news headlines, news sources, publish time, news content, the number of clicks and comments. The dataset comprises 12108 news articles. Since these news articles are on the homepage and aimed at all users, they are more reflective of the quality of the news. At the same time, the research content of this paper focuses on evaluating whether news can get widespread attention and discussion rather than fake news detection. The top news in the dataset comes from different official media organizations, which reduces the possibility of fake news and clickbait, which is suitable for this study.

3.1.2 Data Pre-Processing

This work uses features of the news headline and content for text semantic representation learning. During text pre-processing, this work leverages a popular Chinese word segmentation tool, Jieba3[1], to segment the text, calculate the occurrence frequency of all terms in the dataset, and remove those that occurred fewer than five times or stopwords. Since 98.2% of the news headlines in the dataset are less than 20 words after removing stopwords, the maximum length of the headlines is set to 20. The news content's maximum length is set to 500. Meanwhile, this work pads shorter or truncated longer news headlines and content sequences.

3.2 The Method of Data Labeling

News quality evaluation is a typical classification task, while classification problems usually use supervised learning methods and rely on labeled data. In existing studies, some studies leverage users' dwell time on news content to assess users' satisfaction with news [17,18,27], but the length of news content may affect the dwell time, so it is inaccurate to use dwell time directly to label news quality. During the user's reading process, explicit information dramatically impacts the number of news clicks, and the number of user comments on the news indicates the headline's relevance to the content and the user's attention to the information. In other words, the number of clicks represents the breadth of news dissemination, while the number of comments is more reflective of the publicity effect of the news. Therefore, this section classifies the news quality by the number of clicks and comments. In this way, the collected data can be labeled automatically. Supposing that the number of clicks and comments on the i-th news article in the dataset during its lifetime is B_i and E_i. Then, we scale the values of B_i and E_i to [0, 1] by the normalization method as follows:

$$\begin{cases} E_i = \dfrac{E_i - E_{min}}{E_{max} - E_{min}} \\ \\ B_i = \dfrac{B_i - B_{min}}{B_{max} - B_{min}} \end{cases} \tag{1}$$

---

[1] https://github.com/fxsjy/jieba

where E_{max} and E_{min} represent the maximum and minimum number of comments, B_{max} and B_{min} represent the maximum and minimum number of clicks.

As shown in Fig. 1, this work uses the number of comments and clicks on the news articles to define four quality indicators. The meaning of each indicator is described below:

(1) Indicator C_1: High comment count but low click count. News articles close to this indicator are attractive to users due to their high comment count, but their headlines are not interesting enough to trigger users' click behavior. However, the content of the news sparked discussion among users.

(2) Indicator C_2: High comment count and high click count. News articles close to this indicator show that high-quality news headlines attract users, and users widely discuss news content.

(3) Indicator C_3: High click count but low comment count. News articles close to this indicator show that their headlines interest users, but their content are not.

(4) Indicator C_4: High comment count and high click count. News articles close to this indicator show that neither the headlines nor the content of these news articles effectively attracts users.
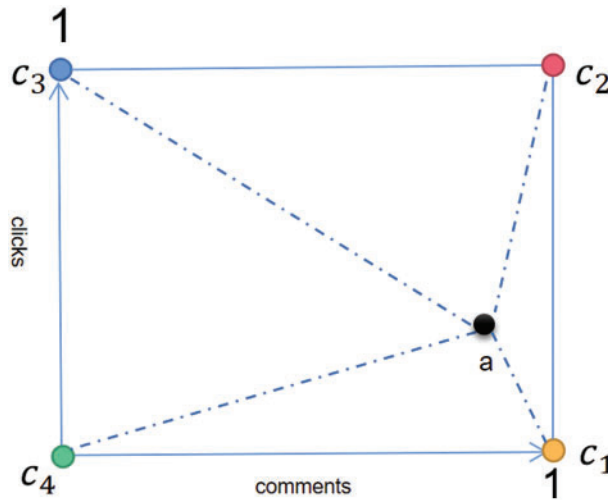

Figure 1: Four news quality indicators

Each news article's proximity to each indicator can be measured by calculating the Euclidean distance to each indicator and finally converting the space into the probability distribution by the softmax function. The calculation method is shown in Eq. (2).

$$\begin{bmatrix} P_{a,C_1} \\ P_{a,C_2} \\ P_{a,C_3} \\ P_{a,C_4} \end{bmatrix} = Softmax \begin{pmatrix} \sqrt{2} - \sqrt{x_a^2 + (1 - y_a)^2} \\ \sqrt{2} - \sqrt{(1 - x_a)^2 + (1 - y_a)^2} \\ \sqrt{2} - \sqrt{(1 - x_a)^2 + y_a^2} \\ \sqrt{2} - \sqrt{x_a^2 + y_a^2} \end{pmatrix} \tag{2}$$

4 The Proposed Model

4.1 Problem Statement

This paper defines the news quality evaluation problem as a multi-classification task; that is, the study outputs news quality class according to the explicit and implicit information of the news. Suppose there are N news articles in dataset $D = \{(T_i, L_i, S_i, A_i)\}_i^N$. The explicit information of i-th news article contains news headline T_i, publish time L_i and news source S_i that users can see before making their decisions to click, while the implicit information refers to the news article's content A_i. Obviously, explicit information has a more significant impact on the users' click behavior. The news quality evaluation is to solve the conditional probability $P(y_j | T_i, L_i, S_i, A_i, \theta)$, and predict the class of news articles by determining the model parameters θ. The calculation method is shown in Eq. (3).

$$\widehat{y_j} = argmax_{j \in \{1,2,3,\dots,m\}} P(y_j | T_i, L_i, S_i, A_i, \theta) \tag{3}$$

