CrowdQM: Learning aspect-level user reliability and comment trustworthiness in discussion forums

Alex Morales*, Kanika Narang*, Hari Sundaram, and Chengxiang Zhai

University of Illinois at Urbana-Champaign, Urbana, IL, USA
{amorale4,knarang2,hs1,czhai}@illinois.edu

Abstract. Community discussion forums are increasingly used to seek advice; however, they often contain conflicting and unreliable information. Truth discovery models estimate source reliability and infer information trustworthiness simultaneously in a mutual reinforcement manner, and can be used to distinguish trustworthy comments with no supervision. However, they do not capture the diversity of word expressions and learn a single reliability score for the user. CrowdQM addresses these limitations by modeling the fine-grained aspect-level reliability of users and incorporate semantic similarity between words to learn a latent trustworthy comment embedding. We apply our latent trustworthy comment for comment ranking for three diverse communities in Reddit and show consistent improvement over non-aspect based approaches. We also show qualitative results on learned reliability scores and word embeddings by our model.

1 Introduction

Users are increasingly turning to community discussion forums to solicit domain expertise, such as querying about inscrutable political events on history forums or posting a health-related issue to seek medical suggestions or diagnosis. While these forums may be useful, due to almost no regulations on post requirements or user background, most responses contain conflicting and unreliable information [10]. This misinformation could lead to severe consequences, especially in health-related forums, that outweigh the positive benefits of these communities. Currently, most of the forums either employ moderators to curate the content or use community voting. However, both of these methods are not scalable [8]. This creates a dire need for an automated mechanism to estimate the trustworthiness of the responses in the online forums.

In general, the answers written by reliable users tend to be more trustworthy, while the users who have written trustworthy answers are more likely to be reliable. This mutual reinforcement, also referred to as the truth discovery principle, is leveraged by previous works that attempt to learn information trustworthiness in the presence of noisy information sources with promising results [28,26,7,6]. This data-driven principle particularly works for community forums as they tend to be of large scale and exhibit redundancy in the posts and comments.

However, a significant deficiency of previous work is the lack of aspect-level modeling of a user's reliability. Community discussion forums usually encompass various

* Equal Contribution.

topics or aspects. This heterogeneity is especially true for discussion forums, like Reddit, with communities catering to broad themes; while within each community, questions span a diverse range of sub-topics. Intuitively, a user's reliability will be limited to only a few topics, for instance, in a science forum, a biologist could be highly knowledgeable, and in turn reliable, when she answers biology or chemistry-related questions but may not be competent enough for linguistic queries.

Another challenge is the diversity of word expressions in the responses. Truth discovery based approaches treat each response as categorical data. However, in discussion forums, users' text responses can include contextually correlated comments [27]. For instance, in the *context* of a post describing symptoms like "headache" and "fever", either of the related responses of a viral fever or an allergic reaction can be a correct diagnosis. On the other hand, unrelated comments in the post should be unreliable; for instance, a comment giving a diagnosis of "bone fracture" for the above symptoms.

CrowdQM addresses both limitations by jointly modeling the aspect-level user reliability and latent trustworthy comment in an optimization framework. In particular, 1) CrowdQM learns user reliability over fine-grained topics discussed in the forum. 2) Our model captures the semantic meaning of comments and posts through word embeddings. We learn a trustworthy comment embedding for each post, such that it is semantically similar to comments of reliable users on the post and also similar to the post's context. Contrary to the earlier approaches [1,2,18], we propose an *unsupervised model* for comment trustworthiness that does not need labeled training data.

We verified our proposed model on the comment ranking task based on trustworthiness for three Ask* *subreddit communities*. Our model outperforms state-of-the-art baselines in identifying the most trustworthy responses, deemed by community experts and community consensus. We also show the effectiveness of our aspect-based user reliability estimation and word embeddings qualitatively. Further, our improved model of reliability enables us to identify reliable users per aspect discussed in the community.

2 Methodology

A challenge in applying truth discovery to discussion forums is capturing the variation in user's reliability and the diversity of word usage in the answers. To address it, we model aspect-level user reliability and use semantic representations for the comments.

2.1 Problem Formulation

Each *submission* is a post, i.e., question, which starts a discussion thread while a *comment* is a response to a submission post. Formally, each submission post, m, is associated with a set of terms, c_m. A user, n, may reply with a comment on submission m, with a set of terms $w_{m,n}$. $\mathcal{V}$ is the vocabulary set comprising of all terms present in our dataset i.e. all submissions and comments. Each term, $\omega \in \mathcal{V}$ has a corresponding word-vector representation, or word embedding, $\boldsymbol{v}_\omega \in \mathbb{R}^D$. Thus, we can represent a post in terms of its constituent terms' embeddings, $\{\boldsymbol{v}_c\}, \forall c \in c_m$. We treat *post embeddings* as static in our model. To capture the semantic meaning, we represent each

comment as the mean word-vector representation of their constituent terms[1]. Formally, we represent the comment given on the post m by user n as the *comment embeddings*, $\boldsymbol{a}_{m,n} = |w_{m,n}|^{-1} \sum_{\omega \in w_{m,n}} \boldsymbol{v}_\omega$. The set of posts user n has commented on is denoted by $\mathcal{M}_n$ and the set of users who have posted on submission m is denoted as $\mathcal{N}_m$.

