electronics

MDPI

Article

The Graph Reasoning Approach Based on the Dynamic Knowledge Auxiliary for Complex Fact Verification

Yongyue Wang [1], Chunhe Xia [1,2], Chengxiang Si [3], Chongyu Zhang [4] and Tianbo Wang [5,*]

[1] Beijing Key Laboratory of Network Technology, School of Computer Science and Engineering, Beihang University, Beijing 100191, China; buaawyy@buaa.edu.cn (Y.W.); xch@buaa.edu.cn (C.X.)
[2] Guangxi Key Lab of Multi-source Information Mining and Security, Guangxi Normal University, Guilin 541004, China
[3] National Computer Network Emergency Response Technical Team/Coordination Center of China, Beijing 100029, China; sichengxiang@cert.org.cn
[4] Talent Introduction Department, JD Group, Beijing 100176, China; zhangchongyu@jd.com
[5] School of Cyber Science and Technology, Beihang University, Beijing 100191, China
* Correspondence: wangtb@buaa.edu.cn

Received: 3 August 2020; Accepted: 7 September 2020; Published: 9 September 2020

Abstract: Complex fact verification (FV) requires fusing scattered sequences and performing multi-hop reasoning over these composed sequences. Recently, by employing some FV models, knowledge is obtained from context to support the reasoning process based on pretrained models (e.g., BERT, XLNET), and this model outperforms previous out-of-the-art FV models. In practice, however, the limited training data cannot provide enough background knowledge for FV tasks. Once the background knowledge changed, the pretrained models' parameters cannot be updated. Additionally, noise against common sense cannot be accurately filtered out due to the lack of necessary knowledge, which may have a negative impact on the reasoning progress. Furthermore, existing models often wrongly label the given claims as 'not enough information' due to the lack of necessary conceptual relationship between pieces of evidence. In the present study, a Dynamic Knowledge Auxiliary Graph Reasoning (DKAR) approach is proposed for incorporating external background knowledge in the current FV model, which explicitly identifies and fills the knowledge gaps between provided sources and the given claims, to enhance the reasoning ability of graph neural networks. Experiments show that DKAR put forward in this study can be combined with specific and discriminative knowledge to guide the FV system to successfully overcome the knowledge-gap challenges and achieve improvement in FV tasks. Furthermore, DKAR is adopted to complete the FV task on the Fake NewsNet dataset, showing outstanding advantages in a small sample and heterogeneous web text source.

Keywords: fact verification; knowledge enhanced; pre-training model; graph neural network; cognitive reasoning

1. Introduction

Fact verification (FV) often requires retrieving a significant number of scattered evidential sequences (documents, paragraphs, or sentences), reasoning over the fused multiple sequences and finally labelling the given claim with 'supported', 'refused', or 'not enough information'. Although the claims do not need to have a specific form, the entity of the claims must be related to textual resources. Since if one given claim is completely unrelated to the given textual context, even if the claim is labeled as "not enough information", the FV process is meaningless. However, the claims can be arbitrarily

complex and allowed for a variety of expressions for the entity (e.g., mutated in various ways or even meaning-altered) and composition of evidence relevant to one claim can be from multiple sentences. The complex FV tasks often require a FV system to have a deeper understanding of the relationship between the given claim and evidence from multiple dimensions (e.g., semantic features, language knowledge, common sense knowledge, world or relevant domain knowledge), which not only needs to use deep learning methods to learn the connections between semantic units, but also needs the support of complex knowledge to understand the illocutionary meaning. Current FV researches [1–5] driven by data focus on simple checking tasks at a semantic level, which only use deep learning methods to construct a unified semantic space to learn the literal meaning. The most common error is caused by failing to match the semantic meaning between phrases that describe the same event. As shown in Figure 1, almost all approaches [1–5] based on pretrained language models (e.g., BERT, XLNet) fail to realize that 'novel' belongs to 'book', and 'Mark Heprin' is one of 'American journalists'. In this case, these advanced FV systems are most likely to label the claim as 'not enough information'. However, for readers, there is no difficulty in verifying this claim on the premise that readers are allowed to refer to background knowledge.

claim	*Winter's Tale is a book, written by an American journalist.*
evidence	*Winter 's Tale is a 1983 novel by Mark Helprin.*

Figure 1. Error case of advanced fact verification (FV) approaches only based on pre-training language models.

In this paper, the human understanding process is mimicked, so that FV systems are able to recognize the sequences with the same semantics but different expressions. When looking for the relationship between the claim and evidence, readers not only need to understand the semantic knowledge, but also make full use of external relevant world knowledge to fill the gap between semantic knowledge. In this way, readers can better understand the concepts expressed in the lexical aspect and the relations between these concepts during the reading process.

Recently, by implicitly adjusting the parameters of FV systems, the pre-trained language model has enhanced the ability to a great extent to represent context information. Thus, FV systems can effectively employ the background knowledge and semantic distribution features of corpora to fulfill FV tasks better. Most language understanding datasets such as FEVER [6], HotPotQA [7], MultiRC [8], and WikiHop [9] require finding relevant knowledge and reasoning over scattered sentences. However, for some simple verification tasks (for example, the claims are non-controversially historical events or the tasks themselves do not require a deep understanding), they can fortunately achieve good performance under partial knowledge. In practice, however, for complex tasks, due to the lack of background knowledge (or knowledge gaps), only extracting limited information from provided corpus cannot satisfy the information needs of reasoning, as shown in Figure 2. Although seemingly provided with all useful and reliable evidence, many claims are wrongly labelled as 'not enough information' because they lack relevant knowledge. This phenomenon also refers to one of important gaps between human race and data-driven FV systems. We know that the given claims are generated relying heavily on datasets. Human beings may use their own prior knowledge unintentionally when annotating claims manually. Even if many datasets do not provide enough background information, a human will automatically fill the knowledge gaps and label the claims correctly and easily. However, for data-driven FV systems, when there is a lack of background information in datasets, they will label all these kind of claims as "not enough information", whether or not there is a real lack of relevant information in datasets. Recently, nearly all the research [1–5] concerning fact verification assume that the used datasets provide FV systems with all knowledge that is necessary to finish tasks. In practice, however, numerous researchers often only have access to partial knowledge when dealing with complex FV tasks requiring multi-hop reasoning and they have clearly demonstrated this limitation in their

