big data and cognitive computing

MDPI

Article

Hierarchical Co-Attention Selection Network for Interpretable Fake News Detection

Xiaoyi Ge [1,†], Shuai Hao [2,†], Yuxiao Li [3], Bin Wei [1] and Mingshu Zhang [1,*]

[1] College of Cryptographic Engineering, Engineering University of PAP, Xi'an 710018, China
[2] Electrical and Computer Engineering, Stevens Institute of Technology, Hoboken, NJ 07030, USA
[3] Mathematics and Statistics, McGill University, Montreal, QC H3A 0G4, Canada
* Correspondence: zms2099@163.com
† These authors contributed equally to this work.

Abstract: Social media fake news has become a pervasive and problematic issue today with the development of the internet. Recent studies have utilized different artificial intelligence technologies to verify the truth of the news and provide explanations for the results, which have shown remarkable success in interpretable fake news detection. However, individuals' judgments of news are usually hierarchical, prioritizing valuable words above essential sentences, which is neglected by existing fake news detection models. In this paper, we propose an interpretable novel neural network-based model, the hierarchical co-attention selection network (HCSN), to predict whether the source post is fake, as well as an explanation that emphasizes important comments and particular words. The key insight of the HCSN model is to incorporate the Gumbel–Max trick in the hierarchical co-attention selection mechanism that captures sentence-level and word-level information from the source post and comments following the sequence of words–sentences–words–event. In addition, HCSN enjoys the additional benefit of interpretability—it provides a conscious explanation of how it reaches certain results by selecting comments and highlighting words. According to the experiments conducted on real-world datasets, our model outperformed state-of-the-art methods and generated reasonable explanations.

Keywords: fake news detection; interpretable AI; co-attention mechanism; hierarchical selection network

check for updates

Citation: Ge, X.; Hao, S.; Li, Y.; Wei, B.; Zhang, M. Hierarchical Co-Attention Selection Network for Interpretable Fake News Detection. *Big Data Cogn. Comput.* **2022**, *6*, 93. https://doi.org/10.3390/bdcc6030093

Academic Editors: Miltiadis D. Lytras and Andreea Claudia Serban

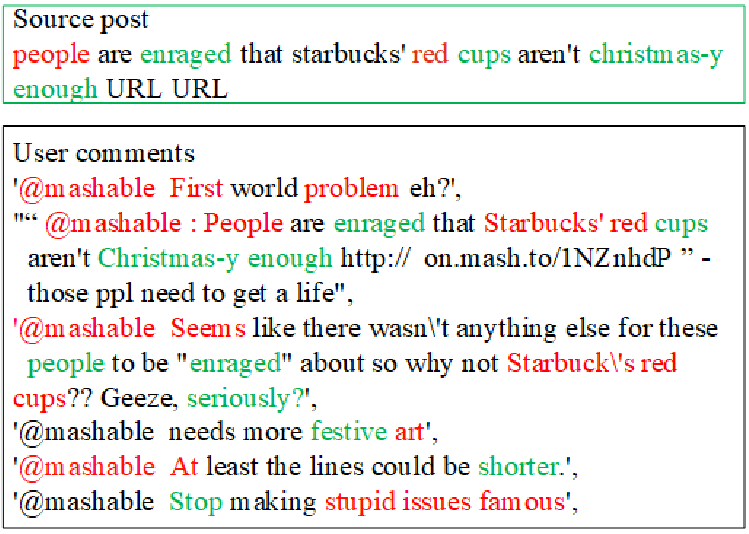
Received: 31 July 2022
Accepted: 1 September 2022
Published: 5 September 2022

1. Introduction

As a consequence of the booming growth of social media platforms, social media fake news has become a pervasive problem in society [1]. Given the ease with which individuals can freely and swiftly share their thoughts and feelings on social media platforms, fake news can spread quickly and easily distort people's assessment of a political [2,3] or economic event [4,5], public health [6–8], etc.

Substantial research, which attempts to use data mining and machine learning techniques, has been conducted in recent years on developing an effective and automated framework for detecting fake news. Relying on classical machine learning approaches, researchers extract special features and utilize supervised learning (e.g., support vector machine and random forest) for the detection of fake news [9–11]. With the advancement of deep learning, user profiles [12,13], user responses [14–16], and the propagation of news [17] have been used for learning the hidden representation of news through neural networks (e.g., recurrent neural networks (RNN) and graph neural networks (GNN)). These methods can improve the detection performance of fake news, but it is challenging to provide a reasonable explanation of the detection results. To address this issue, fake news detection models utilize different mechanisms to provide explanations through user comments [15], user information [18], and retweet sequences [19].

In interpretable fake news detection, while existing efforts have been excellent, grey areas remain. Firstly, the model based on user information and forwarding sequence [18] not only takes time to obtain the user information and forwarding sequence but also involves user privacy. Secondly, the current interpretable models adopt a hierarchical approach to achieve excellent classification performance, such as word–post–subevent–event [20] and word–sentence–event [21], but they ignore the relevance between the source post and user comments, which can cause higher trust in the explanation. Finally, although some current hierarchical interpretable fake news detection models emphasize the correlation between the source posts and user comments to detect fake news, on the one hand, some ignore the effect of post-related tokens. For example, the interpretable model dEFEND [15] only considers the interpretation of sentence-level relevance between source posts and user comments, as shown in Figure 1, red depicts the highlighted words that the dEFEND used to construct sentence vectors. However, it ignores the green, shown in Figure 1, which are post-related words that should be paid more attention to when detecting fake news, and they should also be part of the explanation. On the other hand, some of them do not use a selection process to reduce the irrelevant information. The MAC model [22] uses the multi-head attention mechanism to build a word–document hierarchical model, considering word-level correlation, but it is applicable to the whole document and does not consider the selection process of extracting important information for reducing noise and unrelated sentences.



Source post
people are enraged that starbucks' red cups aren't christmas-y enough URL URL

User comments
'@mashable First world problem eh?',
"“ @mashable : People are enraged that Starbucks' red cups aren't Christmas-y enough http:// on.mash.to/1NZnhdP ” - those ppl need to get a life",
'@mashable Seems like there wasn\'t anything else for these people to be "enraged" about so why not Starbuck\'s red cups?? Geeze, seriously?',
'@mashable needs more festive art',
'@mashable At least the lines could be shorter.',
'@mashable Stop making stupid issues famous',

Figure 1. A fake news example from Twitter 15, where the red words represent the highlighted words in every single sentence by the attention mechanism for current fake news detection models, and the green words are post-related tokens, which should be paid attention to in source posts and user comments when confirming the truthfulness of the news.

