Auditing and Robustifying COVID-19 Misinformation Datasets via Anticontent Sampling

Clay H. Yoo[1], **Ashiqur R. Khudabukhsh**[2]
[1] Carnegie Mellon University
[2] Rochester Institute of Technology
hyungony@andrew.cmu.edu, axkvse@rit.edu

Abstract

This paper makes two key contributions. First, it argues that highly specialized rare content classifiers trained on small data typically have limited exposure to the richness and topical diversity of the negative class (dubbed anticontent) as observed in the wild. As a result, these classifiers' strong performance observed on the test set may not translate into real-world settings. In the context of COVID-19 misinformation detection, we conduct an in-the-wild audit of multiple datasets and demonstrate that models trained with several prominently cited recent datasets are vulnerable to anticontent when evaluated in the wild. Second, we present a novel active learning pipeline that requires zero manual annotation and iteratively augments the training data with challenging anticontent, robustifying these classifiers.

Introduction

During the early phase of COVID-19 pandemic, the scientific community shared several COVID-19 misinformation datasets (see, e.g., (Micallef et al. 2020; Memon and Carley 2020; Alam et al. 2020; Cheng et al. 2021; Cui and Lee 2020)) to tackle the infodemic (Cinelli et al. 2020) within a short turnaround time. While these datasets provided great value in jump-starting the battle against COVID-19 misinformation, follow-on behavioral science studies have started appearing that rely on models trained on these datasets (Verma et al. 2022). A precise understanding of these datasets' effectiveness is thus necessary to trust the broad societal conclusions that depend on the reliability of these models.

How do we conduct external audits of these misinformation datasets to estimate how well models trained on these datasets perform in the wild? In the responsible AI literature, internal audits of datasets presenting guidelines for data dissemination (Gebru et al. 2021b) and resource reproducibility practices (Gebru et al. 2021a) have positively contributed to robust AI efforts. For other rare class classification tasks such as hate speech detection, cross-dataset generalization has been studied before (Arango, Pérez, and Poblete 2019). In-the-wild robustness audits of datasets is an under-explored area. However, instances exist that rare-class classification methods may get completely blindsided when presented with adversarial examples (Sarkar and KhudaBukhsh 2021).

A limiting factor to conducting large-scale in-the-wild audit is the lack of ground truth. In this paper, we present a framework to conduct in-the-wild audit of COVID-19 misinformation datasets. Our framework does not require annotated examples. The heart of the framework lies in a simple yet powerful observation: *before COVID-19, practically no COVID-19 misinformation can exist*. We show that classifiers trained on a broad range of heavily cited COVID-19 misinformation datasets often predict an alarming fraction of social media posts from the pre-COVID-19 era as COVID-19 misinformation, casting serious doubt on how useful they can be when deployed in the wild. Our experiments indicate that testing in the pre-COVID-19 era can present an elegant solution to estimating in-the-wild performance and evaluating real-world deployability.

Anticontent

A realistic assumption in any social web platform with broad participation is that much of the user-generated content will possibly be *non-misinformation*. Hence, even though discussions related to COVID-19 took the center stage during the pandemic, from the kneeling-down controversy of the NFL players to Ruth Bader Ginsburg's failing health; from Novak Djokovic's unfortunate expulsion from the US open to the former president's misspelled Tweet, the diverse set of discussions that are ***not misfinformation***, is what we call anticontent. This anticontent is far too topically diverse to be effectively captured in a small labeled data scenario.

What makes anticontent challenging? Anticontent can pose a challenge to in-the-wild COVID-19 misinformation classification in two possible ways. The first one is caused by shortcut learning (Geirhos et al. 2020). A classifier trained on a small dataset may pick up certain quirks of the dataset and incorrectly generalize. For instance, a model trained on a COVID-19 misinformation dataset where the keyword `hoax` is always associated with positives, will have a hard time when it encounters a linguistic black swan in the form of *this impeachment is a hoax*. A classic example of such shortcut learning happened when hate speech classifiers mistook harmless chess discussions as racist because almost all examples in the train set had `black`, `white`, `kill`,