papers. Especially when background knowledge plays a decisive role in FV, the results are even worse. In addition, as the times change, some world knowledge may be continuously updated (e.g., adding new knowledge or correcting wrong knowledge). The implicitly encoded pretraining models cannot learn this important information. Besides, some noise against common sense in evidential sequences may have a negative effect on the reasoning process if FV systems lack enough background knowledge to filter them. Therefore, this paper intends to introduce a FV model that can explicitly identify and fill the knowledge gaps between provided sources and the given claims. Inspired by human cognition, for the knowledge auxiliary module of the FV model mentioned in this paper, its tuition idea is that firstly, the knowledge gaps are discovered, and then these gaps are filled under the guidance of external knowledge. Fundamentally, the OpenBookQA [10] is used to train our model's ability to find and fill knowledge gaps, since it is the only corpus currently accessible that provides context with annotated knowledge gaps.

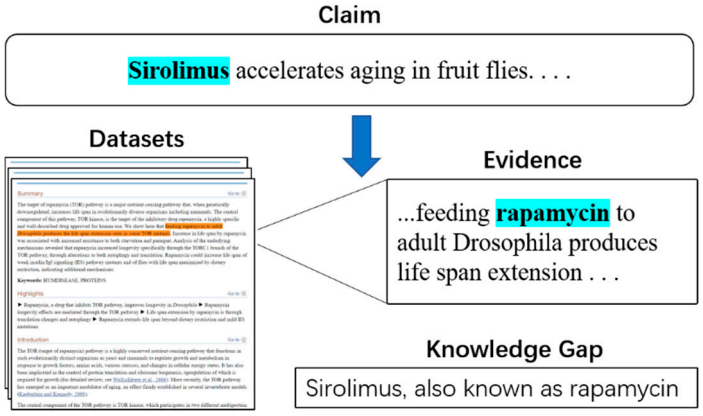

Figure 2. Knowledge gap hinders the reasoning process of FV systems.

Recently, although there is some work to improve the accuracy of open-domain Question Answering (QA) with external knowledge, we are the first to use external knowledge to strengthen the reasoning ability of FV systems. In some models, sentences are directly used as external knowledge, such as in the work of [11–13]. Different from these approaches, the model in this study tends to purposefully find and fill the knowledge gap to assist reasoning based on [14]. Other models, such as [15–18], embed syntactic or semantic knowledge into given context to enrich embeddings. However, this kind of simple syntactic and semantic knowledge supplement may be extremely helpful for disambiguation, but it cannot provide the effective knowledge auxiliary for complex and logical FV tasks. The task of neural explanation retrieval such as in [19,20] is similar to that of semantic knowledge retrieval. Another research idea is based on the use of the structural knowledge base [15,16,21,22], such as Freebase [23], to enhance the understanding of context. Even though these methods finish the task of filling the knowledge gap by using semantic parse [24] and relation lookup [25], they may not find relevant information due to the limitation of the knowledge base. According to [14], the two-step mechanism is introduced to point out and fill the knowledge gap as part of a multi-hop FV model.

To be specific, our FV model operates according to the following steps:

- According to the given claim, the retrieval module will retrieve documents and sentences relative to the claim.
- The auxiliary knowledge module predicts a key span in the retrieved evidence.
- Retrieve knowledge related to the claim and evidence from the external knowledge resources, such as the ConceptNet [26] and large-scale text corpora [27]. For our paper, we adopt four external knowledge sources: ConceptNet [26], WordNet subset (used in [10]), OMCS (Open Mind Common Sense subset), and ARC (AI2 Reasoning Challenge dataset) [27], where both structured and unstructured knowledge resources are included.

- Based on the above steps, predict knowledge gaps and fill the gap with external knowledge.
- Construct the collaborative graph with external knowledge and context information.
- Reason on the collaborative graph and label the given claim as "supported", "refused", or "not enough information".

Experiments demonstrate that DKAR outperforms the previous state-of-the-art FV approaches on FEVER dev sets, and it can also effectively solve the label errors caused by the knowledge gap. Additionally, our method shows outstanding advantages in a small sample and heterogeneous web text sources when checking fake news.

In general, the contribution of our work is as follows:

- Based on [4,14,28], the knowledge gaps for FV under partial knowledge are analyzed, and a new collaborative graph for FV is also proposed to reason over the information with knowledge gaps.
- This paper is the first attempt to introduce a joint knowledge-driven and data-driven mechanism into fact verification, and verify the effectiveness of the approach in a small sample and heterogeneous web text source, which will provide an important reference for further research.

2. Related Work

2.1. Pre-Training Language Processing and Background Knowledge for FV

With the development of deep learning architectures and the large amounts of unlabeled data, especially the emergence of pretrained models in recent years, the ability of machines to understand natural language has achieved significant progress [29–32]. Among the current pre-training mechanisms, BERT [33], which employs the Transformer [34] as the encoder and a bidirectional language model to capture complex linguistic phenomena, is undoubtedly the most advanced mechanism, thus having solved a series of challenging NLU problems (mainly in the field of QA) and being significantly better than other language processing models [35,36]. However, as discussed in the introduction, fact verification or fake news checking requires not only understanding natural language at a semantic level, but also integrating existing background knowledge for supporting complex reasoning process [15,37–39]. Therefore, it is argued that the complex FV models with the language pretraining mechanism, despite their powerfulness in understanding semantics, could be further improved by auxiliary background knowledge.

2.2. Graph Neural Network for FV

The Graph Neural Network (GNN) can provide a powerful model of structural information representation, which shows promising results in NLU tasks requiring reasoning. The network also employs the network-like information passing mechanism to update graph node representation from its neighboring nodes until reaching equilibrium. Recent research aims to automatically verify the given claims using trustworthy datasets, e.g., Wikipedia. By employing the GNN, FV systems [4,5,28] could first retrieve relative evidential sentences from the provided corpus, and then aggregate and reason over the structural information to verify the claim commendably. Complex FV researches are dominated by natural language inference models because the task needs to integrate the scattered evidence and the claim and then infer the semantic relationship between them, which is used in top systems in the FEVER challenge. The GNN takes advantage of its structure to aggregate the features of isolated evidence sentences and the claim on the graph, which will make full use of the structural information of the evidence sentences and the given claim. However, it only aims to investigate how to effectively reason over the graph constructed with the retrieved evidence from the given context [4,5,28], and in this study, external information is first employed to assist the reasoning process and enhance FV performance.

