The Thirty-Second AAAI Conference
on Artificial Intelligence (AAAI-18)

Fact Checking in Community Forums

**Tsvetomila Mihaylova,[1] Preslav Nakov,[2] Lluís Màrquez,[2] Alberto Barrón-Cedeño,[2]
Mitra Mohtarami,[3] Georgi Karadzhov,[1] James Glass [3]**
[1]Sofia University "St. Kliment Ohridski", Sofia, Bulgaria
[2]Qatar Computing Research Institute, Hamad bin Khalifa University, Doha, Qatar
[3]Massachusetts Institute of Technology, Cambridge, MA, USA
tsvetomila.mihaylova@gmail.com, {pnakov, lmarquez, albarron}@hbku.edu.qa,
mitra@csail.mit.edu, georgi.m.karadjov@gmail.com, glass@mit.edu

Abstract

Community Question Answering (cQA) forums are very popular nowadays, as they represent effective means for communities around particular topics to share information. Unfortunately, this information is not always factual. Thus, here we explore a new dimension in the context of cQA, which has been ignored so far: checking the veracity of answers to particular questions in cQA forums. As this is a new problem, we create a specialized dataset for it. We further propose a novel multi-faceted model, which captures information from the answer content (*what is said and how*), from the author profile (*who says it*), from the rest of the community forum (*where it is said*), and from external authoritative sources of information (*external support*). Evaluation results show a MAP value of 86.54, which is 21 points absolute above the baseline.

Introduction

Community Question Answering (cQA) forums such as StackOverflow, Yahoo! Answers, and Quora are very popular nowadays, as they represent effective means for communities around particular topics to share information and to collectively satisfy their information needs. However, the information being shared is not always factual. There are multiple factors explaining the presence of incorrect answers in cQA forums, e.g., misunderstanding, ignorance, or maliciousness of the responder. This is exacerbated by the fact that most cQA forums are barely moderated and lack systematic quality control. Moreover, in our dynamic world of today, truth is often time-sensitive: what was true yesterday may become false today.

We explore a new dimension in the context of cQA: checking the veracity of answers to a given question. This aspect has been ignored so far, e.g., in recent cQA tasks at NTCIR and SemEval (Ishikawa, Sakai, and Kando 2010; Nakov et al. 2015; 2016; 2017a), where an answer is considered as GOOD if it tries to address the question, irrespective of its veracity. Yet, veracity is an important aspect, as high-quality automatic fact checking can offer better user experience for cQA systems. For instance, the user could be presented with veracity scores, where low scores would warn him/her not to completely trust the answer or to double-check it.

q: I have heard its not possible to extend visit visa more than 6 months? Can U please answer me.. Thankzzz...

a_1: Maximum period is 9 Months....

a_2: 6 months maximum

a_3: This has been anwered in QL so many times. Please do search for information regarding this. BTW answer is 6 months.

Figure 1: Example from the Qatar Living forum.

Figure 1 presents an excerpt of an example from the Qatar Living forum, with one question and three answers selected from a longer thread. According to SemEval-2017 Task 3 (Nakov et al. 2017a), all three answers are good since they address the question q. Nevertheless, a_1 contains false information, while a_2 and a_3 are true,[1] as can be checked on an official governmental website.[2]

Determining the veracity of a claim is a very difficult problem, and solving it in full would require language understanding and inference, integration of several sources of information, and world knowledge, among other things. Here, we approach it as a supervised classification task, and we propose a novel model based on multi-faceted modeling of the facts, which integrates knowledge from several complementary sources, such as the answer content (*what is said and how*), the author profile (*who says it*), the rest of the community forum (*where it is said*), and external authoritative sources of information (*external support*).

The main contributions of this paper are as follows: (*i*) first, we are the first to study factuality in cQA, and we create a new high-quality dataset —CQA-QL-2016-fact—, which we release to the research community;[3] to the best of our knowledge, this is the first publicly-available dataset

[1]One could also guess that a_2 and a_3 are more likely to be true from the fact that the *6 months* answer appears many times in the thread, as well as in other threads. While these observations serve the basis for useful features for classification, the real verification for a gold standard annotation requires finding support from a credible external source.

[2]https://www.moi.gov.qa/site/english/departments/PassportDept/news/2011/01/03/23385.html

[3]The dataset and the source code are available online at https://github.com/qcri/QLFactChecking

specifically targeting factuality in a cQA setting; (*ii*) we approach the problem of fact-checking using a multi-faceted model based on a rich input representation, including new features that have not been compared in such a configuration before; (*iii*) this rich representation allows us to obtain strong results that are applicable to supporting in practice the application scenario outlined above; and (*iv*) we perform a qualitative analysis of what works well and what does not.

Related Work

To the best of our knowledge, no previous work has targeted fact-checking of answers in the context of community Question Answering. Yet, there has been work on *credibility* assessment in cQA (Nakov et al. 2017b). However, *credibility* is different from *veracity* (our focus here) as it is a subjective perception about whether a statement is credible, rather than verifying it as true/false as a matter of fact.

