AMIF: A Hybrid Model for Improving Fact Checking in Product Question Answering

1st Hongzhan Lin*
Beijing University of Posts and Telecommunications
Beijing, China
linhongzhan@bupt.edu.cn

2nd Liangliang Chen*
Beijing University of Posts and Telecommunications
Beijing, China
outside@bupt.edu.cn

3rd Jing Ma✉
Hong Kong Baptist University
Hong Kong SAR, China
majing@comp.hkbu.edu.hk

4th Zhiwei Yang
Jilin University
Changchun, China
yangzw18@mails.jlu.edu.cn

5th Guang Chen✉
Beijing University of Posts and Telecommunications
Beijing, China
chenguang@bupt.edu.cn

2022 International Joint Conference on Neural Networks (IJCNN) | 978-1-7281-8671-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/IJCNN55064.2022.9892491

***Abstract*—Fact checking in product-related community question answering is the task of verifying the truthfulness of an answer towards a given question, where the study has just begun. Most existing related work has focused on tailoring solutions to shallow feature fusion for the single-text claim involved with fact-checked evidence, limiting their success and generality in such answer truthfulness prediction task on E-commerce platforms. In this study, we propose an attention-based hybrid framework for multi-feature interaction fusion to determine the truthfulness of the answer towards a product-related question in E-commerce, which could not only support fine-grained semantic calibration between question-answer pairs for better understanding of the target answers, but also substantially cross-check all retrieved evidence to mine coherent opinions towards the pair. In addition, our framework further integrates non-textual features from metadata for improving performance. Extensive experiments conducted on real-world representative benchmark data show that our proposed model achieves superior performance on the task of answer veracity prediction.**

***Index Terms*—Community question answering, fact checking, hybrid model, multi-feature fusion, E-commerce**

I. Introduction

To help potential customers eliminate doubts, E-commerce websites (e.g., Amazon, Taobao, etc) provide community question answering services as shown in Figure 1 to facilitate online shopping via product-related discussions among users. Nevertheless, the user-generated messages on these community forums to answer the online questions are generally less accurate or flawed in terms of veracity, without systematic effort to moderate the messages [1]. Misunderstanding of questions, improper expression in writing, and even malicious attacks from competitors can lead to untruthful answers [2]. However, checking whether the answers are factually correct or not towards a given question is usually ignored by conventional community question answering (CQA) research [3]–[6], where a candidate answer is considered GOOD if it is semantically relevant to or tries to address the question, irrespective of its veracity, accuracy, etc [7].

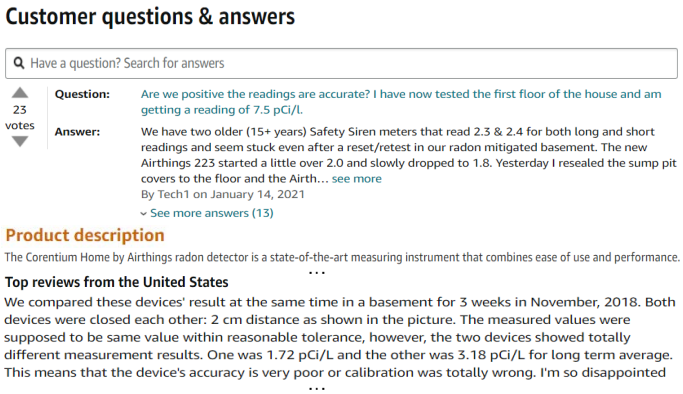


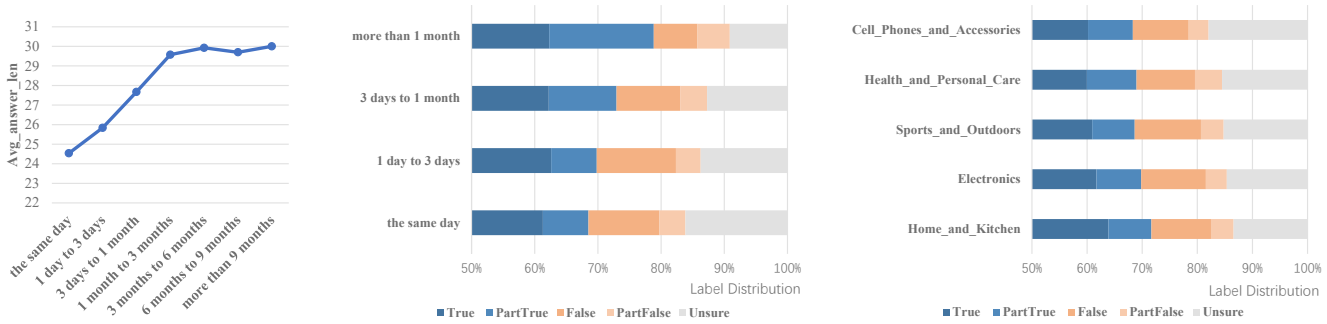
Fig. 1. Example of question answering in E-commerce, with relevant product information like the product description and user reviews.

Fact checking for question answering in E-commerce (FC-QA) aims to predict the veracity of a given answer with respect to a question about a specific product. Traditional approaches for CQA tasks focused on relevance matching between the pair of question and answer [4]–[6]. Recent methods for automatic fact checking were proposed to verify a piece of claim or statement utilizing evidence extracted from various outlets [8]–[10]. However, few previous studies are related to such answer truthfulness prediction task, which has just begun.

To study fact checking in CQA, [7] investigated the potential need of predicting the veracity of answers in community forums and initially formulate the problem of FC-QA. Following that, [11] proposed an evidence-aware model with tailored evidence interpolation in the context of CQA problems. However, the above studies only considered relatively shallow feature fusion strategies like concatenation or simple neural networks for the QA settings, but failed to exploit the close semantic correlations between the question and answer. On the other hand, existing

This work was partially supported by MoE-CMCC "Artificial Intelligence" Project No. MCM20190701, HKBU One-off Tier 2 Start-up Grant (Ref. RCOFSGT2/20-21/SCI/004) and HKBU direct grant (Ref. AIS 21-22/02).

