Article

A Densely Connected GRU Neural Network Based on Coattention Mechanism for Chinese Rice-Related Question Similarity Matching

Haoriqin Wang [1,2,3,4], Huaji Zhu [3,4], Huarui Wu [3,4], Xiaomin Wang [3,4], Xiao Han [3,4] and Tongyu Xu [1,*]

[1] College of Information and Electrical Engineering, Shenyang Agricultural University, Shenyang 110866, China; wanghrq007@nercita.org.cn

[2] College of Computer Science and Technology, Inner Mongolia University for Nationalities, Tongliao 028043, China

[3] National Engineering Research Center for Information Technology in Agriculture, Beijing 100097, China; zhuhj007@nercita.org.cn (H.Z.); wuhr007@nercita.org.cn (H.W.); wangxm007@nercita.org.cn (X.W.); hanx007@nercita.org.cn (X.H.)

[4] Beijing Research Center for Information Technology in Agriculture, Beijing 100097, China

[*] Correspondence: xutongyu@syau.edu.cn

Abstract: In the question-and-answer (Q&A) communities of the "China Agricultural Technology Extension Information Platform", thousands of rice-related Chinese questions are newly added every day. The rapid detection of the same semantic question is the key to the success of a rice-related intelligent Q&A system. To allow the fast and automatic detection of the same semantic rice-related questions, we propose a new method based on the Coattention-DenseGRU (Gated Recurrent Unit). According to the rice-related question characteristics, we applied word2vec with the TF-IDF (Term Frequency–Inverse Document Frequency) method to process and analyze the text data and compare it with the Word2vec, GloVe, and TF-IDF methods. Combined with the agricultural word segmentation dictionary, we applied Word2vec with the TF-IDF method, effectively solving the problem of high dimension and sparse data in the rice-related text. Each network layer employed the connection information of features and all previous recursive layers' hidden features. To alleviate the problem of feature vector size increasing due to dense splicing, an autoencoder was used after dense concatenation. The experimental results show that rice-related question similarity matching based on Coattention-DenseGRU can improve the utilization of text features, reduce the loss of features, and achieve fast and accurate similarity matching of the rice-related question dataset. The precision and F1 values of the proposed model were 96.3% and 96.9%, respectively. Compared with seven other kinds of question similarity matching models, we present a new state-of-the-art method with our rice-related question dataset.

Keywords: rice-related question similarity matching; natural language processing; densely connected GRU; coattention mechanism; question-and-answering communities

1. Introduction

Question-and-answer (Q&A) communities [1] are knowledge service communities based on the Internet, allowing users to ask, answer, and discuss questions. These can meet the users' needs to obtain information and exchange knowledge. They can be used for research with broad development prospects in natural language processing [2] and information retrieval [3]. The China Agricultural Technology Extension Information Platform is a professional platform for agricultural technicians, in which Q&A communities play a vital role in helping farmers find solutions to their problems. As rice is one of the most widely cultivated grain crops in China, users submit more than a thousand questions in the rice-related question-and-answer module every day, and agricultural experts answer the questions quickly. However, due to the complexity of Chinese semantic expression,

there are many questions with different description methods but the same semantics; the experts might answer the same questions repeatedly, which is a waste of human resources. The sparse [4], real-time, and nonstandard text data aggravate the sparseness of keyword features, making it challenging to mine the correlation between features fully. This has become one of the main tasks of text mining in agricultural information classification: finding a method to easily and quickly mine questions with the same semantics from a rice-related text dataset and provide higher quality and intelligent agricultural information services [5]. It is challenging to complete the data processing of classifying similar questions [6] with manual screening using the traditional methods. At present, the traditional and commonly used keyword query and shallow classification model [7] can assist in completing similar questions to judge; however, without automatic extraction from the data and the ability of organization, its excessive reliance on artificial selection features and classifier performance makes the classic text analysis method inapplicable in the short term. Therefore, a significant problem to be solved by the China Agricultural Technology Extension Information Platform is finding an intelligent method to classify rice-related questions automatically. A neural network model with the characteristics of flexibility and diversity shows good performance in natural language processing tasks such as text classification [8], text similarity calculation [9], and emotion analysis [10]. This kind of model can train the data in an end-to-end way, automatically learn specific tasks, and mine many semantic relations in the text; it effectively reduces the traditional statistical machine learning [11], where the researchers set a large number of features manually.

With the rapid development of computer technology, deep learning techniques such as deep convolutional neural networks and recurrent neural networks have become the mainstream text similarity calculation methods. This technology can automatically extract the key features of images and text without complex feature engineering and combine them with the classification process. These models have excellent adaptability and can be migrated easily. Nowadays, many scholars have researched the similarity calculation of English and Chinese texts by using deep learning technology.

The DSSM (Deep Structured Semantic Models) algorithm proposed by Huang [12] was designed to apply a Siamese network architecture for semantic text similarity calculation. The DSSM model achieved outstanding performance in a text-matching task but ignored word order information and contextual information. Shen et al. [13] introduced CNN (Convolutional Neural Networks) networks into the DSSM model to retain more contextual information. The improvement of this method to DSSM occurred mainly in the representation layer, where convolutional and pooling layers were added so that the contextual information was effectively retained. However, the contextual information was still lost at longer distances due to the limitation of convolutional kernels. To retain more contextual information, Palangi et al. [14] introduced the LSTM (Long Short-Term Memory) [15] network, which took into account more distant contextual information and some discourse order information to make the algorithm more practical. Mueller et al. [16] also encoded sentences using Siamese-LSTM based on pre-trained word vectors. It was experimentally demonstrated that the combination of this method and a SVM (Support Vector Machines) [17] for sentiment classification resulted in significant improvement.

With the application of self-attention technology in the field of image and natural language processing, Lin et al. [18] combined BiLSTM (Bi-directional Long Short-Term Memory) with self-attention [19] technology to obtain sentence vector representation and kept a Siamese architecture in the training network, which improved the precision of text matching. Pontes et al. [20] applied the CNN and LSTM models to calculate semantic text similarity, which improved the text similarity calculation. The method based on the Siamese network is independent when coding sentences at the coding layer, and there is no interaction between sentence pairs, which would constrain the model's capability to calculate the semantic similarity of sentence pairs. This limitation can be eliminated by the interaction model, which adds interaction between two parallel networks based on twin networks; thus, more abundant interactive information is extracted between sentence

pairs. Yin et al. [21] proposed the ABCNN (Attention-Based Convolutional Neural Network) model based on word vectors to process sentences through CNN. While convolution and pooling of sentences in sentence pairs are independently carried out, an attention mechanism is used to connect two intermediate steps. Wang et al. [22] proposed a BiMPM (Bilateral Multi-perspective Matching) model based on the BiLSTM network. The input layer utilizes word vector and character vector splicing. Gong et al. [23] proposed the DIIN (Densely Interactive Inference Network) model. The input layer utilizes word embedding, character feature, and syntactic features concatenation; the coding layer utilizes a self-attention mechanism; the interaction layer utilizes dot-product operation to obtain the interaction matrix and then utilizes DenseNet [24] to extract the features. The extracted features are incorporated into a multi-layer perceptron model to obtain the final results. This simple and effective method achieved good performance in the NLI task. The above research shows that the interaction model performs better in text-similarity matching. Compared with the convolutional neural network, the recurrent neural network performs better in dealing with sequence problems. In recurrent neural network, GRU (Gated Recurrent Unit) has the advantages of fewer parameters, simple structure, ease of calculation, and convergence. We utilized the GRU of a multi-layer intensive connection to extract text features and utilized a connection operation to combine the attention mechanism information for the interaction between two question sentences into the repeated features of a dense connection. However, because of the lack of a large-scale dataset available in the agricultural field, there is little research on the similarity calculation for agricultural texts. The main contributions of this paper are as follows.