There are K aspects or topics discussed in the forum, and each post and comment can be composed of multiple *aspects*. We denote submission m's distribution over these aspects as the *post-aspect distribution*, $\boldsymbol{p}_m \in \mathbb{R}^K$. Similarly, we also compute, *user-aspect distribution*, $\boldsymbol{u}_n \in \mathbb{R}^K$, learned over all the comments posted by the user n in the forum. This distribution captures familiarity (or frequency) of user n with each aspect based on their activity in the forum. Each user n also has a *user reliability* vector defined over K aspects, $\boldsymbol{r}_n \in \mathbb{R}^K$. The reliability captures the likelihood of the user providing a trustworthy comment about a specific aspect. Note high familiarity in an aspect does not always imply high reliability in the same aspect.

For each submission post m associated with a set of responses $\{\boldsymbol{a}_{m,n}\}$, our goal is to estimate the real-valued vector representations, or *latent trustworthy comment* embeddings, $\boldsymbol{a}^*_m \in \mathbb{R}^D$. We also simultaneously infer the *user reliability* vector $\{\boldsymbol{r}_n\}$ and update the word embeddings $\{\boldsymbol{v}_\omega\}$. The latent trustworthy comment embeddings, $\boldsymbol{a}^*_m$, can be used to rank current comments on the post.

2.2 Proposed Method

Our model follows the truth discovery principle: trustworthy comment is supported by many reliable users and vice-versa. In other words, the weighted error between the trustworthy comment and the given comments on the post is minimum, where user reliabilities provide the weight. We extend the approach to use an aspect-level user reliability and compute a post-specific reliability weight. We further compute the error in terms of the *embeddings* of posts and comments to capture their semantic meaning.

In particular, we minimize the *embedding error*, $E_{m,n} = ||\boldsymbol{a}^*_m - \boldsymbol{a}_{m,n}||^2$, i.e., mean squared error between learned *trustworthy comment* embeddings, $\boldsymbol{a}^*_m$ and comment embeddings, $\boldsymbol{a}_{m,n}$, on the post m. This error ensures that the trustworthy comment is semantically similar to the comments given for the post.

Next, to ensure context similarity of the comments with the post, we compute the *context error*, $Q_{m,n} = |c_m|^{-1} \sum_{c \in c_m} ||\boldsymbol{a}_{m,n} - \boldsymbol{v}_c||^2$, reducing the difference between the *comment embeddings* and *post embeddings*. The key idea is similar to that of the distributional hypothesis that if two comments co-occur a lot in similar posts, they should be closer in the embedding space.

Further, these errors are weighted by the aspect-level reliability of the user providing the comment. We estimate the reliability of user n for the specific post m through the *user-post reliability* score, $R_{m,n} = \boldsymbol{r}_n \odot s(\boldsymbol{u}_n, \boldsymbol{p}_m) = \sum_k \boldsymbol{r}^{(k)}_n \cdot (\boldsymbol{u}^{(k)}_n \cdot \boldsymbol{p}^{(k)}_m)$. $\odot$ represents the Hadamard product. This scores computes the magnitude of *user reliability* vector, $\boldsymbol{r_n}$, weighted by the similarity function $s(.)$. The similarity function $s(\boldsymbol{u}_n, \boldsymbol{p}_m)$ captures user familiarity with post's context by computing the product of the aspect

[1] Sentence, and furthermore document representation is a complex problem. In our work, we explore a simple aggregation method for comment semantic composition [23].

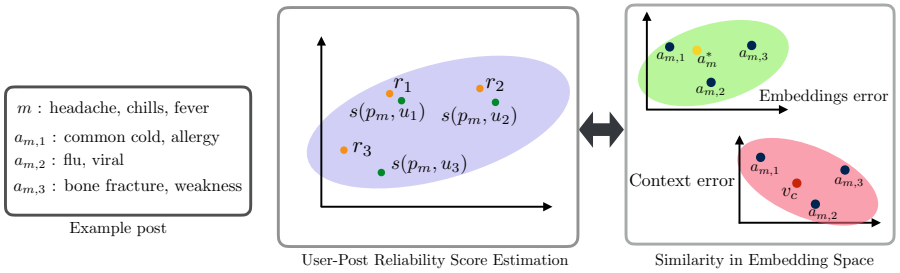


Fig. 1: An illustrative toy example detailing our model components. The left-hand side details the user-post reliability score estimation, $R_{m,n}$, that is a function of similarity function $s(.)$ between the user and post aspect distributions and user aspect reliabilities, $\boldsymbol{r}_n$. In the right-hand, we learn trustworthy comment embedding, $\boldsymbol{a}^*_m$, such that it is similar to user comments, $\boldsymbol{a}_{m,n}$ which are, in turn, similar to the post context $\boldsymbol{v}_c$.

distribution of the user n and the post m. Thus, to get a high *user-post reliability* score, $R_{m,n}$, the user should both be reliable and familiar to the aspects discussed in the post.

