What Are People Asking About COVID-19? A Question Classification Dataset

Jerry Wei[♠] **Chengyu Huang**[♦] **Soroush Vosoughi**[♥] **Jason Wei**[♥]
[♠]Protago Labs [♦]International Monetary Fund [♥]Dartmouth College
jerry.weng.wei@protagolabs.com
huangchengyu24@gmail.com
{soroush, jason.20}@dartmouth.edu

Abstract

We present COVID-Q, a set of 1,690 questions about COVID-19 from 13 sources, which we annotate into 15 question categories and 207 question classes. The most common questions in our dataset asked about transmission, prevention, and societal effects of COVID, and we found that many questions that appeared in multiple sources were not answered by any FAQ websites of reputable organizations such as the CDC and FDA. We post our dataset publicly at https://github.com/JerryWei03/COVID-Q.

For classifying questions into 15 categories, a BERT baseline scored 58.1% accuracy when trained on 20 examples per class, and for classifying questions into 89 question classes, the baseline achieved 54.6% accuracy. We hope COVID-Q can be helpful either for direct use in developing applied systems or as a domain-specific resource for model evaluation.

1 Introduction

A major challenge during fast-developing pandemics such as COVID-19 is keeping people updated with the latest and most relevant information. Since the beginning of COVID, several reputable websites have maintained frequently asked questions (FAQ) pages that they regularly update. But even so, users might struggle to find their questions on FAQ sites, and many common questions about COVID remain unanswered. In this paper, we ask—what are people really asking about COVID, and how can we use NLP to better understand questions and retrieve relevant content?

We present COVID-Q, a dataset of 1,690 questions about COVID from 13 online sources. We annotate COVID-Q by classifying questions into 15 question categories[1] (see Figure 1) and by grouping questions that ask the same thing into question classes, for a total of 207 classes. Throughout §2, we analyze the distribution of COVID-Q in terms of question class, category, and source.

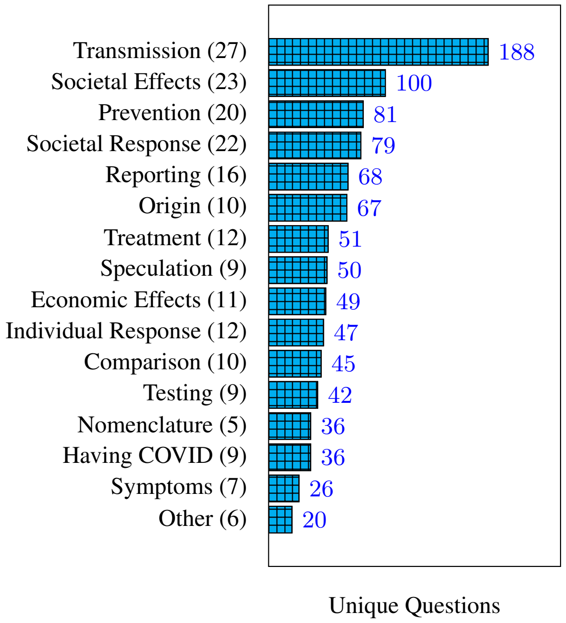

Figure 1: Distribution of question categories, with number of question classes per category shown in parenthesis, in our dataset.

COVID-Q facilitates several question understanding tasks. First, the question categories can be used as a vanilla text classification task to determine the general category of information a question is asking about. Second, the question classes can be used for retrieval question answering, where a system has a database of questions and answers and when given a new question, finds the question in the database that asks the same thing and returns the corresponding answer (Romeo et al., 2016; Sakata et al., 2019). We provide baselines for these two tasks in §3.1 and §3.2. In addition to being directly used for developing an applied system, COVID-Q could also serve as a domain-specific resource for evaluating NLP models trained on COVID data.

[1] We do not count the "other" category.

Source	Total	Questions Matched	Unmatched	Answers	Questions Removed
Quora	675	501 (74.2%)	174 (25.8%)	0	374
Google Search	173	161 (93.1%)	12 (6.9%)	0	174
github.com/deepset-ai/COVID-QA	124	55 (44.4%)	69 (55.6%)	124	71
Yahoo Search	94	87 (92.6%)	7 (7.4%)	0	34
*Center for Disease Control	92	51 (55.4%)	41 (44.6%)	92	1
Bing Search	68	65 (95.6%)	3 (4.4%)	0	29
*Cable News Network	64	48 (75.0%)	16 (25.0%)	64	1
*Food and Drug Administration	57	33 (57.9%)	24 (42.1%)	57	3
Yahoo Answers	28	13 (46.4%)	15 (53.6%)	0	23
*Illinois Department of Public Health	20	18 (90.0%)	2 (10.0%)	20	0
*United Nations	19	18 (94.7%)	1 (5.3%)	19	6
*Washington DC Area Television Station	16	15 (93.8%)	1 (6.2%)	16	0
*Johns Hopkins University	11	10 (90.9%)	1 (9.1%)	11	1
Author Generated	249	249 (100.0%)	0 (0.0%)	0	0
Total	1,690	1,324 (78.3%)	366 (21.7%)	403	717

