Do Not Trust the Trolls: Predicting Credibility in Community Question Answering Forums

Preslav Nakov[1], Tsvetomila Mihaylova[2], Lluís Màrquez[1], Yashkumar Shiroya[3] and Ivan Koychev[2]

[1]Qatar Computing Research Institute, HBKU, Doha, Qatar
{pnakov,lmarquez}@qf.org.qa
[2]Faculty of Mathematics and Informatics, Sofia University "St. Kliment Ohridski", Sofia, Bulgaria
tsvetomila.mihaylova@gmail.com, koychev@fmi.uni-sofia.bg
[3]Purdue University, United States
yshiroya@purdue.edu

Abstract

We address information credibility in community forums, in a setting in which the credibility of an answer posted in a question thread by a particular user has to be predicted. First, we motivate the problem and we create a publicly available annotated English corpus by crowdsourcing. Second, we propose a large set of features to predict the credibility of the answers. The features model the user, the answer, the question, the thread as a whole, and the interaction between them. Our experiments with ranking SVMs show that the credibility labels can be predicted with high performance according to several standard IR ranking metrics, thus supporting the potential usage of this layer of credibility information in practical applications. The features modeling the profile of the user (in particular *trollness*) turn out to be most important, but embedding features modeling the answer and the similarity between the question and the answer are also very relevant. Overall, half of the gap between the baseline performance and the perfect classifier can be covered using the proposed features.

1 Introduction

Community Question Answering (cQA) forums, such as StackOverflow, Yahoo! Answers and Quora are very popular these days, as they represent effective means for communities of users around particular topics to share information and to collectively solve their information needs. Recent research in natural language processing (NLP) and information retrieval (IR) has focused on automatically finding good answers to newly posed questions using preexisting question-answer threads. This typically requires finding related questions in the forum and ranking the answers according to their goodness for a particular question.

This is precisely the setting of the tasks on Community Question Answering at SemEval 2015 and 2016 (Nakov et al., 2015, 2016). These challenges provide benchmark datasets for the above tasks, with one subtask being specifically about classifying the answers in a question–answer thread as *good* or *bad* answers.

Here, we explore a new dimension in the context of cQA—that of the *credibility* of the answers for a particular question. This aspect is ignored, e.g., in recent cQA tasks at SemEval (Nakov et al., 2015, 2016), where the definition of a *Good* answer is very shallow: an answer is considered *Good* if it tries to answer the question, irrespective of its veracity, accuracy, etc. Figure 1 presents an excerpt of a real example from the Qatar Living forum, with one question and three answers selected from a longer thread. In the above SemEval tasks, all three answers are considered *Good* since they are formally answering the question. However, a_1 contains false information, while a_2 and a_3 are correct. In this case, the credibility of the latter two answers can be inferred from the fact that the "6 months" answer appears many times in the thread.

There are multiple factors explaining the presence of non-credible answers in cQA forums, e.g., misunderstanding of the question, ignorance or maliciousness of the responder, etc. In many cases, the forums are barely moderated and there is no quality control established. The interactions and discussions among the users are usually the means to filter out incorrect or inaccurate answers.

We believe that the *credibility* dimension of an answer is complementary to its *goodness*, i.e., as

Proceedings of Recent Advances in Natural Language Processing, pages 551–560,
Varna, Bulgaria, Sep 4–6 2017.
https://doi.org/10.26615/978-954-452-049-6_072

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.

a_1 above shows, an answer can be formally *Good*, but it could contain false information. Combining automatic detection of credibility and goodness would offer better experience to the users of cQA systems, e.g., a possible application scenario would be that in which the user is presented with a ranking of all good answers accompanied by credibility scores, where low scores would warn the user not to completely trust the answer or to double-check it.

Below we start by defining the credibility problem in cQA, and by creating an annotated corpus with data from the Qatar Living forum, extending the current annotation from SemEval-2016 Task 3. We specialize the former *Good* label into *Good-Credible* and *Good-NonCredible*, keeping the *Bad* answers unchanged.[1] We then develop a large variety of features to identify non-credible answers. Finally, we train ranking SVMs and we show that they can learn to rank the non-credible answers with performance that is significantly higher than the baselines and quite close to the theoretical upper bound, on a variety of standard IR measures.