In the context of general QA, there has been work on *credibility* assessment which has been only modeled at the feature level, with the goal of improving GOOD answer identification. For example, Jurczyk and Agichtein (2007) modeled author authority using link analysis, while Agichtein et al. (2008) used PageRank and HITS in addition to intrinsic content quality (e.g., punctuation and typos, syntactic and semantic complexity, and grammaticality), and usage analysis (e.g., number of clicks and dwell time).

In (Lita et al. 2005) the focus was on source credibility, sentiment analysis, and answer contradiction compared to other answers, while in (Su, yun Chen, and Huang 2010) the emphasis was on verbs and adjectives that cast doubt. Other authors used language modeling to validate the reliability of an answer's source (Banerjee and Han 2009) or focused on non-textual features such as click counts, answer activity level, and copy counts (Jeon et al. 2006). There has been also work on curating social media content using syntactic, semantic, and social signals (Pelleg et al. 2016). Unlike this research, we (*i*) target factuality rather than credibility, (*ii*) address it as a task in its own right, (*iii*) use a specialized dataset, and (*iv*) use a much richer text representation.

Information credibility, fact-checking and rumor detection have been also studied in the area of social computing. Castillo, Mendoza, and Poblete (2011) used user reputation, author writing style, and various time-based features. Canini, Suh, and Pirolli (2011) analyzed the interaction of content and social network structure and Morris et al. (2012) and Zubiaga et al. (2016) studied how people handle rumors in social media. Lukasik, Cohn, and Bontcheva (2015) used temporal patterns to detect rumors and to predict their frequency. Ma et al. (2015) further used recurrent neural networks, and Zubiaga et al. (2016) focused on conversational threads. Other authors have gone beyond social media and have been querying the Web to gather support for accepting or refuting a claim (Popat et al. 2016). Finally, there has been also work on studying credibility, trust, and expertise in news communities (Mukherjee and Weikum 2015). However, none of this work was about QA or cQA.

CQA-QL-2016-fact: A Dataset for Fact Checking in cQA

As we have a new problem —fact-checking of answers in the context of cQA—, for which no dataset exists, we had to create our own one. We chose to augment with factuality annotations a pre-existing dataset for cQA, which allows us to stress the difference between (a) distinguishing a GOOD vs. a BAD answer, and (b) distinguishing between a factually-true vs. a factually-false one. In particular, we added annotations for factuality to the CQA-QL-2016 dataset from SemEval-2016 Task 3[4] on Community Question Answering.

In CQA-QA-2016, the data is organized in question–answer threads from the Qatar Living forum.[5] Each question has a subject, a body, and meta information: ID, category (e.g., *Computers and Internet*, *Education*, and *Moving to Qatar*), date and time of posting, user name and ID.

We selected for annotation only the factual questions such as "*What is Ooredoo customer service number?*" In particular, we filtered out all (*i*) socializing, e.g., "*What was your first car?*", (*ii*) requests for opinion/advice/guidance, e.g., "*Which is the best bank around??*", and (*iii*) questions containing multiple sub-questions, e.g., "*Is there a land route from Doha to Abudhabi. If yes; how is the road and how long is the journey?*"

Next, we annotated for veracity the answers to the questions that we retained in the previous step. In CQA-QA-2016, each answer has a subject, a body, meta information (answer ID, user name and ID), and a judgment about how well it answers the question of its thread: GOOD vs. BAD vs. POTENTIALLY USEFUL . We only annotated the GOOD answers, using the following labels:

FACTUAL - TRUE: The answer is True and this can be manually verified using a trusted external resource. (*Q: "I wanted to know if there were any specific shots and vaccinations I should get before coming over [to Doha]."; A: "Yes there are; though it varies depending on which country you come from. In the UK; the doctor has a list of all countries and the vaccinations needed for each."*).[6]

FACTUAL - FALSE: The answer gives a factual response, but it is false. (*Q: "Can I bring my pitbulls to Qatar?", A: "Yes you can bring it but be careful this kind of dog is very dangerous."*).[7]

FACTUAL - PARTIALLY TRUE: We could only verify part of the answer. (*Q: "I will be relocating from the UK to Qatar [...] is there a league or TT clubs / nights in Doha?", A: "Visit Qatar Bowling Center during thursday and friday and you'll find people playing TT there."*).[8]

FACTUAL - CONDITIONALLY TRUE: The answer is True in some cases, and False in others, depending on some con-

[4] http://alt.qcri.org/semeval2016/task3/

[5] http://www.qatarliving.com/forum

[6] This can be verified: https://wwwnc.cdc.gov/travel/destinations/traveler/none/qatar

[7] The answer is not true because pitbulls are included in the list of breeds that are banned in Qatar: http://canvethospital.com/pet-relocation/banned-dog-breed-list-qatar-2015/

[8] The place has table tennis, but we do not know on which days: https://www.qatarbowlingfederation.com/bowling-center/