*The first two authors contributed equally to this work.



(a) The average length of answers with time gap. (b) The label distribution with time gap. (c) The label distribution with product domain.

Fig. 2. The illustration for the impact of metadata on the quality of answers.

TABLE I
AN EXAMPLE QUESTION (Q) AND ANSWER (A) TOGETHER WITH RELEVANT PRODUCT INFORMATION (s_1-s_5) AS RETRIEVED EVIDENCE.

Q: Can i insert a tmobile sim card into this phone? **A**: Sorry. . . That's not gonna work out. . .it's a Verizon phone. No sim.
s_1: The one thing they could do is tell you that it does not come with a sim card and for Page Plus you have to buy separate. s_2: Had to send it back because it didn't come with sim card and would not work with my prepaid plan. s_3: They provided a SIM card and made the change to my account. s_4: plugged in my old Verizon sim card from last phone and it worked in minutes. s_5: put old phones sim card in and the phone worked immediately.
Explanation: "That" and "it" in **A** refers to "insert sim into phone" and "phone" in **Q** respectively. The evidence s_4 and s_5 can further be indicative of the answer veracity. **Verdict:** False

TABLE II
THE EXAMPLES OF THE ANSWERS AT DIFFERENT TIMES TO THE GIVEN QUESTION.

Product Domain: Home_and_Kitchen		
Question: How many holder holes are in the 108" curtain? **Time**: July 3, 2013		
Answers	**Time**	**Verdict**
12.	August 1, 2013	False
No, there are not 12! There are 18 holes in the 108" curtain. I just measured the curtain and counted them to be sure.	September 18, 2013	True

state-of-the-art methods largely ignored or oversimplified the mutual corroboration between pieces of evidence from raw retrieved auxiliary information, as not all evidence sentences are equally useful and reliable, which can be misled by conflicting evidence sentences and makes random predictions.

The example in Table I illustrates our general idea in this paper: given a question-answer pair and its relevant pieces of evidence such as product description and user comments, we try to understand the answer based on the context of the question and attend over the most evidential sentences to verify the answer. Firstly, we observe that antecedents of the referent like "that" and "it" in the answer, are the mentions in the question. We argue that the first key to FC-QA task is to excavate the rich inter-relationship by interpretable reasoning about the interplay of QA pairs. Besides, to eliminate unreliable information (e.g., the marginal evidence s_3) from the evidence set, coherent opinions captured by comparing all shreds of evidence would further enhance automatic verification for the answer [12], [13]. This is because the semantic clues, such as entailment expression (in green) and conflicting snippets (in red), are ubiquitous between the answer and evidence in terms of the question context (in blue).

Then, another shortcoming we look to address is the paucity of metadata utilization for FC-QA. Previous literature [14], [15] showed that the inclusion of non-textual metadata in the E-commerce field will also have an impact on the prediction task related to data mining. Figure 2 illustrates the indicative signals of metadata on the answer veracity from a real-world benchmark (i.e., AnswerFact [11]). From Figure 2(a)-(b), we can see that the time gap between the question and answer can imply the quality of the answer to some extent. For instance, the longer the time gap between the question and answer, the larger the length and the higher the quality of the answer. In particular, the average length of answers more than 1 month after the question was posted tends to level off over time, and nearly 80% of the answers are considered to be positive (i.e., TRUE or PARTTRUE), with the unsure answers less than 10%. To better illustrate our intuition, Table II exemplifies a case that users in E-commerce tend to correct previous wrong answers and give more factually correct answers over time. Not only that, but the product domain of the question is also one of the factors that imply the quality of the answer. As Figure 2(c) demonstrates, the answers in 'Home and Kitchen' have a relatively higher probability ($> 70\%$) to be positive meanwhile with the unsure answers less than 14%, when compared with other product domains. These phenomena suggest that the interaction between metadata (e.g., product domain, time stamp, and text length) can be used to facilitate the answer truthfulness prediction.

To this end, we propose a novel Attention-based hybrid deep model for Multi-feature Interaction Fusion (AMIF) to predict the truthfulness of each answer towards a given question, which not only considers a thorough understanding of the answer evoked by the question but also exploits relevant product information to capture coherent pieces of evidence to verify the answer in a unified framework. Specifically, a QA reasoning calibration module is proposed at first to support the semantic matching between the question and answer. Meanwhile, to explore the relatively informative evidence as a readable explanation, we employ a self-attention mechanism to cross-check all retrieved evidence, and a gate-induced decoupling mechanism is designed for the credibility ranking of each piece of evidence, which disentangles entailment and

conflicting semantics from the feature correlation between the answer and shreds of evidence based on the context of the question. Moreover, we further fuse the non-textual metadata from E-commerce websites with the textual representation via a factorization-machine based neural network [16] to improve our proposed method. Experiments conducted on the real-world FC-QA benchmark [11] show that our proposed model achieves superior performance on FC-QA task and provides rational explanations of the prediction by optimizing token-level QA similarity and extracting informative evidence for prediction. The main contribution of this paper are of three-fold:

- We propose a novel hybrid deep framework to fuse multiple features from questions, answers and auxiliary information for answer truthfulness identification to investigate the problem of fact checking in product question answering.
- Our attention-based framework not only learns a fine-grained understanding of the answer in the question context by the reasoning calibration over the QA pairs, but also explores the coherence of evidence sentences via trust cross-checking and credibility ranking. Moreover, we model interactions of non-textual features from metadata in E-commerce to further improve the performance.
- Experimental results have demonstrated the state-of-the-art performance of our proposed AMIF on real-world E-commerce data.