Table 1: Distribution of questions in COVID-Q by source. The reported number of questions excludes unrelated, vague, and nonsensical questions that were removed. * denotes sources for which questions came from FAQ pages.

2 Dataset Collection and Annotation

Data collection. To collect the data, we scraped questions about COVID from thirteen sources: seven official FAQ websites from recognized organizations such as the Center for Disease Control (CDC) and the Food and Drug Administration (FDA), and six crowd-based sources such as Quora and Yahoo Answers. Table 1 shows the distribution of collected questions from each source. We also post the original scraped websites for each source.

Data cleaning. We performed several preprocessing steps to remove unrelated, low-quality, and nonsensical questions. First, we deleted questions unrelated to COVID and vague questions that have too many interpretations (e.g., "Why COVID?"). Second, we remove location-specific and time-specific versions of questions (e.g., "COVID deaths in New York" and "COVID deaths in California"), since these questions do not contribute linguistic novelty (you could replace "New York" with any state, for example). Questions that only focused on one specific location or time, however, were not removed—for instance, "Was China responsible for COVID?" was not removed because no other questions asked about any other specific country being responsible for the pandemic. Finally, to minimize occurrences of questions that trivially differ, we remove all punctuation and replace synonymous ways of saying COVID, such as "coronavirus," "COVID-19," and "COVID19," with "covid." Table 1 also shows the number of removed questions for each source.

Question Class [#Questions] (Category)	Example Questions
Pandemic Duration [28] (Speculation)	"Will COVID ever go away?" "Will COVID end soon?" "When COVID will end?"
Demographics: General [26] (Transmission)	"Who is at higher risk?" "Are kids more at risk?" "Who is COVID killing?"
Survivability: Surfaces [24] (Transmission)	"Does COVID live on surfaces?" "Can COVID live on paper?" "Can COVID live on objects?"

Table 2: Most common question classes in COVID-Q.

Data annotation. We first annotated our dataset by grouping questions that asked the same thing together into question classes. The first author manually compared each question with existing classes and questions. When it was unclear whether two questions asked the same thing, we used the more formal definition that two questions would belong in the same class if they have the same answer. In other words, two questions were matched to the same question class if and only if they could be answered with a common answer. Because every new example in our dataset is checked against all question classes and all unmatched questions, the time complexity for annotating our dataset is $O(n^2)$, where n is the number of questions.

After all questions were grouped into question classes, the first author gave each question class with at least two questions a name summarizing the questions in that class, and each question class was assigned to one of 15 question categories (as shown in Figure 1), which were generated during a thorough discussion with the last author. In Table 2, we show the question classes with the most questions, along with their assigned question categories and some example questions. In Figure 2, we show the distribution of question classes.

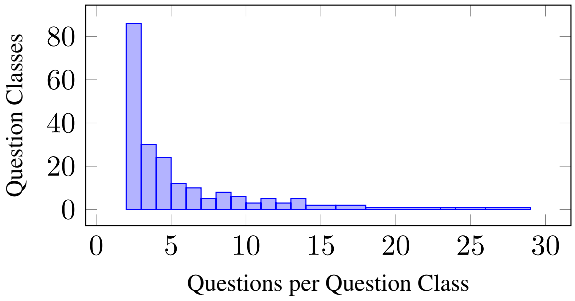

Figure 2: Number of questions per question class for question classes with at least two questions. All questions in a question class asked roughly the same thing. There are 120 question classes with at least 3 questions per class, 66 classes with at least 5 questions per class, and 22 classes with at least 10 questions per class.

Annotation quality. We ran the dataset through multiple annotators to improve the quality of our labels. First, the last author confirmed all labels in the dataset, highlighting any questions that might need to be relabeled and discussing them with the first author. Of the 1, 245 matched questions, 131 questions were highlighted and 67 labels were modified. As a second pass, an external annotator read through the labels in a similar fashion, for which 31 questions were highlighted and 15 labels were modified. Most modifications in questions labels involved separating a single question class that was too broad into more specific classes.