	Label	Answers
+	FACTUAL - TRUE	128
−	FACTUAL - FALSE	22
−	FACTUAL - PARTIALLY TRUE	38
−	FACTUAL - CONDITIONALLY TRUE	16
−	FACTUAL - RESPONDER UNSURE	26
−	NONFACTUAL	19
	+ POSITIVE	**128**
	− NEGATIVE	**121**
	TOTAL	**249**

Table 1: Distribution of the answer labels in the CQA-QL-2016-fact dataset.

ditions that the answer does not mention. (*Q: "My wife does not have NOC from Qatar Airways; but we are married now so can i bring her legally on my family visa as her husband?", A: "Yes you can."*).[9]

FACTUAL - RESPONDER UNSURE: The person giving the answer is not sure about the veracity of his/her statement. (e.g., "*Possible only if government employed. That's what I heard.*")

NONFACTUAL: The answer is not factual. It could be an opinion, advice, etc. that cannot be verified. (e.g., "*Its better to buy a new one.*")

We further discarded answers whose factuality was very time-sensitive (e.g., "*It is Friday tomorrow.*", "*It was raining last week.*")[10], or for which the annotators were unsure.

We considered all questions from the DEV and the TEST partitions of the CQA-QA-2016 dataset. We targeted very high quality, and thus we did not use crowdsourcing for the annotation, as pilot annotations showed that the task was very difficult and that it was not possible to guarantee that *Turkers* would do all the necessary verification; e.g., gather evidence from trusted sources. Instead, all examples were first annotated independently by four annotators, and then they discussed *each example* in detail to come up with a final consensus label. We ended up with 249 GOOD answers[11] to 71 different questions, which we annotated for factuality: 128 POSITIVE and 121 NEGATIVE examples. See Table 1 for more detail.

Modeling Facts

We use a multi-faceted model, based on a rich input representation that models (*i*) the user profile, (*ii*) the language used in the answer, (*iii*) the context in which the answer is located, and (*iv*) external sources of information.

[9]This answer can be true, but this depends upon some conditions: http://www.onlineqatar.com/info/dependent-family-visa.aspx

[10]Arguably, many answers are somewhat time sensitive, e.g., "*There is an IKEA in Doha.*" is true only after IKEA opened, but not before that. In such cases, we just used the present situation (Summer 2017) as a point of reference.

[11]This is comparable in size to other fact-checking datasets, e.g., Ma et al. (2015) experimented with 226 rumors, and Popat et al. (2016) used 100 Wiki hoaxes.

User Profile Features (*who says it*)

These are features characterizing the user who posted the answer, previously proposed for predicting credibility in cQA (Nakov et al. 2017b).

User posts categories (*396 individual features*) We count the answers a user has posted in each of the 197 categories in Qatar Living. We have each feature twice: once raw and once normalized by the total number of answers N the user has posted. We further use as features this N, and the number of distinct categories the user has posted in.

User posts quality (*13 features*) We first use the CQA-QA-2016 data to train a GOOD vs. BAD answer classifier, as described by Barrón-Cedeño et al. (2015). We then run this classifier (which has 80+% accuracy) on the entire unannotated Qatar Living database (2M answers, provided by the SemEval-2016 Task 3 organizers) and we aggregate its predictions to build a user profile: number of GOOD/BAD answers, total number of answers, percentage of GOOD/BAD answers, sum of the classifier's probabilities for GOOD/BAD answers, total sum of the classifier's probabilities over all answers, average score for the probability of GOOD/BAD answers, and highest absolute score for the probability of a GOOD/BAD answer.

User activity (*19 features*) These features describe the overall activity of the user. We include the number of answers posted, number of distinct questions answered, number of questions asked, number of posts in the *Jobs* and in the *Classifieds* sections, number of days since registering in the forum, and number of active days. We also have features modeling the number of answers posted during working hours (7:00-15:00h)[12], after work, at night, early in the morning, and before noon. We also model the day of posting: during a working day vs. during the weekend. Finally, we track the number of answers posted among the first k in a question–answer thread, for $k \in \{1, 3, 5, 10, 20\}$.

Answer Content (*how it is said)*

These features model what the answer says, and how. Such features were previously used by Gencheva et al. (2017).

Linguistic bias, subjectivity and sentiment Forum users (consciously or not), often put linguistic markers in their answers, which can signal the degree of the user's certainty in the veracity of what they say. Table 2 lists some categories of such markers, together with examples.

We use linguistic markers such as *factives* from (Hooper 1974), *assertives* from (Hooper 1974), *implicatives* from (Karttunen 1971), *hedges* from (Hyland 2005), *Wiki-bias* terms from (Recasens, Danescu-Niculescu-Mizil, and Jurafsky 2013), *subjectivity* cues from (Riloff and Wiebe 2003), and *sentiment* cues from (Liu, Hu, and Cheng 2005).[13]

[12]This is forum time, i.e., local Qatar time.