4.2 The Model Architecture

This section presents a deep learning model for online news quality evaluation based on explicit and implicit information. In the model, we first construct embedded representations of explicit and implicit information, and then an attention unit is employed to learn their combinations adaptively. Finally, the model outputs a probability distribution of news quality. The workflow of the proposed model is illustrated in Fig. 2.
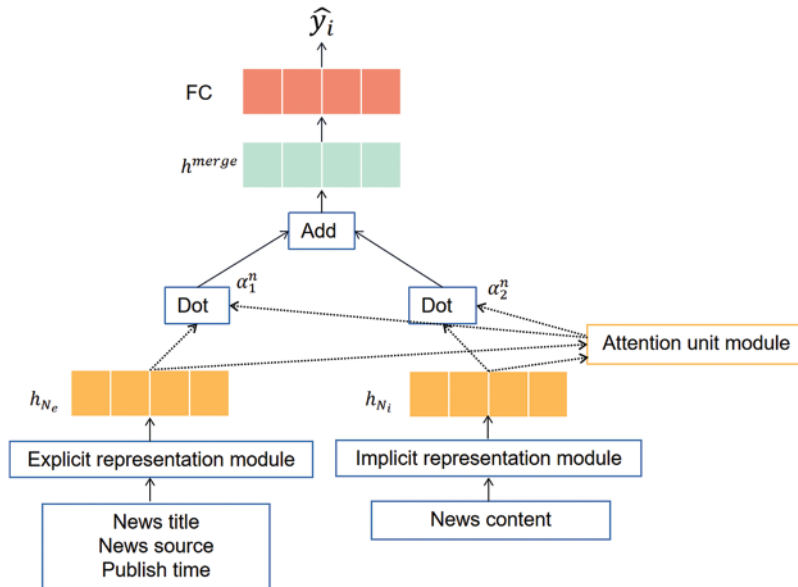

Figure 2: The proposed model for evaluating news quality based on explicit and implicit information

4.2.1 Explicit Information Representation Module

As shown in Fig. 2, the news's headline, source, and publish time are inputs to the explicit information representation module. The architecture of the explicit information representation module is shown in Fig. 3. The first component is embedding news sources and publishing time. We consider them because the source and credibility of news have an essential impact on the quality of news [29]. Meanwhile, the publish time measures different time effects on users' click behavior. Therefore, we

consider the news source and publish time in the explicit information representation. Since the news source and publish time are discrete features, we embed these features into homologous dense vectors. Then we concatenate embedding vectors as e_l and apply fully connected (FC) layers to combine all features to e'_l:

$$e'_l = tanh\left(W_l \times e_l + b_l\right) \tag{4}$$
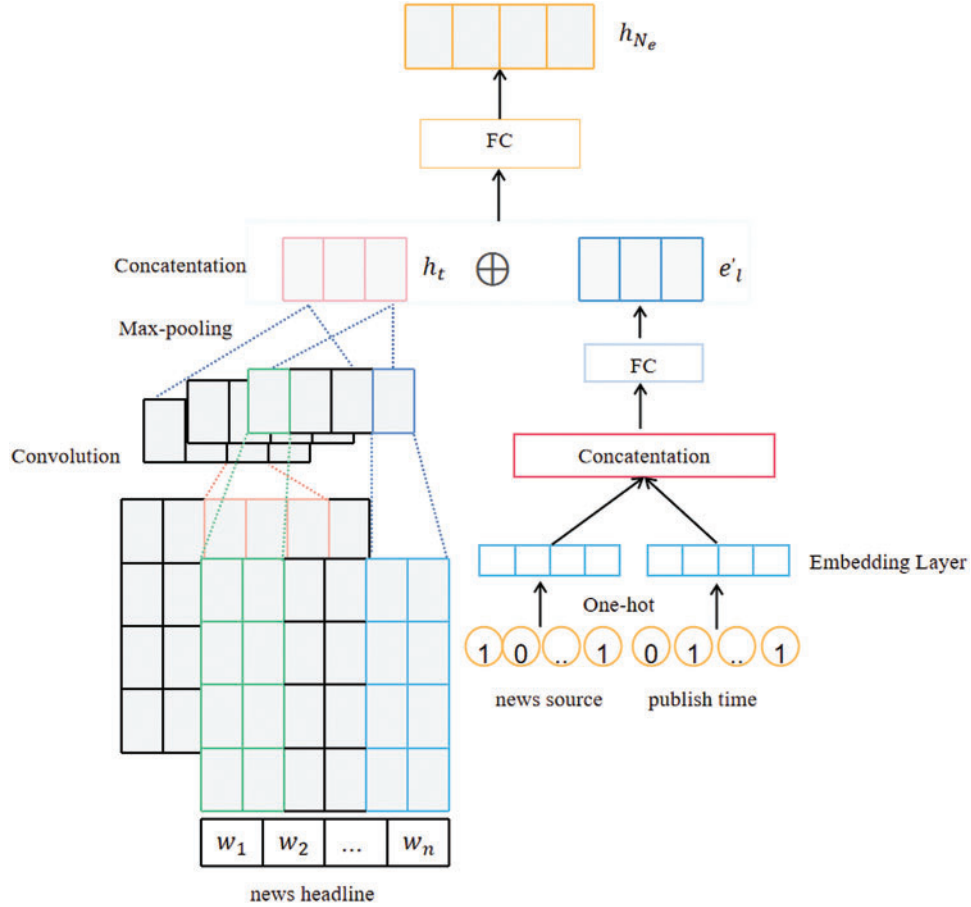
where W_l and b_l are learnable parameters.

