SemEval-2019 Task 8: Fact Checking in Community Question Answering Forums

Tsvetomila Mihaylova,[1] Georgi Karadzhov,[2] Pepa Atanasova,[3]
Ramy Baly,[4] Mitra Mohtarami,[4] Preslav Nakov[5]
[1] Instituto de Telecomunicações, Lisbon, Portugal, [2] SiteGround Hosting EOOD, Bulgaria
[3] University of Copenhagen, Denmark
[4] MIT Computer Science and Artificial Intelligence Laboratory, Cambridge, MA
[5] Qatar Computing Research Institute, HBKU
{tsvetomila.mihaylova, georgi.m.karadjov}@gmail.com,
pepa@di.ku.dk, {baly,mitram}@mit.edu, pnakov@qf.org.qa

Abstract

We present SemEval-2019 Task 8 on *Fact Checking in Community Question Answering Forums*, which features two subtasks. Subtask A is about deciding whether a question asks for factual information vs. an opinion/advice vs. just socializing. Subtask B asks to predict whether an answer to a factual question is true, false or not a proper answer. We received 17 official submissions for subtask A and 11 official submissions for Subtask B. For subtask A, all systems improved over the majority class baseline. For Subtask B, all systems were below a majority class baseline, but several systems were very close to it. The leaderboard and the data from the competition can be found at http://competitions.codalab.org/competitions/20022.

1 Overview

The current coverage of the political landscape in both the press and in social media has led to an unprecedented situation. Like never before, a statement in an interview, a press release, a blog note, or a tweet can spread almost instantaneously. The speed of proliferation leaves little time for double-checking claims against the facts, which has proven critical in politics, e.g., during the 2016 presidential campaign in the USA, which was dominated by fake news in social media and by false claims.

Investigative journalists and volunteers have been working hard to get to the root of a claim and to present solid evidence in favor or against it. Manual fact-checking is very time-consuming, and thus automatic methods have been proposed to speed-up the process, e.g., there has been work on checking the factuality/credibility of a claim, of a news article, or of an information source (Ba et al., 2016; Zubiaga et al., 2016; Ma et al., 2016; Castillo et al., 2011; Baly et al., 2018). The process starts when a document is made public. First, an intrinsic analysis is carried out in which check-worthy text fragments are identified. Then, other documents that might support or rebut a claim in the document are retrieved from various sources. Finally, by comparing a claim against the retrieved evidence, a system can determine whether the claim is likely true or likely false (or unsure, if no strong enough evidence either way could be found). For instance, Ciampaglia et al. (2015) do this using a knowledge graph derived from Wikipedia. The outcome could then be presented to a human expert for final judgement.[1]

For our two subtasks, we explore factuality in the context of Community Question Answering (cQA) forums. Forums such as StackOverflow, Yahoo! Answers, and Quora are very popular these days, as they represent effective means for communities around particular topics to share information. However, the information shared by the users is not always correct or accurate. There are multiple factors explaining the presence of incorrect answers in cQA forums, e.g., misunderstanding of the question, ignorance or maliciousness of the responder. Also, as a result of our dynamic world, the truth is time-sensitive: something that was true yesterday may be false today. Moreover, forums are often barely moderated and thus lack systematic quality control.

Here we focus on checking the factuality of questions and answers in cQA forums. This aspect was ignored in recent cQA tasks (Ishikawa et al., 2010; Nakov et al., 2015, 2016a, 2017a), where an answer is considered GOOD if it addresses the question, irrespective of its veracity, accuracy, etc.

[1] As of present, fully automatic methods for fact checking still lag behind in terms of quality, and thus also of credibility in the eyes of the users, compared to what high-quality manual checking by reputable sources can achieve, which means that a final double-checking by a human expert is needed.

Proceedings of the 13th International Workshop on Semantic Evaluation (SemEval-2019), pages 860–869
Minneapolis, Minnesota, USA, June 6–7, 2019. ©2019 Association for Computational Linguistics

Q: HI ;; IF WIFE IS UNDER HER HUSBAND'S SPONSORSHIP AND IS WILLING TO COME QATAR ON VISIT; HOW LONG SHE CAN STAY AFTER EXTENDING THE VISA EVERY MONTH? 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 answered 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-2016 Task 3 (Nakov et al., 2016a), all three answers would be considered GOOD since they are formally answering the question. Nevertheless, a_1 contains false information, while a_2 and a_3 are correct, as can be established from an official government website.[2]

Checking the veracity of answers in a cQA forum is a hard problem, which requires putting together aspects of language understanding, modelling the context, integrating several information sources, uisng world knowledge and complex inference, among others. Moreover, high-quality automatic fact-checking would offer better experience to users of cQA systems, e.g., the user could be presented with veracity scores, where low scores would warn the user not to completely trust the answer or to double-check it.