(1) A data-set of 21,300 rice-related questions and answers was constructed, and 8000 common high-quality rice-related question data were extracted, and 32,000 rice-related question pairs were divided into five categories.

(2) Combined with agricultural word segmentation dictionary, we utilized Word2vec [25] with the TF-IDF (Term Frequency–Inverse Document Frequency) [26] method, which can effectively solve the problems of high dimensions and sparsity of data in rice-related texts.

(3) The vector representation of input sentences can be obtained using a stacked GRU neural network, which can effectively capture the sentences' semantics. The model combined with an attention mechanism to encode the sentences to obtain the interaction and influence between rice-related question pairs.

2. Corpus Preparation

The data in this study were derived from the Q&A community of China Agricultural Technology Extension Information Platform. We applied Python's Regular Expressions to clean and filter the obtained text data to remove useless information. More than 20,000 pairs of Q&A community data related to rice cultivation, fertilization, weeding, pest control, and other aspects were captured; among them 8000 high-quality pairs were selected for our dataset to be used as our FAQs. Moreover, these 8000 rice-related questions were classified into five categories: diseases and pests, weeds and pesticides, cultivation management, storage and transportation, and other.

The input of the model comprised two sentences and their similarity tags. Firstly, 8000 rice-related frequently asked questions were manually combined, and their similarities were calculated.

(1) Question classification. We applied the set $QS = \{q_1, q_2, q_3, \cdots, q_{8000}\}$ to represent the questions for our 8000 FAQs, where $q_n (1 \leq n \leq 8000)$ represents a specific question. Each question was classified into one of the five categories in QS. After that, we classified similar questions within each category; in total, these 8000 questions were classified into 1200 classes. Then, 1200 types of questions were obtained, namely $QS' = \{Q_1, Q_2, \cdots, Q_{1200}\}$, where $Q_m = \{q_{m1}, q_{m2}, q_{m3}, \cdots, q_{mk}\}$ ($1 \leq m \leq 1200$, $1 \leq k$) represents a set of similar questions with one or more different methods, and q_{mk} stands for problem class Q_m.

(2)　Question combination. For a specific question q_{11} in subset $Q_1 = \{q_{11}, q_{12}, \cdots, q_{1k}\}$ of QS', the number of questions similar to q_{11} is $\{q_{12}, q_{13}, \cdots, q_{1k}\}$, and the number is $k - 1$. The questions that are not similar to q_{11} exist in the complement $C_{QS}Q_1$ of Q_1, which consists of two parts: one is for randomly extracting $(k - 1)/2$ questions from $C_{QS}Q_1$; the other is for extracting the first $(k - 1)/2$ questions that have the highest number of identical keywords as q_{11} from $C_{QS}Q_1$. In this way, we can avoid making the neural network think that, the higher the number of identical keywords, the more similar they are, so as to better learn the features of the two sentences from the semantic level. The specific process is shown in Figure 1 below.
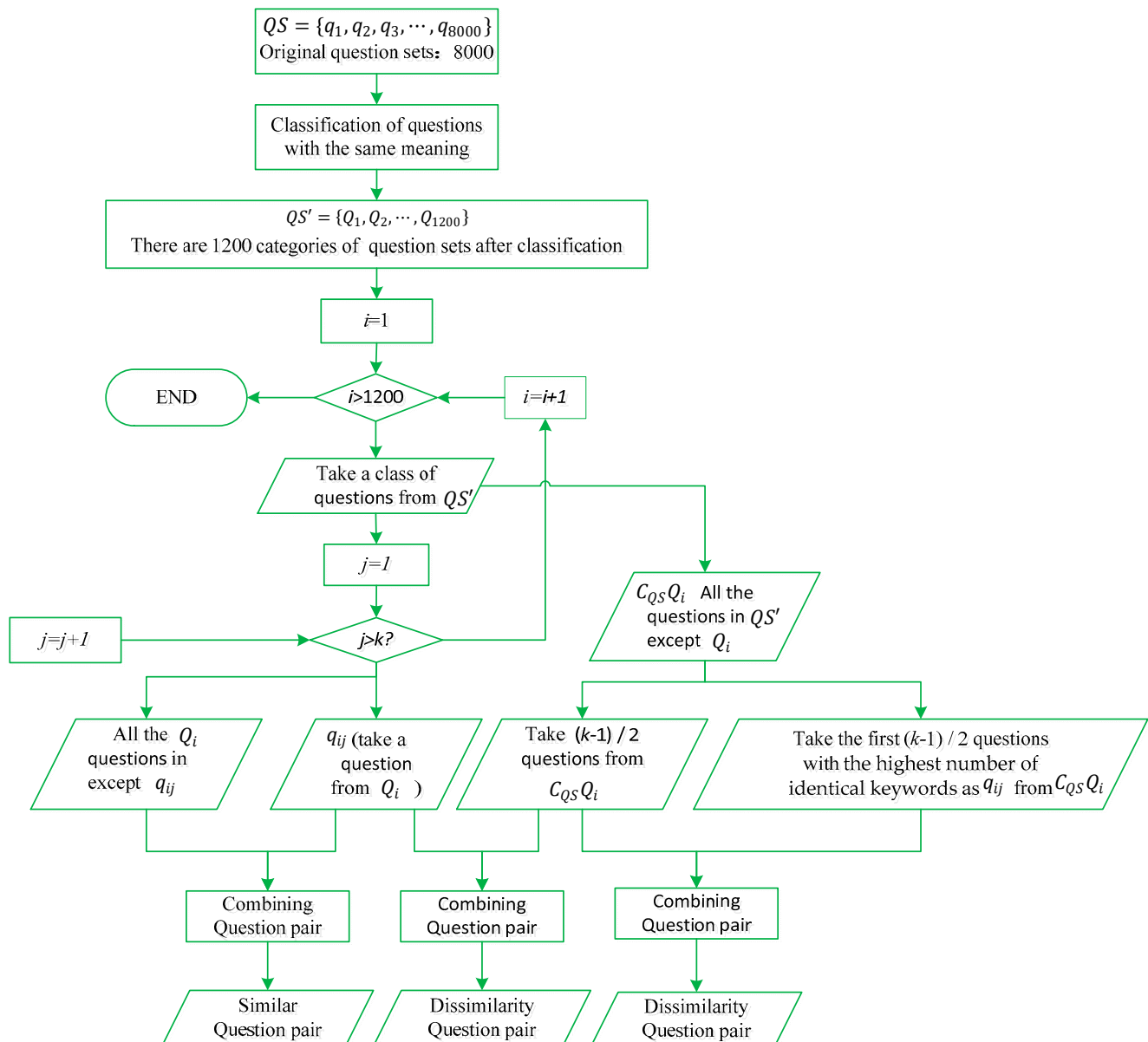

Figure 1. Flow chart of dataset construction.