Figure 3: The architecture of explicit information representation module

The second component is embedding news headlines with a typical one-dimensional CNN structure [31], which has a wide range of short-text semantic extraction applications. We exploit embedding techniques for embedding news headlines into word embedding matrix $T_{1:n} = [w_1, w_2, w_3, w_4 \ldots, w_n] \in \mathbb{R}^{d \times n}$, where w_i represents the i-th word in the headline, n represents the number of words, and d represents the embedding dimension. For each sub-matrix $T_{i:i+q-1}$, we can get a feature representation o_i by convolution operation:

$$o_i = f\left(H * T_{i:i+q-1} + b\right) \tag{5}$$

where $H \in \mathbb{R}^{d \times q}$ is convolution kernel; $q\,(q < n)$ is the window size; $H * T_{i:i+q-1}$ represents the convolution operation; f is the Rectified Linear Unit (ReLu) activation function. After all the

convolution operations, the feature sequences can be expressed as follows:

$$o = [o_1, o_2, o_3 \ldots, o_{n-q+1}]^T \tag{6}$$

where o_i represents i-th result of filter H. Then, we adopt the max-pooling layer to identify the most significant features further:

$$\widetilde{o} = max \{o_1, o_2, o_3 \ldots, o_{n-q+1}\} \tag{7}$$

Then, this work exploits m convolution filters with the same window size for convolution and max-pooling as above to learn complementary features from the same window. As a result, all the parts are concatenated to get the final representation of the news headlines:

$$h_t = [\widetilde{o}_1, \widetilde{o}_2, \widetilde{o}_3 \ldots, \widetilde{o}_m]^T \tag{8}$$

Finally, the vectors e_l' and h_t are concatenated and fed into a one-layer FC layer to get the final explicit information representation h_{N_e}. The FC layer employs ReLU as an activation function.

$$h_{N_e} = FC\left([e_l' ; h_t]\right) \tag{9}$$

4.2.2 Implicit Information Representation Module

In the workflow shown in Fig. 2, news content is used as input to the implicit information representation module. The news content is usually long text that determines the user's comments behavior. To fully extract semantic information of news content, this study divides the semantic representation of news content into two parts. The framework of the implicit information representation module is shown in Fig. 4.
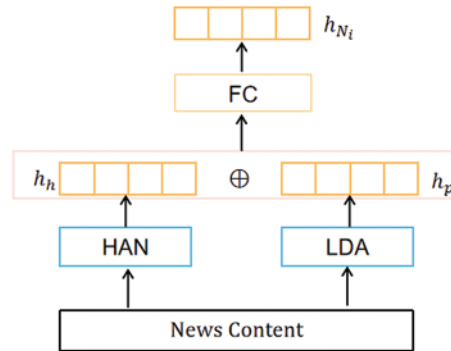

Figure 4: The architecture of implicit information representation

First, this work uses the hierarchical attention networks (HAN) model [32] to obtain the semantics of news content. The HAN model has two characteristics: the first feature is that this hierarchy corresponds to the hierarchy of the document, and the second feature is that it has two attention mechanisms, word level, and sentence level, which allows the network to distinguish essential parts of the document, to generate document representation better. Word and sentence-level encoders are bidirectional gated recurrent units (Bi-GRU) networks. Let A denotes the news article's content. A news article's content semantics can be expressed as:

$$h_h = HAN(A) \tag{10}$$

Second, we exploit the latent Dirichlet allocation (LDA) model [33] to obtain the subject distribution of news content. The LDA topic model can represent each news content as a probability distribution composed of some topics, which can be used as part of the model input to supplement the semantic representation of news content. The calculation process is as follows:

$$h_p = LDA\,(A) \tag{11}$$

Finally, the vectors h_h, h_p are concatenated and fed into the FC layer to get the final content semantic representation. The FC layer employs ReLu as an activation function.

$$h_{N_i} = FC\left([h_h; h_p]\right) \tag{12}$$

By using two models to represent documents separately, not only is the document's hierarchy considered, but it is more conducive to identifying essential parts of the news content. At the same time, better document representation results can be obtained due to the inclusion of the subject distribution of documents. In the semantic representation of news content, the LDA topic model can be trained in advance without the additional computational overhead.

4.2.3 Attention Unit Module

Since explicit and implicit information has different effects on user behavior, news headlines, publish time, and news sources have high visibility to users, significantly impacting user click behavior. In contrast, as implicit information, news content significantly affects user comment behavior after clicking. At the same time, the headline semantics are related to the content semantics and jointly affect user behavior. Therefore, this section introduces an attention mechanism to learn the weights of both parts automatically. The calculation process is shown as follows:

$$\begin{cases} \alpha_1^n = q_n^T \times tanh\left(V_n \times h_{N_e} + b_n\right) \\ \alpha_2^n = q_n^T \times tanh\left(V_n \times h_{N_i} + b_n\right) \\ \widetilde{a}_i^n = \dfrac{exp\left(\alpha_i^n\right)}{\sum_i exp\left(\alpha_i^n\right)}, (i = 1, 2) \end{cases} \tag{13}$$

where $\widetilde{a}_i^n$ is the attention weight assigned to explicit and implicit information. q_n, V_n, b_n are learnable parameters.

Finally, the news embedding is calculated as follows:

$$h^{merge} = \alpha_1^n h_{N_e} + \alpha_2^n h_{N_i} \tag{14}$$

4.2.4 The Model Output

In the model output phase, we input h^{merge} into the FC layer, followed by the probability distribution of news quality output $P_C = \{P_{C_1}, P_{C_2}, P_{C_3}, P_{C_4}\}$ by the *softmax* function. Finally, the *argmax* function is used to output that news falls into the category of news quality. The calculation process is as follows:

$$\begin{cases} P_C = Softmax\left(W_h \times h^{merge} + b_h\right) \\ \hat{y} = argmax\left(P_C\right) \end{cases} \tag{15}$$

where W_h and b_h are learnable parameters.