To solve the issues mentioned above, we propose a novel hierarchical co-attention selection network (HCSN). First, a recurrent neural network (RNN) was utilized to learn word embeddings for source posts and comments, and a sentence embedding was constructed for each sentence based on a self-attention mechanism. Then we adopted a sentence-level co-attention mechanism to learn the correlation between the source post and comments and utilized the Gumbel–Max trick to select certain informative sentences for the next layer. We next utilized the word-level co-attention mechanism to catch the representation of the word-level correlation between the source posts and selected comments. In the end, the binary prediction was produced based on the final learn embedding. In addition, there were two types of explanations for the prediction results, which are sentence-level explanations and word-level explanations. To confirm the interpretability of the model, we also utilized a case study to compare the prediction results with three modification strategies.

The contributions are summarized as follows:

1. We propose an interpretable method HCSN to predict the veracity of news based on a realistic social media scenario.
2. We present a hierarchical co-attention structure that incorporates the Gumble–Max trick for selecting relevant comments and valuable words in accordance with the human judgment order (sentences-to-words) for news to facilitate veracity learning.
3. We compare state-of-the-art models with HCSN on real datasets. In addition to competitive prediction results, the HCSN also provides reasonable sentence-level and word-level explanations, as shown by a case study with three modification strategies.

We organized this paper as follows. We discuss the related fake news detection approaches in Section 2. Then in Section 3, we describe the problem statement. Section 4 details the structure of our proposed HCSN model. The evaluation settings, results, and explanation analysis are in Section 5. We conclude our work and indicate future work in Section 6.

2. Related Work

This section provides an overview of the relevant research on interpretable deep learning and fake news detection.

2.1. Fake News Detection

Automatic fake news detection is usually divided into news content-based, social context-based, and hybrid feature-based methods according to features [23]. For news content-based fake news detection methods, there are two types of news content, textual and visual. Many studies on textual content extract a large number of credibility-indicative features around language style [24], emotion [25,26], writing style [27], and semantics [28]. For example, Cui[24] proposed a novel framework to detect rumors by capturing differences in writing style, as rumors tend to prefer capitalized words that are more appropriate for nouns. In contrast, the different characteristics of fake news in visual features are extracted from images or videos [29]. Social context-based fake news detection methods include user-based and propagation-based methods. The former is modeled according to the characteristics of users who publish and retweet fake news [12,13], and the characteristics mainly include user sex, the number of fans, and the user profile. The latter performs fake news detection through features of retweets or propagation structures in social networks [30–33].In reference [32], they studied a novel method, as an example, which utilized a bidirectional graph neural network model to learn the embedding propagation structure for detecting fake news. Hybrid feature-based methods are fusion multi-models or multiple features for fake news detection [26,34,35].

2.2. The Interpretation of Deep Learning

Machine learning (ML) and artificial intelligence (AI) models have gradually risen in complexity, accuracy, and other quality indicators over the years. However, this growth has often come at the expense of the interpretability of the models' final results. Simultaneously, academics and practitioners have started to come to the realization that greater openness in artificial intelligence and deep learning engines is required if these techniques are to be used in reality [36]. In recent years, interpretable AI (IAI) and explainable AI (or XAI) models have begun to be applied in more domains [37], such as cybersecurity [38], recommender systems [39], healthcare [40], social networks [18], etc. The explanation of deep learning models generally refers to the presentation of model decision results in an understandable manner, which can help the user to understand the inner workings of complex models and the reasons why models make specific decisions. Interpretable AI is based on intrinsic interpretability, which is built by adopting self-explanatory models that incorporate interpretability directly into their structure [15]. In contrast, the explainable AI utilizes post hoc explanations that require the creation of another model to provide explanations of the existing models [41]. Recent studies on detecting fake news have focused on the identification of evidence to make the model interpretable or on the study of results using Interpretable tools. These explainable methods and interpretable methods mainly provide explanations by extracting relevant articles [15], user information [18], and retweet sequences [19].

3. Problem Statement

$S = \{s^1, s^2, \ldots, s^M\}$ is a source post, which contains M sentences, and each sentence $s^m = \{w_1^m, w_2^m, \ldots, w_p^m\}$ contains p words. In fact, there is often only one sentence in the source post. In order to unify the symbols or apply it in long-text fake news, we used multiple sentences to represent the source post. When a source post is published on a social network, some users will share their views or opinions about the source post, forming a large number of comments. $C = \{c^1, c^2, \ldots, c^N\}$ is the set of N comments related to the source post S, where each comment $c^n = \{w_1^n, w_2^n, \ldots, w_q^n\}$ contains q words. We treated fake news detection as a binary classification task, and a binary label $y \in \{0,1\}$ was used to indicate the truthfulness of each source post. In addition, according to the model, we selected certain sentences from the source post content, some comments from the user comments, and then certain words from both to interpret why it was defined as fake news.

4. The Proposed HCSN Model

In this section, we introduce the details of utilizing source posts and user reviews to detect fake news through a Hierarchical Co-attention Selection Network model (HCSN). As shown in Figure 2, the HCSN consisted of the following four components: (1) input encoder, which generated the word-level representation of the source post and comments through the RNN and self-attention mechanism (2) sentence-level co-attention, which selected informative comments through the sentence-level interaction of the source post and user responses; (3) word-level co-attention, which selected informative words or phrases through the word-level interaction of the source post and selected comments; and (4) fake news prediction, which conducted the fake news prediction by concatenating the final learned representations from the source posts and user comments.

4.1. Input Encoding

In fake news detection or text classification tasks, researchers often use self-attention mechanisms to learn word-level or sentence-level representations. Similarly, we also used the same structure to learn sentence-level representations. In particular, we first obtained the word vector $w_t \in \mathbb{R}$ for each word by the embedding matrix. The source posts and user comments on Twitter are usually short texts, so we directly adopted a bidirectional GRU [42] to learn the word sequence representation. Finally, we obtained the sentence vector through the self-attention mechanism.