2 Related Work

Fact-checking of answers was not studied before in the context of community Question Answering, apart from our own recent work (Mihaylova et al., 2018). Yet, in the context of cQA and general QA, there has been work on *credibility* assessment, which has been modelled primarily at the feature level, with the goal of improving GOOD answer identification. A notable exception are (Nakov et al., 2017b; Mihaylov et al., 2018), where credibility was a task on its own right. However, *credibility* is different from *veracity* (our focus here) as it is a subjective perception about whether a statement is credible, rather than actually truthful.

Jurczyk and Agichtein (2007) modelled author authority using link analysis. Agichtein et al. (2008) looked for high-quality answers using PageRank and HITS, in addition to intrinsic content quality, e.g., punctuation and typos, syntactic and semantic complexity, and grammaticality.

Lita et al. (2005) studied three qualitative dimensions for answers: source credibility (e.g., does the document come from a government website), sentiment analysis, and contradiction compared to other answers. Su et al. (2010) looked for verbs and adjectives that cast doubt. Banerjee and Han (2009) used language modelling to validate the reliability of an answer's source. Jeon et al. (2006) focused on non-textual features such as click counts, answer activity level, and copy counts. Pelleg et al. (2016) curated social media content using syntactic, semantic, and social signals. Unlike this research, we (*i*) target factuality rather than credibility, (*ii*) address it as a task in its own right, and on a specialised dataset.

Information credibility was also studied in social computing. Castillo et al. (2011) modeledd user reputation. Canini et al. (2011) analyzed the interaction of content and social network structure. Morris et al. (2012) studied how Twitter users judge truthfulness. Lukasik et al. (2015) used temporal patterns to detect rumors, and Zubiaga et al. (2016) focused on conversations.

Other authors have been querying the Web to gather support for accepting or refuting a claim (Popat et al., 2016; Karadzhov et al., 2017b). In social media, there has been research targeting the user, e.g., finding malicious users (Mihaylov and Nakov, 2016; Mihaylova et al., 2018; Mihaylov et al., 2018), *sockpuppets* (Maity et al., 2017), *Internet water army* (Chen et al., 2013), and *seminar users* (Darwish et al., 2017).

Finally, there has been work on credibility, trust, and expertise in news communities (Mukherjee and Weikum, 2015). Dong et al. (2015) proposed that a trustworthy source is one that contains very few false claims. Recent work has also focused on evaluating the factuality of reporting of entire news outlets (Baly et al., 2018, 2019).[3] However, none of this work was about QA or cQA.

[2] http://portal.moi.gov.qa/wps/portal/MOIInternet/departmentcommittees/visasentrypermeits/

[3] Knowing the reliability of a medium is important when fact-checking a claim (Popat et al., 2017; Nguyen et al., 2018) and when solving article-level tasks such as "fake news" and click-bait detection (Hardalov et al., 2016; Karadzhov et al., 2017a; Pan et al., 2018; Pérez-Rosas et al., 2018).

3 Subtasks and Data Description

SemEval-2019 Task 8 has two subtasks:

- **Subtask A:** Given a question from a cQA forum, predict whether this question asks for factual information vs. opinion/advice vs. just socializing.
- **Subtask B:** Given a factual question from a cQA forum, together with its answer thread, predict whether each answer provides true vs. false vs. non-factual information as a response to the question.

3.1 Data and Resources

We retrieved the data from the Qatar Living web forum[4]. We then cleaned it and we annotated it with the labels described in Sections 3.2 and 3.3.

For subtask A, we annotated the questions using Amazon Mechanical Turk[5]. To ensure high quality of the annotation, we went through all annotations and manually double-checked them.

For subtask B, we did not use an external annotation service, but instead we annotated all the data ourselves. Each answer was processed by three independent annotators, and we made sure we had proof for the label from reliable sources on the Web. Then, the annotations were consolidated after a discussion until agreement was achieved for each example.

All data is freely available under a Creative Commons Attribution 3.0 Unported (CC BY 3.0) license, and is accessible on the competition's website[6].

In addition to the provided annotated data, we also allowed the participants to use unlabelled data from the Qatar Living forum footnotehttp://alt.qcri.org/semeval2016/task3/data/uploads/QL-unannotated-data-subtaskA.xml.zip, as well as additional external resources, which they had to mention explicitly in their submissions.

Note that the class distribution in the training, development and test sets differs, especially for Subtask B. The reason for this is the way the data was prepared. The different datasets (training, development and test) were prepared on stages, because of the very time-consuming data annotation process. For each dataset annotation stage, we had to choose between releasing all the available annotated data or aim at releasing sets with similar label distribution. At the end, we decided to release the available data, although we were aware that this would result in releasing sets with different distribution and, in some cases, unbalanced categories.

3.2 Training Data for Subtask A

To create the dataset for the task, we chose to augment a pre-existing dataset for cQA with factuality annotations; this allowed us to stress the difference between (*a*) distinguishing a good vs. a bad answer, and (*b*) distinguishing 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 on Community Question Answering (Nakov et al., 2016a).