3. Methodology

In the present section, firstly, the document retrieval and sentence selection modules are studied. Then, the analysis on how to incorporate the external knowledge with the retrieved information is made, and a collaborative graph with this external background knowledge is constructed. Next, an inference method is introduced to reason over the collaborative graph. The pipeline of our method is shown in Figure 3, and we use yellow module for input module and red for output module.

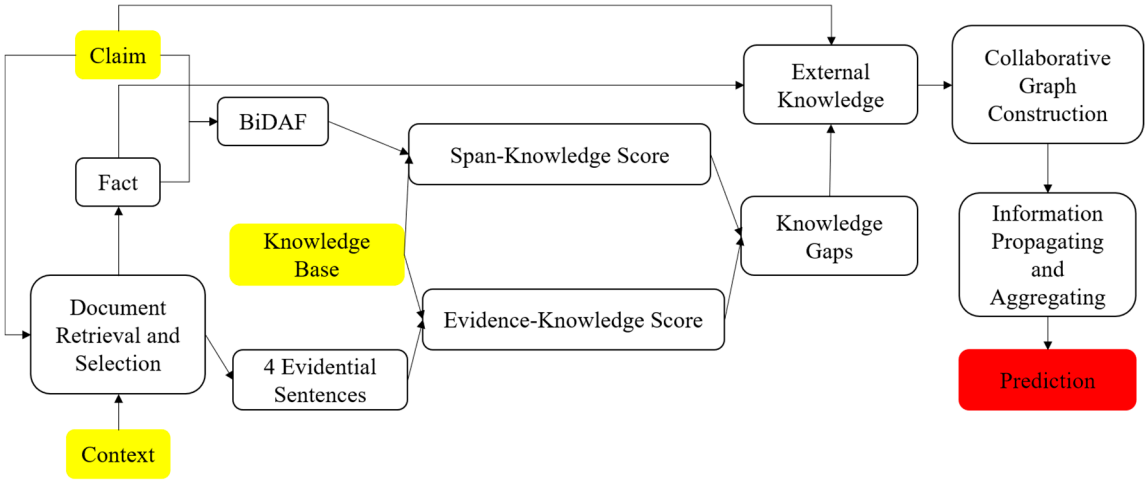

Figure 3. Pipeline of our method 3.1. Document Retrieval and Sentence Selection.

The given claim-unrelated documents and distractions are removed from the original input with a threshold filter. Inspired by the entity linking approach [1], our approach also uses the AllenNLP to extract potential entities from the given claim. Subsequently, the extracted entities are adopted to retrieve the relevant documents, and the eight highest documents are stored to be used for the candidates (although many researches select the top 5 ranked documents, while we find selecting top 8 documents will perform better than top 5 in our experiments). In the end, our approach will filter out the irrelevant documents by the word overlap between their titles and the claim. After the retrieved documents are obtained, the most relevant evidential sentence will be selected from these documents. Besides, the modified ESIM [4] is adopted to compute the relation score between evidence and the given claim. For sentences containing the top 5, the highest relation scores will be chosen as candidates. Then, the sentences are filtered out with a threshold τ, which will effectively alleviate the negative effect on our FV system caused by noise.

3.1. External Knowledge for Retrieved Evidence

In practice, supplementary information may come from various sources, which leads to different kinds of textual structures, such as natural language text, knowledge graph structural triples, and datasets with special structures. Although it is difficult to transform unstructured natural language into structured representation, it is easy to encode structured representation into unstructured natural language by only following simple rules. The next problem is what additional knowledge needs to be involved, given claims and the relevant evidence. To find and fill the knowledge gaps, there are two important modules including evidence relevance and filling gap modules. To deal with the gaps between concepts, the two modules that rely on context representation and external knowledge, respectively. will be focused on. For example, as shown in Figure 2, the FV system finds that there are knowledge gaps between the given claim ("Sirolimus accelerates aging in fruit files ... ") and one key retrieved evidential sentence (" ... feeding rapamycin to adult Drosophila produces life span extension"). The FV system will predict a span in the external knowledge base with the key fact ("Sirolimus" in this example). Then, it will retrieve some relevant knowledge from a knowledge base

("Sirolimus, also known as rapamycin", ...). After finding the key fact information, FV system will predict potential relations between each relevant external knowledge in key span and the evidential sentences. Finally, the FV system will compose the key fact with this filled gap ("Sirolimus, also known as rapamycin", ...).

3.1.1. Identifying the Knowledge Gaps

There are 3 main processes to find the knowledge gaps: our approach first identifies the key span of the evidence, then confirms the relation using retrieved knowledge, and finally retrieves the knowledge gaps.

To select the key span of the given evidence, the BiDirectional Attention Flow model [40] is adopted to make the span prediction. After getting the predicted span, background knowledge is retrieved from the ConceptNet [26] and ARC Corpus [27].

3.1.2. Evidence Relevance

The idea of this module is that relevant evidential sequences often capture the relationship between the given claims and the 4 evidential sentences ranked from 2 to 5 (the blue and red regions in Figure 4). In Figure 4, the "Fact" is the first evidential sentence or the most relevant evidential sentence to the claim.

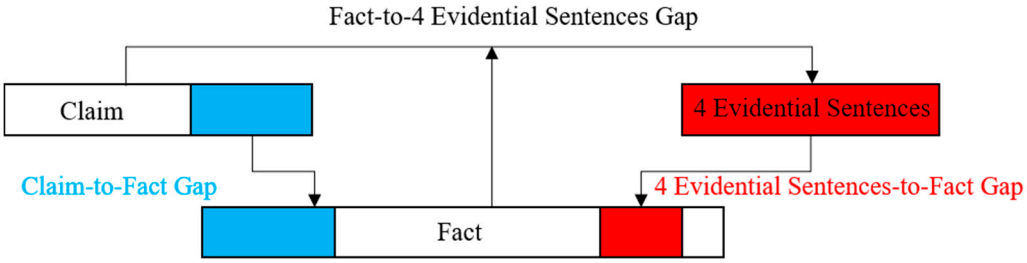

Figure 4. Overview on kinds of knowledge gaps, assuming partial knowledge from evidence.