For a source post consisting of p words, the forward hidden state and the backward hidden state were obtained as follows:

$$\begin{aligned} \overrightarrow{h_t^m} &= \overrightarrow{GRU}(w_t^m, \overrightarrow{h_{t-1}^m}),\ t \in \{1, \ldots, p\} \\ \overleftarrow{h_t^m} &= \overleftarrow{GRU}(w_t^m, \overleftarrow{h_{t-1}^m}),\ t \in \{p, \ldots, 1\} \end{aligned} \tag{1}$$

By connecting the forward hidden state $\overrightarrow{h_t^m}$ and the backward hidden state $\overleftarrow{h_t^m}$, we can obtain the representation of word $h_t^m = [\overrightarrow{h_t^m}, \overleftarrow{h_t^m}]$. In order to find the informative words in the sentence, the importance of each word was measured by the self-attention mechanism [43], and it obtained sentence vectors $s^m \in \mathbb{R}^{2d}$ as follows:

$$s^m = \sum_{t=1}^{p} (\alpha_t^m h_t^m) \tag{2}$$

where the importance of the tth word for the source post s^m was measured by α_t^m, and the calculation method was as follows:

$$\alpha_t^m = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{3}$$

where $Q = h_t^m \times W^Q,\ K = h_t^m \times W^K,$ and $V = h_t^m \times W^V$, W^Q, W^K, W^V are the trainable weight matrix, and d_k is the row of Q and K.

Similarly, given a comment c^n with q words, we can also obtain the representation of word $h_t^q = [\overrightarrow{h_t^q}, \overleftarrow{h_t^q}]$ and the comment vector $c^n \in \mathbb{R}^{2d}$.

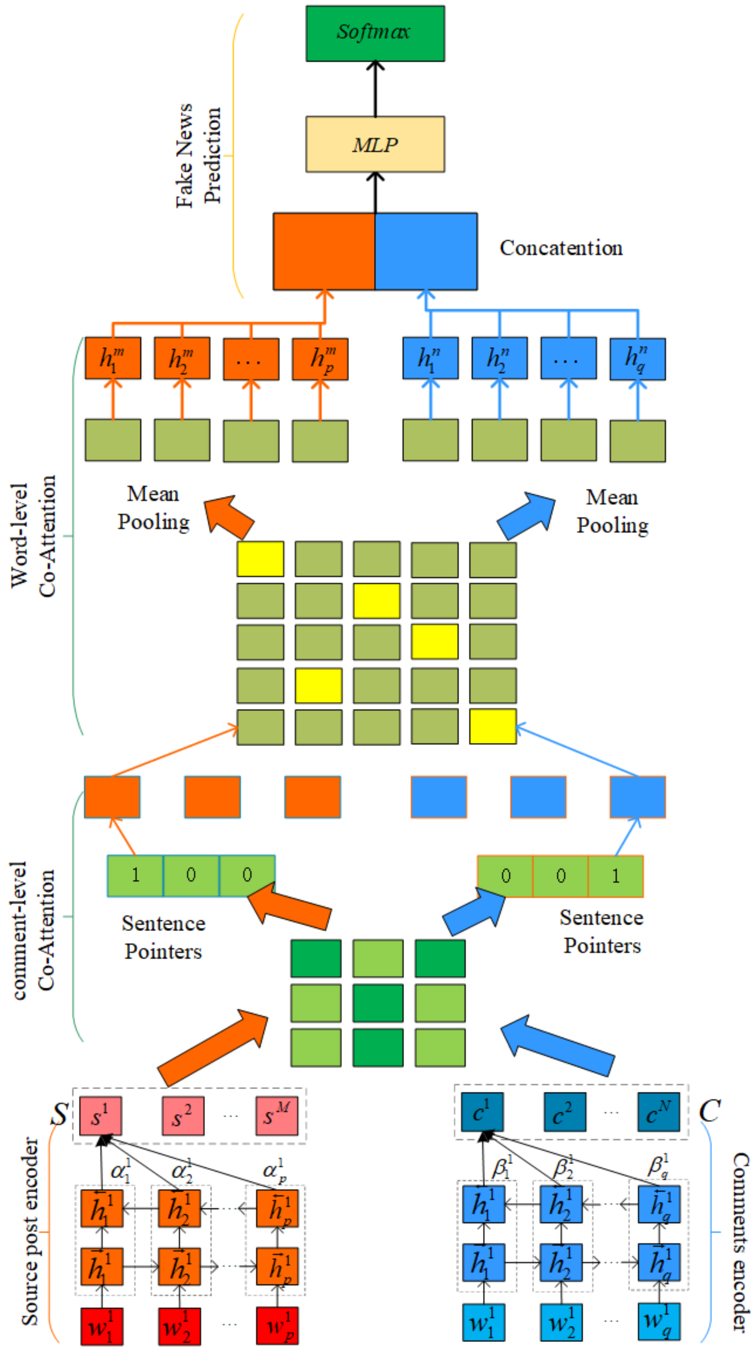


Figure 2. The model of Hierarchical Selection Networks with Co-Attention (HCSN) for interpretable fake news detection following the sequence as word–sentence–word–event.

4.2. Sentence-Level Co-Attention

Since social media platforms allow people to publish the responses to the original post, there are many comments that support or deny the source posts, which can assist the fake news detection models to confirm the authenticity of a piece of news. However, a large amount of information and noise also exist. In this section, we aim to select the most informative comment from the comments. Specifically, we utilized sentence-level co-attention to select comments by the semantic affinity of the source post and user comments. Therefore, we first needed to construct feature matrices for source posts and user comments separately. Similar to the MPCN [44] model, given the source post ($S \in \mathbb{R}^{M\times 2d}$) and corresponding user comments ($C \in \mathbb{R}^{N\times 2d}$), we can easily capture the similarity matrices $X \in \mathbb{R}^{M\times N}$ as follows:

$$X = F(S)^T QF(C) \tag{4}$$

where $Q \in \mathbb{R}^{2d\times 2d}$, and $F(\bullet)$is a feed-forward neural network function. We calculated the row and column maxima of the affinity matrix X and utilized the result to weight source posts and user comments.

In order to select the special sentences from the source posts and comments, we calculated the pointer to sentences as follows:

$$\begin{aligned} p_s &= (Gumbel(max_{col}(X))) \\ p_c &= (Gumbel(max_{row}(X))) \end{aligned} \tag{5}$$

Here, we chose to use max-pooling because it intuitively selects the most influential source post and user comments. Then, we applied the function to obtain $max_{col}(X)$ and $max_{row}(X)$. In this process, the input vector is usually transformed into a probability distribution utilizing the standard Softmax function, and the obtained co-attention vector representation is then fed to the next layer of the framework. However, we did not want to make use of these vector representations but wanted to continue to conduct further operations on the selected comments. Therefore, we used the Gumbel–Max trick to learn pointers based on the co-attentional layer, since the Gumbel–Max trick [45] transforms sampling from a multinomial distribution into a discrete optimization problem; that is, it transforms a sampling problem into an optimization problem.