Key idea: We leverage a simple assumption to conduct in-the-wild audit of COVID-19 misinformation datasets: *before COVID-19, a social media post is highly unlikely to express topics related to COVID-19 misinformation.* We further tap into the richness and diversity of a vast amount of implicitly labeled (as negative, or non-misinformation) pool of social media posts (dubbed anticontent) and design an active learning framework that requires no human annotation. We start with an annotated COVID-19 misinformation dataset drawn from multiple well-known COVID-19 misinformation datasets. At each step, we train a model on labeled data and evaluate on the pool of implicitly labeled anticontent instances. Instances that are classified as positives (i.e. COVID-19 misinformation) with high confidence are added as challenging negatives to augment the train data.

`capture`, `threat` attached to racially charged social media posts (Knight 2021; Sarkar and KhudaBukhsh 2021).

The second situation may arise when the classifier encounters an out-of-domain example and has to spit out a prediction nonetheless. As we already mentioned, the richness and diversity of real-world social media discussions are hard to capture within small-sized misinformation datasets. Precisely due to its sharpened focus on exemplifying *what is misinformation*, *what is* ***not*** *misinformation* often remains under-specified. By letting the models choose their own Achilles heel through our iterative framework, we bolster our models against both types of anticontent.

How can we leverage anticontent to robustify classifiers? Intuitively, an iterative learning framework akin to active learning (Settles 2012) where challenging anticontent instances, guided by model prediction, are added to the training dataset, could be a viable path to robustify content classifiers. In this work, we propose an active learning framework that completely bypasses the annotator requirement of active learning through sampling from an implicitly labeled pool – social media discussions from pre-COVID-19 time. To our knowledge, this is the first active learning framework that requires zero manual annotation.[1]

Contributions

Our contributions are the following:

- ***Robustness Audit:*** We introduce a novel approach to auditing the in-the-wild performance of COVID-19 misinformation classifiers. In this approach, we evaluate these classifiers on a social media dataset predating COVID-19. A high false positive rate indicates vulnerability to anticontent and less suitability for deployment in the wild.
- ***Robust AI Method:*** We present a novel active learning framework using implicit labels (`ALIL`) that requires zero human annotation since it samples from an implicitly labeled anticontent pool. Our method demonstrates substantial performance gain over classifiers trained solely on previously published COVID-19 misinformation datasets.

Active Learning Framework

Background. *Active learning* is a powerful and well-established form of supervised machine learning technique (Settles 2012). It is characterized by the interaction between the learner, aka the classifier, and the teacher (oracle or labeler or annotator) during the learning process. *Pool-based active learning* is a popular variant. In this setting, the learner is initially trained on a small seed set of labeled examples and has access to a large collection of unlabeled samples. At each iteration, the learner employs a sampling strategy to select an unlabeled sample and requests the supervisor to label it (in agreement with the target concept). The dataset is augmented with the newly acquired label, and the classifier is retrained on the augmented dataset. The sequential label-requesting and re-training process continues until some halting condition is reached (e.g., annotation budget is expended or the classifier has reached some target performance). At this point, the algorithm outputs a classifier, and the objective of this classifier is to closely approximate the (unknown) target concept in the future. The key goal of active learning is to reach a strong performance at the cost of fewer labels through the active participation of the learner. Since training and inference on a very large pool of samples can be computationally prohibitive, lines of work have examined the trade-offs of batch active learning (Yang and Carbonell 2013). We follow the pool-based batch active learning pipeline where instead of requesting one label at a time, the learner requests labels in batches.

Active Learning with Implicit Labels. Following traditional pool-based batch active learning, we start with a learner trained on a seed set of labeled examples. Let *batchSize* denote the configurable hyperparameter that specifies the number of samples added in each iteration. The large *unlabeled* pool of samples the learner has access to consists of social media posts from pre-COVID-19 era. As per our assumption, all the comments in this pool are implicitly labeled as negatives. We next sample the top *batchSize* samples with the highest predicted probability (as COVID-19 misinformation) computed by our learner. These samples are essentially highly challenging anticontent that can impact the classifier's performance in the wild. We add these samples to our training dataset as negatives, retrain our model on the augmented dataset, and loop through this cycle till a halting criterion is reached. Note that, in contrast with adversarial attacks (e.g., (He, Ahamad, and Kumar 2021)), our approach adds real-world anticontent instances. Algorithm 1 presents a formal description. Hyperparameter choices are described in the experimental setup section.