[13]Most of these bias cues can be found at https://people.mpi-sws.org/~cristian/Biased_language.html

Bias Type	Sample Cues
Factives	realize, know, discover, learn
Implicatives	cause, manage, hesitate, neglect
Assertives	think, believe, imagine, guarantee
Hedges	approximately, estimate, essentially
Report-verbs	argue, admit, confirm, express
Wiki-bias	capture, create, demand, follow
Modals	can, must, will, shall
Negations	neither, without, against, never, none
Strong-subj	admire, afraid, agreeably, apologist
Weak-subj	abandon, adaptive, champ, consume
Positives	accurate, achievements, affirm
Negatives	abnormal, bankrupt, cheat, conflicts

Table 2: Some cues for various bias types.

Factives (*1 feature*) are verbs that imply the veracity of their complement clause. For example, in *E1* below, *know* suggests that "they will open a second school . . . " and "they provide a qualified french education . . . " are factually true statements.

E1: Q: What do you recommend as a French school; Lycee Voltaire or Lycee Bonaparte?

A: ... About Voltaire; I ***know*** that they ***will*** open a second school; and they are a ***nice*** french school... I ***know*** that they ***provide*** a ***qualified*** french education and add with that the history and arabic language to be adapted to the qatar. I ***think*** that's an ***interesting*** addition.

Assertives (*1 feature*) are verbs that imply the veracity of their complement clause with some level of certainty. For example, in *E1*, *think* indicates some uncertainty, while verbs like *claim* cast doubt on the certainty of their complement clause.

Implicatives (*1 feature*) imply the (un)truthfulness of their complement clause, e.g., *decline* and *succeed*.

Hedges (*1 feature*) reduce commitment to the truth, e.g., *may* and *possibly*.

Reporting verbs (*1 feature*) are used to report a statement from a source, e.g., *argue* and *express*.

Wiki-bias (*1 feature*) This feature involves bias cues extracted from the NPOV Wikipedia corpus (Recasens, Danescu-Niculescu-Mizil, and Jurafsky 2013), e.g., *provide* (in *E1*), and controversial words such as *abortion* and *execute*.

Modals (*1 feature*) can change certainty (e.g., *will* or *can*), make an offer (e.g., *shall*), ask permission (e.g., *may*), or express an obligation or necessity (e.g., *must*).

Negation (*1 feature*) cues are used to deny or make negative statements, e.g., *no*, *never*.

Subjectivity (*2 features*) is used when a question is answered with personal opinions and feelings. There are two types of subjectivity cues: *strong* and *weak*. For example, in *E1*, *nice* and *interesting* are *strong* subjectivity cues, while *qualified* is a *weak* one.

Sentiment cues (*2 features*) We use *positive* and *negative* sentiment cues to model the attitude, thought, and emotions of the person answering. For example, in *E1*, *nice*, *interesting* and *qualified* are positive cues.

The above cues are about single words. We further generate multi-word cues by combining *implicative*, *assertive*, *factive* and *report* verbs with first person pronouns (*I/we*), *modals* and strong subjective *adverbs*, e.g., *I/we+verb* (e.g. "I believe"), *I/we+adverb+verb* (e.g., "I certainly know"), *I/we+modal+verb* (e.g., "we could figure out") and *I/we+modal+adverb+verb* (e.g., "we can obviously see").

Finally, we compute a feature vector for an answer using these cues according to Equation (1), where for each bias type B_i and answer A_j, the frequency of the cues for B_i in A_j is normalized by the total number of words in A_j:

$$B_i(A_j) = \frac{\sum_{cue \in B_i} count(cue, A_j)}{\sum_{w_k \in A_j} count(w_k, A_j)} \quad (1)$$

Quantitative Analysis: Credibility (*31 features*) We use features that have been previously proposed for credibility detection (Castillo, Mendoza, and Poblete 2011): number of URLs/images/emails/phone numbers; number of tokens/sentences; average number of tokens; number of 1st/2nd/3rd person pronouns; number of positive/negative smileys; number of single/double/triple exclamation/interrogation symbols. To this set, we further add number of interrogative sentences; number of nouns/verbs/adjectives/adverbs/pronouns; and number of words not in word2vec's Google News vocabulary (such OOV words could signal slang, foreign language, etc.).

Semantic Analysis: Embeddings$_{Google}$ (*300 features*) We use the pre-trained, 300-dimensional embedding vectors that Mikolov, Yih, and Zweig (2013) trained on 100 billion words from Google News. We compute a vector representation for an answer by simply averaging the embeddings of the words it contains.

Semantic Analysis: Embeddings$_{QL}$ (*100 features*) We also use 100-dimensional word embeddings from (Mihaylov and Nakov 2016), trained on all Qatar Living.

External Evidence (*external support*)

Following Karadzhov et al. (2017), we tried to verify whether an answer's claim is true by searching for support on the Web. We started with the concatenation of an answer to its question. Then, following Potthast et al. (2013), we extracted nouns, verbs and adjectives, sorted by TF-IDF (IDF computed on Qatar Living). We further extracted and added the named entities from the text and we generated a query of 5-10 words. If we did not obtain ten results, we dropped some terms from the query and we tried again.