Consider a k-dimensional categorical distribution with class probabilities described in terms of its unnormalized log probabilities $l_1, \ldots, l_k$:

$$p_i = \frac{exp(log(l_i))}{\sum_{j=1}^{k} exp(log(l_j))} \tag{6}$$

A one-hot sample $e = (e_1, \ldots, e_k) \in \mathbb{R}^{\urcorner}$ from the distribution can be obtained as follows:

$$e_i = \begin{cases} 1, & i = \arg\max_j (log(l_j) + g_j) \\ 0, & otherwise \end{cases} \tag{7}$$

$$\begin{aligned} g_i &= -log(-log(u_i)) \\ u_i &\sim Uniform(0,1) \end{aligned} \tag{8}$$

In this case, the arg max operation is equivalent to taking a sample that is weighted by p_i, p_k, and g_i signifies the Gumbel noise that disturbs each $log(l_i)$ term to the extent of g_i.

By respectively applying p_s, p_c, to S, C, we obtained the selected source post s^i and user comments c^j:

$$\begin{aligned} s^i &= p_s^T S \\ c^j &= p_c^T C \end{aligned} \tag{9}$$

Then, the selected user comments $\bar{C}$ and source post $\bar{S}$ were passed to the next layer, where rich word-level interactions were extracted between them.

4.3. Word-Level Co-Attention

In the previous section, we obtained multiple user comment information through sentence-level co-attention. In this process, we used pointers to obtain the most informative comments one by one. Although in the input encoding, we used the self-attention mechanism to focus on the word information to obtain the vector of the sentence, the sentence information was still redundant for predicting the veracity of the news. So, we adopted the word-level co-attention mechanism for modeling to extract more fine-grained information, which is conducive to richer interactions. According to the method of computing sentence-level co-attention, we computed the affinity matrix between the source post $\bar{S}$ and user comments $\bar{C}$ in the selected ground source, and the affinity matrix $\bar{Y}$ was calculated as follows:

$$Y = F(\bar{S})^T Q_w F(\bar{C}) \tag{10}$$

where $Q_w \in \mathbb{R}^{2d \times 2d}$, and $F(\bullet)$is the same function as in the sentence-level co-attention. Different from the sentence-level co-attention, which relied on pointers to select sentences, we computed the word-level co-attention representations utilizing an affinity matrix with mean pooling as follows:

$$\begin{aligned} s_w^i &= (H(avg_{col}(Y)))^T h_t^m \\ c_w^j &= (H(avg_{row}(Y)))^T h_t^n \end{aligned} \tag{11}$$

where $H(\bullet)$ is the standard softmax function. We used mean pooling and the softmax function to directly achieve the word-level representation obtained by the word-level co-attention. Considering the number of source posts and user comments, we concatenated the feature vectors of all sentences in the source posts, as $\overline{s_w} = [s_w^1,\ s_w^2\ \ldots,\ s_w^p]$, and the feature vectors of all comments in user comments were $\overline{c_w} = [s_w^1,\ s_w^2,\ \ldots,\ s_w^q]$. At the prediction layer, we directly performed classification.

4.4. Prediction Layer

In this layer, the feature vectors of the source posts and the feature vectors in the comments were concatenated and fed into a multilayer perception (MLP) and a Softmax layer for the final prediction of news veracity $\hat{y} = [\hat{y_0}, \hat{y_1}]$, where $\hat{y_0}$ and $\hat{y_0}$ are the label prediction probabilities of 0 and 1, respectively.

$$\hat{y} = softmax(MLP([\overline{s_w}; \overline{c_w}]) \tag{12}$$

We adopted the loss function to minimize the cross entropy value.

$$\mathcal{L}(\theta) = -ylog(\hat{y}_1) - (1-y)log(1-\hat{y}_0) \tag{13}$$

where θ denotes all trainable parameters. In the training process, the Adam optimizer was utilized to learn θ, because the Adam optimizer is very suitable for large-scale data and parameter scenarios and is widely used in neural network training.

5. Experiments

To demonstrate the detection performance and interpretability of the proposed model, in this section, we discuss the design of different experiments to validate and answer the following research questions:

Q1: In terms of fake news detection performance, does our HCSN model outperform state-of-the-art methods?
Q2: What is the performance of the HCSN without the different components?
Q3: Is our model capable of providing a compelling explanation?

5.1. Datasets

We utilized two well-known datasets twitter15 and twitter16 established by [46]. They both contain source posts, user comments, and user information. We utilized source posts and user comments as input and only chose "true" and "false" labels as the groundtruth. The detailed statistics of the twitter15 and twitter16 datasets are shown in Table 1.

Table 1. Statistics of the the twitter15 and twitter16 datasets.

	Twitter15	**Twitter16**
Source tweets	742	412
True tweets	372	205
False tweets	370	207
Comments	9659	4122
Avg. words per source	14	13
Avg. comment per source	13	10

5.2. Compared Methods

We compare our HCSN with the representative state-of-the-art fake news detection methods, as listed below.

- **RNN** [47]: an RNN-based method that models social context information as a variable-length time series for learning continuous representation of microblog events. We utilized a variant of RNN bidirectional GRU.
- **text-CNN** [48]: a fake news detection model based on convolutional neural networks, which utilizes multiple convolutional filters to capture textual features of different granularities.
- **HAN** [21]: a fake news detection model for learning source post representations, based on a hierarchical attention neural network, which utilizes word-level attention and sentence-level attention to learn source post representations.
- **HPA-BLSTM** [20]: a hierarchical attention neural network-based fake news detection model that learns source post representation from word-level, post-level, and sub-event level.
- **dEFEND** [15]: a model that utilizes the co-attention mechanism to learn the correlation representation between source posts and user comments for fake news detection. According to the weight of the co-attention mechanism, the content of source posts and user comments are obtained as the interpretation of the fake news detection results.
- **GCAN** [18]: a fake news detection model based on graph-aware co-attention network, which learns the relationship between source posts, retweets, and user information-related representations through dual co-attention, and connects the features of source posts for fake news detection..
- **PLAN** [19]: an interpretable rumor detection model focusing on user interaction, taking rumors and retweeted comments as the input of Transformer, and using the positional embedding instead of delay time embedding for rumor detection, providing explanations in posts and tags through Attention.
- **Dual emotion** [26]: a fake news detection model based on dual emotional features, which obtains the emotional representation of the source post, the emotional representation of the user comments, and the emotional gap as emotional features through the emotion dictionary, and connects the semantic features of the source post for fake news detection.