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

We analyzed the forum questions and we defined three categories, related to their factuality. We then annotated the questions using Amazon Mechanical Turk. The three factuality categories are as follows:

- ∗ FACTUAL: The question asks for factual information, which can be answered by checking various information sources, and it is not ambiguous (e.g., "*What is Ooredoo customer service number?*").
- ∗ OPINION: The question asks for an opinion or an advice, not for a fact. (e.g., "*Can anyone recommend a good Vet in Doha?*")
- ∗ SOCIALIZING: Not a real question, but rather socializing/chatting. This can also mean expressing an opinion or sharing some information, without really asking anything of general interest. (e.g., "*What was your first car?*")

Table 1 shows the distribution of the labels for the question labels in the training, in the development and in the testing datasets. Overall, there are 1118, 239 and 953 questions annotated with the above-described labels.

[4] http://www.qatarliving.com

[5] http://www.mturk.com/

[6] http://competitions.codalab.org/competitions/20022

Label	Train	Dev	Test
FACTUAL	311	62	299
OPINION	563	126	167
SOCIALIZING	244	51	487
TOTAL	**1118**	**239**	**953**

Table 1: **Subtask A:** Distribution of the factuality labels for the questions.

Label	Train	Dev	Test
TRUE	166	29	34
FALSE	135	31	45
NONFACTUAL	194	52	231
TOTAL	**495**	**112**	**310**

Table 2: **Subtask B:** Distribution of the factuality labels for the answers.

3.3 Training Data for Subtask B

For subtask B, we annotated for veracity the answers to the questions with a FACTUAL label for subtask A. Note that in CQA-QL-2016, each answer has a subject, a body, meta information (answer ID, user name, and ID), the question that it answers, and a judgement about how well it answers the question of its thread (GOOD , BAD or POTENTIALLY USEFUL).

We annotated the GOOD answers for factuality based on the assumption that a GOOD answer means it provides factual information, whether it is true or false. All BAD and POTENTIALLY USEFUL answers are automatically considered as NON-FACTUAL. The factuality labels are described as follows:

∗ FACTUAL – TRUE: The answer is True and can be proven with an 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."*).[7]

∗ FACTUAL – FALSE: The answer gives a factual response, but it is False and this can be proven using an external resource. (**Q:** *"Can I bring my pitbulls to Qatar?"*; **A:** *"Yes you can bring it but be careful this kind of dog is very dangerous."*).[8]

∗ FACTUAL – PARTIALLY TRUE: The answer contains more than one claim, and only some of these claims could be manually verified. (**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."*).[9]

∗ FACTUAL – CONDITIONALLY TRUE: The answer is True in some cases, and False in others, depending on some conditions 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."*).[10]

∗ 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.*")

∗ NON-FACTUAL: When the answer does not provide factual information to the question; it can be an opinion or an advice that cannot be verified. (e.g., "*Its better to buy a new one.*").

We further discarded answers whose factuality was very time-sensitive and it makes no sense to check whether the statements are true or false (e.g., "*It is Friday tomorrow.*", "*It was raining last week.*").

Moreover, many answers are arguably 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 as a point of reference. We further discarded the answers for which the annotators could not find any information.

[7] The answer is factually true and this can be seen at http://wwwnc.cdc.gov/travel/destinations/traveler/none/qatar

[8] The answer is incorrect since pitbulls are included in the list of breeds banned in Qatar. See http://canvethospital.com/pet-relocation/banned-dog-breed-list-qatar-2015/

[9] The place mentioned in the answer has table tennis, but we do not know on which days. See http://www.qatarbowlingfederation.com/bowling-center/

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

Ultimately, we consolidated the above fine-grained labels into the following coarse-grained labels, which we used for subtask B:

- FACTUAL – TRUE: Contains answers with proven true, non-contradictory statements. This includes the answers with the label FACTUAL – TRUE from above. This label is used for answers one can trust as a true statement.

- FACTUAL – FALSE: Contains answers with statements that are proven to be false or not completely true. This includes answers with the following fine-grained factuality labels: FACTUAL – FALSE, FACTUAL – PARTIALLY FALSE, FACTUAL – CONDITIONALLY TRUE, FACTUAL – RESPONDER UNSURE. We also use this label for answers that contain a statement for which the person giving the answer expresses uncertainty in the claim.

- NON-FACTUAL: These are either non-factual statements or statements that could be factual, but no information about them could be found, i.e., we could find no way to check whether the statement was true or false. This category also includes some statements that have been incorrectly annotated as a GOOD answer. It also includes the very time-sensitive statements described before, such as "*It is Friday tomorrow?*". The BAD and the POTENTIALLY USEFUL answers from CQA-QL-2016 also fall in this category.