Finally, these errors are aggregated over all the users and their comments. Thus, we define our objective function as follows,

$$\min_{\{\boldsymbol{a}^*_m\},\{\boldsymbol{v}_\omega\},\{\boldsymbol{r}_n\}} \sum_{n=1}^{N} \sum_{m \in \mathcal{M}_n} \underbrace{R_{m,n}}_{\text{user-post reliability}} \left(\underbrace{E_{m,n}}_{\text{embedding error}} + \beta \odot \underbrace{Q_{m,n}}_{\text{context error}} \right) \quad \text{s.t.} \sum_{n=1}^{N} e^{-\boldsymbol{r}_n^{(k)}} = 1; \forall k \tag{1}$$

where $\odot$ represents the Hadamard product. $R_{m,n} \cdot E_{m,n}$ ensures that the latent trustworthy comment embeddings are most similar to comment embeddings of *reliable* users for post m. While $R_{m,n} \cdot Q_{m,n}$ ensures trust aware learning of contextualized comment embeddings. The hyperparameter β controls the importance of context error in our method. The exponential regularization constraint, $\sum_{n=1}^{N} e^{-\boldsymbol{r}_n^{(k)}} = 1$ for each k, ensures that the reliability across users are nonzero. Figure 1 shows the overview of our model using a toy example of a post in a medical forum with flu-like symptoms. The commenters describing flu-related diagnoses are deemed more reliable for this post.

2.3 Solving the Optimization Problem

We use coordinate descent [3] to solve our optimization problem. In particular, we solve the equation for each variable while keeping the rest fixed.

Case 1: Fixing $\{\boldsymbol{r}_n\}$ and $\{\boldsymbol{v}_\omega\}$, we have the following update equation for $\{\boldsymbol{a}^*_m\}$:

$$\boldsymbol{a}^*_m = \frac{\sum_{n \in \mathcal{N}_m} R_{m,n} \boldsymbol{a}_{m,n}}{\sum_{n \in \mathcal{N}_m} R_{m,n}} \tag{2}$$

Thus, the latent *trustworthy comment* is a weighted combination of comments where weights are provided by the *user-post reliability* score $R_{m,n}$. Alternatively, it can also be interpreted as a reliable summarization of all the comments.

Case 2: Fixing $\{\boldsymbol{a}_m^*\}$, $\{\boldsymbol{v}_\omega\}$, we have the following update equation for $\{\boldsymbol{r}_n^{(k)}\}$:

$$\boldsymbol{r}_n^{(k)} \propto -\ln \sum_{m \in \mathcal{M}_n} s(\boldsymbol{u}_n^{(k)}, \boldsymbol{p}_m^{(k)}) \left(E_{m,n} + \beta Q_{m,n}\right) \tag{3}$$

Reliability of a user in aspect k is inversely proportional to the errors with respect to the latent trustworthy comment $\boldsymbol{a}_m^*$ ($E_{m,n}$) and submission's context $\boldsymbol{v}_c$ ($Q_{m,n}$) over all of her posted comments ($\mathcal{M}_n$). The embedding error ensures that if there is a large difference between the user's comment and the trustworthy comment, her reliability becomes lower. The context error ensures that non-relevant comments to the post's context are penalized heavily. In other words, a reliable user should give trustworthy and contextualized responses to posts.

This error is further weighed by the similarity score, $s(.)$, capturing familiarity of the user with the post's context. Thus, familiar users are penalized higher for their mistakes as compared to unfamiliar users.