For another round of validation, we showed 3 questions from each of the 89 question classes with $N_{class} \geq 4$ to three Mechanical Turk workers, who were asked to select the correct question class from five choices. The majority vote from the three workers agreed with our ground-truth labels 93.3% of the time. The three workers unanimously agreed on 58.1% of the questions, for which 99.4% of these unanimous labels agreed with our ground-truth label. Workers were paid \$0.07 per question.

Unmatched questions. Interestingly, we observe that for the CDC and FDA frequently asked questions websites, a sizable fraction of their questions (44.6% for CDC and 42.1% for FDA) did not match to questions from any other source, suggesting that these sources might want adjust the questions on their websites to question classes that were seen frequently in search engines such as Google or Bing. Moreover, 54.2% of question classes that had questions from at least two non-official sources went unanswered by an official source. Table 3 shows examples of these questions. Conversely, in Table 8 (Supplementary Materials), we show questions from CDC and FDA that were not found in any other source and their closest matches computed using BERT.

Question Class	N_{class}	Example Questions
Number of Cases	21	“Are COVID cases dropping?” “Have COVID cases peaked?” “Are COVID cases decreasing?”
Mutation	19	“Has COVID mutated?” “Did COVID mutate?” “Will COVID mutate?”
Lab Theory	18	“Was COVID made in a lab?” “Was COVID manufactured?” “Did COVID start in a lab?”

Table 3: Questions appearing in multiple sources that were unanswered by official FAQ websites.

3 Question Classification

Here, we provide baselines for *question-category classification*, where each question belongs to one of 15 categories, and *question-class classification*, where questions asking the same thing belong to the same class (of 89 question-classes).

As our dataset is small when split into training and test sets, we manually generate an additional *author-generated* evaluation set of 249 questions. For these questions, the first author wrote new questions for question classes with 4 or 5 questions per class until those question classes had 6 questions per class. These questions were checked in the same fashion as the real questions, but for clarity, we only refer to them in this section (§3) unless explicitly stated.

3.1 Question-Category Classification

The *question-category classification* task assigns each question to one of the 15 categories shown in Figure 1. For the train-test split, we randomly choose 20 questions per category for the training set (as the smallest category has 26 questions), with the remaining questions going into the test set, as shown in Table 4.

We run simple BERT (Devlin et al., 2019) feature-extraction baselines with question representations obtained by average-pooling. For this task, we use two models: (1) SVM and (2) cosine-similarity based k-nearest neighbor classification

Question Categories	15
Training Questions per Category	20
Training Questions	300
Test Questions (Real)	668
Test Questions (Generated)	238

Table 4: Data split for question-category classification.

(k-NN) with $k = 1$. As shown in Table 5, the SVM marginally outperforms k-NN on both the real and generated evaluation sets. Since our dataset is small, we also include results from using simple text data augmentation techniques (Wei and Zou, 2019).

Model	Real Q	Generated Q
BERT-feat: k-NN	47.8	52.1
+ augmentation	47.3	52.5
BERT-feat: SVM	52.2	53.4
+ augmentation	58.1	58.8

Table 5: Performance of BERT baselines (accuracy in %) on question-category classification with 15 classes and 20 examples per class in the training set.

3.2 Question-Class Classification

Of a more granular nature, the *question-class classification* task requires a new test question to be grouped into a question-class that asks the same thing, similar to retrieval QA contexts. For this task, we only consider question classes with at least 4 questions per class, and we split 3 questions from each class into the training set and the remaining questions into the test set, as shown in Table 6.

Question Classes with $N_{class} \geq 4$	89
Training Questions per Class	3
Training Questions	267
Test Questions (Real)	460
Test Questions (Generated)	131

Table 6: Data split for question-class classification.

Because this training dataset has fewer questions per class, we use the k-NN baseline from §3.1. We also evaluate a simple model that uses a triplet loss function to train a two layer neural net on BERT features, a method introduced for facial recognition (Schroff et al., 2015) and now used in NLP for few-shot learning (Yu et al., 2018) and answer selection (Kumar et al., 2019). In Table 7, we show top-1 and top-5 prediction accuracies for these two models.

	Real Q		Generated Q	
Model	Top-1	Top-5	Top-1	Top-5
BERT-feat: k-NN	29.6	50.3	20.8	38.5
+ augmentation	30.5	52.3	20.8	39.2
BERT-feat: triplet loss	42.4	71.5	57.3	80.9
+ augmentation	54.6	78.9	60.3	83.2

Table 7: Performance of BERT baselines (accuracy in %) on question-class classification with 89 classes and 3 examples per class in the training set.