4.3 The Loss Function

In the training stage, this model employs the cross-entropy loss function to measure the classification effect of the model, and the calculation method is shown in Eq. (16).

$$\mathcal{L} = -\sum_{i=1}^{M} y_i \log(P_C) \tag{16}$$

Due to the imbalance problem of each class in the dataset, the CNN model trained with the imbalance data performs poorly in the weak class [34]. Since this work uses the CNN model to represent news headlines, the imbalanced data impacts the model's performance. Currently, the standard methods to deal with sample imbalance include resampling and undersampling. In using CNN for representation learning, resampling may introduce many duplicate samples, increasing training time and being prone to overfitting. However, undersampling may discard essential samples. Therefore, this model uses the re-weighting method proposed by Cui et al. [35] to increase the loss weight of the class with a small sample and reduce the loss weight of the class with a large sample. By weighing the losses of each class in the dataset, the problem of uneven data distribution can be alleviated. In the re-weighting method, the effective size of a class is defined as:

$$E_n = \frac{1 - \beta^n}{1 - \beta} \tag{17}$$

where n is the actual number of samples in a class and β is a hyperparameter whose value is usually close to 1 (for example, 0.99, 0.999, and so on). Thus, the loss function $\mathcal{L}$ considering the class weights, can be defined as:

$$\mathcal{L} = -\sum_{i=1}^{M} \frac{1}{E_{n_{y_i}}} y_i \log(P_C) \tag{18}$$

The EINQ model proposed in this paper consists of three parts: explicit information representation module, implicit information representation module, and attention fusion unit module, briefly described in Algorithm 1.

Algorithm 1: The proposed EINQ model

Input: Dataset $D = \{(T_i, L_i, S_i, A_i)\}_i^N$, initial network parameters $\hat{\theta}$.
Output: The trained parameters θ.
1. For $i = 1, 2, \ldots, num\_epoch$ **do**
2.　　**For** $j = 1, 2, \ldots, num\_batch$ **do**
3.　　　　Get explicit information $h_{N_e} = FC\left(\left[e_i'; h_t\right]\right)$ using Eqs. (4)–(9).
4.　　　　Get implicit information $h_{N_i} = FC\left(\left[h_h; h_p\right]\right)$ using Eqs. (10)–(12).
5.　　　　Input h_{N_e} and h_{N_i} into attention layer $h^{merge} = \alpha_1^n h_{N_e} + \alpha_2^n h_{N_i}$ using Eqs. (13),(14).
6.　　　　Update parameters $\hat{\theta}$ using Eqs. (15)–(18).
7.　　**End**
8. End
9. Return $\theta = \hat{\theta}$

---

5 Experiments

This section examines the effectiveness of EINQ by comparing it with several competitive baselines and conducting extensive experiments to evaluate the impact of different modules on classification performance.

5.1 Dataset Partitioning

The EINQ model uses the proposed data labeling method in Section 3 to automatically label news quality in the dataset and divide the data into four classes. In the experiment, the number of training sets, validation sets, and test sets is divided into 60%, 20%, and 20%, and the distribution of each class is shown in Fig. 5.
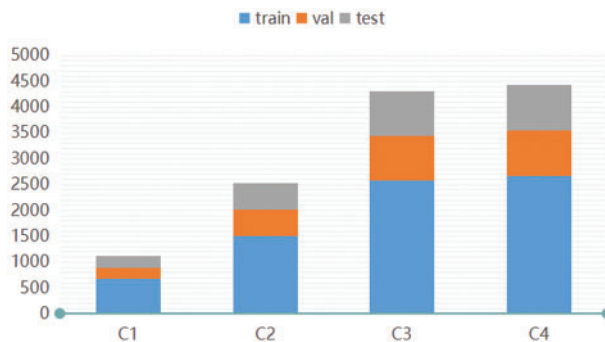

Figure 5: Dataset partitioning

5.2 Experiment Setup

In the explicit information representation phase, this work sets the convolution kernel window q as 3 and sets the number of convolution kernels m as 60 for feature extraction in the CNN model. In the implicit information embedding stage, this work starts with pre-trained HAN and LDA for news content representation and fine-tuning parameters with the dataset. The embedding layer adopts a 100-dimensional word vector pre-trained by Word2vec [36]. All of the FC layers have a suitable hidden size of 512. Through multiple experiments, the value of β is set as 0.999 in the loss function. To reduce dimensionality and avoid overfitting, this works adopts the Adam optimizer for parameters learning, and the learning rate is 0.0001. The intensity of dropout applied to FC layers is 0.2. This model is trained for 20 epochs with a batch size of 32. Training is based on Pytorch 1.11.0 framework, and one NVIDIA A10 is used for calculation.

5.3 Evaluation Metrics

Since this study is a typical classification problem, the widely used accuracy and macro averaged F-Score are employed to evaluate model performance. Accuracy is the ratio of correctly predicted samples to the total samples. Macro averaged F-Score is the mean of F-scores of each class which can evaluate the mode's overall performance in a global sense. The formulas of macro averaged F-Score are given below:

$$F - Score = \frac{1}{C} \sum_{c=1}^{C} \frac{2P_c R_c}{P_c + R_c} \tag{19}$$

where P_c denotes the percentage of all news identified as positive samples that are indeed positive ones in class c; R_c is the percentage of positive predictions that are correctly recognized in class c.

5.4 Baseline Models

This work compares the proposed EINQ method with the following baselines: Bi-LSTM [22], HAN [32], CNN [31], recurrent convolutional neural networks (RCNN) [37], and determining the

quality of news headlines model (DQNH) [18]. We use these models because they are often used for text classification or news quality evaluation. The details of these baseline models used in this work are as follows:

Bi-LSTM: The news headline and content are embedded as 100-dimensional vectors with Word2vec, and the headline and content are entered into the Bi-LSTM module, respectively. In the output stage, the last hidden state of the two Bi-LSTM modules is concatenated and input into the FC layer.