After the above processing, 32,000 question pairs were obtained; based on the first question of the pair, we classified the question pairs into five categories, including 11,650, 2773, 10,767, 3658, and 5152 pairs of data regarding diseases and pests, weeds, and pesticides, cultivation management, storage and transportation, and others. Question 1 and question 2 represent two problems after word segmentation, 1 represents similarity, and 0 represents dissimilarity. Examples of training set samples are shown in Table 1:

Table 1. Sample of training set.

Quesetion1	Quesetion2	Similarity
How to control rice blast?	What are the control methods of rice blast?	1
How to treat rice bacterial streak?	What are the current integrated control measures of rice bacterial streak?	1
What are the characteristics of rice bakanae?	What are the most effective methods to control rice seedling disease?	0
What should be focused on when raising ducks in rice fields?	What should be focused on during fish farming in rice?	0
Are genetic factor responsible for the formation and rate of rice blight?	Do genetic factors account for the formation of rice empty grain rate?	1
What problems should be paid attention to before rice sowing?	What are the conditions of the whole rice field before sowing?	1
How to use imipramine to control rice seedling disease?	The cause of rice bakanae?	0

3. Coattention-DenseGRU Model

This paper utilized the Coattention-DenseGRU model, shown in Figure 2. The model consists of four parts: the text preprocessing layer, DenseGRU layer, coattention layer, and interactive classification layer. Compared with the traditional deep learning classification model, Coattention-DenseGRU added weighted preprocessing to the text. In this study we utilized Word2vec with the TF-IDF algorithm to expand the text feature words and calculated the weighted word vector according to its importance. A variety of methods were used to extract text features, and DenseGRU and coattention were used to extract local features of different granularities of text. Finally, the extracted feature vectors were input into the interactive classification layer.
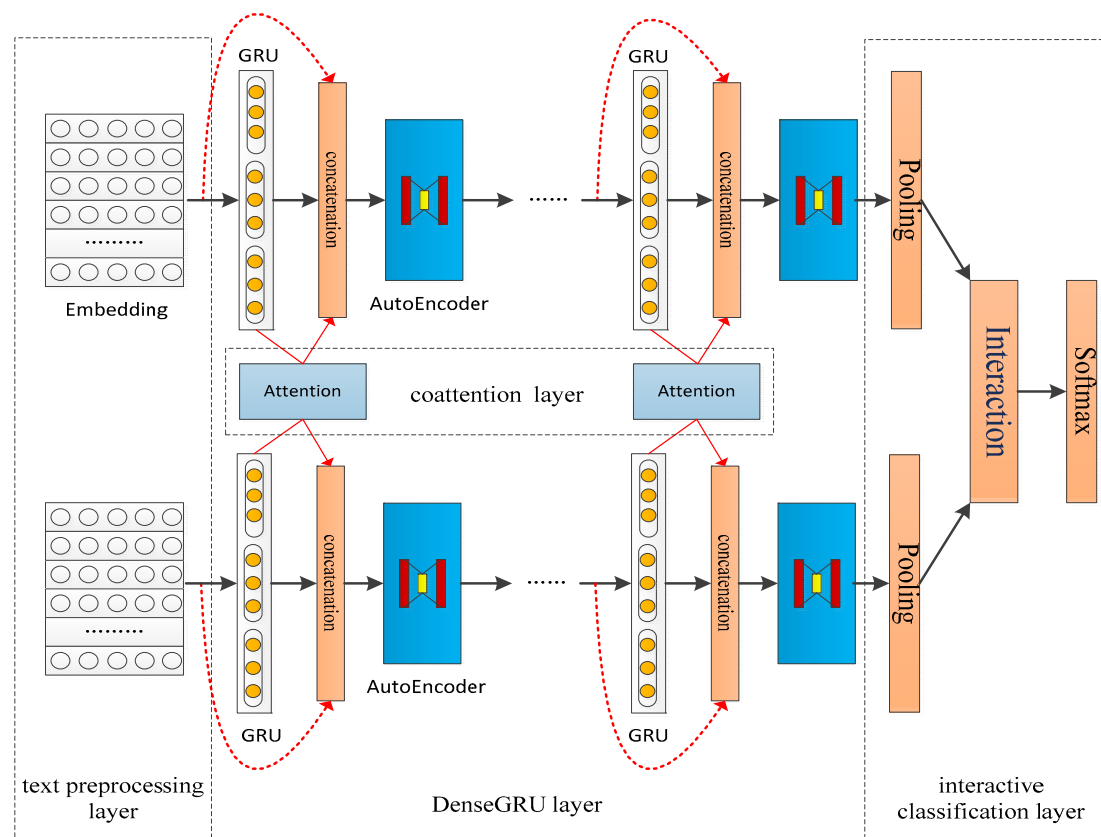
Figure 2. Model architecture diagram.

3.1. Text Preprocessing Layer

The computer cannot classify text directly as the model's input, it is necessary to convert text into a digital vector. In order to keep the text features and semantic information as complete and comprehensive as possible, we first preprocessed the question text, for example by means of noise removal and word segmentation, we utilized Python's jieba to segment text. The segmentation results of Chinese are greatly influenced by semantics and context; in order to improve the precision of segmentation, the stop words table was loaded before segmentation, which can remove the noise of the words, memorable characters, and spaces in the text that are not conducive to feature extraction and reduce the redundant information of the text. Based on the rice-related question-and-answer dataset characteristics, we loaded the Sogou agricultural vocabulary as a word segmentation dictionary instead of the primary vocabulary and improved agricultural vocabulary recognition. And then utilized a word vector transformation tool to convert the segmentation result into a word vector.

Word2vec has become a popular distributed representation method for text in recent years. Word2vec can predict the contextual information according to the input target words and map words with a similar meaning to similar positions in the vector space, which effectively solves word vector isolation and high dimensions. In this study, the skip-gram model of Word2vec was used to train the segmentation results, and words were transformed into low-dimension and continuous word vectors. To further highlight the contribution of the representative feature words to the importance of this paper, the TF-IDF value of words and the word vector represented by Word2vec were weighted.

$$w\_R(t_i) = word2vec(t_i) \times s_{i,j} \tag{1}$$

$w\_R(t_i)$—weighted word vector of word t_i, $word2vec(t_i)$—Word2vec vector of word t_i, $s_{i,j}$—TF-IDF value of the word t_i.

After obtaining the weighted word vector of each word, each word in the text was replaced by its corresponding word vector to form a weighted text vector group. The different lengths of the questions must be unified to input them into the neural network model for training. According to the statistics of our rice-related question data, 99.9% of the questions contained fewer than 100 words, so we set the length of the questions to 100. If the rest of the questions were not long enough, we filled in 0 to complete the text vector. If the length exceeded 100, only the first 100 words were taken.