Related Work

Our work contains flavor of multiple well-established frameworks: domain generalization (Blanchard, Lee, and Scott 2011); domain adaptation (Blitzer, McDonald, and Pereira

[1] As it will be evident later, our active learning framework *does* need a small seed set of labeled examples to generate the initial model. However, the pipeline does not need any manual annotation in the subsequent learning steps.

Algorithm 1:
ALIL($\mathcal{M}_0$, $\mathcal{D}_{train}$, $\mathcal{D}_{valid}$, $\mathcal{D}_{implicit}$, $batchSize$)

Input:
$\mathcal{M}_0$:= baseline classifier
$\mathcal{D}_{train}$:= train data
$\mathcal{D}_{valid}$:= validation data
$\mathcal{D}_{implicit}$:= implicitly labeled data
Initialize:
$prevScore = -1$
$bestValidationScore = 0$
$\mathcal{M}^* = \mathcal{M}_0$
Procedure:
while $bestValidationScore > prevScore$ **do**
  $c = 0$
  $\mathcal{D}_{augment} = \emptyset$
  p = getPositiveProbability($\mathcal{M}^*$, $\mathcal{D}_{implicit}$)
  I_m = argsort(p, ascending=False)
  while $c < batchSize$ **do**
    $\mathcal{D}_{augment} = \mathcal{D}_{augment} \cup \{\mathcal{D}_{implicit}[I_m[c]]\}$
    c += 1
  end
  $\mathcal{D} = \mathcal{D}_{train} \cup \mathcal{D}_{valid} \cup \mathcal{D}_{augment}$
  $\mathcal{D}_{train}, \mathcal{D}_{validation}$ = split($\mathcal{D}$)
  $prevScore = bestValidationScore$
  $\mathcal{M}_1, bestValidationScore$ = train($\mathcal{D}_{train}, \mathcal{D}_{valid}$)
  if *bestValidationScore* > *prevScore* **then**
    $\mathcal{M}^* = \mathcal{M}_1$
    $\mathcal{D}_{implicit} = \mathcal{D}_{implicit} \backslash \mathcal{D}_{augment}$
  end
end
Output: $\mathcal{M}^*$

2006); and active learning (Settles 2012). Given the extensive literature in this field, we do not aim to be extensive and point to high-quality surveys whenever possible.

Domain generalization, also known as out-of-distribution (OOD) generalization, is a widely studied machine learning research area (Krueger et al. 2021; Teney, Abbasnejad, and Hengel 2020; Hendrycks et al. 2021). Shen et al. (2021) present a comprehensive survey. Similar to the extensive empirical evaluation conducted on a broad range of hate speech datasets (Arango, Pérez, and Poblete 2019), we conduct OOD generalization experiments (Secion) on COVID-19 misinformation datasets. Our evaluation framework on pre-COVID-19 content is also essentially geared towards estimating OOD generalization using implicit labels. However, to our knowledge, leveraging the temporal structure of an exogenous shock to construct a vast pool of implicitly labeled samples, and then evaluating OOD generalization, is a novel research direction.

A key distinction between domain adaptation and domain generalization is that the former has access to samples in the target domain (labeled (III 2007; Plank 2011) or unlabeled (Ramponi and Plank 2020)) unlike the latter. Our approach is thus more akin to domain adaptation. Again, what sets us apart is our novel use of pre-COVID-19 data with implicit labels. Following traditional self-training literature (Hearst 1991; Zhou, Kantarcioglu, and Thuraisingham 2012), we indeed select the unlabeled samples that are most confidently classified as positives. However, we flip their labels to negatives when we add them to our training set.