5.3. Experimental Results

To answer Q1, we compared our model with the state-of-the-art model. The evaluation metrics included accuracy, precision, recall, and F1. We randomly chose 60% data for training, 20% for validation, and 20% for testing. The experiment was performed five times, and the average was taken. We ran the source code of all the compared methods, except for GCAN, whose result was cited from the original paper. The experimental results are shown in Tables 2 and 3.

Table 2. Performance comparison between the state-of-the-art models and our model on twitter15.

	Acc	**Pre**	**Rec**	**F1**
RNN	0.720	0.716	0.715	0.713
Text-CNN	0.756	0.732	0.731	0.730
HAN	0.811	0.814	0.813	0.811
HPA-BLSTM	0.842	0.853	0.845	0.844
dEFEND	0.845	0.845	0.846	0.845
GCAN	0.876	0.825	0.829	0.825
PLAN	0.845	0.869	0.863	0.845
Dual Emotion	0.851	0.851	0.851	0.851
Ours	**0.912**	**0.920**	**0.910**	**0.911**

Table 3. Performance comparison between the state-of-the-art models and our model on twitter16.

	Acc	**Pre**	**Rec**	**F1**
RNN	0.653	0.652	0.653	0.653
Text-CNN	0.674	0.672	0.673	0.677
HAN	0.723	0.712	0.712	0.716
HPA-BLSTM	0.817	0.820	0.815	0.816
dEFEND	0.743	0.756	0.774	0.741
GCAN	**0.908**	0.763	0.759	0.759
PLAN	0.824	0.819	0.829	0.827
Dual Emotion	0.812	0.821	0.817	0.812
Ours	0.897	**0.899**	**0.900**	**0.897**

It is clear from the results that our model was significantly competitive on both datasets. On twitter15 and twitter16, it compared favorably with three interpretable fake news detection models. Compared to dEFEND, our model achieved a 6.4% and 15% improvement in f1 and a 6.7% and 15% improvement in accuracy. Compared with the dual co-attention mechanism GCAN, our model achieved a 7% and 4% improvement in f1 and 3.6% and −1.1% improvement in accuracy, respectively. Compared to PLAN, our model achieved a 6.6% and 4% improvement in f1 and a 7.3% and 7.0% in accuracy, respectively. Compared with dual emotion, the state-of-the-art fake news detection model, we achieved a 6.0% and 8.5% improvement in f1 and a 6.1% and 8.5% improvement in the accuracy, respectively. Furthermore, our method (including source posts and user comments) outperformed models that rely only on source posts or user comments, such as HPA-BLSTM.

When comparing fake news detection models, which only utilize source posts, the HAN model obviously outperformed RNN and CNN, indicating that the hierarchical attention mechanism obtained semantic features well. The models based on source post and other information performed better, especially PLAN and GCAN; the former utilizes delay-time embeddings instead of positional embeddings, and the latter utilizes two co-attention mechanisms to fully capture source posts and forward users spread structure. The dEFEND utilizes co-attention to capture the relevance of source posts and user comments, but it is more suitable for long articles when targeting source posts. Dual emotion extracts multiple emotional features, which are used as supplementary features, but their relevance is not considered.

5.4. Ablation Analysis

We studied the contribution of each component of the entire model to answer Q2. We experimented with several models that removed different components, and the results are shown in Figure 3. We removed the sentence encoding structure and the self-attention mechanism as "-se" and "-t". Removing sentence-level co-attention and word-level co-attention was denoted by "-s" and "-w", respectively. Finally, the entailment model with all components (sentence encoding, self-attention mechanism, sentence-level co-attention, and word-level co-attention) is shown.

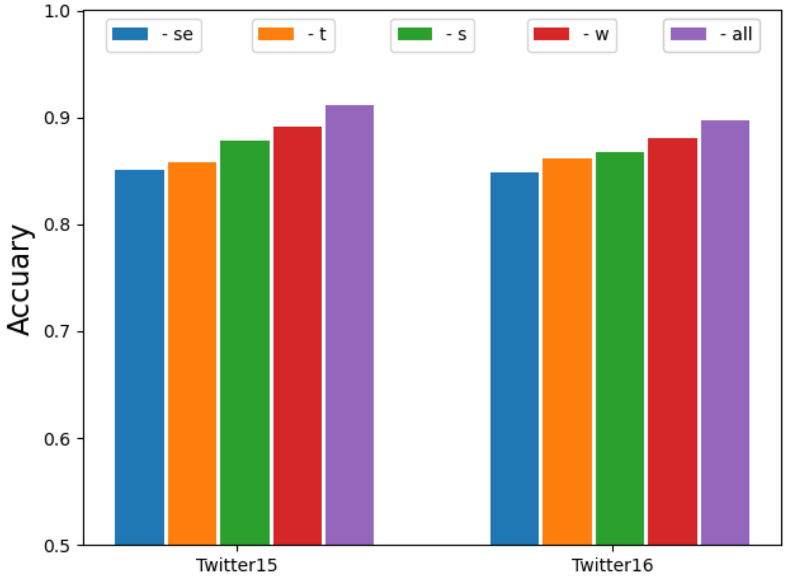


Figure 3. The results of ablation analysis.

We can clearly see that every part of the model is important. When we removed a component of the model, the performance of the model dropped, which shows that the components in our model are all critical. In addition, the performance of the model without the sentence encoding dropped significantly, indicating that the sentence-encoding effect was more obvious at the beginning. We also found that the effect of the word-level co-attention at the end was relatively small. However, in terms of sentence-level co-attention, the performance of the model without it dropped obviously, which showed that the selection reduced redundancy and noise during the learning process.