3.2. DenseGRU Layer

GRU is a special kind of recurrent neural network that can effectively solve the gradient problem in the long-term memory and backpropagation of recurrent neural networks. Compared with LSTM, GRU [27] has fewer parameters, a more straightforward structure, easier calculation, and more substantial convergence.

The GRU structure includes two states and two control gates: hidden state h, candidate state, reset gate r, and update gate z. The update gate decides how much of the previous information will be transferred into the current state, and the reset gate decides how much will be ignored. At the time t, the computation of r_t depends on the input word vectors x_t and h_{t-1}, r_t acts on h_{t-1}, and the degree of preserving the past implied state is controlled according to its importance h_{t-1}. The greater the r_t, the greater the influence of h_{t-1} on the growth rate. Calculation formula of GRU.

x_t is the input vector of the t time step, which will be multiplied by the weight matrix $w_z(w_r)$ through a linear transformation. h_{t-1} stores the information of the previous time step $t-1$, which also goes through a linear transformation; the reset gate and update gate add these two parts of information and put them into a sigmoid activation function.

$$r_t = \sigma_g(W_r x_t + U_r h_{t-1}) \tag{2}$$

$$z_t = \sigma_g(W_z x_t + U_z h_{t-1}) \tag{3}$$

We entered x_t and last time step information h_{t-1} through a linear transformation; the matrixes W and U were right-multiplied, respectively calculating the Hadamard product of reset gate r_t and h_{t-1}. Z_t is the activation result of the update gate. The Hadamard product of z_t and h_{t-1} represents the information retained in the previous time step to the final memory.

$$\widetilde{h}_t = \tanh(Wx_t + U(r_t \odot h_{t-1})) \tag{4}$$

$$h_t = z_t \odot h_{t-1} + (1 - z_t) \odot \widetilde{h}_t \tag{5}$$

The GRU neural network is a one-way output from front to back. This is different from the structure of Chinese semantics, which is related to the context. In the task of question similarity calculation, if the current moment's output can be related to the state of the previous and subsequent moments, it will be more conducive to extract high-level features of the text and highlight the text's critical information. Based on Chinese semantic understanding characteristics, we utilized the BiGRU model to extract the feature vectors of questions. The BiGRU model is a neural network model composed of two multiple unidirectional GRUs with opposite directions. The word vector of the *j*-th word of the *i*-th sentence input at time *t* is c_{tij}, and the state of hidden layer h_t is weighted by forward the hidden layer state $\overrightarrow{h}_{t-1}$ and reverse hidden layer state $\overleftarrow{h}_{t-1}$. GRU(·): nonlinear transformation of word vector, w_t: forward weight matrix, v_t: inverse weight matrix, b_t: offset value.

$$\overrightarrow{h}_t = GRU(c_{tij}, \overrightarrow{h}_{t-1}) \tag{6}$$

$$\overleftarrow{h}_t = GRU(c_{tij}, \overleftarrow{h}_{t-1}) \tag{7}$$

$$h_t = w_t \overrightarrow{h}_t + v_t \overleftarrow{h}_t + b_t \tag{8}$$

This layer is the key to the model. It adopts the structure of multi-layer GRUs stacked together with DenseNet and circulated four times. We employed the bidirectional GRU (BiGRU) as a base block of H_l; L represents the number of GRU layers, t represents the time, and its hidden state value is:

$$h_t^l = H_l\left(x_t^l, h_{t-1}^l\right) \tag{9}$$

$$x_t^l = h_t^{l-1} + x_t^{l-1} \tag{10}$$

However, the incomplete network structure also has some disadvantages, which will hinder the transmission of information between networks. Therefore, DenseNet is used to solve this problem. Its tail is not an additive structure but a splicing structure. In this way, it does not hinder the transmission of information and retains the original information; that is, the output value of the first layer can be effectively transmitted to the last layer, avoiding loss of the gradient. The hidden state is as follows:

$$h_t^l = H(x_t^l, h_{t-1}^l) \tag{11}$$

$$x_t^l = [h_t^{l-1}; x_t^{l-1}] \tag{12}$$

3.3. Coattention Layer

The attention mechanism has achieved great success in many fields. It is an effective technique to learn context vector matching on specific sequences. Given two sentences, in each GRU layer, the context vector is determined using an attention mechanism that focuses on the related parts of the two sentences. The calculated attention information represents the soft alignment between two sentences. We utilized the operation to merge-common concern information into the repeated features of DenseGRU. The dense relation feature series circulation and standard attention features were obtained from the bottom to the top, enriching the collective awareness of lexical and part semantics. The weighted total of

the attention knowledge σ_{pi} of *i*-th word $pi \in P$ of sentence q is determined by h_{hi}, and the weighted value is the Softmax weight, as follows:

$$e_{i,j} = \cos(h_{pi}, h_{hi}) \tag{13}$$

$$\sigma_{i,j} = \frac{\exp(e_{i,j})}{\sum\limits_{k=1}^{J} \exp(e_{i,k})} \tag{14}$$

$$\sigma_{pi} = \sum_{j=1}^{J} \alpha_{i,j} h_{hj} \tag{15}$$

We paid attention to context vector $e_{i,j}$ and trigger vector h_{hi} connection, keeping attention information as the input of the next layer:

$$h_t^l = H_l(x_t^l, h_{t-1}^l) \tag{16}$$

$$x_t^l = [h_t^{l-1}; a_t^{l-1}; x_t^{l-1}] \tag{17}$$

3.4. Interactive Classification Layer

The model presented in this paper treats all layers' production as a group of semantic information. However, the network is a system with input features that increase in number as the layer depth increases, and it has too many parameters, especially in the link layer. An automated encoder was used to minimize the number of features while preserving the original information structure to solve this problem. Additionally, this part served as a regularization in the experiment, which improved the test performance.

To extract each sentence's appropriate representation, we pooled the densely connected GRU with attention features step by step. Specifically, if the GRU layer's final output was a 100-dimensional vector of 30 words in a sentence, a 30 * 100 matrix was obtained so that the size of the synthesized vector M or N was 100. Then, the representation forms of the two sentences M and N were aggregated in various ways in the interaction layer, and finally, the feature vector *D* of semantic sentence matching was obtained as follows:

$$D = [m; n; m + n; m - n; |m - n|] \tag{18}$$

We inferred the relationship between two sentences by performing operations +, −, |. |, from the perspective of elements. Element subtraction $m - n$ is an asymmetric operator for one-way tasks. Two completely linked layers were used after extracting the feature *D*. The activation function was ReLU, and the output layer was completely associated. Finally, the probability distribution of each class was calculated using the Softmax equation.

3.5. Model Training

In this paper, a stochastic gradient descent (SGD) [28] was used to optimize the model parameters. The random gradient descent algorithm trained one sample with one category label each time to update the parameters. The objective function is as follows:

$$\varphi = \varphi - \eta \nabla_{\varphi} J(\varphi; x^{(i)}; y^{(i)}) \tag{19}$$

φ—objective function, η—learning rate, $x^{(i)}$—sample, $y^{(i)}$—category label, $\nabla_{\varphi} J$—parameter gradient.