Overall, the main contributions of this paper are threefold: (*i*) First, we look at credibility in cQA as a problem on its own right, and we create a new dataset that we release to the research community. To the best of our knowledge, this is the first publicly available dataset specifically targeting credibility in a cQA setting. (*ii*) We experiment with a large variety of features for the problem, some of which have not been compared in such a configuration before. Our features target the answer, the question, the thread as a whole, and the interaction between them. We show that the most relevant feature types are the user profile (e.g., *trollness* features) text embeddings and the similarity between the answer and the full answer-thread. (*iii*) We show that ranking-based SVMs can learn to rank non-credible answers with good performance. This supports our idea that modeling credibility on its own right can help cQA systems to refine a search that is based on more shallow answer-quality criteria (as the *goodness* from the cQA tasks at SemEval).

[1] *Bad* answers should be ranked lower than *Good* ones in any reasonable scenario; they probably should not be presented to the user at all. Thus, it does not make sense to further try to distinguish between credible and non-credible *Bad* answers.

2 Related Work

In the context of cQA and general Question Answering (QA), credibility has not been studied on its own right, but rather as a feature to improve good answer identification. Thus, it is typically modeled at the feature level, e.g., Jurczyk and Agichtein (2007) model author authority using link analysis. Similarly, Agichtein et al. (2008) look for high-quality answers in *Yahoo! Answers* by modeling author authority with PageRank and HITS, in addition to other information sources such as 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). Su et al. (2010) use verbs and adjectives that cast doubt on an answer, e.g., *doubt*, *possibly*. Lita et al. (2005) study three qualitative dimensions for answers: source credibility (e.g., does the document come from a government website), sentiment analysis, and potential contradiction compared to other answers. Banerjee and Han (2009) use language modeling for answer validation for QA, which quantifies the reliability of a source document that contains a candidate answer. Jeon et al. (2006) use non-textual features such as click counts, answers activity level, and copy counts. Finally, Pelleg et al. (2016) present a large-scale user study of automatically curating social media content in real time using a combination of syntactic, semantic, and social signals. Unlike this line of research, here we (*i*) study credibility as a task in its own right, (*ii*) using a specialized dataset, and (*iii*) a much richer feature set. As mentioned above, we assume a setting in which credibility is a complementary aspect to answer quality, which can be useful for users in practical application scenarios.

Information credibility has been also studied in the area of social computing. For instance, Castillo et al. (2011) formulate it as a problem of finding false information about a newsworthy event. They compiled their own dataset, focusing on tweets using variety of features including user reputation, author writing style, and various time-based features. We use some of the features they have proposed; yet, their work is not about QA or cQA. Canini et al. (2011) perform a similar study of the interaction of content and social network structure, and Morris et al. (2012) look into how Twitter users judge truthfulness.

Rumor detection in social media represents yet another angle of information credibility. Zubiaga et al. (2015) studied how people handle rumors in social media, and found that users with higher reputation are more trusted, and thus can spread rumors easily. Zubiaga et al. (2016) also studied the spread of rumors in social media but with focus on conversational threads. Lukasik et al. (2015) and Ma et al. (2015) use temporal patterns of rumor dynamics to detect rumors and to predict their frequency. The interested reader can also see (Zaharia et al., 2010) for a review of methods to detect fake news, including linguistic analysis, discourse, linked data, and social network features.

Finally, there is a recent survey on the assessment and ranking methodologies for user-generated content on the Web, which covers credibility and related topics (Momeni et al., 2015). Several truth discovery algorithms are studied and combined in an ensemble method for veracity estimation in the VERA system (Ba et al., 2016).

3 A New Credibility Corpus

We annotated with credibility judgments data from the Qatar Living forum,[2] using questions from the raw unlabeled data that the organizers of SemEval-2016 Task 3 made available,[3] while preserving the original format with all available metadata. This data is organized into question-answer threads, where 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, and user name and ID.

Following the setup of SemEval-2016 Task 3, we selected new questions with at least ten answers. Each answer has a subject, a body, and meta information: answer ID, user ID, and user name. We annotated the first ten answers in a thread with two labels: (*i*) goodness (*Good* vs. *Bad*), i.e., whether this answer tries to answer the question, and (*ii*) credibility (*Credible* vs. *NonCredible*), i.e., whether the answer is credible.