Case 3: Fixing $\{\boldsymbol{a}_m^*\}, \{\boldsymbol{r}_n^{(k)}\}$, we have the following update equation for $\{\boldsymbol{v}_\omega\}$:

$$\boldsymbol{v}_\omega = \frac{\sum_{<m,n> \in D_\omega} R_{m,n} \left(\boldsymbol{a}_m^* + \beta |c_m|^{-1} \sum_{c \in c_m} \boldsymbol{v}_c\right) - R_{m,n}(\beta+1)|c_m|^{-1} \boldsymbol{a}_{m,n}^{-\omega}}{\sum_{<m,n> \in D_\omega} R_{m,n}(\beta+1)} \tag{4}$$

where $< m, n > \in D_\omega = \{(m,n) | \omega \in w_{m,n}\}$ and $\boldsymbol{a}_{m,n}^{-\omega} = |w_{m,n}|^{-1} \sum_{\omega' \in w_{m,n} \setminus \{\omega\}} \boldsymbol{v}_{\omega'}$. To update $\boldsymbol{v}_\omega$, we only consider those comment and submission pairs, D_ω, in which the particular word appears. The update of the embeddings depend on the submission context $\boldsymbol{v}_c$, latent trustworthy comment embedding, $\boldsymbol{a}_m^*$ as well as *user-post reliability* score, $R_{m,n}$. Thus, word embeddings are updated in a trust-aware manner such that reliable user's comments weigh more than those of unreliable users as they can contain noisy text. Note that there is also some negative dependency on the contribution of other terms in the comments.

Implementation Details: We used popular Latent Dirichlet Allocation (LDA) [4] to estimate aspects of the posts in our dataset[2]. Specifically, we combined the title and body text to represent each post. We applied topic model inference to all comments of user n to compute its combined aspect distribution, $\boldsymbol{u}_n$. We randomly initialized the user reliability, $\boldsymbol{r}_n$. We initialized the word embeddings, $\boldsymbol{v}_\omega$, via word2vec [19] trained on our dataset. We used both unigrams and bigrams in our model. We fixed β to 0.15.[3] The model converges after only about six iterations indicating quick approximation. In general, the computational complexity is $O(|\mathcal{V}|NM)$; however, we leverage the data sparsity in the comment-word usage and user-posts for efficient implementation.

3 Experiments

In this section, we first discuss our novel dataset, followed by experiments on the outputs learned by our model. In particular, we evaluate the trustworthy comment embeddings on the comment ranking task while we qualitatively evaluate user reliabilities and word embeddings. For brevity, we focus the qualitative analysis on our largest subreddit, askscience.

[2] We ran LDA with 50 topics for all experiments and examined its sensitivity in Section 3.2.

[3] We did not find a significant change in results for different values of β.

3.1 Dataset

We evaluate our model on a widely popular discussion forum Reddit. Reddit covers diverse topics of discussion and is challenging due to the prevalence of noisy responses. We specifically tested on *Ask** subreddits as they are primarily used to seek answers to a variety of topics from mundane issues to serious medical concerns. In particular, we crawled data from three subreddits, /r/askscience, /r/AskHistorians, and /r/AskDocs from their inception until October 2017 [4]. While these subreddits share the same platform, the communities differ vastly, see Table 1. We preprocessed the data by removing uninformative comments and posts with either less than ten characters or containing only URLs or with a missing title or author information. We removed users who have posted less than two comments and also submissions with three or fewer comments. To handle sparsity, we treated all users with a single comment as "UNK".

Table 1: Dataset statistics for the subreddit communities. N and M denotes total users and posts respectively; N_e: number of experts; $|\boldsymbol{a}_{m,e}|$: number of posts with at least one expert comment; $|w_{m,n}|$: average comment word length.

Dataset	Created	N	N_e	M	$\lvert\boldsymbol{a}_{m,e}\rvert$	$\lvert w_{m,n}\rvert$
*Docs	07/13	3,334	286	17,342	10,389	53.5
*Science	04/10	73,463	2,195	100,237	70,108	74.0
*Historians	08/11	27,264	296	45,650	30,268	103.4

For each submission post, there is an associated flair text denoting the *category* of the post, referred to as the *submission flair* that is either Moderator added or self-annotated,e.g., Physics, Chemistry, Biology. Similarly, users have *author flairs* attributed next to their user-name describing their educational background, e.g., Astrophysicist, Bioengineering. Only users verified by the moderator have *author flairs*, and we denote them as experts in the rest of the paper. AskDocs does not have submission flairs as it is a smaller community. For both subreddits, we observed that around 80% of the users comment on posts from more than two categories.

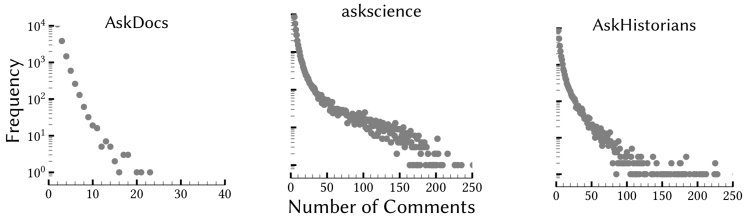


Fig. 2: Frequency plot (log scale) of number of comments per post for three subreddits. A post on AskDocs tend to have fewer comments than the other two communities.