II. Related Work

Community Question Answering. With the development of online community forums, community question answering (CQA) has become an emerging research topic in recent years [5], [6], [17], [18]. Recent neural networks [19]–[21] have been actively applied to the answer selection task and achieved good results, which consider the label of an answer to be positive if the answer is semantically relevant to the corresponding question irrespective of its veracity (our focus here). Yet, in the context of CQA, there has been work on quality assessment of answers, e.g. available answer ranking [22] and answer helpfulness prediction in product-related CQA forums [23]. Fact checking in question answering scenario generally focus on only using article (QA pairs) contents [1], [7]. In this paper, we study the novel problem of fact checking in product question answering, aiming to improve veracity prediction of an answer for a given question making use of evidence and metadata in E-commerce as external sources.

Fact Checking. Previous comprehensive surveys have reviewed the extensive literature on fact checking and credibility assessment [24]–[28]. Hereinafter, we only give a brief review of prior works closely related to the evidence-based fact checking. Deep learning models such as recurrent neural networks (RNN) [29] and convolutional neural networks (CNN) [30] were exploited to learn the claim and evidence representations. [9], [31], [32] built a pipeline to find documents and sentences for fact checking of mutated claims generated from Wikipedia pages. [33] aimed to find webpages related to given fact-checking articles and predict their stances on claims. The word-level [8] and sentence-level [34] attention mechanism on relevant articles were utilized to debunk false claims by learning claim and evidence representations, respectively. Evidence-ranking methods [10], [11] are then proposed to conduct veracity prediction for fact checking. However, due to the insufficient multi-feature fusion in these approaches, the rich context in the question, the coherence cross-checked by pairwise evidence, and the indicative signal from metadata in E-commerce are not fully explored to facilitate the answer truthfulness prediction.

III. Methodology

We define a fact checking question answering dataset as $\{\mathcal{C}\}$, where each instance $\mathcal{C} = (q, a, y, S, M)$ is a tuple representing a question q, an answer a, a ground-truth label y indicating the verdict of a, a set of relevant evidence sentences $S = \{s_i\}_{i=1}^{n}$, and the non-textual metadata set M in E-commerce. Our task is to classify an answer into a pre-defined veracity class, e.g., True, False, etc.

In this section, we introduce our interpretable FC-QA model, which consists of four components: context encoder, QA reasoning calibration, evidence coherence modeling and metadata fusion. Figure 3 gives an overview of our framework, which will be depicted in detail in the subsections.

A. Context Encoder

Given each token in a text sequence that could be either q, a, or s_i, we map it into a representation w initialized with pre-trained word vectors. We then model the context interactions among tokens in the sequence using a Bi-LSTM encoder:

$$x_t = \text{Bi-LSTM}^c\left(w_t, x_{t-1}\right) \tag{1}$$

where x_t is the hidden state of the Bi-LSTM encoder at the t-th time step. We thus denote the representation of the question q, the answer a and the i-th evidence sentence s_i after such context-aware encoding as X_q, X_a and X_{s_i} respectively: $X_* = \left[x_1^*, x_2^*, ..., x_{|*|}^*\right]; * \in \{q, a, s_i\}$, where $X_q \in \mathbb{R}^{|q|\times d}$, $X_a \in \mathbb{R}^{|a|\times d}$ and $X_{s_i} \in \mathbb{R}^{|s_i|\times d}$, d is the dimension of the hidden state of the Bi-LSTM encoder.

B. QA Reasoning Calibration

There is a rich inter-relationship between the answer and question, e.g., answers tend to use pronouns for referring back to mentions in questions. For example, for the QA in Table I, "that" (in A) refers to "insert sim into phone" (in Q). To support token-level semantic calibration between the question and the answer for better context understanding, we first evolve the question representation into that of the answer:

$$\begin{aligned} X_q^a &= \text{ReLU}\left(\mathcal{U}^\top X_q W_{qa}\right) \\ \mathcal{U}_{ij} &= \frac{\exp\left(x_i^{q\top} x_j^a\right)}{\sum\limits_{ij} \exp\left(x_i^{q\top} x_j^a\right)} \end{aligned} \tag{2}$$

where $X_q^a \in \mathbb{R}^{|a|\times d}$ is the attended question representation which represents each answer token by aggregating features

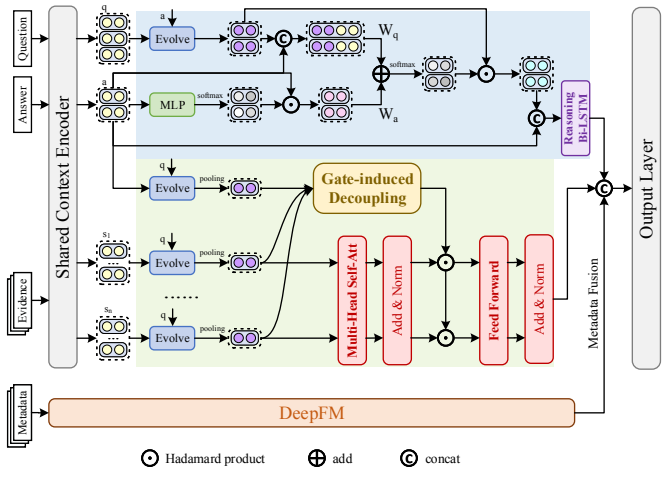


Fig. 3. The architecture of our proposed framework. The blue background portion denotes the QA Reasoning Calibration module and the green background portion denotes the Evidence Coherence Modeling module. The non-textual feature would be fed into the DeepFM module for metadata fusion.

of question, W_{qa} is a trainable matrix. $\mathcal{U} \in \mathbb{R}^{|q|\times|a|}$ is the weighted matrix by calculating similarity of each question token to answer token.