HAN: Since the HAN superimposes the Bi-GRU network structure and attention module, it can better capture the essential words in the sentence. In the comparative experiment, the news headline is input into Bi-GRU, the word sequence of the news content is input into HAN as the input feature, the output of the two models is connected, and the classification result is output through the FC layer.

CNN: In the comparative experiment, the news headline and content are processed according to the method introduced in Section 4.2, and the semantic representations of the headline and content are obtained. Finally, the semantic representation of news headlines and content are connected, and the classification result is output through the FC layer.

RCNN: The model adds an RNN layer to process the intermediate results of the CNN, and the model input is similar to the CNN.

DQNH: The model is a method for evaluating the quality of news headlines. It contains three submodules. CNNs capture the semantic relationship between headline and content in the similarity module. In the semantic module, the Bert model represents the underlying features of the title and article. In the topic module, non-negative matrix factorization (NNMF) describes the thematic characteristics of the headline and content. The headline quality classification result is obtained by inputting the three-module output stitching into the FC layer.

5.5 Performance Analysis

It is evident from Table 1 that the EINQ model yields better results than the baseline models in terms of Accuracy and F-Score. (1) CNN model is the worst-performing of all comparison models. The function of CNN convolution can effectively identify the overall structure of the news headline, but for longer news content, CNN can only process the information in the current window, and the convolution of the latter layer can only fuse the information of adjacent windows. Therefore, CNN significantly depends on parameters such as convolution windows and moving steps when processing news content, resulting in poor results in processing long articles such as news content. (2) Bi-LSTM uses two independent LSTM networks, one processing text from left to right and the other from right to left. It captures more comprehensive contextual information and is more effective when working with news content. Therefore, its news quality evaluation effect is better. (3) Compared with CNNs, RCNN has much improvement in two performance indicators and performs better than Bi-LSTM. Compared with CNN, the two-way loop structure of RCNN can obtain the context information of news content more effectively and retain word order on a large scale when learning text representation. Compared with Bi-LSTM, RCNN can automatically determine the importance of features by extracting text information using the maximum pooling layer. (4) HAN works better than RCNN. Since HAN takes advantage of the hierarchical characteristics of news content and uses attention mechanisms from the sentence and document levels, the importance of words and sentences that influence the final classification decision can be identified. Experiments show that the bidirectional loop structure of RCNN has a relatively poor ability to capture information compared with HAN. (5) Compared with other comparison models, DQNH obtains complete headline and content semantics through multiple

semantic representations compared with other comparison models, which significantly improves performance. (6) Compared with DQNH, the EINQ model proposed in this paper considers the factors affecting user clicks and comments and assigns attention weight to explicit information and implicit information so that the fusion of various information is more flexible than other baseline models, and the model design is more in line with the behavioral characteristics of users, so it has the best performance. Given the applicable scenarios and experimental results of various models, this work uses CNN for explicit news headline representation, and it is reasonable for HAN to express implicit news content.

Table 1: Performance comparison of the proposed model with the baseline models

Model	Accuracy (%)	F-Score (%)	Model size (M)	Inference time (ms)
Bi-LSTM	67.23	61.45	3.5	0.78
HAN	71.76	69.19	5.6	1.07
CNN	63.65	60.38	4.8	0.84
RCNN	69.43	67.93	5.1	1.12
DQNH	79.29	77.42	9.1	1.78
EINQ (ours)	**82.31**	**80.51**	**7.9**	**1.47**

Meanwhile, this study analyzes the model size and inference time. As for the deep neural network, the model size is the storage of the weight matrix. Concretely, the total number of parameters under our experiment setting is 7.9 M. To compare the time cost of the proposed model with baseline models, we conduct one-step forward prediction experiments with batch size 32. The EINQ model takes 1.47 ms and does not significantly increase the inference time.

5.6 Ablation Experiments

This work develops several ablation models with different features to verify the influence of the effectiveness and impact of the model component. As a result, EINQ and the ablation models with different inputs are shown as follows:

EINQ-H*: In the process of explicit information, only the news headline is considered, and the source and publishing time of the news are not included. The design aims to analyze the impact of news sources and publishing time on the model.

EINQ-LDA*: The news content representation uses only the LDA topic model in the implicit information representation process.

EINQ-HAN*: The news content representation uses only the HAN model in the implicit information representation process.

EINQ–AM^: In the model output stage, the implicit and explicit information is concatenated directly and does not assign attention weights.

It can be seen from Table 2 that the effect of using only news headlines in the explicit information representation process is poor, indicating that the publish time and source of news are important factors that trigger user click and comment behavior. Compared with the EINQ-LDA*, the performance of EINQ-HAN* is relatively good, which shows that the semantics of news content has a more

significant impact on the quality of news than the topic. Compared with the EINQ-LDA* and EINQ-HAN*, the model performance of EIQ is improved, mainly because HAN focuses on the structure of news content and LDA focuses on the topic of news content, and the combination of the HAN and LDA makes the semantic representation richer. EINQ-AM^ does not consider the attention weight of explicit and implicit information in the fusion stage, and its performance indicators lag behind other combined modules, indicating that attention weight plays an indispensable role in explicit and implicit information fusion. It can be seen from the results of the ablation experiments that each module has a particular impact on the performance of the model, and the design of explicit and implicit information significantly improves the classification effect compared with the baseline models. The performance of the baseline and ablation models are shown in Fig. 6.