We find that simple data augmentation (perhaps surprisingly) improves performance for most baselines, possibly due to the small size of our dataset and the restricted scope of question categories increasing the utility of augmented data (e.g., there are fewer ways to ask how long COVID will last than ways to write a positive movie review). One drawback of data augmentation, however, is that for n generated questions per original question, evaluation time for k-NN classification increases by up to $O(n)$, and training time for triple loss classification increases by up to $O(n^2)$.

Our baselines for both question-category and question-class classification are also limited in that the word *COVID* is not in the vocabulary of the pre-trained weights (`bert-base-uncased` from Huggingface), and so we suspect that models pre-trained on scientific or COVID-specific data will outperform our baseline.

4 Discussion

We have presented COVID-Q, a dataset of 1,690 COVID questions from 13 sources annotated with 15 category labels and 89 class labels. COVID-Q could directly help train question-answer systems or serve as a domain-specific evaluation resource, for which we have provided simple BERT baselines. Future work could include collecting more questions for the dataset, perhaps in collaboration with search engine companies, or evaluating domain-specific models (e.g., BERT pre-trained on COVID-related scientific papers).

Other COVID datasets. We encourage researchers to also explore other COVID datasets. A few that have already been released include tweets streamed since January 22 (Chen et al., 2020), location-tagged tweets in 65 languages (Abdul-Mageed et al., 2020), tweets of COVID symptoms (Sarker et al., 2020), a multi-lingual Twitter and Weibo dataset (Gao et al., 2020), an Instagram dataset (Zarei et al., 2020), emotional responses to COVID (Kleinberg et al., 2020), and annotated research abstracts (Huang et al., 2020).

References

Muhammad Abdul-Mageed, AbdelRahim Elmadany, Dinesh Pabbi, Kunal Verma, and Rannie Lin. 2020. Mega-cov: A billion-scale dataset of 65 languages for covid-19. *ArXiv*, abs/2005.06012. https://arxiv.org/pdf/2005.06012.pdf.

Emily Chen, Kristina Lerman, and Emilio Ferrara. 2020. Covid-19: The first public coronavirus twitter dataset. *ArXiv*, abs/2003.07372. https://arxiv.org/pdf/2003.07372.pdf.

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 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics. https://www.aclweb.org/anthology/N19-1423.pdf.

Zhiwei Gao, Shuntaro Yada, Shoko Wakamiya, and Eiji Aramaki. 2020. Naist covid: Multilingual covid-19 twitter and weibo dataset. *ArXiv*, abs/2004.08145. https://arxiv.org/pdf/2004.08145.pdf.

Ting-Hao Huang, Chieh-Yang Huang, Chien-Kuang Cornelia Ding, Yen-Chia Hsu, and C. Lee Giles. 2020. Coda-19: Reliably annotating research aspects on 10, 000+ cord-19 abstracts using non-expert crowd. *ArXiv*, abs/2005.02367. https://arxiv.org/pdf/2005.02367.pdf.

Bennett Kleinberg, Isabelle van der Vegt, and Maximilian Mozes. 2020. Measuring emotions in the covid-19 real world worry dataset. *ArXiv*, abs/2004.04225. https://arxiv.org/pdf/2004.04225.pdf.

Sawan Kumar, Shweta Garg, Kartik Mehta, and Nikhil Rasiwasia. 2019. Improving answer selection and answer triggering using hard negatives. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5911–5917. Association for Computational Linguistics. "https://www.aclweb.org/anthology/D19-1604".

Salvatore Romeo, Giovanni Da San Martino, Alberto Barrón-Cedeño, Alessandro Moschitti, Yonatan Belinkov, Wei-Ning Hsu, Yu Zhang, Mitra Mohtarami, and James Glass. 2016. Neural attention for learning to rank questions in community question answering. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1734–1745, Osaka, Japan. The COLING 2016 Organizing Committee. https://www.aclweb.org/anthology/C16-1163.pdf.

Wataru Sakata, Tomohide Shibata, Ribeka Tanaka, and Sadao Kurohashi. 2019. FAQ retrieval using query-question similarity and bert-based query-answer relevance. *CoRR*, abs/1905.02851. https://arxiv.org/pdf/1905.02851.pdf.

A. Sarker, S. Lakamana, William E. Hogg, Allen Xie, Mohammed Ali Al-garadi, and Yc Yang. 2020. Self-reported covid-19 symptoms on twitter: An analysis and a research resource. In *medRxiv*. https://www.medrxiv.org/content/10.1101/2020.04.16.20067421v3.full.pdf.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. *CoRR*, abs/1503.03832. http://arxiv.org/abs/1503.03832.