It should be noticed that the weighted representation of claim-fact and 4 evidential sentences-to-fact pairs is calculated to capture the relationship between evidence with the given claim and annotation, respectively. To be specific, the method of Fact Relevant Attention [14] is first followed to get the claim-evidence weight representation:

$$R_c(e) = softmax_{e_m}\left(\max_{c_m}(E_c \cdot E_e)\right) \in \mathbb{R}^{1 \times e_m} \tag{1}$$

where E_c and E_e are the encodings of the claim and evidence respectively, and $E_c \cdot E_e \in \mathbb{R}^{c_m \times e_m}$.

Besides, the 4 evidential sentences-to-fact $R_l(e)$ representation is obtained in a similar way.

Finally, a feedforward neural network will be adopted to get a scalar score for each evidential sequence

$$score_e(c) = feedforward\left(\left[\frac{(R_c(e) + R_l(e))}{2} - avg(E_e); \frac{(R_c(e) + R_l(e))}{2} * avg(E_e)\right]\right) \tag{2}$$

3.1.3. Filling the Gap

In this module, background knowledge is employed to focus on the 4 evidential sentences-to-fact pair gaps. Based on [14], there are two main steps to fulfill the gap filling task:

1. The module predicts the relation between the fact span and the 4 evidential sentences information representations for each evidential sentence s_j, as shown in Equations (3) and (4):

$$R_{\hat{s}}(s_j) = softmax_{e_m}\left(\max_{c_m}\left(E_{\hat{s}} \cdot E_{s_j}\right)\right) \cdot E_{s_j} \tag{3}$$

$$R_l(s_j) = softmax_{e_m}\left(\max_{c_m}\left(E_l \cdot E_{s_j}\right)\right) \cdot E_{s_j} \tag{4}$$

where $\hat{s}$, s_j denote the predicted span and the j-th evidential sentence. These two representations obtain the contextual embedding of the words in the s_j that is most relevant to $\hat{s}$ and c_i respectively.

2. Combine the predicted relation with evidence to score the additional knowledge by the feedforward neural network

$$R'_j(\hat{s}, l) = feedforward\left(\left[S_{\hat{s}}(s_j) - S_l(s_j); S_{\hat{s}}(s_j) \cdot S_l(s_j)\right]\right) \tag{5}$$

Then, the information without knowledge gaps, called knowledge context will be gained.

3.2. Constructing the Entity Graph

After the Knowledge Context (KC) is acquired, the Stanford Corenlp Toolkit [41] is used to recognize entities from the KC. The number of extracted entities is denoted as N, while the entity graph is constructed with entities as nodes. The edge, which is the same with the DFGN [42], is built because this link ensures that entities across multiple documents are connected. Different from the DFGN, additional background knowledge is adopted to enhance the nodes' relations, which will make our entity graph more exact.

3.3. Claim Verification with the GNN Reasoning Method

In this section, the graph-based reasoning method is elaborated to fulfil FV tasks, which is another main part of this paper. Given the sequence of the claims and evidence, our model will label the claims as 'supported', 'refused', or 'not enough information'. The tuition idea of this approach is to use semantic-level structure information of knowledge evidence to predict labels of the given claims. Besides, in the present section, firstly, an encoder is used to get representation for the claim and the knowledge evidence. Then, information is propagated among knowledge evidence and reasoned over the collaborative graph. Finally, the module of aggregation is utilized to predict the label of the given claim. In the process of information propagation and aggregation, the setting of the GEAR [4] is followed. However, distinct from our baseline model GEAR, we first construct the entity graph, which contains rich information and knowledge on its edges, instead of the simple fully-connected graph as the baseline model GEAR. In addition, the common GNN reasoning approaches assume that the feature of each node is a vector. In order to retain the information of the graph nodes as much as possible, by following the GSN [28], the representation of sequence features is learnt directly, instead of transforming the sequence feature into a fixed dimensional feature vector through a summary module.

Besides, $G = (V, E)$ is employed to represent the constructed graph, where V refers to a set of N nodes, and the i-th node Vi indicates a sequence of the feature vector, which can be denoted as $V_i = \left[v_i^1, v_i^2 \cdots v_i^{l_i}\right]$. l_i is the sequence length of node i; Different from the GNN, each element of the feature vector is a D-dimensional vector. E denotes a set of edges, linking two nodes.

In the k-th hop, based on the GSN's aggregation f_{agg} and combination f_{com} function [28], it is easy to calculate the structure-aware feature representation and the current feature representation fused with neighboring information, as shown in Equations (6) and (7), respectively.

$$Z_i^k = f_{agg}\left(\left\{V_j^{k-1} : \forall j \in N(i)\right\}\right) \tag{6}$$

$$V_i^k = f_{com}\left(V_i^{k-1}, Z_i^k\right) \tag{7}$$

where $N(i)$ is the neighboring nodes of node i, and V_i^k is the node representation learned after the k-th hop. Putting these two functions together, we can obtain a formula specifically designed for this sequential application [28], as shown in Equation (8).

$$V_i^k = f_{com}\left(g\left(V_i^{k-1}, V_j^{k-1}\right)\right) \tag{8}$$

where g denotes the co-attention function, and the Bidirectional Attention Flow [40] is also selected as g for the aggregation function. For the combination function, the mean pooling function is chosen. After learning neighbor-aware representations of the current node, our model has the strengthened ability to reason over multiple evidential sentences. Once the final state is obtained, a one-layer MLP is adopted to get final prediction l.

4. Experiments

4.1. Experimental Setting

4.1.1. Dataset

Our experiments are performed on the large-scale dataset FEVER. The dataset consists of 185 K annotated claims with a set of 5416 k Wikipedia documents, which is developed to extract evidence and verify synthetic claims. In this section, FV is emphasized. In addition, the setting of GEAR [4] is followed to extract the evidence. The training set and the dev set, which contain about 145 K and 20 K samples accordingly, are also used.