Table 2: Ablation model performance

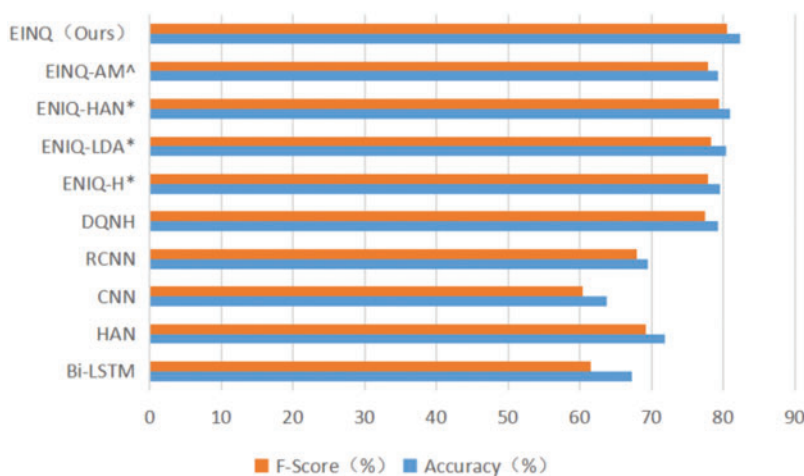| Model	Accuracy (%)	F-Score (%)
EINQ-H*	79.45	77.89
EINQ-LDA*	80.35	78.32
EINQ-HAN*	80.95	79.32
EINQ-AM^	79.23	77.78
EINQ (ours)	**82.31**	**80.51**

Figure 6: The performance comparison of ablation and baseline models

5.7 Hyperparameters Analysis

5.7.1 The Effective Size in the Loss Function

In the loss function, the hyperparameter β is used to calculate the effective size of the class, whose value is a number close to 1. It can be seen from Fig. 7 that when $\beta = 0.999$, the model performs best on both performance indicators. It should be noted that when $\beta = 0$, the loss function degrades the traditional cross-entropy loss function. Experimental results show that increasing the category weight in the loss function affects performance improvement for unbalanced datasets.
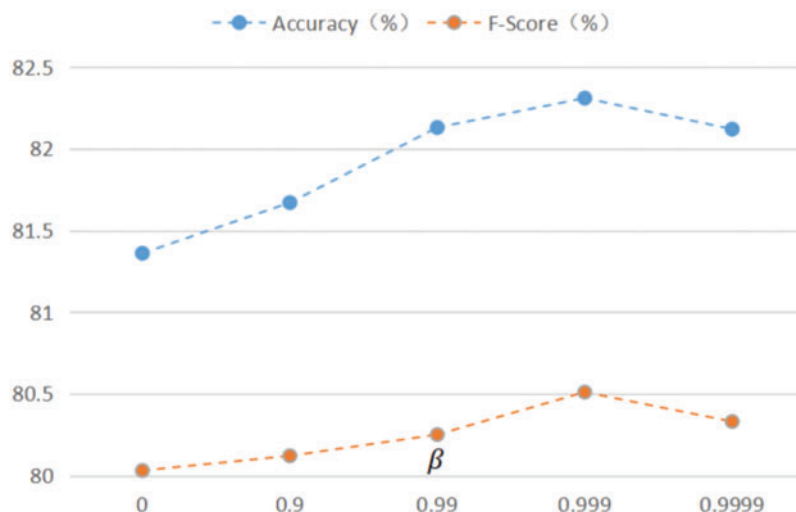
Figure 7: The effect of β on model performance

5.7.2 Investigating the Influence of Convolution Kernels

Since EINQ uses CNN to extract the semantic features of news headlines in the explicit information representation stage, the feature processing at this stage is more sensitive to the window size q of the convolution kernel and the number m of convolution kernels, so different sizes and numbers of convolution kernels are used for verification in the experiment. This paper conducted multiple experiments with different $q \in \{2, 3, 4, 5\}$ and $m \in \{20, 30, 40, 50, 60, 79, 80, 90, 100\}$ values to find the appropriate window size and the number of convolution kernels. The experimental results show that the model works best when $q = 3$ and $m = 60$. In the case of $q = 3$, the value of m is adjusted in the experiment to observe the effect of the number of convolution kernels on the model's performance. As shown in Fig. 8, the optimal model performance can be achieved at $m = 60$.
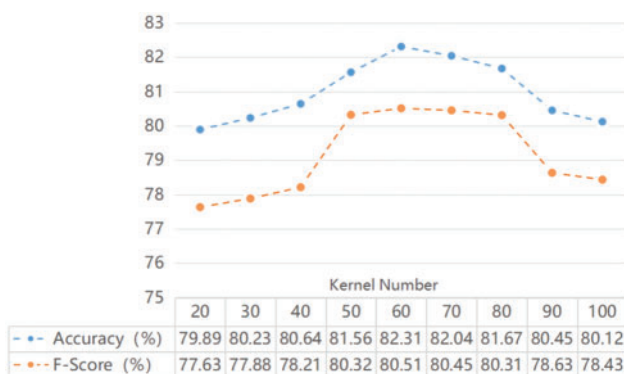

Kernel Number	20	30	40	50	60	70	80	90	100
Accuracy (%)	79.89	80.23	80.64	81.56	82.31	82.04	81.67	80.45	80.12
F-Score (%)	77.63	77.88	78.21	80.32	80.51	80.45	80.31	78.63	78.43

Figure 8: Model performance under different numbers of convolution kernels when $q = 3$

As shown in Table 3, different window sizes of the convolution kernel cause inevitable fluctuations in model performance. In the case of $m = 60$, the CNN module in the explicit information representation can better consider the relationship between the current convolution window and the information of adjacent windows and more effectively identify the overall structure of news headlines when $q = 3$.