Experts are highly active in the community answering around 60-70% of the posts (Table 1). askscience and AskHistorians have significantly higher (Figure 2) and more detailed comments ($|w_{m,n}|$ in Table 1) per post than AskDocs. Due to the prevalence of a large number of comments, manual curation is very expensive, thus necessitating the need for an automatic tool to infer comments trustworthiness.

[4] praw.readthedocs.io/en/latest/

3.2 Trustworthy Comment Embedding Analysis

We evaluate latent trustworthy comment learned by our model on a trustworthy comment ranking task. That is, given a submission post, our goal is to rank the posted comment based on their trustworthiness. For this experiment, we treat expert users' comment as the most trustworthy comment of the post.[5] Besides, we also report results using the highest upvoted comment as the gold standard. Highest upvoted comments represent community consensus on the most trustworthy response for the post [16]. In particular, we rank comments for each post m, in the order of descending cosine similarity between their embedding, $\boldsymbol{a}_{m,n}$, and the latent trustworthy comment embeddings, $\boldsymbol{a}^*_m$. We then report average Precison@k values over all the posts, where k denotes the position in the output ranked list of comments.

Baselines: We compare our model with state-of-the-art truth discovery methods proposed for continuous and text data and non-aspect version of our model [6].

Mean Bag of Answers (MBoA): In this baseline, we represent the trustworthy comment for a post as the mean comment embedding and thus assume uniform user reliability.

CRH: is a popular truth discovery-based model for numerical data [12]. CRH minimizes the weighted deviation of the trustworthy comment embedding from the individual comment embeddings with user reliabilities providing the weights.

CATD: is an extension of CRH that learns a confidence interval over user reliabilities to handle data skewness [11]. For both the above models, we represent each comment as the average word embeddings of its constituent terms.

TrustAnswer: Li et al. [14] modeled semantic similarity between comments by representing each comment with embeddings of its key phrase.

CrowdQM-no-aspect: In this baseline, we condense the user's aspect reliabilities to a single $\boldsymbol{r}_n$. This model acts as a control to gauge the performance of our proposed model.

Results: Table 2a reports the Precision@1 results using expert's comments as the gold standard. MBoA, with uniform source reliability, outperforms the CRH method that estimates reliability for each user separately. Thus, simple mean embeddings provide a robust representation for the trustworthy comment.

We also observe that CrowdQM-no-aspect performs consistently better than TrustAnswer. Note that both approaches do not model aspect-level user reliability but use semantic representations of comments. However, while TrustAnswer assigns a single reliability score for each comment, CrowdQM-no-aspect additionally takes into account the user's familiarity with the post's context (*similarity* function, $s(.)$) to compute her reliability for the post. Finally, CrowdQM consistently outperforms both the models, indicating that aspect modeling is beneficial.

CATD uses a confidence-aware approach to handle data skewness and performs the best among the baselines. This skewness is especially helpful in Reddit as experts are the most active users (Table 1); and, CATD likely assigns them high reliability.

[5] While human judgment would be the most precise; it is also the most challenging to collect. For instance, in askscience we would need experts in over 35 science fields, reading up to 250 comments for a single post.

[6] Note that there is no label information used, so we cannot compare to other supervised CQA models [1,24,21] which need this supervision. Our *unsupervised model* is complementary to these approaches, and thus, a rigorous comparison is impossible.

Table 2: Precision@1 for all three Ask* subreddits, with (2a) the experts' comments and (2b) upvotes used to identify trustworthy comments.

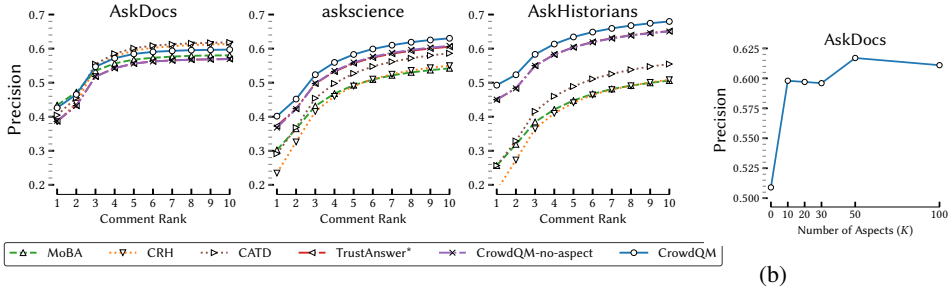
(a)

Model	*Docs	*Science	*Historians
MBoA	0.592	0.633	0.602
CRH [12]	0.585	0.597	0.556
CATD [11]	**0.635**	0.700	0.669
TrustAnswer [14]	0.501	0.657	0.637
CrowdQM-no-aspect	0.509	0.666	0.640
CrowdQM	0.617	**0.734**	**0.753**

(b)

Model	*Docs	*Science	*Historians
MBoA	**0.434**	0.302	0.257
CRH [12]	0.386	0.234	0.183
CATD [11]	0.405	0.291	0.257
TrustAnswer [14]	0.386	0.373	0.449
CrowdQM-no-aspect	0.388	0.368	0.450
CrowdQM	0.426	**0.402**	**0.493**

Our model achieves competitive precision as CATD for AskDocs while outperforming for the others. This indicates that our data-driven model works better for communities which are less sparse (Section 3.1 and Figure 2).