In this way, each word in the answer is represented based on the related words in the question, indicating the reference relationship. To measure the importance of each token in the answer, we get an enhanced answer representation leveraging a token-level attention mechanism:

$$\begin{aligned} \gamma_a &= \mathrm{softmax}\left(\mathrm{MLP}\left(X_a\right)\right) \\ \tilde{X}_a &= \gamma_a \odot X_a \end{aligned} \tag{3}$$

where $\mathrm{MLP}(\cdot)$ is a multilayer perceptron to encode the answer feature, γ_a is the token-level attention weight matrix normalized among all token-level linguistic representation of the answer via a softmax function. We again utilize another attention mechanism to improve the question representation on top of the enhanced vectors:

$$\begin{aligned} \gamma_q &= \mathrm{softmax}\left(\left[X_a, X_q^a\right] W_q + \tilde{X}_a W_a\right) \\ \tilde{X}_q^a &= \gamma_q \odot X_q^a \end{aligned} \tag{4}$$

where W_q and W_a are both trainable transformation matrices to map the different embedding features into a common space, $[\cdot,\cdot]$ means the concatenation for each position (token). We accomplish the final reasoning step over the QA pair using another Bi-LSTM to attain the interplay representation:

$$\mathcal{T}^{qa} = \text{Bi-LSTM}^r\left(\left[X_a, \tilde{X}_q^a\right]\right) \tag{5}$$

where $\mathcal{T}^{qa} \in \mathbb{R}^{|a|\times d}$. We concatenate $\mathcal{T}^{qa}$ along with the representations of QA pair for each position (token) and the resulting sequence is max-pooled to get the final question-to-answer guided representation as $\bar{\mathcal{T}}^{qa} = \text{max-pooling}\left(\left[\mathcal{T}^{qa}, X_a, \tilde{X}_q^a\right]\right)$.

C. Evidence Coherence Modeling

Previous literature has generally found that the truth of any (true) proposition consists in its coherence with some specified set of propositions [12]. This module (ECM) fully exploits the reliability of evidence to enhance the coherence modeling from trust and credibility perspectives [35] for debunking or confirming the factuality in the answer. First of all, to prevent off-topic coherence which deviates from the question's focus, we represent each evidence s_i as $X_{s_i}^q = \mathrm{f}(X_{s_i}, X_q)$, where the function $\mathrm{f}(\cdot)$ is a shorthand of Eq. 2.

Trust Cross-Checking. For measuring the trust that is the relative likelihood for one evidence sentence being coherent with another one, we utilize self-attention mechanism to learn dependencies and semantics between any two evidence sentences. We obtain the sentence-level representation $\bar{X}_{s_i}$ for s_i by max-pooling the involved word vectors, i.e., $X_{s_i}^q$. Therefore, we define the query, key and value as $\{Q, K, V\} = \{\bar{X}_S \cdot W_Q, \bar{X}_S \cdot W_K, \bar{X}_S \cdot W_V\}$, where $\bar{X}_S = \left[\bar{X}_{s_1}, \ldots, \bar{X}_{s_n}\right]^\top \in \mathbb{R}^{n\times d}$ denotes the representations of all evidence sentences, $\{W_Q, W_K, W_V\} \in \mathbb{R}^{d\times d_k}$ are trainable weights. Then, attention functions are applied to generate the output states:

$$O = \mathrm{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V \tag{6}$$

In consistent with the setting of standard transformer encoder [36], we conduct multi-head self-attention and concatenate the vectors to generate the final output, followed by a normalization layer to represent all evidence sentences as $\bar{\mathbf{O}} = \left[\bar{O}_1, \bar{O}_2, \ldots, \bar{O}_n\right] \in \mathbb{R}^{n\times d}$.

Credibility Ranking. The credibility targets to measure the consistency of each evidence sentence regarding the entire set as a whole. We assume that the conflicting and entailed semantics between answer and evidence could respectively contribute to discovering false and true answers. Inspired by the gate mechanism [37], we propose a gate-induced decoupling mechanism to identify the features of conflicting and entailed text snippets as follows:

$$\begin{aligned} dis_i &= \mathrm{sigmoid}\left(\left[\bar{X}_a, \bar{X}_{s_i}\right] W_d\right) \\ sim_i &= \mathrm{sigmoid}\left(\left[\bar{X}_a, \bar{X}_{s_i}\right] W_s\right) \end{aligned} \tag{7}$$

where dis_i is the discrepancy gate, sim_i is the similarity gate, and $\bar{X}_a$ is the sentence-level representation of the answer a by max-pooling word vectors obtained from $\mathrm{f}(X_a, X_q)$. On top of the gates, we obtain the features of conflicting and entailed text snippets from each answer-evidence pair as:

$$\begin{aligned} f_i^d &= \tanh\left(\left[dis_i \odot \bar{X}_a, (1 - dis_i) \odot \bar{X}_{s_i}\right] W_c\right) \\ f_i^s &= \tanh\left(\left[sim_i \odot \bar{X}_a, sim_i \odot \bar{X}_{s_i}\right] W_r\right) \end{aligned} \tag{8}$$

where W_* in Eq. 7-8 are trainable parameters. Based on the decoupling features of conflict and entailment, i.e., f_i^d and f_i^s, we generate a probability for each s_i indicating the coherence with the answer:

$$\begin{aligned} \alpha_i &= \tanh\left(\left[f_i^d, f_i^s\right] W_g\right) \\ \beta_i &= \frac{\exp(\alpha_i)}{\sum_i \exp(\alpha_i)} \end{aligned} \tag{9}$$

where W_g is a parameter turning the gate-induced decoupling features to a credibility score α_i. We omit the bias to avoid notation clutter. β_i is the normalized credibility weight of s_i towards the answer. To highlight informative evidence

for the answer verdict, we apply the normalized weights $\beta = [\beta_1, ..., \beta_n]$ over the middle evidence representations $\bar{\mathbf{O}}$ to produce the weighted evidence representations: $T^{eqa} = \beta \odot \bar{\mathbf{O}}$. Then it would be passed through a feed-forward and a normalization layer to get the final evidence representations $\mathcal{T}^{eqa} \in \mathbb{R}^{n \times d}$. Here we employ max-pooling to jointly capture the coherent opinions expressed in the whole evidence as $\bar{\mathcal{T}}^{eqa} = \text{max-pooling}\left(\mathcal{T}^{eqa}\right)$.