Table 3: Model performance under different window sizes when $m = 60$

Kernel size (q)	Accuracy (%)	F-Score (%)
2	81.78	80.89
3	**82.31**	**80.51**
4	81.67	80.03
5	81.02	79.67

6 Conclusion and Future Work

Recommending high-quality news to users is vital in improving user stickiness and news platforms' reputation. This study proposes a deep learning model based on explicit and implicit information for news quality evaluation. An empirical evaluation that involved real-world news datasets demonstrates the significant positive impact of the proposed explicit and implicit information on classification performance. This study makes several research contributions. First, this work defines four news quality indicators representing whether the article can attract widespread attention and discussion. Second, this paper presents a deep learning model named EINQ for news quality evaluation, which divides the features that determine the quality of news into explicit and implicit information. Third, this study conducts various experiments to verify the effectiveness of explicit and implicit information dynamic fusion. The results demonstrate performance improvements over various baseline models in news quality evaluation. Meanwhile, this work verifies the rationality of module combination and parameter design through ablation experiments and parameter analysis. Last, this research's findings imply that combining text and meta-features (e.g., news source and publishing time) would benefit evaluating news quality.

This work has some limitations that offer potential opportunities for future study. First, this paper evaluates whether news articles can get widespread attention and discussion rather than clickbait or fake news detection. Meanwhile, the data are collected from Toutiao, and the news sources are different official media organizations of China, which reduces the possibility of fake news. However, news articles with false information might get many comments and views in practical applications. Therefore, It is beneficial to enhance the ability to identify fake news through feature extraction [19] and auxiliary information (e.g., news comments and user-news interactions) [21] on different datasets in future work. Second, this study focuses on the task of categorizing news quality. However, how to generate high-quality news is also an important issue. We plan to study the generation methods of high-quality news from the perspectives of word attractiveness [10,25], sentence importance [26], and news timeliness [10]. Third, this study uses classical natural language processing (NLP) models such as CNN, HAN, and LDA because they have been widely used in research. But using other NLP models, such as BERT [38], may lead to different model performances. Therefore, we plan to explore the impact of the BERT models for semantic representation on news quality evaluation.

Author Contributions: Conceptualization, Y. W.; methodology, G. S., and Y. W.; model design, G. S., J. L., and H. H.; formal analysis, Y. W.; writing—original draft preparation, G. S.; writing—review and editing, J. L., and H. H.; visualization, G. S., J. L., and H. H.; project administration, Y. W.; funding acquisition, Y. W. All authors have read and agreed to the published version of the manuscript.

Availability of Data and Materials: The data for this study is extracted from toutiao.com, and can be requested from the corresponding author upon request.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] D. Özdemir and N. N. Arslan, "Analysis of deep transfer learning methods for early diagnosis of the COVID-19 disease with chest X-ray images," *Duzce üniversity Journal of Science & Technology*, vol. 10, no. 2, pp. 628–640, 2022.

[2] M. Ala'raj, M. F. Abbod, M. Majdalawieh and L. Jum'a, "A deep learning model for behavioral credit scoring in banks," *Neural Computing and Applications*, vol. 34, pp. 5839–5866, 2022.

[3] H. H. Chen, L. Wu, J. P. Chen, W. Lu and H. J. Ding, "A comparative study of automated legal text classification using random forests and deep learning," *Information Processing & Management*, vol. 59, no. 2, pp. 102798, 2022.

[4] A. A. El-Demerdash, S. E. Hussein and J. F. W. Zaki, "Course evaluation based on deep learning and ssa hyperparameters optimization," *Computers, Materials & Continua*, vol. 71, no. 1, pp. 941–959, 2022.

[5] D. Özdemir and M. S. Kunduraci, "Comparison of deep learning techniques for classification of the insects in order level with mobile software application," *IEEE Access*, vol. 10, pp. 35675–35684, 2022.

[6] C. Liang, Z. Q. Liu, B. Liu, J. Zhou, X. L. Li *et al.,* "Uncovering insurance fraud conspiracy with network learning," in *Proc. of the 42nd Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*, Paris, France, pp. 1181–1184, 2019.

[7] X. Y. Zhang, Q. Yang and D. L. Xu, "Combining explicit entity graph with implicit text information for news recommendation," in *Companion Proc. of the Web Conf. 2021*, Ljubljana, Slovenia, pp. 412–416, 2021.

[8] C. H. Wu, F. Z. Wu, T. Qi, Q. Liu, X. Tian *et al.,* "Feedrec: News feed recommendation with various user feedbacks," in *Proc. of the ACM Web Conf. 2022*, Lyon, France, pp. 2088–2097, 2022.

[9] C. H. Wu, F. Z. Wu, T. Qi, C. L. Li and Y. F. Huang, "Is news recommendation a sequential recommendation task," in *Proc. of the 45th Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*, Madrid, Spain, pp. 2382–2386, 2022.

[10] J. Xiong, L. Yu, D. S. Zhang and Y. F. Leng, "DNCP: An attention-based deep learning approach enhanced with attractiveness and timeliness of news for online news click prediction," *Information & Management*, vol. 58, no. 2, pp. 103428, 2021.

[11] R. Saeed, H. Abbas, S. Asif, S. Rubab, M. M. Khan *et al.,* "A framework to predict early news popularity using deep temporal propagation patterns," *Expert Systems with Applications*, vol. 195, pp. 116496, 2022.

[12] X. G. Sun, J. Y. Zhou, L. Liu and W. Q. Wei, "Explicit time embedding based cascade attention network for information popularity prediction," *Information Processing & Management*, vol. 60, no. 3, pp. 103278, 2023.

[13] S. Kaur, P. Kumar and P. Kumaraguru, "Detecting clickbaits using two-phase hybrid CNN-LSTM biterm model," *Expert Systems with Applications*, vol. 151, pp. 11350, 2020.