We drew inspiration from several existing lines of active learning research for constructing our misinformation classifier (Roy and McCallum 2001; Baram, Yaniv, and Luz 2004; Donmez, Carbonell, and Bennett 2007; Settles 2012). Since sequentially labeling and retraining models may not be practically feasible, following Yang and Carbonell (2013), we adopted a batch active learning setting to expand our pool of labeled samples. Among the extensive literature of active learning query strategies (Lewis and Gale 1994; Zhu et al. 2009; Sindhwani, Melville, and Lawrence 2009; Attenberg, Melville, and Provost 2010), we consider minority class certainty sampling (Sindhwani, Melville, and Lawrence 2009; Attenberg, Melville, and Provost 2010) in our framework. Unlike more traditional applications of this sampling technique to address class imbalance (Palakodety, KhudaBukhsh, and Carbonell 2020; Dutta et al. 2022), we use certainty sampling to detect challenging anticontent.

Data

COVID-19 Misinformation Datasets. We audit five publicly available COVID-19 misinformation datasets. These datasets are: $\mathcal{D}_{infodemic}$ (Micallef et al. 2020); $\mathcal{D}_{miscov19}$ (Memon and Carley 2020); $\mathcal{D}_{rumor}$ (Cheng et al. 2021); $\mathcal{D}_{coaid}$ (Cui and Lee 2020); and $\mathcal{D}_{disinfo}$ (Alam et al. 2020). Table 1 presents the overall statistics. We collapse fine-grained labels into two broad categories: *misinformation* and *non-misinformation* and use standard preprocessing steps. We denote train and validation sets as $\mathcal{D}^{train}$ and test set as $\mathcal{D}^{test}$. Moreover, a model trained with dataset $\mathcal{D}$ is denoted as $\mathcal{M}(\mathcal{D})$.

YouTube Video Comments. For our in-the-wild audit, we consider a dataset of YouTube comments from the official channels of four prominent US cable news networks: CNN, Fox News, MSNBC, and One American News Network (OANN)[2] (KhudaBukhsh et al. 2022, 2021). We consider this dataset for the following three reasons: broad participation, topical diversity, and substantial presence of discussions relevant to COVID-19.

Overall, the evaluation dataset consists of 33.3 million comments on news videos starting from 25 January 2014 to 2 November 2020. We focus on two non-overlapping partitions of the dataset – pre-COVID-19 ($\mathcal{D}_{pre}$) and post-COVID-19 ($\mathcal{D}_{post}$) – separating before and after the outbreak of COVID-19. Following the Centers for Disease Control and Prevention (CDC) timeline[3], we mark 12 December 2019 as the start date of COVID-19. To summarise, $\mathcal{D}_{pre}$ spans between 25 January 2014 and 11 December

[2]Compared to the other three networks, OANN is a fringe network often harboring outlandish views on science and democracy. Through using OANN, we are in no way legitimizing OANN as a mainstream news network. We include OANN due to its well-documented history of propagating COVID-19 misinformation.

[3]https://www.cdc.gov/museum/timeline/covid19.html

Dataset	# Citations	Source	Statistics Collection Period	# Positives/ # Negatives/ # Total	Avg. Token Length
$\mathcal{D}_{infodemic}$	23	Twitter	Jan. – May 2020	1,195 / 2,784 / 3,979	$26.5_{13.4}$
$\mathcal{D}_{miscov19}$	107	Twitter	Mar. – Jun. 2020	615 / 2,853 / 3,468	$28.7_{13.1}$
$\mathcal{D}_{rumor}$	22	News Media & Twitter	Jan. – Apr. 2020	3,652 / 1,844 / 5,496	$19.8_{13.6}$
$\mathcal{D}_{coaid}$	109	News Media	Dec. 2019 – Nov. 2020	1,723 / 590 / 2,313	$11.2_{6.3}$
$\mathcal{D}_{disinfo}$	56	Twitter	Jan. 2020 – Mar. 2021	61 / 1,041 / 1,102	$33.5_{10.7}$
Combined	–	–	–	6,113 / 10,245 / 16,358	$23.0_{14.0}$

Table 1: Statistics of datasets used for our research. Positives correspond to misinformation samples, while negatives to non-misinformation. The average and standard deviation (in subscript) of token length are measured by splitting preprocessed texts with a whitespace character. Citation counts are measured on 16 August 2022.