For the *goodness* labels we stick to the definition from SemEval-2016 Task 3, which is agnostic with respect to answer's credibility or veracity: an answer is considered *Good* if "the answer or a portion of it directly answers at least one subquestion of the target question".[4] Regarding *credibility*, we define an answer *Credible* if "the information in the question's thread and/or world knowledge and/or our common sense tells us that the answer is (somewhat) credible". Otherwise, we consider it *NonCredible*.

We used CrowdFlower[5] to obtain five annotations per example. In order to stress the difference between *goodness* and *credibility*, we adopted a three-label annotation schema: *Good-Credible*, *Good-NonCredible*, and *Bad*. In this way, we made sure that the annotators did not confuse credibility and goodness. We annotated a total of 476 questions and 4,760 answers; we further used 24 questions and 240 answers as hidden tests to ensure quality[6]. The inter-annotator agreement in terms of Fleiss' Kappa (Fleiss, 1971) was 0.6245, which corresponds to substantial agreement (Landis and Koch, 1977)

Finally, we converted the 3-way annotations into (*i*) goodness and (*ii*) credibility labels. For goodness, if there were three or more out of five votes for *Bad*, we assigned *Bad*; otherwise, we assigned *Good*. For those examples that were labeled *Good*, we further assigned a credibility label as follows: we set the value to *NonCredible* if at least two annotators assigned *Good-NonCredible*; otherwise, we assigned *Credible*. The rationale here is that since in our application scenario we do not envision to use the credibility information to filter out non-credible answers but to provide extra information to the user, we are interested in characterizing any answer that has a reasonable

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

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

[4] Questions in the Qatar Living forums can present long stories with multiple embedded subquestions.

[5] CrowdFlower offers a service which allows users to access an online workforce to clean, label and enrich data: https://www.crowdflower.com/

[6] CrowdFlower allows importing of gold-label examples in order to verify that the crowd-annotated labels are of good quality

	Ques-tions	Good Answers		Bad answ.
		Credible	*NonCredible*	
TRAIN	376	1,733	73	1,954
DEV	50	137	15	348
TEST	50	213	19	268

Table 1: Statistics about our credibility datasets.

Q: "I need to renew my passport very soon but the Qatari visa stamped on it will expire in 2008. I wonder, what happens to the visa when I get a new passport? Do they need to duplicate it on the new passport or can I just carry old and new passport together when traveling? Does anyone know?"

a_1: "It did happen to me.. but I honestly dont remeber what I did... well I'm not sure I think you keep your old passport with you, for the time being. thats it.. u still need to check on this. You could pop this question to the embassy.. I'm sure the'll help." *NonCredible*

a_2: "They will ussually clip the old and new passport together, that haapened to me when i have to get a new passport coz my old one is full, so everytime i travel my old and new passport are cliped together..." *Credible*

a_3: "You will have to get a new visa stamped on teh new passport. You cannot use your old passport." *Credible*

Figure 2: Example for answers annotated as *Credible* and *NonCredible* from the Qatar Living forum.

chance to be non-credible.[7] We selected randomly 50 questions for dev and for test and used their answers as examples for the classification. The answers of the remaining questions are used for training. Table 1 shows some statistics about the resulting credibility datasets [8] and Figure3 shows an example of credible and non-credible answers.

4 Features

Below we describe the types of features we use.

4.1 Answer Features

CREDIBILITY. (*31 features*) We have features that model the contents of the answer, most of which have been previously used for credibility detection (Castillo et al., 2011): number of URLs/images/emails/phone numbers; number of tokens/sentences; average number of tokens; number of nouns/verbs/adjectives/adverbs/pronouns; number of 1st/2nd/3rd person pronouns; number of positive/negative smileys; number of single/double/triple exclamation/ interrogation symbols; number of interrogative sentences (based on syntactic analysis); number of words that are not in word2vec's Google News vocabulary (this can signal slang, foreign language, etc.)