Jason Wei and Kai Zou. 2019. EDA: easy data augmentation techniques for boosting performance on text classification tasks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. http://dx.doi.org/10.18653/v1/D19-1670.

Mo Yu, Xiaoxiao Guo, Jinfeng Yi, Shiyu Chang, Saloni Potdar, Yu Cheng, Gerald Tesauro, Haoyu Wang, and Bowen Zhou. 2018. Diverse few-shot text classification with multiple metrics. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1206–1215. Association for Computational Linguistics. https://www.aclweb.org/anthology/N18-1109.

Koosha Zarei, Reza Farahbakhsh, Noel Crespi, and Gareth Tyson. 2020. A first instagram dataset on covid-19. *ArXiv*, abs/2004.12226. https://arxiv.org/pdf/2004.12226.pdf.

	Food and Drug Administration
Question	Closest Matches (BERT)
“Can I donate convalescent plasma?”	“Why is convalescent plasma being investigated to treat COVID?” “Can I make my own hand sanitizer?” “What are suggestions for things to do in the COVID quarantine?”
“Where can I report websites selling fraudulent medical products?”	“What kind of masks are recommended to protect healthcare workers from COVID exposure?” “Where can I get tested for COVID?” “How do testing kits for COVID detect the virus?”
	Center for Disease Control
Question	Closest Matches (BERT)
“What is the difference between cleaning and disinfecting?”	“How effective are alternative disinfection methods?” “Why has Trump stated that injecting disinfectant will kill COVID in a minute?” “Should I spray myself or my kids with disinfectant?”
“How frequently should facilities be cleaned to reduce the potential spread of COVID?”	“What is the survival rate of those infected by COVID who are put on a ventilator?” “What kind of masks are recommended to protect healthcare workers from COVID exposure?” “Will warm weather stop the outbreak of COVID?”

Table 8: Questions from the Food and Drug Administration (FDA) and Center for Disease Control (CDC) frequently asked questions websites that were not matched to any questions from other sources.

Supplementary Materials

In Table 8, we show questions from the FDA and CDC FAQ websites that were not found elsewhere in our dataset. In Table 9, we show sample questions from each of the 15 question categories.

Corresponding Answers. The FAQ websites from reputable sources (denoted with * in Table 1) also provide answers to their questions, and so we also provide them as an auxiliary resource. Using these answers, 23.8% of question classes have at least one corresponding answer. We caution against using these answers in applied settings, however, because information on COVID changes rapidly.

Category	Example Questions
Transmission	“Can COVID spread through food?” “Can COVID spread through water?” “Is COVID airborne?”
Societal Effects	“In what way have people been affected by COVID?” “How will COVID change the world?” “Do you think there will be more racism during COVID?”
Prevention	“Should I wear a facemask?” “How can I prevent COVID?” “What disinfectants kill the COVID virus?”
Societal Response	“Have COVID checks been issued?” “What are the steps that a hospital should take after COVID outbreak?” “Are we blowing COVID out of proportion?”
Reporting	“Is COVID worse than we are being told?” “What is the COVID fatality rate?” “What is the most reliable COVID model right now?”
Origin	“Where did COVID originate?” “Did COVID start in a lab?” “Was COVID a bioweapon?”
Treatment	“What treatments are available for COVID?” “Should COVID patients be ventilated?” “Should I spray myself or my kids with disinfectant?”
Speculation	“Was COVID predicted?” “Will COVID return next year?” “How long will we be on lockdown for COVID?”
Economic Effects	“What is the impact of COVID on the global economy?” “What industries will never be the same because of COVID?” “Why are stock markets dipping in response to COVID?”
Individual Response	“How do I stay positive with COVID?” “What are suggestions for things to do in the COVID quarantine?” “Can I still travel?”
Comparison	“How are COVID and SARS-COV similar?” “How can I tell if I have the flu or COVID?” “How does COVID compare to other viruses?”
Testing	“How COVID test is done?” “Are COVID tests accurate?” “Should I be tested for COVID?”
Nomenclature	“Should COVID be capitalized?” “What COVID stands for?” “What is the genus of the SARS-COVID?”
Having COVID	“How long does it take to recover?” “How COVID attacks the body?” “How long is the incubation period for COVID?”
Symptoms	“What are the symptoms of COVID?” “Which COVID symptoms come first?” “Do COVID symptoms come on quickly?”

Table 9: Sample questions from each of the 15 question categories.