Model	Test dataset $\mathcal{D}_{infodemic}$	$\mathcal{D}_{miscov19}$	$\mathcal{D}_{rumor}$	$\mathcal{D}_{coaid}$	$\mathcal{D}_{disinfo}$	Misinformation Rate $\mathcal{D}^{1m}_{pre}$	$\mathcal{D}^{1m}_{post}$
$\mathcal{M}(\mathcal{D}^{train}_{infodemic})$	$89.7_{0.5}/78.7_{0.9}$	$\mathbf{86.2_{0.6}}/23.4_{1.2}$	$\mathbf{55.1_{0.6}}/39.7_{2.9}$	$83.0_{1.1}/31.9_{0.9}$	$\mathbf{94.2_{0.7}}/4.6_{1.7}$	$1.7_{1.5}$	$2.0_{1.4}$
$\mathcal{M}(\mathcal{D}^{train}_{miscov19})$	$\mathbf{81.3_{0.4}}/30.8_{5.0}$	$93.9_{0.6}/69.4_{2.8}$	$51.0_{0.3}/19.3_{3.0}$	$\mathbf{83.7_{1.0}}/17.7_{1.3}$	$91.8_{2.6}/5.4_{2.3}$	$7.6_{2.8}$	$6.9_{2.4}$
$\mathcal{M}(\mathcal{D}^{train}_{rumor})$	$76.8_{2.0}/\mathbf{52.7_{1.1}}$	$79.3_{2.6}/25.2_{1.1}$	$77.5_{1.9}/89.9_{1.3}$	$59.0_{4.9}/\mathbf{45.1_{1.4}}$	$88.3_{1.8}/\mathbf{12.2_{2.6}}$	$18.3_{4.6}$	$18.8_{4.8}$
$\mathcal{M}(\mathcal{D}^{train}_{coaid})$	$48.0_{6.1}/47.3_{1.0}$	$52.5_{7.0}/\mathbf{29.6_{0.9}}$	$38.7_{2.7}/\mathbf{76.5_{0.4}}$	$97.0_{0.6}/91.2_{1.8}$	$44.9_{6.0}/11.5_{0.5}$	$62.2_{6.1}$	$60.5_{6.4}$

Table 2: Mean and standard deviation (in subscript) of F-1 scores of two labels (non-/misinformation) trained and evaluated on each dataset over five seeds. When the train and test datasets are the same (diagonal entries), performance on the test set is reported. Otherwise, performance on the entire dataset is reported (non-diagonal entries). The best out-of-domain F-1 score of each label is boldfaced. We also report misinformation rates on 1 million randomly selected pre-COVID-19 comments ($\mathcal{D}^{1m}_{pre}$) and post-COVID-19 comments ($\mathcal{D}^{1m}_{post}$).

2019, while $\mathcal{D}_{post}$ spans between 12 December 2019 and 2 November 2020.

How rare is COVID-19 misinformation in the wild? We randomly sample 5,000 comments from $\mathcal{D}_{post}$ and annotate them as one of misinformation, non-misinformation, and unverifiable. We confirm the veracity of comments from the official websites of credible sources such as CDC, WHO, IFCN's CoronaVirusFacts / DatosCorona-Virus alliance database, Reuters, WebMD and PolitiFact.

Out of 5,000 comments, 4,904 are marked as non-misinformation, 83 as misinformation, and 13 as indeterminate. We discard 13 indeterminate comments and construct a test set of 4,904 comments ($\mathcal{D}^{test}_{youtube}$), yielding 1.7% misinformation rate. Two annotators independently annotated with a post-annotation adjudication step to dissolve the disagreements with $\kappa = 0.82$ (Cohen 1960).

These annotated 4,987 samples also serve another critical purpose. We now have a dataset that is markedly different both in terms of the relative proportion of misinformation and non-misinformation and (possibly) richness in anticontent, allowing us to estimate OOD performance.