D. Metadata Fusion

Apart from the textual features, we assume the metadata field (e.g., question/answer time, product domain, and answer text length) implicitly improve the answer veracity prediction. Intuitively, the truthful answers in E-commerce are usually generated by the users who are familiar with the product, obtained from metadata such as question/answer time [1] and different product domains [23]. The core idea is to capture the hidden correlations among the metadata features, which are generally difficult to be clearly defined and hand-crafted (e.g., the classic association rules "diapers and beer" [38] is mined from the data, instead of pre-defined by experts). Inspired by [16], we utilize a DeepFM module to learn both low- and high-order features from the metadata which consists of two components, i.e., factorization machine (FM) and deep neural network (DNN). We first input each field $m_i \in M$ into a sparse input layer and embed sparse original features into a low-dimensional embedding V_i. The FM part calculates the dot product of each pair of features to capture the hidden correlations:

$$Z_{FM} = \sum_{i=1}^{N} \sum_{j=i+1}^{N} \langle V_i, V_j \rangle m_i \cdot m_j \tag{10}$$

where N means the total number of metadata fields for each instance $\mathcal{C}$. The DNN part obtains the high-order feature interactions through an MLP layer:

$$Z_{DNN} = \text{MLP}\left([V_1, V_2, ..., V_N]\right) \tag{11}$$

Then we get the features from metadata:$Z_{DeepFM} = [Z_{FM}, Z_{DNN}]$, to combine the low-order features with high-order features, which are complementary due to the shared embeddings.

E. The Overall Model

For our base model *AMIF-Base*, we integrate $\bar{\mathcal{T}}^{qa}$ with $\bar{\mathcal{T}}^{eqa}$ to predict the probability distribution over the veracity classes:

$$\begin{aligned} Z &= \text{MLP}\left(\left[\bar{\mathcal{T}}^{qa}, \bar{\mathcal{T}}^{eqa}\right]\right) \\ p &= \text{softmax}(Z) \end{aligned} \tag{12}$$

where p is a low-dimensional vector for answer veracity prediction.

To integrate the metadata for *AMIF-DeepFM*, we employ a fusion strategy with adaptive adjustability. For conciseness, the low-dimensional veracity prediction vector is obtained with the final fusion for the multiple relevant features as:

$$p = \text{softmax}\left(\text{MLP}\left(\left[\lambda Z, (1-\lambda) Z_{DeepFM}\right]\right)\right) \tag{13}$$

where $\lambda \in (0, 1)$ is the trade-off coefficient randomly initialized.

TABLE III
SUMMARY STATISTICS OF THE ANSWERFACT DATASET

	Electronics	Home	Sports	Health	Phones	Total
# Answers per Label						
TRUE	13,054	10,592	4,539	6,879	2,467	37,531
PARTTRUE	1,737	1,297	581	1,035	336	4,986
UNSURE	3,116	2,228	1,134	1,782	738	8,998
PARTFALSE	822	683	308	564	151	2,528
FALSE	2,491	1,797	897	1,211	415	6,821
#Answers	21,220	16,597	7,459	11,481	4,107	60,864
#Questions	11,554	8,210	3,918	5,816	2,245	31,743

During training, we exploit the cross-entropy loss over training data with the L2-norm. Since the focus in this paper is primarily on better fusion strategy on multiple features from the question, answer, and supporting evidence plus metadata to improve the FC-QA task, we represent input words using GloVe word embeddings [39]. We set hidden dimension d to 256, head number H to 8, evidence sentences number n to 5. Parameters are updated through back-propagation [40] with the AdamW optimizer [41]. The learning rate is initialized as 0.01, and the dropout rate is 0.2. Early stopping [42] is applied to avoid overfitting.

IV. EXPERIMENTS

A. Dataset

We carry out extensive experiments on AnswerFact [11] benchmark, the representative dataset for FC-QA task so far*, with 60,864 QA pairs in total. We consider the label settings from two perspectives: 1) Finer-grained labels: TRUE, PARTTRUE, UNSURE, PARTFALSE and FALSE, which lead to a more challenging classification problem [29]; 2) Following [34], we merge PARTTRUE, UNSURE and PARTFALSE into MIXED, resulting of a more practical classification on AnswerFact, i.e., TRUE, FALSE and MIXED. The dataset is challenging because of the QA-claim setting, multiple product domains and product-related auxiliary information to be exploited. The statistics of the dataset are shown in Table III.

B. Experimental Setup

We compare our proposed model with the following baseline and state-of-the-art models: 1) **CNN-claim** and 2) **LSTM-claim**: The CNN-based detection model [30] and LSTM-based RNN model for representation learning from word sequences [29], respectively, both using only claim content without considering external resources; 3) **DeClarE**: The evidence-based fact-checking model [8] with word-level neural attention to capture world-level evidence from relevant articles; 4) **NSMN**: A pipeline method [9] based on Neural Semantic Matching Network (NSMN), which ranked first in the FEVER shared task [43]. Its single-text claim verification module is used here for our task as one of the representative baselines. 5) **MultiFC**: A joint model for evidence ranking and veracity prediction conduted by [10]. 6) **AVER**: the state-of-the-art

*We didn't evaluate the SemEval-2019 Task 8 [7] due to its limited training data with only 495 QA pairs.

TABLE IV
RESULTS OF ANSWER VERACITY PREDICTION ON ANSWERFACT. F_{TRUE}, F_{MIXED} AND F_{FALSE} DENOTE THE F1 SCORES OF THE THREE CLASSES.

Model	3-CLASS					5-CLASS	
	Mac-F1	Mic-F1	F_{TRUE}	F_{MIXED}	F_{FALSE}	Mac-F1	Mic-F1
CNN-claim	0.442	0.648	0.791	0.144	0.390	0.249	0.649
LSTM-claim	0.492	0.649	0.785	0.302	0.390	0.253	0.653
DeClarE	0.450	0.635	0.785	0.153	0.413	0.243	0.635
NSMN	0.504	0.663	0.799	0.284	0.429	0.279	0.651
MultiFC	0.513	0.655	0.787	0.300	0.453	0.299	0.655
AVER	0.534	0.673	0.802	0.314	0.486	0.330	0.665
AVER*	0.515	0.660	0.790	0.322	0.427	0.306	0.650
AMIF-Base	0.548	0.678	0.802	0.352	**0.493**	**0.342**	0.666
AMIF-DeepFM	**0.551**	**0.705**	**0.817**	**0.353**	0.482	0.340	**0.672**

TABLE V
ABLATION STUDIES ON OUR PROPOSED MODEL.