5.5. Interpretability Case Study

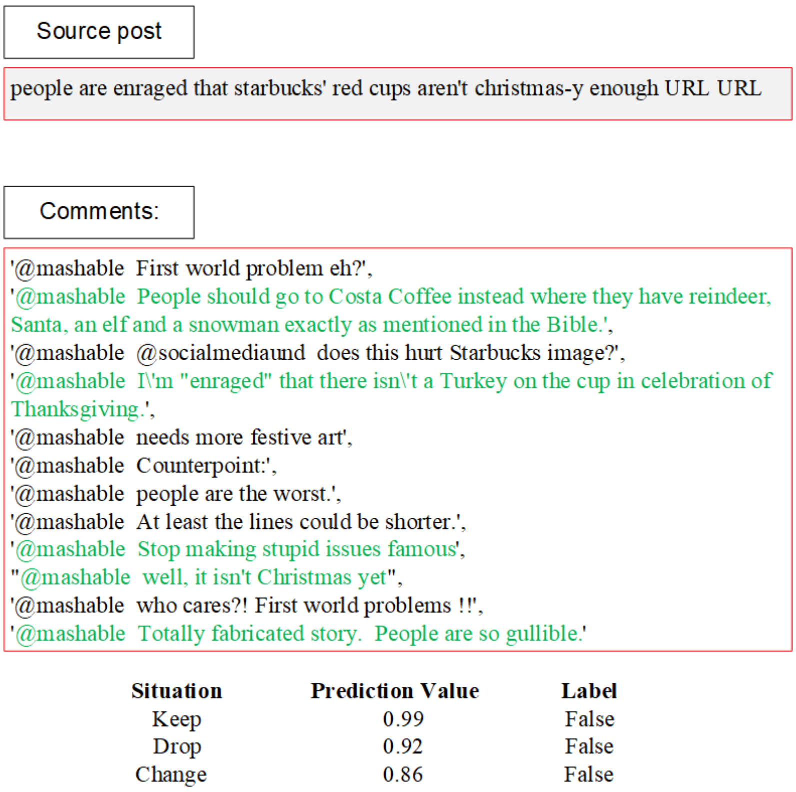
In this subsection, to answer Q3, we show a case study to illustrate the interpretability performance of the HCSN framework through the sentence-level co-attention and word-level co-attention, respectively (Figures 4 and 5). The prediction value was larger than 0.5 representing that the news was false; otherwise, it was true.

Interpretability case study on sentence-level co-attention.

Utilizing the one-hot vector in sentence-level co-attention, user comments related to the source post were selected as the explanation. To show the interpretability of the selected comments, we adopted two types of sentence-level modification techniques to compare the prediction performance of the HCSN under different situations:

- **-Keep**: keep the selected comments.
- **-Drop**: delete the selected comments.
- **-Change**: replace the selected comments with comments in different data randomly.

We selected a fake post and related comments in the test data on twitter15 to verify its explainability, in which the source post had only one sentence (“People are enraged that Starbucks’s red cups aren’t christmasy enough URL URL”), and there were a total of 12 comments. We utilized the one-hot vector produced by the sentence-level co-attention to locate the appropriate comment, as seen in Figure 4. Among the selected comments, we found that each one was extremely closely related to the source post. Then, we showed the prediction results of the model under different situations, and the prediction values changed when we dropped or changed selected comments. In particular, if we replace selected sentences using random comments from another post, the prediction results value dropped. This case study shows that selected comments played a significant role in the prediction process, and they could be regarded as the sentence-level interpretation.



Source post

people are enraged that starbucks' red cups aren't christmas-y enough URL URL

Comments:

'@mashable First world problem eh?',
'@mashable People should go to Costa Coffee instead where they have reindeer, Santa, an elf and a snowman exactly as mentioned in the Bible.',
'@mashable @socialmediaund does this hurt Starbucks image?',
'@mashable I\'m "enraged" that there isn\'t a Turkey on the cup in celebration of Thanksgiving.',
'@mashable needs more festive art',
'@mashable Counterpoint:',
'@mashable people are the worst.',
'@mashable At least the lines could be shorter.',
'@mashable Stop making stupid issues famous',
"@mashable well, it isn't Christmas yet",
'@mashable who cares?! First world problems !!',
'@mashable Totally fabricated story. People are so gullible.'

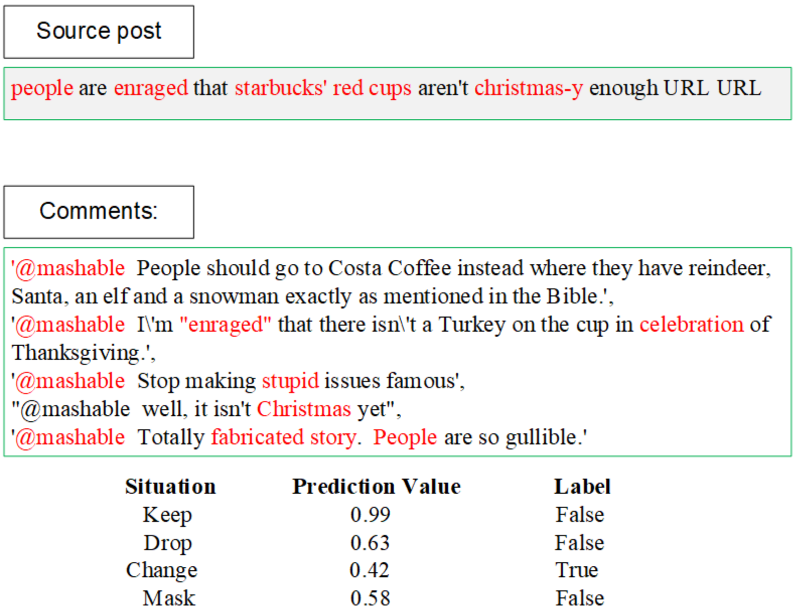
Situation	Prediction Value	Label
Keep	0.99	False
Drop	0.92	False
Change	0.86	False

Figure 4. The case study of interpretability on sentence-level co-attention. The green sentences are selected comments, and it also shows the value of prediction under different situations.

Interpretability case study on word-level co-attention. By examining the attention weight associated with word-level co-attention, it was possible to determine the predictive power of informative words in detecting fake news. To test the interpretability of the highlighted words, we adopted three types of word-level modification techniques on news posts and comments to compare the performance of HCSN under different situations:

- **-Keep**: keep the top 10 attention weighted words in comments, posts, and all.
- **-Drop**: delete the top 10 attention weighted words in comments, posts, and all.
- **-Change**: randomly replace the top 10 attention weighted words in comments, posts, and all with words in different data.
- **-Mask**: replace the top 10 attention weighted words in comments, posts, and all with a special token [MASK].

The word-level co-attention of our model can further explain the words the model cares about. After obtaining the comments pointed to by the one-hot vector, we continued to utilize the word-level co-attention to obtain the relevant words in each comment. As shown in Figure 5, among the five comments pointed to by the one-hot vector, the model highlighted some words in the word-level co-attention layer. We compared the prediction value of true or false and found that when we dropped or masked these tokens, the prediction value dropped dramatically. Even when we replaced these tokens with other tokens randomly, the prediction result changed. This case study shows the importance of these tokens, which are the reasons the model gives us such decisions and confirms the word-level interpretability.