The discrepancy between the probability distribution obtained by current training and the real distribution was evaluated using the cross-entropy loss function. This was 1 if

the semantics of the query pairs were the same; otherwise, it was 0. The cross-entropy loss function has the following formula:

$$L(p,y) = -\sum_{c=1}^{M} y_c \log(p_c) \tag{20}$$

M—number of categories, *y*—indicating variable (0 or 1), and *p*—the prediction probability of observation samples belonging to category C.

4. Experiments

4.1. Hardware, Software Environment, and Evaluation Indicators

The experimental software environment was Python 3.6.2 and TensorFlow 1.13.1, and the server's hardware environment was an NVIDIA Corporation device 1e04 (Rev A1); the GPU was an NVIDIA GeForce RTX 2080ti. In this study, the TensorFlow neural network framework was used to construct the neural network. A total of 32,000 question pairs were divided into the training set, verification set, and test set according to the ratio 7:2:1. The random gradient descent algorithm was used to update the model weight. There were 22,400 training sets, 6400 verification sets, and 3200 test sets. Precision (P), recall (R), and F1-score (F1) were used as evaluation indexes. The formulas are as follows:

$$P = \frac{TP}{TP + FP} \tag{21}$$

$$R = \frac{TP}{TP + TN} \tag{22}$$

$$F_1 = \frac{2PR}{P + R} \times 100\% \tag{23}$$

4.2. Text Vectorization Processing and Analysis

In this paper, we applied word2vec with TF-IDF to vectorize the rice-related question data. The word vector dimension was set to 300, and the training window size was set to 5. We compared the GloVe [29], TF-IDF, Word2vec and Word2vec with TF-IDF models. The results for precision, recall, and F1 values are shown in Table 2.

Table 2. Model classification effect under different embedding layers.

Word Embedding	P (%)	R (%)	F1 (%)
TF-IDF	81.7	69.9	75.3
GloVe	83.8	73.6	78.4
Word2vec	84.6	74.3	79.1
Word2vec + TF-IDF	86.3	77.3	81.6

It can be seen from Table 2 that among the four external neural networks based on text vectorization methods, word2vec with the TF-IDF method had the highest precision and F1 values compared with the other methods, with a precision of 86.3% and F1 value of 81.6%. The TF-IDF method had the worst outcomes. Although TF-IDF considers the semantic information between words, it does not solve the problems of high vector dimensions and sparse data; with the increase in extracting continuous words, the dimension will become higher. Compared with Word2vec, the precision and F1 values of Word2vec weighted by TF-IDF were improved by 1.7% and 2.5%, respectively, which shows that the precision and F1 values of the neural network can be improved by combining Word2vec with TF-IDF weighted representative feature words.

4.3. Parameter Setting

We set the number of training rounds of the model to 50, and we set the learning rate to 0.01. Moreover, we set the dense connection recurrent neural network to 5 layers; each layer had 100 hidden units, and the hidden units of the whole connection layer were set to 1000. After the word and character embedding layer, we set dropout to 0.5. For the autoencoder, 200 hidden units are set as the auto encoder's encoding features, and dropout was set to 0.2. We applied the rmsprop optimizer with an initial learning rate of 0.001. We test Coattention-DenseGRU and BiLSTM [30], Selfattention-BiLSTM [31], TextCNN [32],ABCNN, BiGRU [33], Attention-BiGRU [34], DenseGRU on the rice-related question similarity pair dataset.

5. Results and Discussion

Table 3 shows the comparison of precision, recall, and F1 values of the eight different deep learning models. The proposed model, Coattention-DenseGRU, achieved the highest F1 value and precision, and the precision and F1 value reached 96.3% and 96.9%, respectively, which shows that the dense connected GRU can enhance the transmission and extraction of features and reduce feature loss, and is conducive to the final matching effect. Compared with traditional BiLSTM, BiLSTM based on the self-attention mechanism had better precision and F1 values, but it performed slightly worse than DenseGRU, which indicates that the attention mechanism can better express feature information through weight reset during training. The DenseGRU model was better than BiLSTM model in the training outcomes. In the DenseGRU model, the features were extracted through five GRUs connected densely. Compared with other models, only the previous layer's output was used as the input of this layer when features were transferred. DenseGRU takes the output of all previous layers instead of the previous one when features are transferred to reduce the loss of text features effectively. Through densely connected GRUs, text features can be better transferred and expressed and improve the text matching effect.

Table 3. Outcome of different models on rice-related question dataset.

Models	P (%)	R (%)	F1 (%)
BiLSTM	93.7	94.6	94.1
Selfattention-BiLSTM	94.6	95.6	95.1
TextCNN	94.7	95.4	95.0
ABCNN	95.6	96.3	95.9
BiGRU	92.3	93.7	93.0
Attention-BiGRU	93.3	94.9	94.1
DenseGRU	94.7	95.6	95.1
Coattention-DenseGRU	96.3	97.6	96.9

Figure 3 shows the text matching precision of the eight experimental models under Word2vec text representation and Word2vec with TF-IDF weighted text representation. As shown in Figure 3, the classification precision of the TF-IDF + Word2vec text representation method proposed in this paper is significantly higher than that of the word2vec text representation method in eight experimental models. The Coattention-DenseGRU model achieved the best results in word2vec text representation and TF-IDF + word2vec weighted text representation, and the precision were 96.3% and 91.5%, respectively. Compared with the other six comparative models, the Coattention-DenseGRU model has significant advantages. It can be seen from Figure 3 that the weighted text representation method of TF-IDF + word2vec improved the precision in each group of the comparative experiments. Therefore, the weighted method of TF-IDF + word2vec can improve the importance of keywords and the precision of question similarity matching.
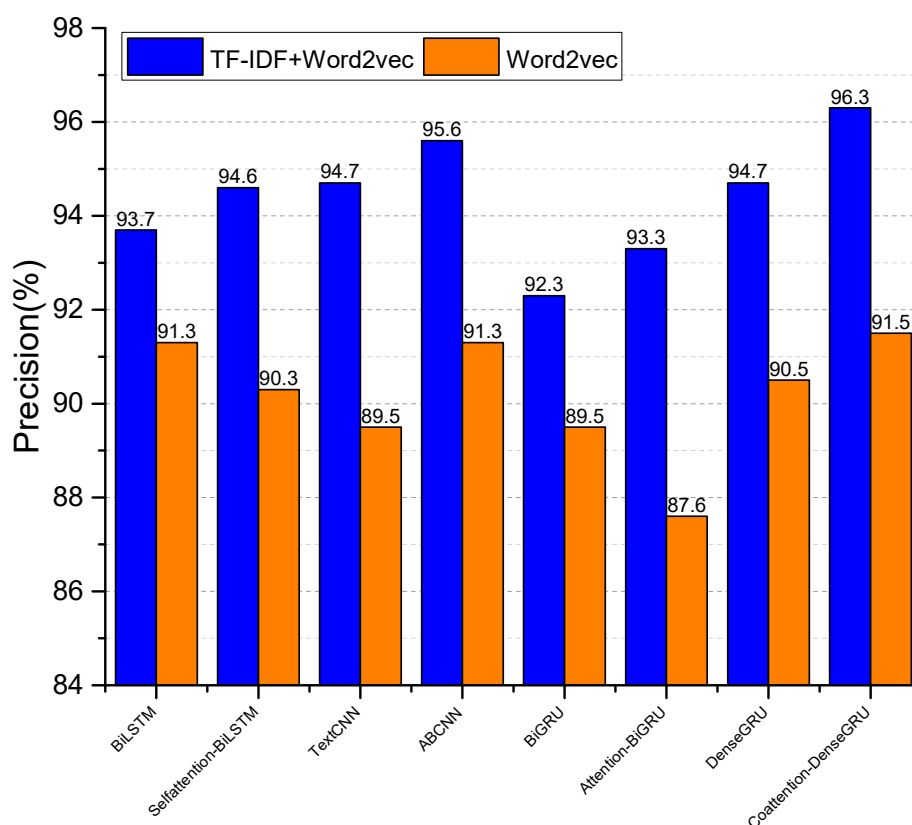
Figure 3. Precision of question similarity matching under different models.