[14] V. Indurthi, B. Syed, M. Gupta and V. Varma, "Predicting clickbait strength in online social media," in *Proc. of the 28th Int. Conf. on Computational Linguistics*, Barcelona, Spain, pp. 4835–4846, 2020.

[15] T. Liu, K. Yu, L. Wang, X. Y. Zhang, H. Zhou *et al.,* "Clickbait detection on WeChat: A deep model integrating semantic and syntactic information," *Knowledge-Based Systems*, vol. 245, pp. 108605, 2022.

[16] A. Pujahari and D. S. Sisodia, "Clickbait detection using multiple categorisation techniques," *Journal of Information Science*, vol. 47, no. 1, pp. 118–128, 2021.

[17] C. H. Wu, F. Z. Wu, T. Qi and Y. F. Huang, "Quality-aware news recommendation," arXiv:2202.13605, 2022.

[18] A. Omidvar, H. Pourmodheji, A. An and G. Edall, "A novel approach to determining the quality of news headlines," in *Proc. of Natural Language Processing in Artificial Intelligence*, Berlin, German, pp. 227–245, 2021.

[19] Q. Zhang, Z. W. Guo, Y. Y. Zhu, P. Vijayakumar, A. Castiglione *et al.,* "A deep learning-based fast fake news detection model for cyber-physical social services," *Pattern Recognition Letters*, vol. 168, pp. 31–38, 2023.

[20] A. Romanou, P. Smeros and K. Aberer, "On representation learning for scientific news articles using heterogeneous knowledge graphs," in *Companion Proc. of the Web Conf. 2021*, Ljubljana, Slovenia, pp. 422–425, 2021.

[21] A. Mosallanezhad, M. Karami, K. Shu, M. V. Mancenido and H. Liu, "Domain adaptive fake news detection via reinforcement learning," in *Proc. of the ACM Web Conf. 2022*, Lyon, France, pp. 3632–3640, 2022.

[22] W. Stokowiec, T. Trzcinski, K. Wołk, K. Marasek and P. Rokita, "Shallow reading with deep learning: Predicting popularity of online content using only its title," in *Foundations of Intelligent Systems: 23rd Int. Symp.*, Warsaw, Poland, pp. 136–145, 2017.

[23] E. Weissburg, A. Kumar and P. S. Dhillon, "Judging a book by its cover: Predicting the marginal impact of title on reddit post popularity," in *Proc. of the Int. AAAI Conf. on Web and Social Media*, Atlanta, USA, pp. 1098–1108, 2022.

[24] J. Rieis, F. Souza, P. V. Melo, R. Partes, H. Kwak *et al.,* "Breaking the news: First impressions matter on online news," in *Proc. of the Int. AAAI Conf. on Web and Social Media*, Oxford, UK, pp. 357–366, 2015.

[25] J. H. Kim, A. Mantrach, A. Jaimes and A. Oh, "How to compete online for news audience: Modeling words that attract clicks," in *Proc. of the 22nd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, San Francisco, USA, pp. 1645–1654, 2016.

[26] Q. Y. Wu, L. Li, H. Zhou, Y. Zeng and Z. Yu, "Importance-aware learning for neural headline editing," in *Proc. of the AAAI Conf. on Artificial Intelligence*, New York, USA, pp. 9282–9289, 2020.

[27] T. F. Zhou, H. Qian, Z. B. Shen, C. Zhang, C. W. Wang *et al.,* "Jump: A joint predictor for user click and dwell time," in *Proc. of the 27th Int. Joint Conf. on Artificial Intelligence*, Stockholm, Sweden, pp. 3704–3710, 2018.

[28] A. Romanou, P. Smeros, C. Castillo and K. Aberer, "Scilens news platform: A system for real-time evaluation of news articles," *Proceedings of the Vldb Endowment*, vol. 13, no. 12, pp. 2969–2972, 2020.

[29] S. Choi, H. Shin and S. Kang, "Predicting audience-rated news quality: Using survey, text mining, and neural network methods," *Digital Journalism*, vol. 9, no. 1, pp. 84–105, 2021.

[30] S. M. Alam, E. Asevska, M. Roche and M. Teisseire, "A data-driven score model to assess online news articles in event-based surveillance system," in *Annual Int. Conf. on Information Management and Big Data*, Virtual Event, pp. 264–280, 2021.

[31] Y. Chen, "Convolutional neural network for sentence classification," M.S. dissertation, University of Waterloo, Waterloo, Canada, 2015.

[32] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola *et al.,* "Hierarchical attention networks for document classification," in *2016 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, San Diego, USA, pp. 1480–1489, 2016.

[33] D. M. Blei, A. Y. Ng and M. I. Jordan, "Latent dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[34] M. Buda, A. Maki and M. A. Mazurowski, "A systematic study of the class imbalance problem in convolutional neural networks," *Neural Networks*, vol. 106, pp. 246–259, 2018.

[35] Y. Cui, M. Jia, T. Lin, Y. Song and S. Belongie, "Class-balanced loss based on effective number of samples," in *Proc. of CVPR*, Long Beach, CA, USA, pp. 9268–9277, 2019.

[36] S. Li, Z. Zhao, R. Hu, W. Li, T. Liu *et al.,* "Analogical reasoning on Chinese morphological and semantic relations," in *56th Annual Meeting of the Association for Computational Linguistics (ACL)*, Melbourne, Australia, pp. 138–143, 2018.

[37] S. W. Lai, L. H. Xu, K. Liu and J. Zhao, "Recurrent convolutional neural networks for text classification," in *Twenty-Ninth AAAI Conf. on Artificial Intelligence*, Austin, USA, pp. 2267–2273, 2015.

[38] J. Devlin, M. W. Chang, K. Lee and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. of the 2019 Annual Conf. of the North American Chapter of the Association for Computational Linguistics*, Minneapolis, USA, pp. 4171–4186, 2019.