Support from the Web (*180 features*): We automatically queried Bing[14] and extracted features from the resulting webpages, excluding those that are not related to Qatar. In particular, we calculated similarities: (*i*) cosine with TF-IDF weighting, (*ii*) cosine using Qatar Living embeddings, and (*iii*) containment (Lyon, Malcolm, and Dickerson 2001).

[14]We also experimented with Google and the aggregation of Bing and Google, with slightly worse results.

We calculated these similarities between, on the one hand, (*i*) the question or (*ii*) the answer or (*iii*) the question-answer pair, vs. on the other hand, (a) the snippets or (b) the web pages. In order to calculate the similarity against a webpage, we first converted that webpage into a list of rolling sentence triplets. Then we calculated the score of the Q/A/Q-A vs. this triplet, and finally we took the average and also the maximum similarity over these triplets. Now, as we had up to ten Web results, we further took the maximum and the average over all the above features over the returned Qatar-related pages.

We created three copies of each feature, depending on whether it came (*i*) from a reputed source (e.g., news, government websites, official sites of companies), (*ii*) from a forum-type site (forums, reviews, social media), or (*iii*) from some other type of websites.

Intra-forum Evidence *(where it is said)*

Intra-thread Analysis: Support from the current thread (*3 features*) :We use the cosine similarity between an answer- and a thread-vector of all GOOD answers using Embeddings$_{Google}$ and Embeddings$_{QL}$. The idea is that if an answer is similar to other answers in the thread, it is more likely to be true. To this, we add a feature for the reciprocal rank of the answer in the thread, assuming that more recent answers are more likely to be up-to-date and factually true.

Forum-Level Evidence: Support from all of Qatar Living (*60 features*) We further collect supporting evidence from all threads in the Qatar Living forum. We use a search engine as for the external evidence features above, but this time we limit the search to the Qatar Living forum only.

Forum-Level Evidence: Support from high-quality posts in Qatar Living (*10 features*) Among the 60,000 active users of the Qatar Living forum, there is a community of 38 trusted users who have written 5,230 high-quality articles on topics that attract a lot of interest, e.g., visas, work legislation, etc. We try to verify the answers against these high-quality posts. (*i*) Since an answer can combine both relevant and irrelevant information with respect to its question, we first generate a query as explained above for each Q&A. (*ii*) We then compute cosines between the query and the sentences in the high-quality posts, and we select the k-best matches. (*iii*) Finally, we compute textual entailment scores (Kouylekov and Negri 2010) for the answer given the k-best matches, which we then use as features.

Evaluation and Results

Settings

We train an SVM classifier (Joachims 1999) on the 249 examples as described above, where each example is one question–answer pair. For the evaluation, we use leave-one-thread-out cross validation, where each time we exclude and use for testing one of the 71 questions together with all its answers. We do so in order to respect the structure of the threads when splitting the data. We report Accuracy, Precision, Recall, and F_1 for the classification setting. We also calculate Mean Average Precision (MAP).

Results

Table 3 shows results for each of the above-described feature groups, further grouped by type of evidence —external, internal, answer-based, or user-related—, as well as for ensemble systems and for some baselines.

We can see that the best-performing feature group, both in terms of accuracy and MAP (65.46 and 83.97, respectively), is the one looking for intra-forum evidence based on search for similar answers in Qatar Living. It is closely followed by the feature group looking for external evidence in Qatar-related web sites, excluding Qatar Living, which achieved accuracy of 63.45, and the best overall F_1 score of 71.65.

Evidence from high-quality posts in Qatar Living ranks 4th with accuracy of 60.24, and support from the current thread only comes 7th with accuracy of just 53.41. These results show the importance of forum-level evidence that goes beyond the target thread and beyond known high-quality posts in the forum.

Answer-related features are the third most important feature family. In particular, linguistic features rank third overall with accuracy of 60.64; this should not be surprising as such features have been shown to be important in previous work (Popat et al. 2016). We can also see the strong performance of using knowledge about the domain in the form of word embeddings trained on Qatar Living, which are ranked 5th with accuracy of 59.44. However, general word embeddings, e.g., those trained on Google News, do not work well: with accuracy of 52.61, they are barely above the majority class baseline, which has an accuracy of 51.41.

The answer content feature family also contains a group of features that have been previously proposed for modeling credibility. This group achieves an accuracy of 56.23, and we also use it as one of the baselines in the bottom of the table. There are two reasons for its modest performance: (*i*) credibility is different from veracity as the former is subjective while the latter is not, and (*ii*) these features are generally not strong enough by themselves, as they have been originally proposed to work together with features modeling the user (age, followers, friends, etc.), a target topic, and propagation (spreading tree) on Twitter (Castillo, Mendoza, and Poblete 2011).