Source post

people are enraged that starbucks' red cups aren't christmas-y enough URL URL

Comments:

'@mashable People should go to Costa Coffee instead where they have reindeer, Santa, an elf and a snowman exactly as mentioned in the Bible.',
'@mashable I\'m "enraged" that there isn\'t a Turkey on the cup in celebration of Thanksgiving.',
'@mashable Stop making stupid issues famous',
"@mashable well, it isn't Christmas yet",
'@mashable Totally fabricated story. People are so gullible.'

Situation	Prediction Value	Label
Keep	0.99	False
Drop	0.63	False
Change	0.42	True
Mask	0.58	False

Figure 5. The case study of interpretability on word-level co-attention. The red words are the top 10 attention words and it also shows the value of prediction under different situations.

6. Conclusion and Future Work

In recent years, interpretable fake news detection has received increasing attention. However, few researchers directly filter out interpretable information and usually only focus on a certain part. We solved the problem of interpretable fake news detection by filtering user comments and the words in them. The purposes were to (1) significantly improve the detection performance; and to (2) screen interpretable news sentences and user comments and relocate words to explain why they were deemed false. We proposed a Hierarchical Co-attention Selection Network for fake news detection and explanatory sentence/comment discovery, and the experiments demonstrated the model has satisfying detection performance and reasonable explainability. In addition, we believe that our model can also be used for other explainability classification tasks on social media, such as position detection, hate detection, and malicious comment prediction. During the experiment, we found that comments that often contain emotional words were selected, so for future work, we will conduct tnterpretable fake news detection from the perspective of emotion features, especially using the relationship between the user comment emotion and the source posts to further improve fake news detection performance and explainability.

Author Contributions: Conceptualization, X.G. and S.H.; methodology, X.G., S.H. and M.Z.; software, X.G., Y.L. and S.H.; validation, X.G., S.H. and B.W.; resources, X.G. and S.H.; data curation, X.G., S.H. and Y.L.; writing—original draft preparation, X.G. and S.H.; writing—review and editing, X.G., S.H., M.Z., Y.L. and B.W.; visualization, X.G., S.H. and B.W.; supervision, M.Z.; project administration, M.Z.; funding acquisition, M.Z. and B.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Social Science Foundation of China (No. 20BXW101, No. 18XWW015).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All relevant datasets are publicly available on the web, and you can also obtain them from our public GitHub repository; our model is available on our public GitHub repository: https://github.com/wj-gxy/HCSN.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Burkhardt, J.M. *Combating Fake News in the Digital Age*; American Library Association: Chicago, IL, USA, 2017; Volume 53.
2. Fisher, M.; Cox, J.W.; Hermann, P. Pizzagate: From rumor, to hashtag, to gunfire in DC. *Wash. Post* **2016**, *6*, 8410–8415.
3. Rashkin, H.; Choi, E.; Jang, J.Y.; Volkova, S.; Choi, Y. Truth of varying shades: Analyzing language in fake news and political fact-checking. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Copenhagen, Denmark, 7–11 September 2017; pp. 2931–2937.
4. ElBoghdady, D. Market quavers after fake AP tweet says Obama was hurt in White House explosions. *The Washington Post*, 23 April 2013.
5. Kshetri, N.; Voas, J. The economics of "fake news". *IT Prof.* **2017**, *19*, 8–12. [CrossRef]
6. Naeem, S.B.; Bhatti, R. The COVID-19 'infodemic': A new front for information professionals. *Health Inf. Libr. J.* **2020**, *37*, 233–239. [CrossRef] [PubMed]
7. Wani, A.; Joshi, I.; Khandve, S.; Wagh, V.; Joshi, R. Evaluating deep learning approaches for covid19 fake news detection. In Proceedings of the International Workshop on Combating Online Hostile Posts in Regional Languages during Emergency Situation, Virtual, 8 February 2021; pp. 153–163.
8. Yang, W.; Wang, S.; Peng, Z.; Shi, C.; Ma, X.; Yang, D. Know it to Defeat it: Exploring Health Rumor Characteristics and Debunking Efforts on Chinese Social Media during COVID-19 Crisis. *arXiv* **2021**, arXiv:2109.12372.
9. Yang, F.; Liu, Y.; Yu, X.; Yang, M. Automatic detection of rumor on Sina Weibo. In Proceedings of the ACM SIGKDD Workshop on Mining Data Semantics, Beijing, China, 12–16 August 2012.
10. Rubin, V.L.; Conroy, N.J.; Chen, Y. Towards news verification: Deception detection methods for news discourse. In Proceedings of the Hawaii International Conference on System Sciences, Kauai, HI, USA, 5–8 January 2015; pp. 5–8.
11. Ma, J.; Gao, W.; Wei, Z.; Lu, Y.; Wong, K.F. Detect rumors using time series of social context information on microblogging websites. In Proceedings of the 24th ACM International on Conference on Information and Knowledge Management, Melbourne, Australia, 19–23 October 2015; pp. 1751–1754.
12. Castillo, C.; Mendoza, M.; Poblete, B. Information credibility on Twitter. In Proceedings of the 20th International Conference on World Wide Web, WWW 2011, Hyderabad, India, 28 March–1 April 2011.
13. Shu, K.; Zhou, X.; Wang, S.; Zafarani, R.; Liu, H. The Role of User Profile for Fake News Detection. In Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), Vancouver, BC, Canada, 27–30 August 2019.
14. Ruchansky, N.; Seo, S.; Liu, Y. CSI: A Hybrid Deep Model for Fake News Detection. In Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, Singapore, 6–10 November 2017.
15. Shu, K.; Cui, L.; Wang, S.; Lee, D.; Liu, H. DEFEND: Explainable Fake News Detection. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD'19), Anchorage, AK, USA, 4–8 August 2019; Association for Computing Machinery: New York, NY, USA, 2019; pp. 395–405. [CrossRef]
16. Susaiyah, A.; Härmä, A.; Reiter, E.; Petković, M. Neural scoring of logical inferences from data using feedback. *Int. J. Interact. Multimed. Artif. Intell.* **2021**, *6*, 90–100. [CrossRef]
17. Liu, Y.; Wu, Y.F.B. Early detection of fake news on social media through propagation path classification with recurrent and convolutional networks. In Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018.
18. Lu, Y.J.; Li, C.T. GCAN: Graph-aware Co-Attention Networks for Explainable Fake News Detection on Social Media. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 5–10 July 2020; pp. 505–514. [CrossRef]
19. Khoo, L.M.S.; Chieu, H.L.; Qian, Z.; Jiang, J. Interpretable rumor detection in microblogs by attending to user interactions. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 8783–8790.
20. Guo, H.; Cao, J.; Zhang, Y.; Guo, J.; Li, J. Rumor Detection with Hierarchical Social Attention Network. In Proceedings of the 27th ACM International Conference on Information and Knowledge Management, Turin, Italy, 22–26 October 2018.
21. Yang, Z.; Yang, D.; Dyer, C.; He, X.; Smola, A.; Hovy, E. Hierarchical Attention Networks for Document Classification. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego, CA, USA, 12–17 June 2016; pp. 1480–1489. [CrossRef]
22. Vo, N.; Lee, K. Hierarchical Multi-head Attentive Network for Evidence-aware Fake News Detection. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, Online, 19–23 April 2021; pp. 965–975.
23. Shu, K.; Sliva, A.; Wang, S.; Tang, J.; Liu, H. Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explor. Newsl.* **2017**, *19*, 22–36. [CrossRef]
24. Horne, B.; Adali, S. This just in: Fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news. In Proceedings of the International AAAI Conference on Web and Social Media, Montreal, QC, Canada, 15–18 May 2017; Volume 11, pp. 759–766.
25. Cui, L.; Wang, S.; Lee, D. Same: Sentiment-aware multi-modal embedding for detecting fake news. In Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, Vancouver, BC, Canada, 27–30 August 2019; pp. 41–48.