As can be seen from Table 4, compared with BiLSTM, Selfattention-BiLSTM, TextCNN, ABCNN, BiGRU, Attention-BiGRU and DenseGRU, Coattention-DenseGRU had the highest matching performance in the dataset of rice-related question pairs (five categories) of diseases and pests, weeds and pesticides, cultivation management, storage and transportation, and other. The precision, recall, and F1 value of matching rice-related question pairs are greater than 93.6%, 92.7%, and 94.9%, respectively, and the overall classification effect was better than other models. The F1 value of this model was slightly higher than that of other models in the dataset with sufficient data of diseases and pests, cultivation management experiments. The F1 value of this model is significantly higher than that of other models in the data sets with fewer data relating to weeds and pesticides, storage and transportation, and other, three categories, which indicates that the Coattention-DenseGRU model can still effectively extract the features of a short text in the case of insufficient data.

Table 4. Outcomes of different models on rice-related question dataset.

Model	P (%)					R (%)					F1 (%)				
	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
BiLSTM	80.1	93.7	94.2	89.7	79.7	95.2	96.6	80.1	63.7	77.8	87.0	95.1	86.6	74.5	78.7
Selfattention-BiLSTM	92.5	91.9	92.7	85.5	80.1	92.1	92.6	93.1	85.6	82.3	92.3	92.2	92.9	85.6	81.1
TextCNN	94.7	92.4	94.0	94.7	88.4	94.0	94.7	92.4	94.0	87.6	94.3	93.5	93.1	94.3	88.0
ABCNN	96.6	95.3	96.9	92.6	90.3	94.9	95.6	94.3	94.9	95.6	95.7	95.4	95.6	93.7	92.9
BiGRU	92.3	87.7	91.5	92.3	87.7	91.5	92.3	87.7	91.5	92.3	91.9	89.9	89.6	91.9	89.9
Attention-BiGRU	93.3	89.9	92.0	93.3	89.9	92.0	93.3	89.9	92.0	93.3	92.6	91.6	90.9	92.6	91.6
DenseGRU	94.7	91.7	92.4	90.8	91.1	95.1	92.7	93.6	92.1	90.7	94.9	92.2	93.0	91.4	90.9
Coattention-DenseGRU	97.2	96.6	98.7	97.3	93.6	96.7	95.3	94.6	92.7	94.3	96.9	95.9	96.6	94.9	93.9

Note: 1, 2, 3, 4 and 5 represent the data (five categories) of diseases and pests, weeds and pesticides, cultivation management, storage and transportation, and other, respectively.

Table 5 shows that a set of experiments was undertaken to investigate the effectiveness of each module in the Coattention-DenseGRU model. Firstly, model 2 was obtained after deleting the autoencoder in the model. It can be seen that the precision and recall rate of model 2 was decreased, which verified the effectiveness of the self-encoder. Then, we deleted the dense connection and collaborative attention mechanism between GRUs and obtained models 3 and 4. It can be seen that the precision and F1 value of models 3 and 4 decreased by 0.6% and 0.7%; the results show that the dense connection between GRUs can improve the effectiveness of the model more than the collaborative attention mechanism. Models 5 and 6 are a five-layer GRU model based on the attention mechanism and a five-layer GRU model without attention mechanism. From the Table 5, we can see that the attention mechanism can improve the model's effect by paying good attention to keywords in question similarity matching.

Table 5. Outcomes of different models on rice-related question dataset.

Label	Model	P (%)	R (%)	F1 (%)
1	Coattention-DenseGRU	96.3	97.6	96.9
2	-autoencoder	94.9	95.1	95.0
3	-Dense(Att)	93.6	94.8	94.2
4	-Dense(Rec)	94.6	96.1	95.3
5	-Dense(ATT+Rec)	91.7	92.3	92.0
6	GRU+Attention	89.9	91.1	90.5
7	GRU	88.1	89.5	88.8

Figure 4 shows the classification effect of models 1–7 on the rice-related question dataset at different GRU levels. From the figure, it can be seen that Coattention-DenseGRU had the highest precision in a five-layer GRU, which shows that the text feature extraction can be effectively improved, and feature loss and classification efficiency can be improved by increasing the number of layers and dense connection. Models 6 and 7 had the highest precision at the second layer and then gradually declined, which indicates the feature loss will be caused by the multi-layer extraction of features without intensive connection of the GRU.
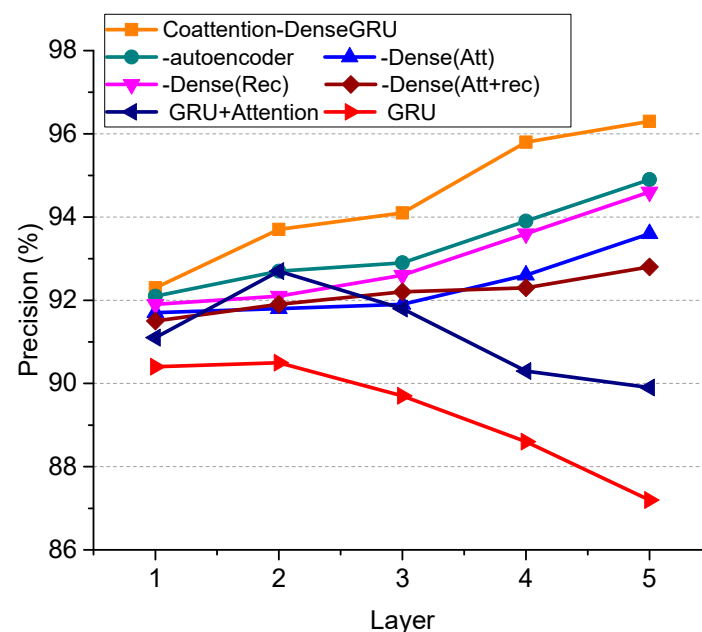

Figure 4. Matching outcomes of models in each layer network under different conditions.