Interestingly, the feature types about the user profile perform the worst. They are also below the majority class baseline in terms of accuracy; however, they outperform the baselines in terms of MAP. We believe that the poor performance is due to modeling a user based on her activity, posting categories, and goodness (whether she tries to answer the question irrespective of the veracity of the given answer) of her posts in the past, which do not target factuality directly. In future work, we could run our factuality classifier over all of Qatar Living, and we can then characterize a user based on our predicted veracity of his/her answers.

The bottom of the table shows the results for two ensemble systems that combine the above feature groups, yielding accuracy of 72.29 (19 points of improvement over the majority class baseline, absolute) and MAP of 86.54 (23 points of improvement over the chronological baseline, absolute). These results indicate that our system might already be usable in real applications.

Rank	Feature Group / System	Acc	P	R	F_1	MAP
	External Evidence					
2	Support from the Web	63.45	59.59	89.84	71.65	67.71
	Intra-Forum Evidence					
1	Support from all of Qatar Living	65.46	66.41	66.41	66.41	83.97
4	Support from high-quality posts in Qatar Living	60.24	61.60	60.16	60.87	74.50
7	Support from the current thread	53.41	53.53	71.09	61.07	64.15
	Answer Content					
3	Linguistic bias, subjectivity and sentiment	60.64	60.42	67.97	63.97	78.81
5	Embeddings$_{QL}$	59.44	59.71	64.84	62.17	75.63
6	Credibility	56.23	56.21	67.19	61.21	64.92
8	Embeddings$_{Google}$	52.61	53.62	57.81	55.64	69.23
	User Profile					
9	User activity	42.57	46.67	82.03	59.49	69.04
10	User posts categories	42.57	46.67	82.03	59.49	68.50
11	User posts quality	28.92	31.01	31.25	31.13	67.43
	Ensemble Systems					
	Optimizing for Accuracy	**72.29**	70.63	78.91	**74.54**	74.32
	Optimizing for MAP	69.88	**70.87**	70.31	70.59	**86.54**
	Baselines					
	Credibility (Castillo, Mendoza, and Poblete 2011)	56.23	56.21	67.19	61.21	64.92
	All POSITIVE (majority class)	51.41	51.41	100.00	67.91	—
	Thread order (chronological)	—	—	—	—	63.75

Table 3: Experimental results for different feature groups as well as for ensemble systems and for some baselines. The first column shows the rank of each feature group, based on accuracy. The following columns describe the feature group and report accuracy (Acc), precision (P), recall (R), F_1, and mean-average precision (MAP).

Discussion

High-Quality Posts As explained above, we use a three-step approach to extract supporting evidence from the high-quality posts, namely query generation (Step 1), evidence retrieval using vector-based similarity (Step 2), and re-ranking based on entailment (Step 3). We conducted an ablation experiment in order to investigate the individual contribution of steps 1 and 3. We considered the following settings:

$S1$**:** *Full system*. All features from the three steps are used.

$S2$**:** *No re-ranking*. Only steps 1 and 2 are applied.

$S3$**:** *No query generation*. The entire answer is used to extract evidence instead of using the generated query, i.e., only steps 2 and 3 are applied.

$S4$**:** *No query generation and no re-ranking*. Only step 2 is applied. As in $S3$, the entire answer is used to retrieve evidence.

The results confirmed (*i*) *the importance of generating a good query*: discarding step 1 yields sizable drop in performance by 12 accuracy points when comparing $S4$ to $S2$, and by 4 accuracy points when comparing $S3$ to $S1$; and (*ii*) *the importance of re-ranking based on textual entailment*: discarding step 3 yields 11 accuracy points decrease in performance when comparing $S4$ to $S3$, and 3 accuracy points when comparing $S2$ to $S1$.[15]

[15]More detailed results are omitted for the sake of brevity.

Table 4 illustrates the effect of the entailment-based re-ranking (step 3). It shows a question (Q), an answer to verify (A), and the top-4 supporting sentences retrieved by our system, sorted according to the entailment-based re-ranking scores ($R1$). Column $R2$ shows the ranking for the same sentences using vector-based similarity (i.e., without applying step 3). We can see that using re-ranking yields better results. For example, the first piece of support in $R1$'s ranking is the best overall, while the same sentence is ranked 10th by $R2$. Moreover, the top-ranked evidence in $R2$, although clearly pertinent, is not better than the best one in $R1$.

Linguistic Bias We further investigated the effectiveness of the linguistic features. The experimental results show that the top-5 linguistic features are (in this order) *strong subjectivity cues*, *implicatives*, *modals*, *negatives*, and *assertives*.

External Sources Features The query generated from the question–answer pair provides enough context for a quality Web search. The results returned by the search engine are mostly relevant, which indicates that the query generation works well. More importantly, as Table 5 shows, the results returned by the search engine are relevant with respect to both the query and the question–answer pair. Note also that, as expected, the results that are Qatar-related and also from a reputed or a forum source tend to be generally more relevant.