Combining Datasets. Seeking to build a robust classifier capturing a wide range of topics in COVID-19 misinformation, we combine five misinformation datasets into a single set. We first split each dataset using 80/10/10 (train/-validation/test) ratio. We then concatenate five individual test sets and $\mathcal{D}^{test}_{youtube}$ to obtain 5,634 non-misinformation and 558 misinformation samples as a combined test set ($\mathcal{D}^{test}_{combined}$). Since our goal is to test the OOD performance, $\mathcal{D}^{train}_{combined}$ does not contain any of the labeled YouTube comments.

Experimental Setup

We fine-tune `BERT` (Devlin et al. 2019) when training all classifiers.

`ALIL` Setup. For augmentation, we construct a dataset of 1 million randomly sampled YouTube comments from pre-COVID-19 era ($\mathcal{D}^{1m}_{pre}$) such that none of the comments in $\mathcal{D}^{test}_{youtube}$ is included. $\mathcal{D}^{1m}_{pre}$ is used as $\mathcal{D}_{implicit}$ in Algorithm 1. We set $batchSize = 500$ ($\approx 4\%$ of the combined train data). As a stopping criterion, we stop training if we observe a decrease in the validation F-1 score.

Self-training Setup. Since we do not assume there exists a temporal structure in YouTube comments, we randomly sample 1 million comments from $\mathcal{D}_{post}$ such that it does not overlap with any comment in $\mathcal{D}^{1m}_{post}$ or $\mathcal{D}^{test}_{youtube}$ and use it as an unlabeled pool. At each self-training iteration, we add 250 samples with the lowest misinformation probability as non-misinformation and 250 with the highest probability as misinformation to match the size of added samples same as `ALIL`. We avoid using $\mathcal{D}_{pre}$ as a pool since adding

$\mathcal{M}(\mathcal{D}^{train}_{combined})$	$\mathcal{M}(\mathcal{D}^{train}_{combined} \cup \mathcal{D}^{ALIL}_{500})$
at the ohio state university they limited domestic students and started to give those seats to the chinese	*there is now ample evidence that the united states of america government through the national institute of health have committed an act of bio terror by funding an experiment in 2015 at the university of north carolina chapel hill that created sars covid 2 covid 19 ...*
obama was fully aware of covid 19 and also enabled the funding and creation aswell as selling it to china	*the plan is to infect as many americans with corona virus to flood the streets with regeneron*
joe biden and hussein obama directly funded the missiles that were launched at our servicemen with the pallets of cash they sent to iran	*no one is safe from this bioweapon virus it now has over 40 mutations young and old r now in great danger human depopulation is happening at a rapid pace hopefully there will be a vaccine created soon to eradicate it*
john bolton said there are interest groups who profit on wars bombs and bullets etc are a source of money on wartime	*would not be surprised if we hear the dems went to china and came back with the china bioweapon virus attack on purpoae*
ronald reagan said if facism ever comes to the us it ll come in the form of liberalism	*this is propaganda they eat those types of animals for years this was predicted by billy meier back in 1995 and he says quote a lung disease will also break out in humans through the guilt of china where bioweapons are being researched and a carelessness is releasing pathogens*

Table 3: Top five misinformation with the highest misinformation probability selected by the baseline model ($\mathcal{M}(\mathcal{D}^{train}_{combined})$) versus the top comments selected by $\mathcal{M}(\mathcal{D}^{train}_{combined} \cup \mathcal{D}^{ALIL}_{500})$.

250 pre-COVID-19 comments as misinformation will likely harm the model performance.

Baselines. We compare our ALIL against (1) a model trained with self-training and (2) a model trained on the seed set ($\mathcal{D}^{train}_{combined}$) without any data augmentation.

Results

In-the-wild Audit and OOD Generalization

We first examine OOD generalization through training models on a single dataset and testing on another. We could not obtain $\mathcal{M}(\mathcal{D}_{disinfo})$ since training didn't converge, possibly due to the scarcity of positive samples (5.5%) obtained during rehydration. Table 2 summarizes the OOD generalization of the four datasets. We observe the in-domain performances are significantly better than the OOD performances.