Table 6 shows the response time and precision of four neural network models based on the attention mechanism on 3200 test sets, which meets the requirements for quick classification of rice-related question pairs. ABCNN is the fastest in response time due to the simple structure of the ABCNN model, fewer training layers, and fewer model parameters. The model proposed in this paper, Coattention-DenseGRU model was able to accurately judge rice-related question sentences' similarity in the test set of 3200 question pairs in 12 s; the precision rate reached 93.6%.

Table 6. Response time and precision of four network models.

Model	Response Time (S)	P (%)
ABCNN	10	92.8
Selfattention-BiLSTM	14	91.7
Attention-BiGRU	14	90.9
Coattention-DenseGRU	12	93.6

The features obtained through the GRU network's dense connection and collaborative attention mechanism are linked to the classification layer via the max-pooling layer, causing the loss function to be affected by the features of each layer, and the deep supervised learning is carried out. Therefore, we applied attention to weight and maximum pooling position to explain the classification results. Attention weight includes the information relating to question pairs and the max-pooling position information in each dimension. Attention weight plays a vital role in classification. Figure 5 shows the visualization of the attention weight value of the model in different layers. Except for duck and fish, most of the words in question 1 and question 2 exist simultaneously. In the first layer of the attention weight graph, the corresponding degree of the same or similar words in each sentence was higher. However, with the increase in the level, the attention weight of ducks and fish also increased. There were apparent differences between ducks and fish; in the fifth layer, the attention weights of other words in question 1 and question 2 except ducks and fish became very small. As there are noticeable semantic differences between "fish" and "ducks", the model judged that the question pairs were semantically dissimilar; that is, the label was 0.
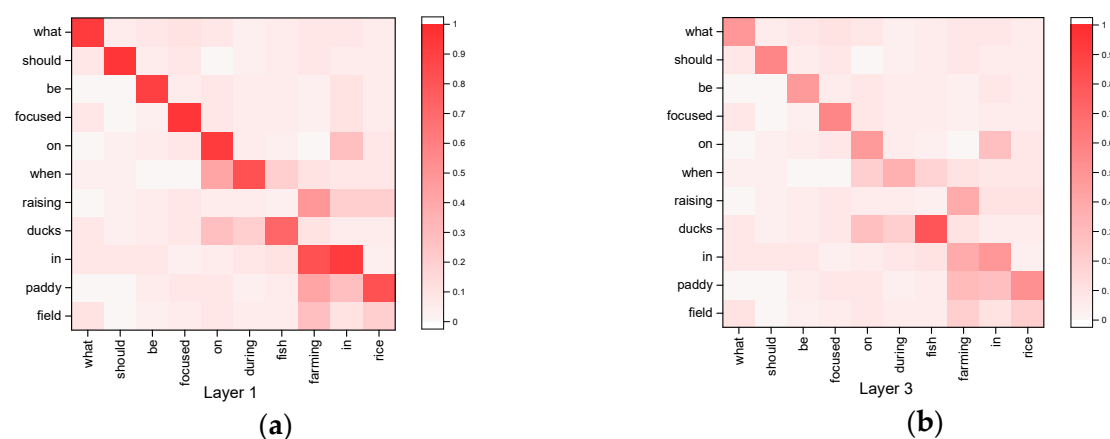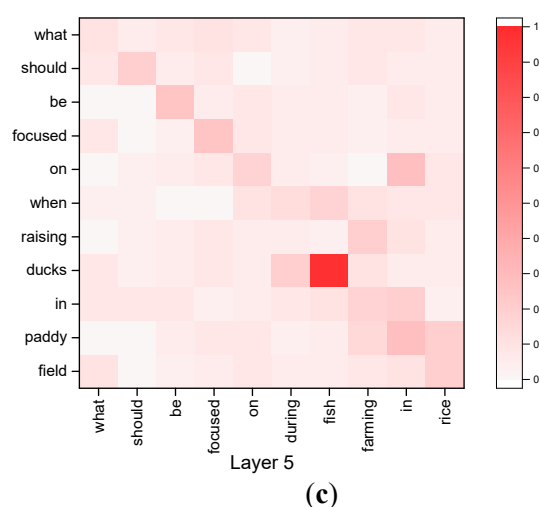

Figure 5. *Cont.*

Figure 5. Visualization of attention weight of rice-related question similarity with different layers. The darker the color, the higher the value. (**a**): Visualization of attention weight of rice-related question similarity with layer 1. (**b**): Visualization of attention of rice-related question similarity weight with layer 3. (**c**): Visualization of attention of rice-related question similarity weight with layer 5.

6. Conclusions

To solve the problem that Chinese Agricultural Technology Extension Q&A communities not being able to automatically and accurately detect repetitive semantic questions, a corpus with five categories and 32,000 pairs of rice-related questions was constructed. A densely connected GRU model based on the coattention mechanism was introduced to solve the problems of rice-related question matching. We utilized the dense connection GRU model based on the collaborative attention mechanism to carry out rapid and automatic repetitive semantic detection of rice-related Q&A community query data. We introduced the agricultural word segmentation dictionary to word segmentation and word vector representation. We utilized the DenseGRU network to extract texts' emotional expression as a text feature vector used for question similarity matching. Furthermore, we optimized and improved its important structural parameters and training strategies and built a rice-related text similarity matching algorithm based on Coattention-DenseGRU to realize the precise and efficient identification of the rice-related questions in a question-and-answer community. The proposed model achieved the best performance on the rice-related question similarity dataset compared to other models. Future work will focus on the following three aspects:

(1) The noise errors and limited amount of data for rice-related questions will be revised and expanded.

(2) On the China Agricultural Technology Extension Information Platform, there are corresponding pictures uploaded with all question-answering data. Nowadays, the multimodal fusion question-answering system that integrates image and text representation has achieved good results. In future work, using the multimodal question-answering model of images and text will be our focus.

(3) Some useful features and advanced pre-trained models, such as BERT, will be used to further improve the model outcomes.

Author Contributions: Conceptualization, T.X. and H.W. (Haoriqin Wang); methodology, H.W. (Haoriqin Wang); software, H.W. (Haoriqin Wang); validation, H.W. (Huarui Wu), H.Z. and T.X.; formal analysis, H.W. (Haoriqin Wang); investigation, X.W.; resources, T.X.; data curation, H.W.; writing—original draft preparation, H.W. (Huarui Wu), H.Z.; writing—review and editing, H.W. (Haoriqin Wang), H.W. (Huarui Wu); visualization, H.W. (Haoriqin Wang); supervision, X.H., H.Z.; project administration, X.H.; funding acquisition, X.W. All authors have read and agreed to the published version of the manuscript.