(a) (b)

Fig. 3: Precision of our model (3a) vs. comment rank computed by user's upvotes and (3b) vs. number of aspects. Our model outperforms the baselines for askscience and AskHistorians while performs similarly for AskDocs. Value of K does not have much impact on the precision value.

Table 2b reports Precision@1 results using community upvoted comments as the gold standard, while Figure 3a plots the precision values against the size of the output ranked comment list. In general, there is a drop in performance for all models on this metric because it is harder to predict upvotes as they are inherently noisy [8]. TrustAnswer and CrowdQM-no-aspect perform best among the baselines indicating that modeling semantic representation is essential for forums. CrowdQM again consistently outperforms the non-aspect based models verifying that aspect modeling is needed to identify trustworthy comments in forums. CrowdQM remains competitive in the smaller AskDocs dataset, where the best performing model is MoBA. Thus, for AskDocs, the comment summarizing all other comments tends to get the highest votes.

Parameter Sensitivity In Figure 3b, we plot our model's precision with varying number of aspects. Although there is an optimal range around 50 aspects, the precision remains relatively stable indicating that our model is not sensitive to aspects.[7] We also did similar analysis with β and did not find any significant changes to the Precision.

[7] We also observed similar results for the other datasets and omitted those figures for lack of space.

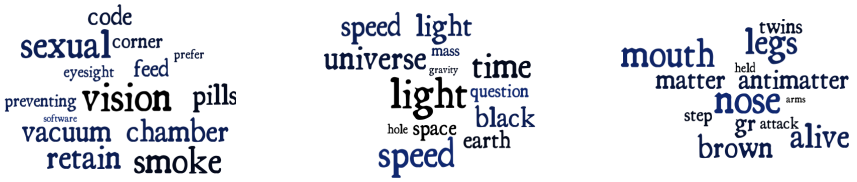
3.3 Aspect Reliability Analysis

We evaluate learned user reliabilities through users commenting on a post with a *submission flair*. Note that a submission flair is manually curated and denotes post's category, and this information is not used in our model. Specifically, for each post m, we compute the *user-post reliability* score, $R_{m,n}$, for every user n who commented on the post. We then ranked these scores for each category and report top *author flairs* for few categories in Table 3. The top *author flairs* for each category are domain experts.

Table 3: Top author flairs with their corresponding post categories.

Post Category: Computing	**Post Category:Linguistics**
Embedded Systems, Software Engineering, Robotics	Linguistics, Hispanic Sociolinguistics
Computer Science	Comparative Political Behaviour
Quantum Optics, Singular Optics	Historical Linguistics, Language Documentation
Robotics, Machine Learning, Computer Vision, Manipulators	Linguistics, Hispanic Sociolinguistics
Computer Science	Historical Linguistics, Language Documentation
Biomechanical Engineering, Biomaterials	Nanostructured Materials, Heterogeneous Catalysis
Post Category: Biology	**Post Category: Psychology**
Animal Cognition	Clinical Psychology, Psychotherapy, Behavior Analysis
Cell and Developmental Biology	International Relations, Comparative Politics
Biochemistry, Molecular Biology, Enzymology	Neuropsychology
Genetics, Cell biology, Bioengineering	Psychology, PTSD, Trauma, and Resilience
Computational Physics, Biological Physics	Cognitive Neuroscience, Neuroimaging, fMRI
Aquatic Ecology and Evolution, Active Acoustics	Psychology, Legal psychology, Eyewitness testimonies

For instance, for the Computing category highly reliable users have author flairs like Software Engineering and Machine Learning, while for Linguistics authors with flairs Hispanic Sociolinguistics and Language Documentation rank high. These results align with our hypothesis that in-domain experts should have higher reliabilities. We also observe out of domain authors with flairs like Comparative Political Behavior and Nanostructured Materials in the Linguistic category. This diversity could be due to the interdisciplinary nature of the domain. Our model, thus, can be used by the moderators of the discussion forum to identify and recommend potential reliable users to respond to new submission posts of a particular category.



(a) Health (b) Cosmos (c) Oceanography

Fig. 4: Top words for highly correlated aspects between user reliability and user karma.

To further analyze the user reliability, we qualitatively examine the aspects with the largest reliability value of highly upvoted users in a post category. First, we identify users deemed reliable by the community for a category through a *karma* score. Category-specific user *karma* is given by the average upvotes the user's comments have received in the category. We then correlate the category-specific user *karma* with her reliability score in each $k \in K$ aspect, $\boldsymbol{r}_n^{(k)}$ to identify aspects relevant for that category.