As we have mentioned above, we have annotated the answers to the FACTUAL questions selected from the Qatar Living forum. We targeted very high quality annotation, and thus we did not use crowd-sourcing, as a pilot experiment has shown that the task was very difficult and that it was not possible to guarantee that Turkers would do all the necessary verification and gather evidence from trusted sources. Instead, all examples were first annotated independently by three of us, and then, we carefully discussed *each example* to come up with a final label. The distribution of the labels on the training, on the development, and on the testind dataset are shown in Table 2[11].

3.4 Evaluation

Both subtasks are three-way classification problems. In subtask A, the questions were to be classified as FACTUAL, OPINION, or SOCIALIZING. Similarly, in subtask B there were also three target categories for the answers: FACTUAL - TRUE, FACTUAL - FALSE, and NON-FACTUAL.

We further scored the submissions based on Accuracy, macro-F1, and average recall (AvgRec).[12] For subtask B, we also report mean average precision (MAP), where the FACTUAL - TRUE instances were considered to be positive, and the remaining ones were negative. The official evaluation measure for both subtasks was Accuracy.

4 Participants and Results

We received 17 official submissions for Subtask A and 11 official submissions for Subtask B. Below we report the evaluation results.

Table 3 presents the results for subtask A on question classification. The results are based the official submissions in the evaluation phase. In this subtask, all of the submitted systems managed to improve over the majority class baseline, and several teams achieved similarly good results. Whenever a number of teams achieve the same result with respect to the main evaluation measure, i.e., Accuracy, we rank them according to the F1 score, and then by AvgRec if a tie still appears.

Table 4 presents the results based on the evaluation phase on the test set for predicting answer factuality labels. This subtask was more difficult as the majority class baseline was very high due to label unbalance. No team managed to improve over that baseline, but several teams had results that were very close to it.

5 Discussion

In the evaluation phase of the competition, the participants had to specify one official submission and were allowed up to two contrastive submissions. In the post-evaluation phase, they could upload an unlimited number of contrastive submissions. Below, we will only discuss the official submissions. The contrastive submissions, the ablation studies, and the experiments with different techniques are described by the participants in their respective system description papers.

[11] Although not very big, our dataset is larger than datasets used for similar problems, e.g., Ma et al. (2015) experimented with 226 rumors for rumor detection, and Popat et al. (2016) used 100 Wiki hoaxes for credibility assessment of textual claims.

[12] Average recall has some attractive properties and has been used in previous SemEval tasks, e.g., (Nakov et al., 2016b; Rosenthal et al., 2017).

Team ID	Affiliation	Accuracy	F1	AvgRec
Fermi (Syed et al., 2019)	IIIT Hyderabad, Microsoft, Teradata	**0.840**	0.718_{2}	0.735_{3}
TMLab (Niewiński et al., 2019)	Samsung R&D Institute, Warsaw, Poland	**0.834**	0.725_{1}	0.764_{1}
SolomonLab (Gupta et al., 2019)	Samsung R&D Institute India, Bangalore	**0.831**	0.709_{4}	0.728_{4}
ColumbiaNLP (Chakrabarty and Muresan, 2019)	Columbia University, Department Of Computer Science and Data Science Institute	**0.828**	0.645_{7}	0.662_{9}
DOMLIN (Stammbach et al., 2019)	Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI), Saarbrucken, Germany	**0.823**	0.710_{3}	0.755_{2}
BLCU_NLP (Xie et al., 2019)	Beijing Language and Culture University, Beijing, China	**0.820**	0.696_{5}	0.723_{5}
pjetro	Warsaw University of Technology	**0.790**	0.661_{6}	0.698_{6}
LP0606		**0.768**	0.637_{8}	0.679_{8}
PP08		**0.766**	0.637_{9}	0.684_{7}
AUTOHOME-ORCA (Lv et al., 2019)	Autohome Inc., Beijing, China and Beijing University of Posts and Telecommunications, Beijing, China	**0.745**	0.583_{10}	0.596_{11}
DUTH (Bairaktaris et al., 2019)	Democritus University of Thrace, Xanthi, Greece	**0.711**	0.563_{11}	0.604_{10}
cococold		**0.702**	0.543_{12}	0.594_{12}
nothing		**0.702**	0.543_{12}	0.594_{12}
chchao		**0.630**	0.454_{13}	0.523_{13}
CodeForTheChange (Avvaru and Pandey, 2019)	International Institute of Information Technology, Hyderabad, Teradata and Qubole	**0.630**	0.442_{14}	0.513_{14}
Tuefact (Juhasz et al., 2019)	University of Tübingen, Tübingen, Germany	**0.599**	0.360_{15}	0.348_{15}
Reem06		**0.549**	0.263_{16}	0.343_{16}
Majority Class Baseline		**0.450**	*0.009*	*0.333*

Table 3: **Subtask A:** Results for question classification based on the official submissions, evaluated on the test set. (Some teams did not submit system description papers, and thus we have no citations for their systems.)