4.1.2. Baselines

In our experiments, the current out-of-the-art FV methods are adopted as baselines, including three previous top approaches (Athene [1], UCL MRG [2], and UNC NLP [3]) and the approaches based on pre-trained language models (GEAR [4] and KGAT by [5]).

4.1.3. Evaluation Metrics

In order to facilitate the comparison with baseline models, the official metrics, including Label Accuracy (LA) and FEVER scores are employed to evaluate our FV model. LA is to measure claim classification accuracy, and FEVER is to measure whether the FV system can provide at least one complete set of golden evidence.

4.1.4. Data Processing

The identification of knowledge span and the extraction of knowledge are extremely important for this work. We first evaluate the key span identification model with the annotated spans in knowledge gap dataset for training, and then trained it on the SQuAD dataset. Unfortunately, the performance is quite poor. We found that knowledge gaps dataset [14] can improve the accuracy, F1 and EM scores of our model, which was pretrained on SQuAD. Therefore, all the experiments in the present study use this fine-tuned method. In addition, we adopt four external knowledge sources: ConceptNet [26], WordNet subset (used in [10]), OMCS (Open Mind Common Sense subset), and ARC (AI2 Reasoning Challenge dataset) [27]. In order to avoid the noise caused by human subjective factors as much as possible, we use the method in [14] to identify a subset of fact verification and split the fact-to-4 evidential sentences candidates gap annotation into two steps, respectively, core term identification and relation extraction.

4.2. Performance

In this section, our framework is compared with baseline models on FEVER dev sets to evaluate its performance, including its advantages in various reasoning scenarios, and the effectiveness under

the condition of knowledge gaps. Table 1 displays the performance of DKAR and the baselines on the FEVER dev set, showing that DKAR outperforms previous systems with 79.64% and 77.12% in terms of LA and FEVER scores, respectively, on the FEVER dev set.

Table 1. Results on the FEVER dev set.

Model	LA (%)	FEVER Score (%)
Athene [1]	68.49	64.74
UCL MRG [2]	69.66	65.41
UNC NLP [3]	69.72	66.49
GEAR [4]	74.84	70.69
KGAT [5]	78.02	75.86
DKAR	**79.64**	**77.12**

Table 2 presents the confusion matrix for the dev set prediction, where "prediction" is denoted as P and "ground truth" is indicated as G. For the baselines, it is easy to wrongly label the claims as "not enough information", and it is also a great challenge for them to correctly recognize the "not enough information" claims. Based on Table 2, it can be seen that our DKAR with auxiliary knowledge can effectively enhance accuracy (these improvements were highlighted with **bold** and ↑) and reduce the number of label errors corresponding to "not enough information" (these improvements were highlighted with **bold** and ↓).

Table 2. Confusion matrix on the FEVER dev set.

	Supported (G)	Refused (G)	Not Enough Information (G)	Total
Supported (P)	**5942 (↑8%)**	625	398	6965
Refused (P)	957	**4397 (↑4.6%)**	1031	6385
Not enough information (P)	561 **(↓ 9.7%)**	499 **(↓ 3.6%)**	**5588 (↑17%)**	6648
total	7460	5521	7017	19,998

4.3. Further Expeeriments of FV System on Diverse Web Information

Although the above experiments demonstrate that DKAR can effectively solve difficult problems caused by the knowledge gap in the synthetic dataset, this paper continues to study the model performance in a more practical dataset with multiple web sources in this section. The dataset of the FakeNewsNet contains labelled news and social context information from two platforms: PolitiFact (politifact.com) and GossipCop (gossipcop.com). Besides, this dataset involves both meta attributes (e.g., all the tweets and news body text) and social context (e.g., comments for each item of news).

In order to test that our DKAR can retrieve and fill the knowledge gaps in diverse text resources effectively, we use the whole and part of the datasets respectively in the following experiments. Specifically, our model is trained on the FakeNewsNet (PolitiFact and GossipCop, respectively), and the partial dataset involves 100 tweets for each piece of news, respectively. We follow the practice of training GNN to train and test our module: randomly select 75% of the data as the training data, the rest as the test data, and the report is the average result of 5 replicates. In order to support comparison with baseline methods for fake news detection (e.g., HAN [43], TCNN-URG [44], HPA-BLSTM [45], Csi [46], dEFEND [47], and the GNN with Continual Learning [48]), the following metrics, namely accuracy, precision, recall, and the F1 score are employed and defined as follows.

Accuracy: Accuracy is simply a ratio of correct predictions to the total number of predictions.
Precision: Precision is the ratio of true positives to the total predicted positive observations.
Recall: Recall is the ratio of true positives to actual positives.
F1: F1 is the harmonic mean of precision and recall.

The experimental results are shown in Figures 5 and 6, where the first 6 bars of each set in each figure are the performance results of the baseline models in the FakeNewsNet, while the second last bar is the result of our approach on the whole dataset, and the last bar is the result of our approach on the partial dataset. For "the whole dataset", there may be less knowledge gap than in "partial dataset" due to the mutual support between sentences, which is not conductive to testing whether our approach has the ability to fill knowledge gaps. In order to test this ability of our method, we use part of "the whole dataset" as "partial dataset". Recently, almost all the FV systems are based on data-driven methods, and the performance of these systems relies heavily on the size of datasets. If the experimental performance does not get worse in an obvious manner when the size of the dataset decreases significantly, we can argue that the knowledge-driven mechanism can effectively improve the robustness of data-driven FV systems. According to the results, the model put forward in this study has the comparable performance on PolitiFact and GossipCop, both on the complete and partial dataset. In addition, in order to test the ability of our model under the conditions of knowledge gaps, partial datasets with 200, 500, 1000, and 1500 tweets for each piece of news, respectively, are used to test our model, and the results are shown in Table 3, presenting that with gradual increase of data, the performance improves slowly, which shows that the reasoning ability of our model does not highly rely on the data, and the internal knowledge-driven function plays an important role.