Model	3-CLASS		5-CLASS	
	Mac-F1	Mic-F1	Mac-F1	Mic-F1
AMIF	0.551	0.705	0.340	0.672
QARC+ECM(w/o DeepFM)	0.548	0.678	0.342	0.666
QARC+DeepFM(w/o ECM)	0.538	0.658	0.324	0.663
Vanilla QA+ECM+DeepFM	0.543	0.674	0.336	0.665
Vanilla QA+ECM	0.538	0.667	0.334	0.664
Vanilla QA	0.505	0.618	0.257	0.631

TABLE VI
COMPARISON OF DIFFERENT PARTS IN ECM.

Model	3-CLASS		5-CLASS	
	Mac-F1	Mic-F1	Mac-F1	Mic-F1
AMIF	0.551	0.705	0.340	0.672
w/o discrepancy	0.543	0.671	0.333	0.670
w/o similarity	0.545	0.679	0.336	0.674
w/o decoupling	0.540	0.665	0.326	0.669
w/o self-attention	0.548	0.678	0.328	0.667

neural model for fact checking in QA settings [11] with a tailored evidence ranking module.

For our proposed model AMIF, we consider the following two variants: **AMIF-Base**: without DeepFM module. **AMIF-DeepFM**: with DeepFM module involved.

C. Implementation Details

As known, BERT [44] is not a ready-to-use model to generate embeddings in its original form for specific tasks on E-commerce CQA platforms, so it is rather a model that can be tuned for a task. However, considering the relatively small scale of the benchmark dataset for FC-QA task (60,864 QA pairs in total where a lot of identical questions are paired with different answers), we represent input words using 300-dim pre-trained GloVe [39] Wikipedia 6B word embeddings widely used in the state-of-the-art baselines, for a fair comparison. We hold out 10% of the datasets for tuning the hyperparameters and conduct 10-fold cross-validation on the rest of the datasets. We use micro-/macro-averaged F1, class-specific F-measure as evaluation metrics, where macro-averaged F1 can capture competitive performance beyond the majority class for AnswerFact owing to the imbalanced class prevalence. We implement our model with pytorch.

D. Answer Truthfulness Prediction

Table IV demonstrates the performance of all the compared methods respectively based on two label settings of the benchmark. As AVER is not yet open-source, we post both the results referred from AVER and those implemented by ourselves (*). It is observed that the performances of the baselines in the first group only relying on claim text are obviously poor. Since the inductive bias of most existing claim verification models could prefer single-text claims to QA-setting claims, they largely ignored or oversimplified the feature interactions between the question context and evidence sentences, whose application in this real-world E-commerce scenario remains to be explored. We just compared DeClareE, NSMN and MultiFC because they are all classical and effective representative of single-text claim verification modules that can be applied to this answer truthfulness prediction task flexibly, though not designed for this question-aware task. AVER performs best among all evidence-based models in the second group because it utilizes the shallow features from QA text to guide the interpolation of the evidence sentences instead of the single-text claim used in the other methods. However, AMIF-Base can already achieve better results compared with other baselines, which demonstrates the effectiveness of considering the fine-granularity understanding between QA pairs and the necessity of modeling coherence of evidence from both trust and credibility perspectives. We can also notice that our proposed AMIF-DeepFM consistently outperforms all baselines and further improves the prediction performance of AMIF-Base in general, suggesting that equipping the network with features from metadata in E-commerce can provide positive guidance for more accurate truthfulness predictions.

E. Ablation Study

We perform ablation studies by discarding some important components of AMIF. As demonstrated in Table V, the models suffer different degrees of such performance degradation by discarding some important components of AMIF, indicating the effectiveness of our proposed components for predicting the veracity of answers. AMIF makes mild improvements over our base model discarding the DeepFM module (QARC+ECM), reflecting the promoting role of DeepFM for fact checking in CQA forums, which excavates appropriate low- and high-order interactions from raw features. Neglecting the evidence coherence modeling (QARC+DeepFM) also leads to performance degradation, which implies exploring the coherence of evidence from trust and credibility perspectives enables our model hardly compromised when not all evidence is reliable and contributes to the final answer veracity prediction performance. We also replace the QA reasoning calibration component with the naive QA representation used in AVER [11] and it (Vanilla QA+ECM+DeepFM) results in performance degradation on both label settings, since the performance may suffer from limited context information of answers caused by the insufficient understanding over the interplay of QA pairs. Substituting the evidence ranking module in AVER by our ECM (Vanilla QA+ECM) still achieves superior performance, which reaffirms the strengths of the ECM component.

For details in ECM, we also conduct an ablation study as shown in Table VI. We can observe that ablating any part of ECM could decrease the performance, which further confirms the effectiveness of the self-attention and gate-induced decoupling mechanisms in ECM. An interesting point is that our model without discrepancy gate obtains an inferior performance compared with that without similarity gate. For that, we investigate the importance of discrepancy and similarity gates to the final prediction in the following subsection.