Table 2 conducts an in-the-wild audit through evaluating on 1 million randomly sampled YouTube comments from pre-COVID-19 ($\mathcal{D}^{1m}_{pre}$) and post-COVID-19 ($\mathcal{D}^{1m}_{post}$) eras, respectively. We expect robust models to have (1) close to zero pre-COVID-19 misinformation rate; and (2) close to 1.7% post-COVID-19 misinformation rate (estimated through manual inspection). We observe that two models trained on these datasets exhibit a pre-COVID-19 misinformation rate higher than 10%, indicating susceptibility to anticontent and making them unsuitable for real-world deployment.

Evaluating Active Learning with Implicit Labels

Qualitative Evaluations. Table 3 shows the top five misinformation comments from $\mathcal{D}^{1m}_{post}$ assigned the highest probability by two different models. The left column contains predictions by a model trained on $\mathcal{D}^{train}_{combined}$ (baseline) without any augmentation of anticontent. The right column contains predictions by $\mathcal{M}(\mathcal{D}^{train}_{combined} \cup \mathcal{D}^{ALIL}_{500})$ obtained after the first iteration of ALIL. On the left side, we notice that one comment is both COVID-19-related and COVID-19 misinformation, while the other comments are unrelated to COVID-19. On the right side, however, we notice that the model can sift through anticontent far more effectively, detecting several misinformation posts.

Quantitative Evaluations. In order to demonstrate that iterative augmentation of the train data is more effective than augmenting with a larger *batchSize* of samples added all at once, we construct two datasets with increasing number of $N \in \{500, 1000\}$ implicitly labeled pre-COVID-19 comments. $\mathcal{D}^{baseline}_{N}$ is constructed by selecting pre-COVID-19 comments assigned the highest misinformation probabilities computed by the baseline model, i.e., $\mathcal{M}(\mathcal{D}^{train}_{combined})$, while $\mathcal{D}^{ALIL}_{N}$ is constructed with ALIL (Algorithm 1). Note that $\mathcal{D}^{baseline}_{500} \subset \mathcal{D}^{baseline}_{1000}$, $\mathcal{D}^{ALIL}_{500} \subset \mathcal{D}^{ALIL}_{1000}$ and $\mathcal{D}^{baseline}_{500} = \mathcal{D}^{ALIL}_{500}$.

Table 4 summarizes the performance of ALIL. We observe that the initial addition of 500 pre-COVID-19 comments achieves a substantial gain in performance, improving a weighted F-1 score from 87.7% to 94.0%, when evaluated on $\mathcal{D}^{test}_{combined}$. Pre-COVID-19 misinformation rate drastically decreases from 14.3% to 1.7% as well as the post-COVID-19 misinformation rate from 13.6% to 3.4%. In the second step of augmentation, models trained with $\mathcal{D}^{ALIL}_{1000}$ achieves a better performance (95.2% v.s. 94.9%) and significantly lower pre-COVID-19 misinformation rate (0.2% v.s. 0.7%) compared to the models trained with $\mathcal{D}^{baseline}_{1000}$. We stop augmenting at 1,000 pre-COVID-19 comments since the average validation performance drops with $\mathcal{D}^{ALIL}_{1500}$. With the addition of 1,000 samples, we observe that the post-