Team ID	Affiliation	Accuracy	F1	AvgRec	MAP
AUTOHOME-ORCA	Autohome Inc., Beijing, China and Beijing University of Posts and Telecommunications, Beijing, China	**0.815**	0.511_{2}	0.512_{2}	0.155_{7}
ColumbiaNLP	Columbia University, Department Of Computer Science and Data Science Institute	**0.791**	0.524_{1}	0.635_{1}	0.134_{8}
DOMLIN	Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI), Saarbrucken, Germany	**0.718**	0.402_{3}	0.445_{3}	0.267_{3}
SolomonLab	Samsung R&D Institute India, Bangalore	**0.686**	0.375_{4}	0.403_{4}	0.333_{2}
CodeForTheChange	International Institute of Information Technology, Hyderabad, Teradata and Qubole	**0.654**	0.325_{5}	0.326_{5}	0.156_{6}
BLCU_NLP	Beijing Language and Culture University, Beijing, China	**0.611**	0.296_{6}	0.317_{6}	0.222_{4}
LP0606		**0.548**	0.271_{7}	0.341_{7}	0.121_{9}
PP08		**0.548**	0.271_{7}	0.341_{7}	0.121_{9}
Tuefact	University of Tübingen, Tübingen, Germany	**0.527**	0.260_{8}	0.347_{8}	0.571_{1}
cococold		**0.439**	0.133_{9}	0.241_{9}	0.208_{5}
nothing		**0.439**	0.133_{9}	0.241_{9}	0.208_{5}
Majority Class Baseline		**0.830**	*0.285*	*0.333*	*0.156*

Table 4: **Subtask B:** Results for answer classification based on the official submissions, evaluated on the test set.

The best system for Subtask A was by team Fermi (IIIT Hyderabad). They used Google's Universal Sentence representation (Cer et al., 2018), and XGBoost (Chen and Guestrin, 2016).

The best system for Subtask B was by team AUTOHOME-ORCA (Autohome Inc. and Beijing University of Posts and Telecommunications), who used BERT (Devlin et al., 2019).

They achieved their best results by using an ensemble, and by also using question meta-information (category and subject) in addition to the question and the answer text. They concatenated the category, the subject and the body of the questions into the first part separated by [SEP]. The replier's username and statement were concatenated as the second part. The two parts separated by [SEP] were pushed into the BERT model for answer classification. Then, based on the sequential outputs of the BERT model, some variant methods such as average-pooling, and bi-LSTM were adopted to produce the final results. To tackle the problem with insufficient training data, they further used data augmentation based on translation with Google Translate: in particular, they performed consecutive English-Chinese and Chinese-English translation to generate more synthetic training data.

Overall, the submitted systems for the two subtasks used a number of pre-processing steps to clean the text of the question and of the answer. As shown by the DOMLIN team, the pre-processing of the data turns out to be crucial. They reported up to 5% improvement in terms of accuracy when cleaning the unannotated forum data before fine-tuning a BERT model. Common preprocessing steps included removing or replacing the URLs, the numbers, the punctuation, the symbols, spell-checking, expansion of contractions, HTML tags, etc. DUTH also used lemmatization and stopword removal.

The submitted systems used a wide range of strategies for training their models. A sizable part of the systems used manually crafted features such as linguistic, syntactic, stylistic, and semantic features. Moreover, the systems used task-specific information such as answer ranking and rating. ColumbiaNLP also computed an average cosine similarity of one answer with respect to the other answers in the thread for subtask B, assuming that bad answers would differ substantially from the remaining answers.

While some of the approaches used character and word n-gram information, the teams also used word- and sentence-level embeddings. CodeForTheChange evaluated different classification algorithms fed with Skip-Thought vectors, and ultimately found that neural networks performed best for both subtasks with either concatenation or averaging over the vectors of the available texts. Fermi performed evaluation of different embedding models - InferSent, Concatenated Power Mean Word Embedding, Lexical Vectors, ELMo and The Universal Sentence Encoder, used in subtask A to feed an XGBoost classifier. ColumbiaNLP used ULMFiT, but performed additional unsupervised tuning of the language model on questions, answers and question-answer pairs from the Qatar Living Forum. TMLab's system used the Universal Sentence Encoder.

A common neural network architecture was LSTM, where YNU-HPCC combined LSTM with an attention mechanism. TueFact used comment chain embeddings. Other machine learning algorithms that participants tried include Random Forest, Adaboost, Perceptron, and SVM, inter alia.

While for question classification (subtask A), all the necessary information was contained in the question text and in the metadata, subtask B required additional resources. Most teams used the provided additional unannotated forum data in order to pre-train their language models or to extract more data with weak supervision (DOMLIN). Furthermore, several teams used other means for data augmentation such as SQuAD (BLCU NLP) or external Web information (SolomonLab).