SENTIMENT (*36 features*) We extract features modeling the sentiment polarity of the answer, which has been previously proposed as a useful feature for credibility (Castillo et al., 2011). We use two sentiment polarity lexicons (Mohammad et al., 2013): the *NRC Hashtag Sentiment Lexicon*, which contains 54,129 words and 316,531 bigrams, and the *Sentiment140 Lexicon*, with 62,468 words and 677,698 bigrams. In these lexicons, for each term there is a real number representing the strength of association of the term with positive/negative sentiment. We use as features the number of positive/negative terms in the answer, both as absolute numbers and normalized by the total number of sentiment-bearing terms in the answer. We also have as features the sum of the scores for the positive/negative/all sentiment-bearing terms. Finally, we have the maximum absolute value for a positive/negative term. We have four copies of these nine features: for words vs. bigrams, and for each of the two lexicons.

GOODNESS (*9 features*) Similarly, we build goodness polarity lexicons that contain 41,633 words, each associated with a real number representing its strength of association with *Good* or *Bad* answers. Following (Balchev et al., 2016), we build this lexicon using pointwise mutual information, starting with the training data from SemEval-2016 task 3, and then extending this to words from the Qatar Living dump. We use the same nine features as for sentiment, but this time we only have one lexicon and we only use words (no bigrams).

GOOGLE_VEC (*300 features*) We use the pre-trained, 300-dimensional embedding vectors that Tomas Mikolov trained on 100 billion words from Google News (Mikolov et al., 2013). We compute a vector representation for an answer by simply averaging the embeddings of the words it contains.

QL_VEC (*100 features*) We train 100-dimensional in-domain word embeddings using WORD2VEC on all the available Qatar Living data, which we then use to produce embeddings for the

[7] Note that the annotations from all annotators are included in the corpus as complementary information. Thus, other more strict mappings can be considered from the users' annotations to the credibility labels depending on the final application objective.

[8] The full corpus can be found at the following address: https://bitbucket.org/cqa-credibility/cqa-credibility-corpus

answers by averaging the embedding vectors of the answer's words.

SYNTAX_VEC (*25 features*) We parse the answer using the Stanford neural parser (Socher et al., 2013), and we use the final 25-dimensional syntactic embedding vector that is produced internally as a by-product of parsing as a representation for the answer.

4.2 Question-Answer Features

These features measure the similarity between the question and the answer.

MTFEATS (*6 features*) We use the following six machine translation evaluation features: (*i*) BLEU: This is the most commonly used measure for machine translation evaluation, which is based on n-gram overlap and length ratios (Papineni et al., 2002). (*ii*) NIST: This measure is similar to BLEU, and is used at evaluation campaigns run by NIST (Doddington, 2002). (*iii*) TER: Translation error rate; it is based on the edit distance between a translation hypothesis and the reference (Snover et al., 2006). (*v*) Unigram PRECISION and RECALL, which originally come from information retrieval.

BLEUCOMP (*17 features*) We further use as features various components that are involved in the computation of BLEU: n-gram precisions, n-gram matches, total number of n-grams (n=1,2,3,4), lengths of the hypotheses and of the reference, length ratio between them, and BLEU's brevity penalty.

VEC_COSINES (*3 features*) We calculate pairwise similarity features between an answer and the corresponding question using their GOOGLE_VEC, QL_VEC, and SYNTAX_VEC vectors.

4.3 Thread-Answer Features

RANK (*4 features*) We have thread-level features related to the rank of the answer in the thread: (*i*) reciprocal rank of the answer in the thread; (*ii*) percentile of the answer in the thread, calculated as follows: the first answer gets the score of 1.0, the second one gets 0.9, the next one gets 0.8, and so on. We calculate these two features twice: once for the full list of answers, and once for the list of *Good* answers only.

VEC_COS_THREAD (*3 features*) We further use embeddings at the thread-level, which we calculate over the concatenation of all *Good* answers in the thread. The idea is that if a *Good* answer is similar to other *Good* answers, it is likely to be credible; conversely, if it is dissimilar, it is likely to be an outlier, and thus less credible. We use as features the cosines between an answer- and a thread-vector using GOOGLE_VEC, QL_VEC, and SYNTAX_VEC vectors.

4.4 User Profile Features

We further have some features characterizing the user who has posted an answer.

CATEGORIES (*396 features*) We build a vector of the number of answers a user has posted in each of the 197 categories. We have each feature twice: once as a raw feature and once normalized by the total number of answers the user has posted. We further add as features the total number of answers and the number of distinct categories the user has posted in.