F. Analysis of Gate-induced Decoupling

We design an experiment to make deeper analysis on the importance of differential and consistent features inherent in semantic correlation decoupled from the holistic context by discrepancy and similarity gates, respectively. Figure 4 shows Macro-averaged F1 scores of two test sets we choose to verify the impact of our proposed gate-induced decoupling. We choose two sets of test samples from the whole original test set: the test samples with TRUE label and PARTTRUE label are merged into one set, and those with FALSE label and PARTFALSE label are merged into another set. We can observe that ablating similarity or discrepancy gate could decrease the performance. Combining them makes further improvements and implies their complementary. However, the model without similarity gate achieves a little worse performance on the positive test set with fact (TRUE+PARTTRUE) compared with the model without discrepancy gate. We conjecture that the reason why the difference is not obvious is that the proportion of answers with fact (TRUE+PARTTRUE) in the training set is much larger than that with misinformation (FALSE + PARTFALSE), so the variant models tend to fit the positive training data in the training process, resulting in performance on the test data can not reflect a significant difference. Nevertheless, the model without discrepancy gate leads to a relatively large margin performance degradation on the negative test set (FALSE+PARTFALSE) compared with the model without similarity gate, which explicitly explains that it is effective to disentangle the semantic conflicts between the answer and evidence for the identification of the answer with misinformation, to which discrepancy gate contributes more compared with similarity gate.

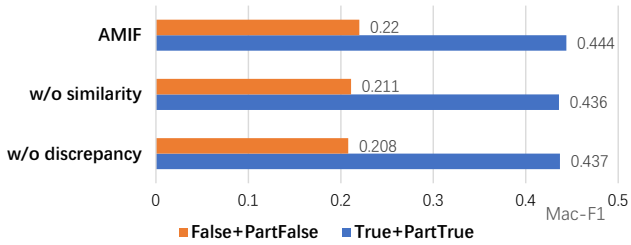


Fig. 4. Comparison between discrepancy and similarity gates of gate-induced decoupling mechanism.

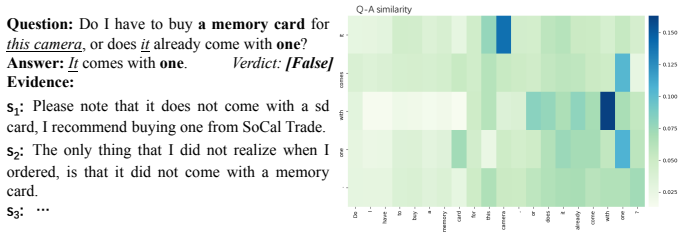


Fig. 5. Example of correctly predicted false answers.

G. Case study

One key advantage of our model is that we could retrieve both token-level and sentence-level explanations for our predictions, as presented in Figure 5. In the simple example, one interesting phenomenon is that the question contains two sub-questions. However, our model can provide powerful clues that the answer targets the second sub-question through reasoning over the heatmap of the QA similarity after proper fine-grained semantic calibration between QA pairs, which includes resolving the referent 'it', 'one' and understanding which salient units in the question should be taken into account. Also, drawing on the practice of [45], we top the indicative evidence sentences based on the comprehensive consideration of both self-attention weights and credibility scores. The efficient FC-QA model hinges on the reliability and helpfulness of the answer [1], [23], and explanations (e.g., coherent opinions) can strengthen this by communicating fidelity in predictive models and assist users for better understanding the predictions.

V. CONCLUSION

We propose a novel attention-based hybrid deep framework AMIF to predict the answer truthfulness for community question answering in E-commerce, which not only supports fine-grained semantic reasoning calibration between QA pairs but also models the opinion coherence between the answer and evidence from trust and credibility perspectives. Moreover,

features in metadata that describe E-commerce QA are incorporated for better prediction. Extensive experiments show that our model achieves superior and explainable performance on the task of fact checking for question answering in E-commerce. In our future work, we plan to dive into the research for applying more single-text claim verification models to our FC-QA task.

REFERENCES

[1] Tsvetomila Mihaylova, Preslav Nakov, Lluís Màrquez, Alberto Barrón-Cedeño, Mitra Mohtarami, Georgi Karadzhov, and James Glass, "Fact checking in community forums," in *AAAI*, 2018.

[2] David Carmel, Liane Lewin-Eytan, and Yoelle Maarek, "Product question answering using customer generated content-research challenges," in *SIGIR*, 2018.

[3] Daisuke Ishikawa, Tetsuya Sakai, and Noriko Kando, "Overview of the ntcir-8 community qa pilot task (part i): The test collection and the task.," in *NTCIR*, 2010.

[4] Preslav Nakov, Lluís Màrquez, Walid Magdy, Alessandro Moschitti, Jim Glass, and Bilal Randeree, "Semeval-2015 task 3: Answer selection in community question answering," in *Proceedings of the 9th International Workshop on Semantic Evaluation*, 2015.

[5] Preslav Nakov, Lluís Màrquez, Alessandro Moschitti, Walid Magdy, Hamdy Mubarak, Abed Alhakim Freihat, James Glass, and Bilal Randeree, "Semeval-2017 task 3: Community question answering," in *Proceedings of the 10th International Workshop on Semantic Evaluation*, 2016.

[6] Preslav Nakov, Doris Hoogeveen, Lluís Màrquez, Alessandro Moschitti, Hamdy Mubarak, Timothy Baldwin, and Karin Verspoor, "Semeval-2017 task 3: Community question answering," in *Proceedings of the 11th International Workshop on Semantic Evaluation*, 2017.

[7] Tsvetomila Mihaylova, Georgi Karadzhov, Pepa Atanasova, Ramy Baly, Mitra Mohtarami, and Preslav Nakov, "Semeval-2019 task 8: Fact checking in community question answering forums," in *Proceedings of the 13th International Workshop on Semantic Evaluation*, 2019.

[8] Kashyap Popat, Subhabrata Mukherjee, Andrew Yates, and Gerhard Weikum, "Declare: Debunking fake news and false claims using evidence-aware deep learning," in *EMNLP*, 2018.

[9] Yixin Nie, Haonan Chen, and Mohit Bansal, "Combining fact extraction and verification with neural semantic matching networks," in *AAAI*, 2019.

[10] Isabelle Augenstein, Christina Lioma, Dongsheng Wang, Lucas Chaves Lima, Casper Hansen, Christian Hansen, and Jakob Grue Simonsen, "Multifc: A real-world multi-domain dataset for evidence-based fact checking of claims," in *EMNLP-IJCNLP*, 2019.

[11] Wenxuan Zhang, Yang Deng, Jing Ma, and Wai Lam, "Answerfact: Fact checking in product question answering," in *EMNLP*, 2020.