6 Conclusion

We have described SemEval 2019 Task 8 on Fact Checking in Community Question Answering Forums. We received 17 and 11 submissions for Subtask A and B, respectively. Overall, subtask A (question classification) was easier and all submitted systems managed to improve over the majority class baseline. However, Subtask B (answer classification) proved to be much more challenging, and no team managed to improve over the majority class baseline, even though several teams came very close. For this latter subtask, using external resources and preprocessing proved to be crucial.

Acknowledgments

This research is part of the Tanbih project,[13] which aims to limit the effect of "fake news", propaganda and media bias by making users aware of what they are reading. The project is developed in collaboration between the Qatar Computing Research Institute (QCRI), HBKU and the MIT Computer Science and Artificial Intelligence Laboratory (CSAIL).

[13] http://tanbih.qcri.org/

References

Eugene Agichtein, Carlos Castillo, Debora Donato, Aristides Gionis, and Gilad Mishne. 2008. Finding high-quality content in social media. In *Proceedings of the 2008 International Conference on Web Search and Data Mining*, WSDM '08, pages 183–194, Palo Alto, CA, USA.

Adithya Avvaru and Anupam Pandey. 2019. CodeForTheChange at SemEval-2019 task 8: Skip-thoughts for fact checking in community question answering. In *Proceedings of the International Workshop on Semantic Evaluation*, SemEval '19, Minneapolis, MN, USA.

Mouhamadou Lamine Ba, Laure Berti-Equille, Kushal Shah, and Hossam M Hammady. 2016. VERA: A platform for veracity estimation over web data. In *Proceedings of the 25th International Conference Companion on World Wide Web*, pages 159–162, Montreal, Canada.

Anastasios Bairaktaris, Symeon Symeonidis, and Avi Arampatzis. 2019. DUTH at SemEval-2019 task 8: Part-of-speech features for question classification. In *Proceedings of the International Workshop on Semantic Evaluation*, SemEval '19, Minneapolis, MN, USA.

Ramy Baly, Georgi Karadzhov, Dimitar Alexandrov, James Glass, and Preslav Nakov. 2018. Predicting factuality of reporting and bias of news media sources. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, EMNLP '18, pages 3528–3539, Brussels, Belgium.

Ramy Baly, Georgi Karadzhov, Abdelrhman Saleh, James Glass, and Preslav Nakov. 2019. Multi-task ordinal regression for jointly predicting the trustworthiness and the leading political ideology of news media. In *Proceedings of the 17th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL-HLT '19, Minneapolis, MN, USA.

Protima Banerjee and Hyoil Han. 2009. Answer credibility: A language modeling approach to answer validation. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, NAACL-HLT '09, pages 157–160, Boulder, CO, USA.

Kevin R. Canini, Bongwon Suh, and Peter L. Pirolli. 2011. Finding credible information sources in social networks based on content and social structure. In *Proceedings of the IEEE International Conference on Privacy, Security, Risk, and Trust, and the IEEE International Conference on Social Computing*, SocialCom/PASSAT '11, pages 1–8, Boston, MA, USA.

Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. 2011. Information credibility on Twitter. In *Proceedings of the 20th International Conference on World Wide Web*, WWW '11, pages 675–684, Hyderabad, India.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, et al. 2018. Universal sentence encoder. *arXiv preprint arXiv:1803.11175*.

Tuhin Chakrabarty and Smaranda Muresan. 2019. ColumbiaNLP at SemEval-2019 task 8: The answer is language model fine-tuning. In *Proceedings of the International Workshop on Semantic Evaluation*, SemEval '19, Minneapolis, MN, USA.

Cheng Chen, Kui Wu, Venkatesh Srinivasan, and Xudong Zhang. 2013. Battling the Internet Water Army: detection of hidden paid posters. In *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, ASONAM '13, pages 116–120, Niagara, Canada.

Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A scalable tree boosting system. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, San Francisco, California, USA.

Giovanni Luca Ciampaglia, Prashant Shiralkar, Luis M. Rocha, Johan Bollen, Filippo Menczer, and Alessandro Flammini. 2015. Computational fact checking from knowledge networks. *PLOS ONE*, 10(6):1–13.

Kareem Darwish, Dimitar Alexandrov, Preslav Nakov, and Yelena Mejova. 2017. Seminar users in the Arabic Twitter sphere. In *Proceedings of the 9th International Conference on Social Informatics*, SocInfo '17, pages 91–108, Oxford, UK.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, NAACL-HLT '19, Minneapolis, MN, USA.

Xin Luna Dong, Evgeniy Gabrilovich, Kevin Murphy, Van Dang, Wilko Horn, Camillo Lugaresi, Shaohua Sun, and Wei Zhang. 2015. Knowledge-based trust: Estimating the trustworthiness of web sources. *Proc. VLDB Endow.*, 8(9):938–949.

Ankita Gupta, Sudeep Kumar Sahoo, Divya Prakash, Rohit R R, Vertika Srivastava, and YEON HYANG KIM. 2019. SolomonLab at SemEval-2019 task 8: Question factuality and answer veracity prediction in community forums. In *Proceedings of the International Workshop on Semantic Evaluation*, SemEval '19, Minneapolis, MN, USA.