QUALITY (*13 features*) We model the quality of the posts by the authors. We first use the SemEval-2016 Task 3 data (Nakov et al., 2016) to train a classifier that predicts whether a given answer is a *Good* answer to the question heading its thread or not. We then run this classifier (which has 80+% accuracy) on the entire Qatar Living dataset dump, and we aggregate its predictions to estimate whether a given user tends to give *Good* answers or not. We have the following features for each user: number of *Good*/*Bad* answers, total number of answers, percentage of *Good*/*Bad* answers, sum of the classifier probabilities for *Good*/*Bad* answers, total sum of the classifier probabilities over all answers, average score for the probability of *Good*/*Bad* answer, and highest absolute score for the probability of *Good*/*Bad* answer.

TROLLNESS (*27 features*) These features model the likelihood that the author of the answer is a troll. They are inspired by the trollness definition proposed by Mihaylov et al. (2015), namely that a person who is called a troll by other users is likely to be one. In particular, we have the following features: number of answers that are exactly k (k=1,2,. . .,10) answers before a troll mention, i.e., an answer that contains words like *troll*, *trolls*, *trolling*, and the number of answers that come within [0;n] answers before a troll

mention ($n = 3, 5$). These 12 features have two versions each: once as absolute numbers and once normalized by the total number N of answers the user has posted in a thread and they were followed later by a troll mention. This total number is also a feature. Another feature is the average distance of the user's post to a troll mention that comes somewhere below in the thread. Finally, we have a feature that measures the average distance not in terms of number of answers but in terms of average time (days). For users who have never posted answers in a trollness context, the values of these features are zero.

ACTIVITY (*19 features*) These features describe the overall activity of the user (regardless of the quality of the answers and without the requirement for them to appear in a troll context). We use features such as number of answers posted, number of distinct questions to which an answer was posted, number of questions asked, number of posts in the *Job*s, 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 in different hourly periods (note that these intervals overlap): during working hours (7:00-17:00h), after work, at night, early in the morning, and before noon. We further 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 $k \in \{1, 3, 5, 10, 20\}$.

5 Experiments

General setup. We experimented with each feature type individually, where the different feature types as well as the features inside larger groups are ordered by their relative MAP (Mean Average Precision) scores. We further combined the best k feature types. The results are shown in Table 2.

Scoring. As we imagine a ranking application scenario, we are interested in ranking evaluation metrics, such as Mean Average Precision (MAP), Mean Reciprocal Rank (MRR) and Average Recall (AvgRec), which were used at SemEval-2016 Task 3. Note that we use the minority class, i.e., *NonCredible*, as the positive class.

Baselines. The last rows of Table 2 show the performance of two baselines. The first one is the chronological ranking, where the answers are ordered by their time of posting; the rationale here is that later answers might be more credible as over time people tend to gradually converge towards consensus answers. The second baseline classifies all answers as *NonCredible*.

Upper bound. Coming back to ranking, we can see that the scores for MAP and MRR seem quite low, both for the baselines and for the systems trained using various features. The reason for that is shown in the *Oracle* row of the table: we can see that an oracle system (i.e., one that assigns the correct *Credible*/*NonCredible* label and also assigns such scores that rank all *NonCredible* answers above all *Credible* ones) only achieves a MAP of 0.2273 and an MRR of 22.73. These numbers are an upper bound of what we could possibly achieve. The reason for MAP and MRR being so low is that they are zero by definition when a thread has no *NonCredible* answers, and in our test dataset 34 of the 50 questions have no *NonCredible* answers (moreover, six of the questions have only *Bad* answers), which pushes the scores down.

Learning algorithm. We used an SVM-rank (Joachims, 2002). We scaled the feature weights, and we experimented with linear and RBF kernels, using grid search to find the best values for the SVM hyper-parameters C and γ.