References

1. Li, M.; Li, Y.; Peng, Q.; Wang, J.; Yu, C. Evaluating community question-answering websites using interval-valued intuitionistic fuzzy DANP and TODIM methods. *Appl. Soft Comput.* **2021**, *99*. [CrossRef]
2. Young, T.; Hazarika, D.; Poria, S.; Cambria, E. Recent trends in deep learning based natural language processing. *IEEE Comput. Intell. Mag.* **2018**, *13*, 55–75. [CrossRef]
3. Selvalakshmi, B.; Subramaniam, M. Intelligent ontology based semantic information retrieval using feature selection and classification. *Clust. Comput.* **2018**, *22*, 12871–12881. [CrossRef]
4. Yogatama, D.; Smith, N.A. Linguistic structured sparsity in text categorization. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Baltimore, MD, USA, 22–27 June 2014; pp. 786–796.
5. Matous, P.; Todo, Y.; Pratiwi, A. The role of motorized transport and mobile phones in the diffusion of agricultural information in Tanggamus Regency, Indonesia. *Transportation* **2015**, *42*, 771–790. [CrossRef]
6. Liu, Y.; Tang, A.; Sun, Z.; Tang, W.; Cai, F.; Wang, C. An integrated retrieval framework for similar questions: Word-semantic embedded label clustering—LDA with question life cycle. *Inf. Sci.* **2020**, *537*, 227–245. [CrossRef]
7. Liu, L.; Sun, X.; Li, C.; Lei, Y. Classification of Medical Text Data Using Convolutional Neural Network-Support Vector Machine Method. *J. Med. Imaging Health Inform.* **2020**, *10*, 1746–1753. [CrossRef]
8. Altınel, B.; Ganiz, M.C. Semantic text classification: A survey of past and recent advances. *Inf. Process. Manag.* **2018**, *54*, 1129–1153. [CrossRef]
9. Li, C.; Liu, F.; Li, P. Text Similarity Computation Model for Identifying Rumor Based on Bayesian Network in Microblog. *Int. Arab. J. Inf. Technol.* **2020**, *17*, 731–741. [CrossRef]
10. Wang, Z.; Ho, S.-B.; Cambria, E. A review of emotion sensing: Categorization models and algorithms. *Multimed. Tools Appl.* **2020**, *79*, 35553–35582. [CrossRef]
11. Sun, S.; Cao, Z.; Zhu, H.; Zhao, J. A Survey of Optimization Methods From a Machine Learning Perspective. *IEEE Trans. Cybern.* **2020**, *50*, 3668–3681. [CrossRef] [PubMed]
12. Huang, P.-S.; He, X.; Gao, J.; Deng, L.; Acero, A.; Heck, L. Learning deep structured semantic models for web search using clickthrough data. In Proceedings of the 22nd ACM International Conference on Information & Knowledge Management, Online, 19–23 October 2020; pp. 2333–2338.
13. Shen, Y.; He, X.; Gao, J.; Deng, L.; Mesnil, G. A Latent Semantic Model with Convolutional-Pooling Structure for Information Retrieval. In Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management, Online, 19–23 October 2020; pp. 101–110.
14. Palangi, H.; Deng, L.; Shen, Y.; Gao, J.; He, X.; Chen, J.; Song, X.; Ward, R. Semantic modelling with long-short-term memory for information retrieval. *arXiv* **2014**, arXiv:1412.6629 2014.
15. Yao, L.; Pan, Z.; Ning, H. Unlabeled Short Text Similarity With LSTM Encoder. *IEEE Access* **2019**, *7*, 3430–3437. [CrossRef]
16. Mueller, J.; Thyagarajan, A. Siamese recurrent architectures for learning sentence similarity. In Proceedings of the AAAI Conference on Artificial Intelligence, Phoenix, AZ, USA, 12–17 February 2016.
17. Chauhan, V.K.; Dahiya, K.; Sharma, A. Problem formulations and solvers in linear SVM: A review. *Artif. Intell. Rev.* **2018**, *52*, 803–855. [CrossRef]
18. Lin, Z.; Feng, M.; Santos, C.N.d.; Yu, M.; Xiang, B.; Zhou, B.; Bengio, Y. A structured self-attentive sentence embedding. *arXiv* **2017**, arXiv:1703.03130 2017.
19. Xie, J.; Chen, B.; Gu, X.; Liang, F.; Xu, X. Self-Attention-Based BiLSTM Model for Short Text Fine-Grained Sentiment Classification. *IEEE Access* **2019**, *7*, 180558–180570. [CrossRef]

20. Pontes, E.L.; Huet, S.; Linhares, A.C.; Torres-Moreno, J.-M. Predicting the semantic textual similarity with siamese CNN and LSTM. *arXiv* **2018**, arXiv:1810.10641.
21. Yin, W.; Schütze, H.; Xiang, B.; Zhou, B. Abcnn: Attention-based convolutional neural network for modeling sentence pairs. *Trans. Assoc. Comput. Linguist.* **2016**, *4*, 259–272. [CrossRef]
22. Wang, Z.; Hamza, W.; Florian, R. Bilateral multi-perspective matching for natural language sentences. *arXiv* **2017**, arXiv:1702.03814.
23. Gong, Y.; Luo, H.; Zhang, J. Natural Language Inference over Interaction Space. *arXiv* **2017**, arXiv:1709.04348.
24. Huang, G.; Liu, S.; Laurens, V.; Weinberger, K.Q. CondenseNet: An Efficient DenseNet using Learned Group Convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2017.
25. Rong, X. word2vec Parameter Learning Explained. *arXiv* **2014**, arXiv:1411.2738.
26. Zhou, Z.; Qin, J.; Xiang, X.; Tan, Y.; Liu, Q.; Xiong, N.N. News Text Topic Clustering Optimized Method Based on TF-IDF Algorithm on Spark. *Comput. Mater. Contin.* **2020**, *62*, 217–231. [CrossRef]
27. Chung, J.; Gulcehre, C.; Cho, K.H.; Bengio, Y. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. *arXiv* **2014**, arXiv:1412.3555.
28. L Goyal, P.; Dollár, P.; Girshick, R.; Noordhuis, P.; Wesolowski, L.; Kyrola, A.; Tulloch, A.; Jia, Y.; He, K. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv* **2017**, arXiv:1706.02677.
29. Pennington, J.; Socher, R.; Manning, C. Glove: Global Vectors for Word Representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 25–29 October 2014; pp. 1532–1543.
30. Tan, M.; Santos, C.D.; Xiang, B.; Zhou, B. LSTM-based Deep Learning Models for non-factoid answer selection. *arXiv* **2016**, arXiv:1511.04108.
31. Li, W.; Qi, F.; Tang, M.; Yu, Z. Bidirectional LSTM with self-attention mechanism and multi-channel features for sentiment classification. *Neurocomputing* **2020**, *387*, 63–77. [CrossRef]
32. He, T.; Huang, W.; Qiao, Y.; Yao, J. Text-Attentional Convolutional Neural Networks for Scene Text Detection. *arXiv* **2016**, arXiv:1510.03283. [CrossRef] [PubMed]
33. She, D.; Jia, M. A BiGRU method for remaining useful life prediction of machinery. *Measurement* **2021**, *167*. [CrossRef]
34. Wang, W.; Sun, Y.; Qi, Q.; Meng, X. Text sentiment classification model based on BiGRU-attention neural network. *Appl. Res. Comput.* **2019**, *36*, 126–137.