[12] James O. Young, "The Coherence Theory of Truth," in *The Stanford Encyclopedia of Philosophy*, Edward N. Zalta, Ed. Metaphysics Research Lab, Stanford University, fall 2018 edition, 2018.

[13] Hongzhan Lin, Jing Ma, Mingfei Cheng, Zhiwei Yang, Liangliang Chen, and Guang Chen, "Rumor detection on twitter with claim-guided hierarchical graph attention networks," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021, pp. 10035–10047.

[14] Nikos Manouselis and Constantina Costopoulou, "Quality in metadata: a schema for e-commerce," *Online Information Review*, 2006.

[15] Suhail Ansari, Ron Kohavi, Llew Mason, and Zijian Zheng, "Integrating e-commerce and data mining: Architecture and challenges," in *Proceedings 2001 IEEE International Conference on Data Mining*. IEEE, 2001.

[16] Huifeng Guo, Ruiming Tang, Yunming Ye, Zhenguo Li, and Xiuqiang He, "Deepfm: A factorization-machine based neural network for ctr prediction," 2017.

[17] Ivan Srba and Maria Bielikova, "A comprehensive survey and classification of approaches for community question answering," *ACM Transactions on the Web (TWEB)*, 2016.

[18] Guangyou Zhou, Tingting He, Jun Zhao, and Po Hu, "Learning continuous word embedding with metadata for question retrieval in community question answering," in *ACL-IJCNLP*, 2015.

[19] Yi Tay, Minh C Phan, Luu Anh Tuan, and Siu Cheung Hui, "Learning to rank question answer pairs with holographic dual lstm architecture," in *SIGIR*, 2017.

[20] Xiao Yang, Madian Khabsa, Miaosen Wang, Wei Wang, Ahmed Hassan Awadallah, Daniel Kifer, and C Lee Giles, "Adversarial training for community question answer selection based on multi-scale matching," in *AAAI*, 2019.

[21] Yang Deng, Wai Lam, Yuexiang Xie, Daoyuan Chen, Yaliang Li, Min Yang, and Ying Shen, "Joint learning of answer selection and answer summary generation in community question answering," in *AAAI*, 2020.

[22] Wenxuan Zhang, Yang Deng, and Wai Lam, "Answer ranking for product-related questions via multiple semantic relations modeling," in *SIGIR*, 2020.

[23] Wenxuan Zhang, Wai Lam, Yang Deng, and Jing Ma, "Review-guided helpful answer identification in e-commerce," in *Proceedings of The Web Conference 2020*, 2020.

[24] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu, "Fake news detection on social media: A data mining perspective," *ACM SIGKDD explorations newsletter*, 2017.

[25] Arkaitz Zubiaga, Ahmet Aker, Kalina Bontcheva, Maria Liakata, and Rob Procter, "Detection and resolution of rumours in social media: A survey," *ACM Computing Surveys (CSUR)*, 2018.

[26] Srijan Kumar and Neil Shah, "False information on web and social media: A survey," *arXiv preprint arXiv:1804.08559*, 2018.

[27] Karishma Sharma, Feng Qian, He Jiang, Natali Ruchansky, Ming Zhang, and Yan Liu, "Combating fake news: A survey on identification and mitigation techniques," *ACM Transactions on Intelligent Systems and Technology*, 2019.

[28] Neema Kotonya and Francesca Toni, "Explainable automated fact-checking: A survey," in *COLING*, 2020.

[29] Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi, "Truth of varying shades: Analyzing language in fake news and political fact-checking," in *EMNLP*, 2017.

[30] William Yang Wang, ""liar, liar pants on fire": A new benchmark dataset for fake news detection," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, 2017.

[31] James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal, "Fever: a large-scale dataset for fact extraction and verification," in *NAACL-HLT (1)*, 2018.

[32] Canasai Kruengkrai, Junichi Yamagishi, and Xin Wang, "A multi-level attention model for evidence-based fact checking," *arXiv preprint arXiv:2106.00950*, 2021.

[33] Xuezhi Wang, Cong Yu, Simon Baumgartner, and Flip Korn, "Relevant document discovery for fact-checking articles," in *Companion Proceedings of the The Web Conference 2018*, 2018.

[34] Jing Ma, Wei Gao, Shafiq Joty, and Kam-Fai Wong, "Sentence-level evidence embedding for claim verification with hierarchical attention networks," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019.

[35] Ortwin Renn and Debra Levine, "Credibility and trust in risk communication," in *Communicating risks to the public*. Springer, 1991.

[36] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *NIPS*, 2017.

[37] Sepp Hochreiter and Jürgen Schmidhuber, "Long short-term memory," *Neural computation*, 1997.

[38] Dick Tsur, Jeffrey D Ullman, Serge Abiteboul, Chris Clifton, Rajeev Motwani, Svetlozar Nestorov, and Arnon Rosenthal, "Query flocks: A generalization of association-rule mining," *Acm sigmod record*, 1998.

[39] Jeffrey Pennington, Richard Socher, and Christopher D Manning, "Glove: Global vectors for word representation," in *EMNLP*, 2014.

[40] Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa, "Natural language processing (almost) from scratch," *Journal of machine learning research*, 2011.

[41] Ilya Loshchilov and Frank Hutter, "Fixing weight decay regularization in adam," 2018.

[42] Yuan Yao, Lorenzo Rosasco, and Andrea Caponnetto, "On early stopping in gradient descent learning," *Constructive Approximation*, 2007.

[43] James Thorne, Andreas Vlachos, Oana Cocarascu, Christos Christodoulopoulos, and Arpit Mittal, "The fact extraction and verification (fever) shared task," in *FEVER*, 2018.

[44] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *NAACL-HLT (1)*, 2019.

[45] Ling Min Serena Khoo, Hai Leong Chieu, Zhong Qian, and Jing Jiang, "Interpretable rumor detection in microblogs by attending to user interactions," in *AAAI*, 2020.