Table 3. Performance of our approach on the partial datasets.

Dataset	Metrics	100	500	1000	1500
PolitiFact	Accuracy	0.879	0.892	0.903	0.908
	Precision	0.897	0.913	0.921	0.925
	Recall	0.891	0.909	0.910	0.913
	F1	0.895	0.903	0.915	0.919
GossipCop	Accuracy	0.890	0.896	0.900	0.903
	Precision	0.906	0.910	0.915	0.916
	Recall	0.898	0.901	0.905	0.907
	F1	0.903	0.906	0.913	0.914

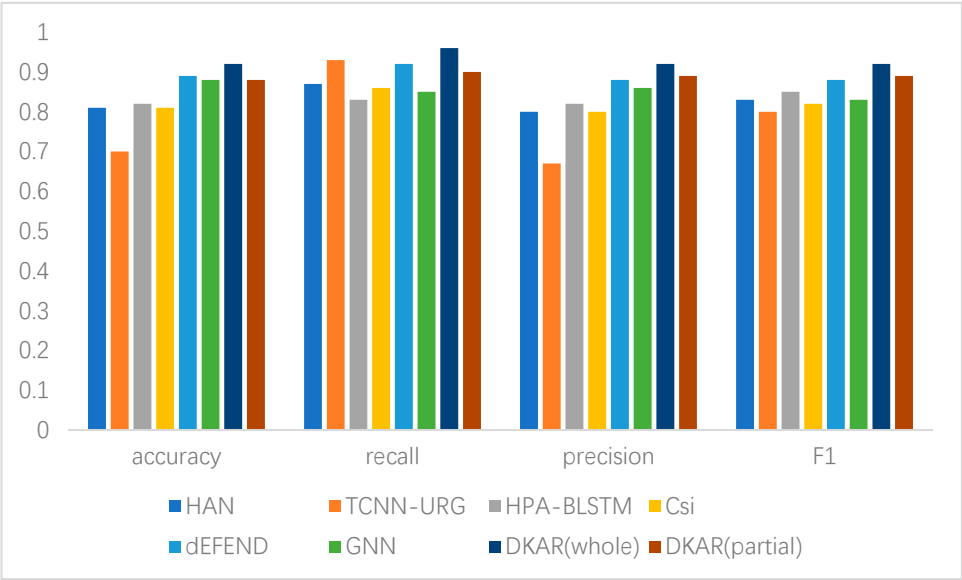

Figure 5. Performance comparison on the dataset of PolitiFact.

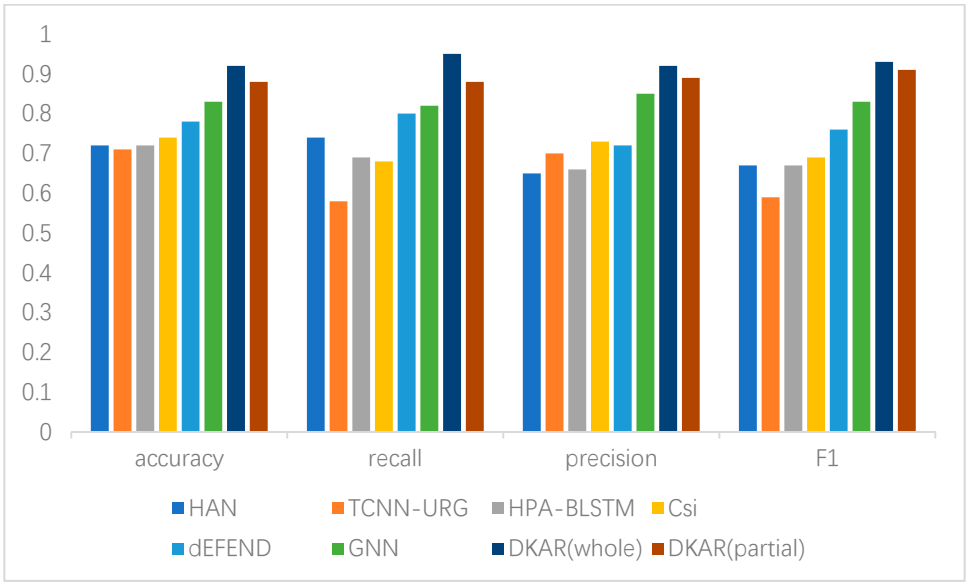

Figure 6. Performance comparison on the dataset of GossipCop.

Although the above experiments have confirmed that our model can overcome the knowledge gap obstacle on a single dataset, whether our method works well on new datasets still should be tested. If the model continues to perform well on a new dataset, it reveals that our method has stronger learning ability and robustness. Besides, the model is performed on the datasets with 100 tweets for each item of news (train on the PolitiFact and test on the GossipCop). As presented in Figure 7, the results show that compared with a baseline model, although the graph representation of PolitiFact and GossipCop is vastly different (heterogeneous information), including numbers of nodes and edges, our model has a relatively stable performance. In this case, it can be proved that our model is featured with strong robustness.

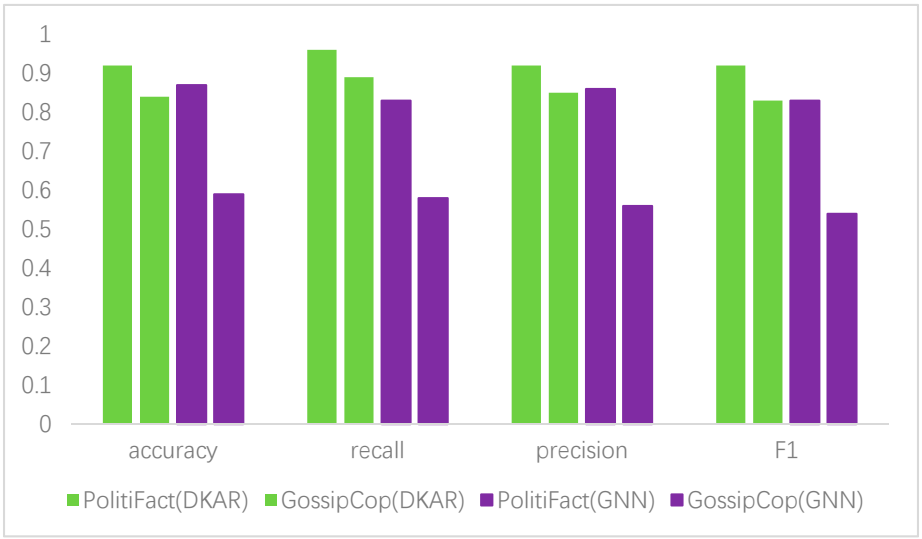

Figure 7. Models trained on the dataset of the PolitiFact perform comparably on the dataset of GossipCop, compared with the state-of-the-art model (GNN).