Q: does anyone know if there is a french speaking nursery in doha?
A: there is a french school here. don't know the ages but my neighbor's 3 yr old goes there...

Best Matched Sentence for Q&A: there is a french school here.

Post Id	sId	R1	R2	Sentence
35639076	15	1	10	the pre-school follows the english program but also gives french and arabic lessons.
32448901	4	2	11	france bought the property in 1952 and since 1981 it has been home to the french institute.
31704366	7	3	1	they include one indian school, two french, seven following the british curriculum...
27971261	6	4	4	the new schools include six qatari, four indian, two british, two american and a finnish...

Table 4: Sample of sentences from high-quality posts automatically extracted to support the answer *A* to question *Q*. *sId* is the sentence id in the post, *R1* is the ranking based on entailment, and *R2* is the similarity ranking.

Q: Hi; Just wanted to confirm Qatar's National Day. Is it 18th of December? Thanks.
A: yes; it is 18th Dec.

Query generated from Q&A: `"National Day" "Qatar" National December Day confirm wanted`

URL	Qatar-related?	Source type	Snippet
qppstudio.net	No	Other	Public holidays and national ... the world's source of Public holidays information
dohanews.co	Yes	Reputed	culture and more in and around Qatar ... The documentary features human interest pieces that incorporate the day-to-day lives of Qatar residents
iloveqatar.net	Yes	Forum	Qatar National Day - Short Info ... the date of December 18 is celebrated each year as the National Day of Qatar...
cnn.com	No	Reputed	The 2022 World Cup final in Qatar will be held on December 18 ... Qatar will be held on December 18 – the Gulf state's national day. Confirm. U.S ...
icassociation.co.uk	No	Other	In partnership with ProEvent Qatar, ICA can confirm that the World Stars will be led on the 17 December, World Stars vs Qatar Stars - Qatar National Day.

Table 5: Sample snippets returned by a search engine for a given query generated from a Q&A pair.

Conclusion and Future Work

We have explored a new dimension in the context of community question answering, which has been ignored so far: checking the veracity of forum answers. As this is a new problem we created CQA-QL-2016-fact, a specialized dataset which we are releasing freely to the research community. We further proposed a novel multi-faceted model, which captures information from the answer content (*what is said and how*), from the author profile (*who says it*), from the rest of the community forum (*where it is said*), and from external authoritative sources of information (*external support*). The evaluation results have shown very strong performance.

In future work, we plan to extend our dataset with additional examples. We would also like to try distant supervision based on known facts, e.g., from high-quality posts, which would allow us to use more training data, thus enabling more sophisticated learning architectures, e.g., based on deep learning. We also want to improve user modeling, e.g., by predicting factuality for the user's answers and then building a user profile based on that. Finally, we want to explore the possibility of providing justifications for the verified answers and to integrate our system in a real application.

Acknowledgments

This research is developed by the Arabic Language Technologies (ALT) group at Qatar Computing Research, HBKU in collaboration with MIT-CSAIL. It is part of the Interactive sYstems for Answer Search (Iyas) project.

References

Agichtein, E.; Castillo, C.; Donato, D.; Gionis, A.; and Mishne, G. 2008. Finding high-quality content in social media. In *Proceedings of the International Conference on Web Search and Data Mining*, 183–194.

Banerjee, P., and Han, H. 2009. Answer credibility: A language modeling approach to answer validation. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 157–160.

Barrón-Cedeño, A.; Filice, S.; Da San Martino, G.; Joty, S.; Màrquez, L.; Nakov, P.; and Moschitti, A. 2015. Thread-level information for comment classification in community question answering. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, 687–693.

Canini, K. R.; Suh, B.; and Pirolli, P. L. 2011. Finding credible information sources in social networks based on content and social structure. In *Proceedings of the IEEE Third International Conference on Privacy, Security, Risk and Trust and the IEEE Third International Conference on Social Computing*, 1–8.

Castillo, C.; Mendoza, M.; and Poblete, B. 2011. Information credibility on Twitter. In *Proceedings of the 20th International Conference on World Wide Web*, 675–684.

Gencheva, P.; Nakov, P.; Màrquez, L.; Barrón-Cedeño, A.; and Koychev, I. 2017. A context-aware approach for detecting worth-checking claims in political debates. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing*, 267–276.

Hooper, J. 1974. *On Assertive Predicates*. Indiana University Linguistics Club.

Hyland, K. 2005. *Metadiscourse: Exploring Interaction in Writing*. Continuum Discourse. Bloomsbury Publishing.

Ishikawa, D.; Sakai, T.; and Kando, N. 2010. Overview of the NTCIR-8 community QA pilot task (part i): The test collection and the task. In *Proceedings of NTCIR-8 Workshop Meeting*, 421–432.

Jeon, J.; Croft, W. B.; Lee, J. H.; and Park, S. 2006. A framework to predict the quality of answers with non-textual features. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 228–235.