Figure 4 shows the top words of the highest correlated aspects for some categories. The identified words are topically relevant thus our model associates aspect level user reliability coherently. Interestingly, the aspects themselves tend to encompass several themes, for example, in the Health category, the themes are software and health.

3.4 Word Embedding Analysis

The CrowdQM model updates word embeddings to better model semantic meaning of the comments. For each category, we identify the frequent terms and find its most similar keywords using cosine distance between the learned word embeddings.

Table 4: Similar words using embeddings learned using CrowdQM for askscience.

Liquid		**Cancer**		**Quantum**		**Life**	
Initial	CrowdQM	Initial	CrowdQM	Initial	CrowdQM	Initial	CrowdQM
unimaginably	gas	mg	disease	search results	model	molaison	species
bigger so	chemical	curie	white	sis	energy	around	natural
two lenses	solid	wobbly	cell	shallower water	particle	machos	nature
orbiting around	air	subject	food	starts rolling	mechanics	brain	production
fire itself	material	”yes” then	complete	antimatter galaxies	mathematical	”dark” matter	size

The left column for each term in Table 4 are the most similar terms returned by the initial embeddings while the right column reports the results from updated embeddings $\{\boldsymbol{v}_\omega\}$ from our CrowdQM model. We observe that there is a lot of noise in words returned by the initial model as they are just co-occurrence based while words returned by our model are semantically similar and describe similar concepts. This improvement is because our model updates word embeddings in a trust aware manner such that they are similar to terms used in responses from reliable users.

4 Related Work

Our work is related to two main themes of research, truth discovery and community question answering (CQA).

Truth Discovery : Truth discovery has attracted much attention recently. Different approaches have been proposed to address different scenarios [29,13,20]. Most of the truth discovery approaches are tailored to categorical data and thus assume there is a single objective truth that can be derived from the claims of different sources [15]. Faitcrowd [17] assumes an objective truth in the answer set and uses a probabilistic generative model to perform fine-grained truth discovery. On the other hand, Wan et al. [22] propose trustworthy *opinion* discovery where the true value of an entity is modeled as a random variable with a probability density function instead of a single value. However, it still fails to capture the semantic similarity between the textual responses. Some truth discovery approaches also leverage text data to identify correct responses effectively. Li et al. [14] proposed a model for capturing semantic meanings of crowd provided diagnosis in a Chinese medical forum. Zhang et al. [27] also leveraged semantic representation of answers and proposed a Bayesian approach to capture the multifactorial property of text answers. These approaches only use certain keywords to represent each answer and are thus, limited in their scope. Also, they learn a scalar

user reliability score. To the best of our knowledge, there has been no work that models both fine-grained user reliability with semantic representations of the text to discover trustworthy comments from community responses.

Community Question Answering: Typically CQA is framed as a classification problem to predict correct responses for a post. Most of the previous work can be categorized into feature-based or text relevance-based approaches. Feature-driven models [5,1,9] extract content or user based features that are fed into classifiers to identify the best comment. CQARank leverages voting information as well as user history and estimates user interests and expertise on different topics [25]. Barron-Cedeno et al. [2] also look at the relationship between the answers, measuring textual and structural similarities between them to classify useful and relevant answers. Text-based deep learning models learn an optimal representation of question and answer pairs to identify the most relevant answer [24]. In SemEval 2017 task on CQA, Nakov et al. [21] developed a task to recommend related answers to a new question in the forum. SemEval 2019 further extends this line of work by proposing fact checking in community question answering [18]. It is not only expensive to curate each reply manually to train these models, but also unsustainable. On the contrary, CrowdQM is an unsupervised method and thus does not require any labeled data. Also, we estimate the comments' *trustworthiness* that implicitly assumes relevance to the post (modeled by these works).

5 Conclusion

We proposed an unsupervised model to learn a trustworthy comment embedding from all the given comments for each post in a discussion forum. The learned embedding can be further used to rank the comments for that post. We explored Reddit, a novel community discussion forum dataset for this task. Reddit is challenging as posts typically receive a large number of responses from a diverse set of users and each user engages in a wide range of topics. Our model estimates aspect-level user reliability and semantic representation of each comment simultaneously. Experiments show that modeling aspect level user reliability improves the prediction performance compared to the non-aspect version of our model. We also show that the estimated user-post reliability can be used to identify trustworthy users for particular post categories.