6 Results and Discussion

Overall, the most important category of features are those modeling the user profile. In particular, this category contains the TROLLNESS feature, which achieves the best results in terms of MAP, AvgRec, and MRR. This shows that users that are seen as trolls by other users in one context, tend to give generally noncredible answers. The success of this feature is somewhat surprising. First, there is no guarantee that a mere mention of words such as *troll*, *trolls*, or *trolling* means that one user accuses some of the previous users who posted in the same thread to be trolls; the word might refer to users in some other thread. Yet, a quick manual analysis of threads containing a troll word shows that most mentions are indeed troll accusations. The problem is that not all users who posted before such an accusation are its target; there are many innocents.[9] Apparently, this is not a big problem,

[9]For an illustration, see an example of a thread discussing trolls in Qatar Living here:
http://www.qatarliving.com/forum/qatar-living-lounge/posts/beware-trolls

Features	MAP	AvgRec	MRR	P	R	F1	Acc
USER PROFILE FEATURES							
TROLLNESS	**0.1739**	**0.9119**	**19.5076**	**0.3333**	0.1579	**0.2143**	0.9052
QUALITY	**0.1598**	**0.9024**	**17.0455**	**0.5000**	0.0526	0.0952	**0.9181**
ACTIVITY	**0.1391**	0.8331	**14.3939**	0	0	0	**0.9181**
CATEGORIES	0.1230	0.8168	12.0265	**0.1250**	0.1053	0.1143	0.8664
ANSWER FEATURES							
SYNTAX_VEC	**0.1657**	**0.8814**	**17.4242**	**0.2400**	0.3158	**0.2727**	0.8621
CREDIBILITY	**0.1538**	**0.8829**	**15.5682**	**0.0938**	0.1579	0.1176	0.8060
SENTIMENT	**0.1447**	**0.8373**	**16.0227**	**0.1176**	0.2105	0.1509	0.8060
GOODNESS	**0.1337**	0.8068	**13.4280**	**0.2000**	0.0526	0.0833	0.9052
GOOGLE_VEC	**0.1331**	**0.8436**	**13.7121**	0.0714	0.1053	0.0851	0.8147
QL_VEC	0.1234	0.7895	**13.5417**	0.0606	0.1053	0.0769	0.7931
QUESTION-ANSWER FEATURES							
VEC_COSINES	**0.1631**	**0.8478**	**16.5404**	0	0	0	**0.9181**
BLEUCOMP	**0.1426**	**0.8436**	**14.5833**	0	0	0	**0.9181**
MTFEATS	**0.1341**	0.8162	**15.7197**	0	0	0	**0.9181**
ANSWER-THREAD FEATURES							
RANK	**0.1512**	**0.8484**	**15.6061**	0	0	0	**0.9181**
VEC_COSINES_THREAD	**0.1433**	0.8173	**14.2424**	0	0	0	0.9138
THREAD FEATURES							
GOOGLE_VEC_THREAD	**0.1307**	**0.8384**	13.1439	**0.1875**	0.3158	**0.2353**	0.8319
SYNTAX_VEC_THREAD	**0.1307**	**0.8384**	13.1439	**0.1163**	0.2632	**0.1613**	0.7759
QL_VEC_THREAD	**0.1307**	**0.8384**	13.1439	**0.0909**	0.2105	0.1270	0.7629
COMBINATIONS							
TOP-2	**0.1857**	**0.9230**	**18.5606**	**0.1912**	0.6842	**0.2989**	0.7371
TOP-4	**0.1698**	**0.9024**	**17.8030**	**0.2571**	0.4737	**0.3333**	0.8448
TOP-6	**0.1888**	**0.9345**	**19.3182**	0	0	0	**0.9181**
UPPER BOUND							
Oracle	0.2273	1.0000	22.7273	1.0000	1.0000	1.0000	1.0000
BASELINES							
Chronological	0.1307	0.8384	13.1400	—	—	—	—
Random	0.1263	0.8111	13.2386	0.0726	0.4737	0.1259	0.4612
All-Credible	—	—	—	0	0	0	**0.9181**
All-NonCredible	—	—	—	0.0819	**1.0000**	0.1514	0.0819

Table 2: **Evaluation results.** We show the performance for each group of features in isolation, and for the combination of the top-k features groups, as well as for four baselines and an oracle upper bound. For each evaluation measure, we underline the best baseline score, and we mark in bold all system results that are higher than or equal to that score. We further underline the best overall result for each column.

as we model trollness using a number of features, and these only get high values if a user appears many times in a troll-accusation context. Another problem is that we have less than 3,000 users who appeared before a trollness accusation, while there are close to 70,000 users in QatarLiving. Yet, many of those for which we have trollness features, are also those that are among the most active users and thus are likely to be the authors of our test-time answers. Overall, the TROLLNESS feature group has the second-highest precision of 0.3333 among all feature groups we experimented with.