Momchil Hardalov, Ivan Koychev, and Preslav Nakov. 2016. In search of credible news. In *Proceedings of the 17th International Conference on Artificial Intelligence: Methodology, Systems, and Applications*, AIMSA '16, pages 172–180, Varna, Bulgaria.

Daisuke Ishikawa, Tetsuya Sakai, and Noriko Kando. 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*, pages 421–432, Tokyo, Japan.

Jiwoon Jeon, W. Bruce Croft, Joon Ho Lee, and Soyeon Park. 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*, SIGIR '06, pages 228–235, Seattle, WA, USA.

Reka Juhasz, Franziska-Barbara Linnenschmidt, and Teslin Roys. 2019. TueFact at SemEval 2019 Task 8: Fact checking in community question answering forums: context matters. In *Proceedings of the International Workshop on Semantic Evaluation*, SemEval '19, Minneapolis, MN, USA.

Pawel Jurczyk and Eugene Agichtein. 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*, CIKM '07, pages 919–922, Lisbon, Portugal.

Georgi Karadzhov, Pepa Gencheva, Preslav Nakov, and Ivan Koychev. 2017a. We built a fake news & clickbait filter: What happened next will blow your mind! In *Proceedings of the International Conference on Recent Advances in Natural Language Processing*, RANLP '17, pages 334–343, Varna, Bulgaria.

Georgi Karadzhov, Preslav Nakov, Lluís Màrquez, Alberto Barrón-Cedeño, and Ivan Koychev. 2017b. Fully automated fact checking using external sources. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing*, RANLP '17, pages 344–353, Varna, Bulgaria.

Lucian Vlad Lita, Andrew Hazen Schlaikjer, WeiChang Hong, and Eric Nyberg. 2005. Qualitative dimensions in question answering: Extending the definitional QA task. In *Proceedings of the National Conference on Artificial Intelligence*, volume 20 of *AAAI '05*, pages 1616–1617, Pittsburgh, PA, USA.

Michal Lukasik, Trevor Cohn, and Kalina Bontcheva. 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*, ACL-IJCNLP '15, pages 518–523, Beijing, China.

Zhengwei Lv, Duoxing Liu, Haifeng Sun, Xiao Liang, Tao Lei, Zhizhong Shi, Feng Zhu, and Lei Yang. 2019. AUTOHOME-ORCA at SemEval-2019 task 8: Application of BERT for fact-checking in community forums. In *Proceedings of the International Workshop on Semantic Evaluation*, SemEval '19, Minneapolis, MN, USA.

Jing Ma, Wei Gao, Prasenjit Mitra, Sejeong Kwon, Bernard J. Jansen, Kam-Fai Wong, and Meeyoung Cha. 2016. Detecting rumors from microblogs with recurrent neural networks. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, IJCAI '16, pages 3818–3824, New York, NY, USA.

Jing Ma, Wei Gao, Zhongyu Wei, Yueming Lu, and Kam-Fai Wong. 2015. Detect rumors using time series of social context information on microblogging websites. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, CIKM '15, pages 1751–1754, Melbourne, Australia.

Suman Kalyan Maity, Aishik Chakraborty, Pawan Goyal, and Animesh Mukherjee. 2017. Detection of sockpuppets in social media. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work and Social Computing*, CSCW '17, pages 243–246, Portland, OR, USA.

Todor Mihaylov, Tsvetomila Mihaylova, Preslav Nakov, Lluís Màrquez, Georgi Georgiev, and Ivan Koychev. 2018. The dark side of news community forums: Opinion manipulation trolls. *Internet Research*, 28(5):1292–1312.

Todor Mihaylov and Preslav Nakov. 2016. Hunting for troll comments in news community forums. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, ACL '16, pages 399–405, Berlin, Germany.

Tsvetomila Mihaylova, Preslav Nakov, Lluís Màrquez, Alberto Barrón-Cedeño, Mitra Mohtarami, Georgi Karadjov, and James Glass. 2018. Fact checking in community forums. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, AAAI '18, pages 879–886, New Orleans, LA, USA.

Meredith Ringel Morris, Scott Counts, Asta Roseway, Aaron Hoff, and Julia Schwarz. 2012. Tweeting is believing?: Understanding microblog credibility perceptions. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work*, CSCW '12, pages 441–450, Seattle, WA, USA.

Subhabrata Mukherjee and Gerhard Weikum. 2015. Leveraging joint interactions for credibility analysis in news communities. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, CIKM '15, pages 353–362, Melbourne, Australia.

Preslav Nakov, Doris Hoogeveen, Lluís Màrquez, Alessandro Moschitti, Hamdy Mubarak, Timothy Baldwin, and Karin Verspoor. 2017a. SemEval-2017 task 3: Community question answering. In *Proceedings of the 11th International Workshop on Semantic Evaluation*, SemEval '17, pages 27–48, Vancouver, Canada.