5. Limitation of the Study

Limited by the size of knowledge bases, some knowledge gaps cannot be filled effectively in practice. We are studying a creative reasoning mechanism to solve this problem. In addition,

fact verification is a very open and challenging task. It needs not only the support of linguistic features and background knowledge, but also the support of more complex multi-dimension information, such as social content and spatiotemporal information. For example, claims evolve over time, and what was fake yesterday is true today.

6. Conclusions

In this study, a novel graph-based reasoning framework was proposed for complex fact verification (FV), which can dynamically supplement useful knowledge in the case of knowledge gaps. The framework retrieves and fills the knowledge gaps between the given claim and evidence to construct the collaborative graph before propagating and aggregating sequential information. Experiments have shown that DKAR can effectively solve the "not enough information" mislabeling problem in the FV task and outperform other baselines. In addition, our approach shows outstanding advantages in a small sample and heterogeneous web text sources. Our research first illustrates that dynamic knowledge supplementation plays an important role in complex FV tasks, which contributes to the study of reasoning methods driven by data and knowledge for fact verification. It is expected that our first exploration encourages others to expand upon our work, and to further shed light on the broader and more challenging goal of complex and practical FV tasks with joint data and knowledge.

Author Contributions: C.X. and T.W. planned and supervised the whole project; Y.W. developed the main theory and wrote the manuscript; C.S., C.Z., and T.W. contributed themselves to doing the experiments and discussing the results. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China (Grant U1636208, No. 61862008 and No. 61902013) and Beihang Youth Top Talent Support Program (Grant No. YWF-20-BJ-J-1038).

Acknowledgments: We greatly appreciated all the anonymous reviewers' hard work.

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

References

1. Hanselowski, A.; Zhang, H.; Li, Z.; Sorokin, D.; Schiller, B.; Schulz, C.; Gurevych, I. Ukp-athene: Multi-sentence textual entailment for claim verification. *arXiv* **2018**, arXiv:1809.01479.
2. Yoneda, T.; Mitchell, J.; Welbl, J.; Stenetorp, P.; Riedel, S. Ucl machine reading group: Four factor framework for fact finding (hexaf). In Proceedings of the First Workshop on Fact Extraction and VERification (FEVER), Brussels, Belgium, 1 November 2018; pp. 97–102.
3. Nie, Y.; Chen, H.; Bansal, M. Combining fact extraction and verification with neural semantic matching networks. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; pp. 6859–6866.
4. Zhou, J.; Han, X.; Yang, C.; Liu, Z.; Wang, L.; Li, C.; Sun, M. GEAR: Graph-based evidence aggregating and reasoning for fact verification. *arXiv* **2019**, arXiv:1908.01843.
5. Liu, Z.; Xiong, C.; Sun, M.; Liu, Z. Fine-grained Fact Verification with Kernel Graph Attention Network. *arXiv* **2019**, arXiv:1910.09796.
6. Thorne, J.; Vlachos, A.; Christodoulopoulos, C.; Mittal, A. Fever: A large-scale dataset for fact extraction and verification. *arXiv* **2018**, arXiv:1803.05355.
7. Yang, Z.; Qi, P.; Zhang, S.; Bengio, Y.; Cohen, W.W.; Salakhutdinov, R.; Manning, C.D. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. *arXiv* **2018**, arXiv:1809.09600.
8. Khashabi, D.; Chaturvedi, S.; Roth, M.; Upadhyay, S.; Roth, D. Looking beyond the surface: A challenge set for reading comprehension over multiple sentences. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, New Orleans, LA, USA, 1–6 June 2018; Volume 1 (Long Papers), pp. 252–262.
9. Welbl, J.; Stenetorp, P.; Riedel, S. Constructing datasets for multi-hop reading comprehension across documents. *Trans. Assoc. Comput. Linguist.* **2018**, *6*, 287–302. [CrossRef]
10. Mihaylov, T.; Clark, P.; Khot, T.; Sabharwal, A. Can a suit of armor conduct electricity? a new dataset for open book question answering. *arXiv* **2018**, arXiv:1809.02789.