The highest overall precision is achieved by another user profile feature group: QUALITY. The goal here is again to find unreliable users, but the way this is achieved is more indirect: it is hypothesized that users who gave mostly bad answers in the past (bad in the sense that they did not try to answer the question, e.g., because they instead engaged in conversation with users, changed topic, started asking new questions, etc.), should not be believed when they actually give seemingly good answers to some question later. This feature has very low recall though.

The second most important feature category is that of answer features. Interestingly, we see at the top SYNTAX_VEC, which is second-best overall on MAP and MRR, but also notably the best in F_1. This suggests that the syntactic structure of an answer is important for human judges when suggesting that a answer is not credible. Indeed, previous work on finding high-quality content in social media has made use of grammaticality as a

feature (Agichtein et al., 2008). However, there it was modeled using part-of-speech n-grams, while here we use syntactic answer embeddings.

Naturally, among the top-performing features in this category we find the CREDIBILITY group, which contains some features that have been previously proposed for credibility, but in social media (Castillo et al., 2011). Another strong feature group from this category is SENTIMENT, which has been shown to be useful for credibility.

The third most important feature category in terms of performance is that of the question-answer features. This is to be expected as answers are posted with respect to a question and thus their credibility should take the question into account. The best feature group here is VEC_COSINES; this should not be surprising given the strong performance of SYNTAX_VEC, which is used for one of the three cosines. BLEUCOMP is also relatively strong, which is an indicator of the importance of modeling n-gram overlaps between the question and the answer directly (modeling similarity indirectly as in MTFEATS performs somewhat worse). Finally note that, even though relatively good at ranking, the feature groups in this category never predict NonCredible as a label.

Next in terms of importance comes the answer-thread feature category. We can see that modeling RANK is somewhat important, e.g., maybe because early answers are more likely to be credible as they are more likely to be on topic. Another indication of the importance of the relative ranking of a answer in the thread is the fact that the chronological baselines is a bit better than the random one. The cosines between the vector of a answer and of the corresponding thread, or VEC_COSINES_THREAD, performs relatively well, as it models whether the answer is similar to the set of the other good answers in the thread. The idea is that if several answers say similar things, they should reinforce each other's credibility. Finally, this feature category also cannot predict NonCredible as a label.

The last group of features is that of thread-level feature vectors. Obviously, they are not strong enough in isolation, and perform roughly at the baseline level in terms of ranking measures; yet, they are above the baseline in terms of precision and F_1.

Finally, as we mentioned above, we further combined the prediction scores for the best k feature groups in a meta classifier. This yielded additional improvements, e.g., MAP improved from 0.1739 to 0.1888, AvgRec from 0.9119 to 0.9345, and F_1 from 0.2727 to 0.3333.

7 Conclusion and Future Work

In this paper, we have addressed *information credibility in community Question Answering* as a problem on its own right. To the best of our knowledge, this is done for a first time. We have motivated the problem in the context of answer-quality ranking, and we have created a publicly available corpus, again for the first time for this task. We have also proposed a large set of relatively cheap features, which we used to train ranking SVM classifiers to predict the credibility of an answer with respect to a question in the context of a question-answer thread. The features model the user, the answer, the question, the thread as a whole, and the interaction between them. Our experimental evaluation demonstrate sizable improvements over the baselines across several standard IR ranking-based metrics, which shows that the credibility annotation is indeed learnable. The results further show that features modeling the profile of the user (in particular *trollness*) are the most important for detecting answer credibility. The feature groups based on *semantic similarity of the answer to its associated question as well as to the entire thread* also proved to be relevant. Overall, more than 70% of the gap between the baseline performance and the perfect Oracle classifier (in terms of MAP scores) could be covered by a combination of the most productive feature types. This results support the idea of using the credibility prediction layer in a real-world cQA scenario.

In future work, we plan to enlarge the training and the testing datasets in order to avoid overfitting and to get more reliable conclusions about the utility of our large set of features. In doing so, we also plan to use more complex feature selection algorithms. From machine learning perspective, we are also interested in exploring other approaches, such as deep convolutional neural networks or long short-term memory (LSTM), in order to obtain better embedded representations and to model the structure of the data more adequately, and semi-supervised learning, e.g., exploiting self-training from the entire Qatar Living forum, as a way to partially avoid the need for costly supervision.