Preslav Nakov, Lluís Màrquez, Walid Magdy, Alessandro Moschitti, Jim Glass, and Bilal Randeree. 2015. SemEval-2015 task 3: Answer selection in community question answering. In *Proceedings of the 9th International Workshop on Semantic Evaluation*, pages 269–281, Denver, CO, USA.

Preslav Nakov, Lluís Màrquez, Alessandro Moschitti, Walid Magdy, Hamdy Mubarak, Abed Alhakim Freihat, Jim Glass, and Bilal Randeree. 2016a. SemEval-2016 task 3: Community question answering. In *Proceedings of the 10th International Workshop on Semantic Evaluation*, SemEval '16, pages 525–545, San Diego, CA, USA.

Preslav Nakov, Tsvetomila Mihaylova, Lluís Màrquez, Yashkumar Shiroya, and Ivan Koychev. 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*, RANLP '17, pages 551–560, Varna, Bulgaria.

Preslav Nakov, Alan Ritter, Sara Rosenthal, Fabrizio Sebastiani, and Veselin Stoyanov. 2016b. SemEval-2016 task 4: Sentiment analysis in Twitter. In *Proceedings of the 10th International Workshop on Semantic Evaluation*, SemEval '16, pages 1–18, San Diego, CA, USA.

An T. Nguyen, Aditya Kharosekar, Matthew Lease, and Byron C. Wallace. 2018. An interpretable joint graphical model for fact-checking from crowds. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, AAAI '18, New Orleans, LA, USA.

Piotr Niewiński, Aleksander Wawer, Maria Pszona, and Maria Janicka. 2019. TMLab SRPOL at SemEval-2019 Task 8: Fact checking in community question answering forums. In *Proceedings of the International Workshop on Semantic Evaluation*, SemEval '19, Minneapolis, MN, USA.

Jeff Z. Pan, Siyana Pavlova, Chenxi Li, Ningxi Li, Yangmei Li, and Jinshuo Liu. 2018. Content based fake news detection using knowledge graphs. In *Proceedings of the International Semantic Web Conference*, ISWC '18, pages 669–683, Monterey, CA, USA.

Dan Pelleg, Oleg Rokhlenko, Idan Szpektor, Eugene Agichtein, and Ido Guy. 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*, CSCW '16, pages 1080–1090, San Francisco, CA, USA.

Verónica Pérez-Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. 2018. Automatic detection of fake news. In *Proceedings of the 27th International Conference on Computational Linguistics*, COLING '18, pages 3391–3401, Santa Fe, NM, USA.

Kashyap Popat, Subhabrata Mukherjee, Jannik Strötgen, and Gerhard Weikum. 2016. Credibility assessment of textual claims on the web. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, CIKM '16, pages 2173–2178, Indianapolis, IN, USA.

Kashyap Popat, Subhabrata Mukherjee, Jannik Strötgen, and Gerhard Weikum. 2017. Where the truth lies: Explaining the credibility of emerging claims on the Web and social media. In *Proceedings of the 26th International Conference on World Wide Web Companion*, WWW '17, pages 1003–1012, Perth, Australia.

Sara Rosenthal, Noura Farra, and Preslav Nakov. 2017. SemEval-2017 task 4: Sentiment analysis in Twitter. In *Proceedings of the 11th Internationafon*, SemEval '17, pages 502–518, Vancouver, Canada.

Dominik Stammbach, Stalin Varanasi, and Guenter Neumann. 2019. DOMLIN at SemEval-2019 Task 8: Automated fact checking exploiting user ratings in community question answering forums. In *Proceedings of the International Workshop on Semantic Evaluation*, SemEval '19, Minneapolis, MN, USA.

Qi Su, Helen Kai yun Chen, and Chu-Ren Huang. 2010. Incorporate credibility into context for the best social media answers. In *Proceedings of the 24th Pacific Asia Conference on Language, Information and Computation*, PACLIC '10, pages 535–541, Sendai, Japan.

Bakhtiyar Syed, Vijayasaradhi Indurthi, Manish Shrivastava, Manish Gupta, and Vasudeva Varma. 2019. Fermi at SemEval-2019 task 8: An elementary but effective approach to question discernment in community qa forums. In *Proceedings of the International Workshop on Semantic Evaluation*, SemEval '19, Minneapolis, MN, USA.

Wanying Xie, Mengxi Que, Ruoyao Yang, Chunhua Liu, and Dong Yu. 2019. BLCU_NLP at SemEval-2019 task 8: A contextual knowledge-enhanced GPT model for fact checking. In *Proceedings of the International Workshop on Semantic Evaluation*, SemEval '19, Minneapolis, MN, USA.

Arkaitz Zubiaga, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Peter Tolmie. 2016. Analysing how people orient to and spread rumours in social media by looking at conversational threads. *PloS one*, 11(3):e0150989.