26. Zhang, X.; Cao, J.; Li, X.; Sheng, Q.; Zhong, L.; Shu, K. Mining dual emotion for fake news detection. In Proceedings of the Web Conference 2021, Ljubljana, Slovenia, 19–23 April 2021; pp. 3465–3476.
27. Potthast, M.; Kiesel, J.; Reinartz, K.; Bevendorff, J.; Stein, B. A stylometric inquiry into hyperpartisan and fake news. *arXiv* **2017**, arXiv:1702.05638.
28. Braşoveanu, A.M.; Andonie, R. Semantic fake news detection: A machine learning perspective. In Proceedings of the International Work-Conference on Artificial Neural Networks, Gran Canaria, Spain, 12–14 June 2019; pp. 656–667.
29. Cao, J.; Qi, P.; Sheng, Q.; Yang, T.; Guo, J.; Li, J. Exploring the role of visual content in fake news detection. In *Disinformation, Misinformation, and Fake News in Social Media*; Springer: Cham, Switzerland, 2020; pp. 141–161.
30. Wu, L.; Liu, H. Tracing fake-news footprints: Characterizing social media messages by how they propagate. In Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining, Los Angeles, CA, USA, 5–9 February 2018; pp. 637–645.
31. Shu, K.; Wang, S.; Liu, H. Beyond news contents: The role of social context for fake news detection. In Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining, Melbourne, VIC, Australia, 11–15 February 2019; pp. 312–320.
32. Bian, T.; Xiao, X.; Xu, T.; Zhao, P.; Huang, W.; Rong, Y.; Huang, J. Rumor detection on social media with bi-directional graph convolutional networks. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 549–556.
33. Xu, W.; Wu, J.; Liu, Q.; Wu, S.; Wang, L. Evidence-aware Fake News Detection with Graph Neural Networks. In Proceedings of the ACM Web Conference 2022, Virtual, 25–29 April 2022; pp. 2501–2510.
34. Wu, Y.; Zhan, P.; Zhang, Y.; Wang, L.; Xu, Z. Multimodal Fusion with Co-Attention Networks for Fake News Detection. In Proceedings of the Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, Online, 1–6 August 2021; pp. 2560–2569.
35. Chen, M.; Wang, N.; Subbalakshmi, K. Explainable Rumor Detection using Inter and Intra-feature Attention Networks. *arXiv* **2020**, arXiv:2007.11057.
36. Bai, X.; Wang, X.; Liu, X.; Liu, Q.; Song, J.; Sebe, N.; Kim, B. Explainable deep learning for efficient and robust pattern recognition: A survey of recent developments. *Pattern Recognit.* **2021**, *120*, 108102. [CrossRef]
37. Du, M.; Liu, N.; Hu, X. Techniques for interpretable machine learning. *Commun. ACM* **2019**, *63*, 68–77. [CrossRef]
38. Xu, X.; Zheng, Q.; Yan, Z.; Fan, M.; Jia, A.; Liu, T. Interpretation-enabled Software Reuse Detection Based on a Multi-Level Birthmark Model. In Proceedings of the 2021 IEEE/ACM 43rd International Conference on Software Engineering (ICSE), Madrid, Spain, 22–30 May 2021; pp. 873–884.
39. Ye, L.; Yang, Y.; Zeng, J.X. An interpretable mechanism for personalized recommendation based on cross feature. *J. Intell. Fuzzy Syst.* **2021**, *40*, 9787–9798. [CrossRef]
40. Pintelas, E.; Liaskos, M.; Livieris, I.E.; Kotsiantis, S.; Pintelas, P. A novel explainable image classification framework: Case study on skin cancer and plant disease prediction. *Neural Comput. Appl.* **2021**, *33*, 15171–15189. [CrossRef]
41. Wu, L.; Rao, Y.; Zhao, Y.; Liang, H.; Nazir, A. DTCA: Decision tree-based co-attention networks for explainable claim verification. *arXiv* **2020**, arXiv:2004.13455.
42. Bahdanau, D.; Cho, K.; Bengio, Y. Neural machine translation by jointly learning to align and translate. *arXiv* **2014**, arXiv:1409.0473.
43. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 1–11.
44. Tay, Y.; Luu, A.T.; Hui, S.C. Multi-Pointer Co-Attention Networks for Recommendation. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, London, UK, 19–23 August 2018.
45. Jang, E.; Gu, S.; Poole, B. Categorical reparameterization with gumbel-softmax. *arXiv* **2016**, arXiv:1611.01144.
46. Jing, M.; Wei, G.; Wong, K.F. Detect Rumors in Microblog Posts Using Propagation Structure via Kernel Learning. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL 2017), Vancouver, BC, Canada, 30 July–4 August 2017.
47. Ma, J.; Gao, W.; Mitra, P.; Kwon, S.; Cha, M. Detecting rumors from microblogs with recurrent neural networks. In Proceedings of the International Joint Conference on Artificial Intelligence, New York, NY, USA, 9–15 July 2016.
48. Kim, Y. Convolutional Neural Networks for Sentence Classification. *arXiv* **2014**, arXiv:1408.5882.