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*. Palo Alto, California, USA, WSDM '08, pages 183–194.

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*. Montréal, Québec, Canada, WWW '16 Companion, pages 159–162.

Daniel Balchev, Yasen Kiprov, Ivan Koychev, and Preslav Nakov. 2016. PMI-cool at SemEval-2016 task 3: Experiments with PMI and goodness polarity lexicons for community question answering. In *Proceedings of the 10th International Workshop on Semantic Evaluation*. San Diego, California, SemEval '2016, pages 844–850.

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, Companion Volume: Short Papers*. Association for Computational Linguistics, Boulder, Colorado, pages 157–160.

K. R. Canini, B. Suh, and P. L. Pirolli. 2011. Finding credible information sources in social networks based on content and social structure. In *Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on*. pages 1–8.

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

George Doddington. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the Second International Conference on Human Language Technology Research*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, HLT '02, pages 138–145.

Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin* 76(5):378–382.

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*. ACM, New York, NY, USA, SIGIR '06, pages 228–235.

Thorsten Joachims. 2002. Optimizing search engines using clickthrough data. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Edmonton, Alberta, Canada, KDD '02, pages 133–142.

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*. ACM, New York, NY, USA, CIKM '07, pages 919–922.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics* 33(1):159–174.

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*. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, volume 20, page 1616.

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 (Volume 2: Short Papers)*. Association for Computational Linguistics, Beijing, China, pages 518–523.

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*. Melbourne, Australia, CIKM '15, pages 1751–1754.

Todor Mihaylov, Georgi D Georgiev, AD Ontotext, and Preslav Nakov. 2015. Finding opinion manipulation trolls in news community forums. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning, CoNLL*. volume 15, pages 310–314.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 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*. Atlanta, Georgia, USA, NAACL-HLT '13, pages 746–751.

Saif Mohammad, Svetlana Kiritchenko, and Xiaodan Zhu. 2013. NRC-Canada: Building the state-of-the-art in sentiment analysis of tweets. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation*. Atlanta, Georgia, USA, SemEval '2013, pages 321–327.

Elaheh Momeni, Claire Cardie, and Nicholas Diakopoulos. 2015. A survey on assessment and ranking methodologies for user-generated content on the web. *ACM Comput. Surv.* 48(3):41:1–41:49.

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*. ACM, New York, NY, USA, CSCW '12, pages 441–450.

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 (SemEval 2015)*. Association for Computational Linguistics, Denver, Colorado, pages 269–281.

Preslav Nakov, Lluís Màrquez, Alessandro Moschitti, Walid Magdy, Hamdy Mubarak, Abed Alhakim Freihat, Jim Glass, and Bilal Randeree. 2016. SemEval-2016 Task 3: Community Question Answering. In *Proceedings of SemEval-2016*. Association for Computational Linguistics, San Diego, California, SemEval '16.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of 40th Annual Meting of the Association for Computational Linguistics*. Philadelphia, Pennsylvania, USA, ACL '02, pages 311–318.

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*. ACM, New York, NY, USA, CSCW '16, pages 1080–1090.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Biennial Conference of the Association for Machine Translation in the Americas*. Cambridge, Massachusetts, USA, AMTA '06.

Richard Socher, John Bauer, Christopher D. Manning, and Ng Andrew Y. 2013. Parsing with compositional vector grammars. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Sofia, Bulgaria, pages 455–465.

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*. Institute of Digital Enhancement of Cognitive Processing, Waseda University, Tohoku University, Sendai, Japan, pages 535–541.

Matei Zaharia, Mosharaf Chowdhury, Michael J. Franklin, Scott Shenker, and Ion Stoica. 2010. Spark: Cluster computing with working sets. In *Proceedings of the 2nd USENIX Conference on Hot Topics in Cloud Computing*. Boston, MA, HotCloud '10, pages 10–10.

Arkaitz Zubiaga, Geraldine Wong Sak Hoi, Maria Liakata, Rob Procter, and Peter Tolmie. 2015. Analysing how people orient to and spread rumours in social media by looking at conversational threads. *arXiv preprint arXiv:1511.07487* .

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):1–29.