Joachims, T. 1999. Making large-scale support vector machine learning practical. In Schölkopf, B.; Burges, C. J. C.; and Smola, A. J., eds., *Advances in Kernel Methods*. 169–184.

Jurczyk, P., and Agichtein, E. 2007. Discovering authorities in question answer communities by using link analysis. In *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management*, 919–922.

Karadzhov, G.; Nakov, P.; Màrquez, L.; Barrón-Cedeño, A.; and Koychev, I. 2017. Fully automated fact checking using external sources. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing*, 344–353.

Karttunen, L. 1971. Implicative verbs. *Language* 47(2):340–358.

Kouylekov, M., and Negri, M. 2010. An open-source package for recognizing textual entailment. In *Proceedings of the ACL 2010 System Demonstrations*, 42–47.

Lita, L. V.; Schlaikjer, A. H.; Hong, W.; and Nyberg, E. 2005. Qualitative dimensions in question answering: Extending the definitional QA task. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 1616–1617.

Liu, B.; Hu, M.; and Cheng, J. 2005. Opinion observer: Analyzing and comparing opinions on the web. In *Proceedings of the 14th International Conference on World Wide Web*, 342–351.

Lukasik, M.; Cohn, T.; and Bontcheva, K. 2015. Point process modelling of rumour dynamics in social media. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, 518–523.

Lyon, C.; Malcolm, J.; and Dickerson, B. 2001. Detecting short passages of similar text in large document collections. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 118–125.

Ma, J.; Gao, W.; Wei, Z.; Lu, Y.; and Wong, K.-F. 2015. Detect rumors using time noseries of social context information on microblogging websites. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, 1751–1754.

Mihaylov, T., and Nakov, P. 2016. SemanticZ at SemEval-2016 Task 3: Ranking relevant answers in community question answering using semantic similarity based on fine-tuned word embeddings. In *Proceedings of the 10th International Workshop on Semantic Evaluation*, 879–886.

Mikolov, T.; Yih, W.-t.; and Zweig, G. 2013. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 746–751.

Morris, M. R.; Counts, S.; Roseway, A.; Hoff, A.; and Schwarz, J. 2012. Tweeting is believing?: Understanding microblog credibility perceptions. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work*, 441–450.

Mukherjee, S., and Weikum, G. 2015. Leveraging joint interactions for credibility analysis in news communities. In *Proceedings of the 24th ACM International Conference on Information and Knowledge Management*, 353–362.

Nakov, P.; Màrquez, L.; Magdy, W.; Moschitti, A.; Glass, J.; and Randeree, B. 2015. SemEval-2015 task 3: Answer selection in community question answering. In *Proceedings 9th International Workshop on Semantic Evaluation*, 269–281.

Nakov, P.; Màrquez, L.; Moschitti, A.; Magdy, W.; Mubarak, H.; Freihat, A. A.; Glass, J.; and Randeree, B. 2016. SemEval-2016 task 3: Community question answering. In *Proceedings of the 10th International Workshop on Semantic Evaluation*, 525–545.

Nakov, P.; Hoogeveen, D.; Màrquez, L.; Moschitti, A.; Mubarak, H.; Baldwin, T.; and Verspoor, K. 2017a. SemEval-2017 task 3: Community question answering. In *Proceedings of the 11th International Workshop on Semantic Evaluation*, 27–48.

Nakov, P.; Mihaylova, T.; Màrquez, L.; Shiroya, Y.; and Koychev, I. 2017b. Do not trust the trolls: Predicting credibility in community question answering forums. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing*, 551–560.

Pelleg, D.; Rokhlenko, O.; Szpektor, I.; Agichtein, E.; and Guy, I. 2016. When the crowd is not enough: Improving user experience with social media through automatic quality analysis. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, 1080–1090.

Popat, K.; Mukherjee, S.; Strötgen, J.; and Weikum, G. 2016. Credibility assessment of textual claims on the web. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, 2173–2178.

Potthast, M.; Hagen, M.; Gollub, T.; Tippmann, M.; Kiesel, J.; Rosso, P.; Stamatatos, E.; and Stein, B. 2013. Overview of the 5th international competition on plagiarism detection. In *Proceedings of the CLEF Conference on Multilingual and Multimodal Information Access Evaluation*, 301–331.

Recasens, M.; Danescu-Niculescu-Mizil, C.; and Jurafsky, D. 2013. Linguistic models for analyzing and detecting biased language. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, 1650–1659.

Riloff, E., and Wiebe, J. 2003. Learning extraction patterns for subjective expressions. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 105–112.

Su, Q.; yun Chen, H. K.; and Huang, C.-R. 2010. Incorporate credibility into context for the best social media answers. In *Proceedings of the 24th Pacific Asia Conference on Language, Information and Computation*, 535–541.

Zubiaga, A.; Liakata, M.; Procter, R.; Wong Sak Hoi, G.; and Tolmie, P. 2016. Analysing how people orient to and spread rumours in social media by looking at conversational threads. *PLoS ONE* 11(3):1–29.