References

1. Agichtein, E., Castillo, C., Donato, D., Gionis, A., Mishne, G.: Finding high-quality content in social media. In: WSDM (2008)
2. Barrón-Cedeno, A., Filice, S., Da San Martino, G., Joty, S.R., Màrquez, L., Nakov, P., Moschitti, A.: Thread-level information for comment classification in community question answering. In: ACL (2015)
3. Bertsekas, D.P.: Nonlinear programming. Athena scientific Belmont (1999)
4. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. JMLR **3**(Jan) (2003)
5. Burel, G., Mulholland, P., Alani, H.: Structural normalisation methods for improving best answer identification in question answering communities. In: WWW (2016)
6. Dong, X.L., Berti-Equille, L., Srivastava, D.: Integrating conflicting data: The role of source dependence. Proc. VLDB Endow. **2**(1) (Aug 2009)

7. Galland, A., Abiteboul, S., Marian, A., Senellart, P.: Corroborating information from disagreeing views. In: WSDM (2010)
8. Gilbert, E.: Widespread underprovision on reddit. In: CSCW (2013)
9. Jenders, M., Krestel, R., Naumann, F.: Which answer is best?: Predicting accepted answers in MOOC forums. In: International Conference on World Wide Web (2016)
10. Li, G., Wang, J., Zheng, Y., Franklin, M.: Crowdsourced data management: A survey. In: IEEE ICDE (2017)
11. Li, Q., Li, Y., Gao, J., Su, L., Zhao, B., Demirbas, M., Fan, W., Han, J.: A confidence-aware approach for truth discovery on long-tail data. VLDB **8**(4) (2014)
12. Li, Q., Li, Y., Gao, J., Zhao, B., Fan, W., Han, J.: Resolving conflicts in heterogeneous data by truth discovery and source reliability estimation. In: ACM SIGMOD (2014)
13. Li, Q., Ma, F., Gao, J., Su, L., Quinn, C.J.: Crowdsourcing high quality labels with a tight budget. In: WSDM (2016)
14. Li, Y., Du, N., Liu, C., Xie, Y., Fan, W., Li, Q., Gao, J., Sun, H.: Reliable medical diagnosis from crowdsourcing: Discover trustworthy answers from non-experts. In: WSDM (2017)
15. Li, Y., Gao, J., Meng, C., Li, Q., Su, L., Zhao, B., Fan, W., Han, J.: A survey on truth discovery. ACM Sigkdd Explorations Newsletter **17**(2) (2016)
16. Lyu, S., Ouyang, W., Wang, Y., Shen, H., Cheng, X.: What we vote for? answer selection from user expertise view in community question answering. In: WWW (2019)
17. Ma, F., Li, Y., Li, Q., Qiu, M., Gao, J., Zhi, S., Su, L., Zhao, B., Ji, H., Han, J.: Faitcrowd: Fine grained truth discovery for crowdsourced data aggregation. In: KDD (2015)
18. Mihaylova, T., Karadzhov, G., Atanasova, P., Baly, R., Mohtarami, M., Nakov, P.: Semeval-2019 task 8: Fact checking in community question answering. In: International Workshop on Semantic Evaluation (2019)
19. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: NIPS (2013)
20. Mukherjee, T., Parajuli, B., Kumar, P., Pasiliao, E.: Truthcore: Non-parametric estimation of truth from a collection of authoritative sources. In: IEE BigData (2016)
21. Nakov, P., Hoogeveen, D., Màrquez, L., Moschitti, A., Mubarak, H., Baldwin, T., Verspoor, K.: Semeval-2017 task 3: Community question answering. In: International Workshop on Semantic Evaluation (2017)
22. Wan, M., Chen, X., Kaplan, L., Han, J., Gao, J., Zhao, B.: From truth discovery to trustworthy opinion discovery: An uncertainty-aware quantitative modeling approach. In: KDD (2016)
23. Wang, Z., Mi, H., Ittycheriah, A.: Sentence similarity learning by lexical decomposition and composition. arXiv preprint (2016)
24. Wen, J., Ma, J., Feng, Y., Zhong, M.: Hybrid attentive answer selection in cqa with deep users modelling. In: AAAI Conference on Artificial Intelligence (2018)
25. Yang, L., Qiu, M., Gottipati, S., Zhu, F., Jiang, J., Sun, H., Chen, Z.: Cqarank: jointly model topics and expertise in community question answering. In: CIKM (2013)
26. Yin, X., Han, J., Yu, P.S.: Truth discovery with multiple conflicting information providers on the web. In: KDD (2007)
27. Zhang, H., Li, Y., Ma, F., Gao, J., Su, L.: Texttruth: An unsupervised approach to discover trustworthy information from multi-sourced text data. In: KDD (2018)
28. Zhao, B., Han, J.: A probabilistic model for estimating real-valued truth from conflicting sources. In: Proc. of QDB (2012)
29. Zheng, Y., Li, G., Li, Y., Shan, C., Cheng, R.: Truth inference in crowdsourcing: is the problem solved? Proceedings of the VLDB Endowment **10**(5) (2017)