11. Pîrtoacă, G.-S.; Rebedea, T.; Ruseti, S. Answering questions by learning to rank–Learning to rank by answering questions. *arXiv* **2019**, arXiv:1909.00596.
12. Banerjee, P.; Pal, K.K.; Mitra, A.; Baral, C. Careful selection of knowledge to solve open book question answering. *arXiv* **2019**, arXiv:1907.10738.
13. Mitra, A.; Banerjee, P.; Pal, K.K.; Mishra, S.; Baral, C. Exploring ways to incorporate additional knowledge to improve natural language commonsense question answering. *arXiv* **2019**, arXiv:1909.08855.
14. Khot, T.; Sabharwal, A.; Clark, P. What's Missing: A Knowledge Gap Guided Approach for Multi-hop Question Answering. In Proceedings of the International Joint Conference on Natural Language Processing, Hong Kong, China, 3–7 November 2019; pp. 2814–2828.
15. Mihaylov, T.; Frank, A. Knowledgeable Reader: Enhancing Cloze-Style Reading Comprehension with External Commonsense Knowledge. In Proceedings of the Meeting of the Association for Computational Linguistics, Melbourne, Australia, 15–20 July 2018; pp. 821–832.
16. Chen, Q.; Zhu, X.; Ling, Z.; Inkpen, D.; Wei, S. Neural Natural Language Inference Models Enhanced with External Knowledge. In Proceedings of the Meeting of the Association for Computational Linguistics, Melbourne, Australia, 15–20 July 2018; pp. 2406–2417.
17. Yang, A.; Wang, Q.; Liu, J.; Liu, K.; Lyu, Y.; Wu, H.; She, Q.; Li, S. Enhancing pre-trained language representations with rich knowledge for machine reading comprehension. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019; pp. 2346–2357.
18. Wang, C.; Jiang, H. Explicit utilization of general knowledge in machine reading comprehension. *arXiv* **2018**, arXiv:1809.03449.
19. Jansen, P.; Ustalov, D. TextGraphs 2019 shared task on multi-hop inference for explanation regeneration. In Proceedings of the Thirteenth Workshop on Graph-Based Methods for Natural Language Processing (TextGraphs-13), Hong Kong, China, 4 November 2019; pp. 63–77.
20. Banerjee, P. ASU at TextGraphs 2019 shared task: Explanation ReGeneration using language models and iterative re-ranking. *arXiv* **2019**, arXiv:1909.08863.
21. Kadlec, R.; Schmid, M.; Bajgar, O.; Kleindienst, J. Text Understanding with the Attention Sum Reader Network. *arXiv* **2016**, arXiv:1603.01547.
22. Weissenborn, D.; Kocisky, T.; Dyer, C. Dynamic Integration of Background Knowledge in Neural NLU Systems. *arXiv* **2018**, arXiv:1706.02596.
23. Bollacker, K.; Evans, C.; Paritosh, P.; Sturge, T.; Taylor, J. Freebase: A collaboratively created graph database for structuring human knowledge. In Proceedings of the International Conference on Management of Data, Vancouver, BC, Canada, 10–12 June 2008; pp. 1247–1250.
24. Krishnamurthy, J.; Dasigi, P.; Gardner, M. Neural Semantic Parsing with Type Constraints for Semi-Structured Tables. In Proceedings of the Empirical Methods in Natural Language Processing, Copenhagen, Denmark, 9–11 September 2017; pp. 1516–1526.
25. Petrochuk, M.; Zettlemoyer, L. SimpleQuestions Nearly Solved: A New Upperbound and Baseline Approach. In Proceedings of the Empirical Methods in Natural Language Processing, Brussels, Belgium, 31 October–4 November 2018; pp. 554–558.
26. Speer, R.; Chin, J.; Havasi, C. ConceptNet 5.5: An Open Multilingual Graph of General Knowledge. In Proceedings of the National Conference on Artificial Intelligence, Copenhagen, Denmark, 9–11 September 2017; pp. 4444–4451.
27. Clark, P.; Cowhey, I.; Etzioni, O.; Khot, T.; Sabharwal, A.; Schoenick, C.; Tafjord, O. Think you have Solved Question Answering? Try ARC, the AI2 Reasoning Challenge. *arXiv* **2018**, arXiv:1803.05457.
28. Tu, M.; Huang, J.; He, X.; Zhou, B. Graph sequential network for reasoning over sequences. *arXiv* **2020**, arXiv:2004.02001.
29. Rajpurkar, P.; Zhang, J.; Lopyrev, K.; Liang, P. Squad: 100,000+ questions for machine comprehension of text. *arXiv* **2016**, arXiv:1606.05250.
30. Rajpurkar, P.; Jia, R.; Liang, P. Know what you don't know: Unanswerable questions for SQuAD. *arXiv* **2018**, arXiv:1806.03822.
31. Joshi, M.; Choi, E.; Weld, D.S.; Zettlemoyer, L. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *arXiv* **2018**, arXiv:1705.03551.

32. Nguyen, T.; Rosenberg, M.; Song, X.; Gao, J.; Tiwary, S.; Majumder, R.; Deng, L. Ms marco: A human-generated machine reading comprehension dataset. *arXiv* **2016**, arXiv:1611.09268.
33. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
34. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 5998–6008.
35. Goldberg, Y. Assessing BERT's syntactic abilities. *arXiv* **2019**, arXiv:1901.05287.
36. Peters, M.E.; Neumann, M.; Zettlemoyer, L.; Yih, W.-T. Dissecting contextual word embeddings: Architecture and representation. *arXiv* **2018**, arXiv:1808.08949.
37. Chen, D.; Bolton, J.; Manning, C.D. A thorough examination of the cnn/daily mail reading comprehension task. *arXiv* **2016**, arXiv:1606.02858.
38. Bauer, L.; Wang, Y.; Bansal, M. Commonsense for generative multi-hop question answering tasks. *arXiv* **2018**, arXiv:1809.06309.
39. Zhong, W.; Tang, D.; Duan, N.; Zhou, M.; Wang, J.; Yin, J. Improving question answering by commonsense-based pre-training. In Proceedings of the CCF International Conference on Natural Language Processing and Chinese Computing, Dunhuang, China, 9–14 October 2019; pp. 16–28.
40. Seo, M.; Kembhavi, A.; Farhadi, A.; Hajishirzi, H. Bidirectional attention flow for machine comprehension. *arXiv* **2016**, arXiv:1611.01603.
41. Manning, C.D.; Surdeanu, M.; Bauer, J.; Finkel, J.R.; Bethard, S.; Mcclosky, D. The Stanford CoreNLP Natural Language Processing Toolkit. In Proceedings of the Meeting of the Association for Computational Linguistics, Baltimore, MD, USA, 22–27 June 2014; pp. 55–60.
42. Qiu, L.; Xiao, Y.; Qu, Y.; Zhou, H.; Li, L.; Zhang, W.; Yu, Y. Dynamically fused graph network for multi-hop reasoning. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019; pp. 6140–6150.
43. 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.
44. Qian, F.; Gong, C.; Sharma, K.; Liu, Y. Neural User Response Generator: Fake News Detection with Collective User Intelligence. In Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence Main track (IJCAI), Stockholm, Sweden, 13–19 July 2018; pp. 3834–3840.
45. 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; pp. 943–951.
46. 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; pp. 797–806.
47. 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, Anchorage, AK, USA, 4–8 August 2019; pp. 395–405.
48. Han, Y.; Karunasekera, S.; Leckie, C. Graph Neural Networks with Continual Learning for Fake News Detection from Social Media. *arXiv* **2020**, arXiv:2007.03316.