Model	Performance on $\mathcal{D}^{test}_{combined}$ Precision	Recall	F-1	Misinformation Rate $\mathcal{D}^{1m}_{pre}$	$\mathcal{D}^{1m}_{post}$
$\mathcal{M}(\mathcal{D}^{train}_{combined})$	$91.3_{0.3}$	$85.7_{2.6}$	$87.7_{1.8}$	$14.3_{3.4}$	$13.6_{3.1}$
$\mathcal{M}(\mathcal{D}^{train}_{combined} \cup \mathcal{D}^{baseline}_{500})$	$94.1_{0.2}$	$93.9_{0.5}$	$94.0_{0.4}$	$1.7_{0.5}$	$3.4_{0.6}$
$\mathcal{M}(\mathcal{D}^{train}_{combined} \cup \mathcal{D}^{baseline}_{1000})$	$94.9_{0.3}$	$95.0_{0.4}$	$94.9_{0.4}$	$0.7_{0.4}$	$1.9_{0.7}$
$\mathcal{M}(\mathcal{D}^{train}_{combined} \cup \mathcal{D}^{ALIL}_{500})$	$94.1_{0.2}$	$93.9_{0.5}$	$94.0_{0.4}$	$1.7_{0.5}$	$3.4_{0.6}$
$\mathcal{M}(\mathcal{D}^{train}_{combined} \cup \mathcal{D}^{ALIL}_{1000})$	$\mathbf{95.2}_{0.2}$	$\mathbf{95.3}_{0.2}$	$\mathbf{95.2}_{0.2}$	$\mathbf{0.2_{0.1}}$	$\mathbf{1.4_{0.4}}$
$\mathcal{M}(\mathcal{D}^{train}_{combined} \cup \mathcal{D}^{random,pre}_{1000})$	$93.9_{0.3}$	$92.9_{1.1}$	$93.3_{0.8}$	$3.7_{1.2}$	$5.0_{1.6}$
$\mathcal{M}(\mathcal{D}^{train}_{combined} \cup \mathcal{D}^{random,post}_{1000})$	$94.0_{0.2}$	$93.5_{0.4}$	$93.7_{0.3}$	$3.4_{0.4}$	$4.0_{0.4}$
$\mathcal{M}(\mathcal{D}^{train}_{combined} \cup \mathcal{D}^{self\text{-}train}_{500})$	$91.1_{0.2}$	$84.3_{0.7}$	$86.7_{0.5}$	$16.7_{0.0}$	$16.3_{0.0}$
$\mathcal{M}(\mathcal{D}^{train}_{combined} \cup \mathcal{D}^{self\text{-}train}_{1000})$	$90.8_{0.1}$	$81.6_{1.1}$	$84.8_{0.8}$	$20.8_{0.0}$	$19.7_{0.0}$

Table 4: Mean and standard deviation (in subscript) of precision, recall, and F-1 scores of two labels (non-/misinformation) evaluated on $\mathcal{D}^{test}_{combined}$ over five seeds. We also report misinformation rates on fixed sets of 1 million randomly selected pre-COVID-19 comments ($\mathcal{D}^{1m}_{pre}$) and post-COVID-19 comments ($\mathcal{D}^{1m}_{post}$).

COVID-19 misinformation rate gets closer to the approximated population misinformation rate of 1.7%.

To test the effect of *batchSize*, we ran the same experiment with $batchSize = 100$. ALIL halted at the fifth iteration, and it showed a similar pattern as reported in Table 4; improvement in performance and reduction in pre/post-COVID-19 misinformation rate.

We denote $\mathcal{D}^{self\text{-}train}_{N}, N \in \{500, 1000\}$, as the augmentation sets found by self-training. We observe that performance rather decreases, and the misinformation rates for both pre-COVID-19 and post-COVID-19 eras increase. We tried (1) decreasing the batch size from 500 to 250, i.e., add 125 positives and 125 negatives at each iteration, and (2) changing the distribution to match that of $\mathcal{D}_{combined}$, i.e., adding 190 positives and 310 negatives at each iteration, but the story stayed the same.

Conclusions and Discussions

While dataset audit is still in its infancy mainly focusing on fairness, transparency, and accountability (Gebru et al. 2021b), we highlight a relatively under-explored area of auditing in-the-wild reliability of COVID-19 misinformation datasets. As the field of information mining is moving towards dense inter-dependence, where a dataset proposed by one research group facilitates social investigations by other groups, testing the in-the-wild effectiveness of datasets will become increasingly important. We believe our work will open the gates for further innovative research in reliability audits for datasets.

We also raise an important point that given misinformation is (hopefully) rare, and to operate successfully in a real-world setting with high topical diversity, a COVID-19 misinformation content classifier needs to both understand what **is** and **is not** misinformation. While a dataset of a few 1,000 samples can do a decent job in describing what is misinformation, the complement of the set, what we dub anticontent, has a richness extremely challenging to be captured well within a small dataset scenario. Arguing that before COVID-19, no social media content can express COVID-19 misinformation, we propose that testing on a large-scale dataset of pre-COVID-19 era social media discussions can shed light on content classifiers' vulnerability to anticontent. Based on this insight, we further propose an active learning framework that remarkably improves the content classifiers' performance in handling anticontent. And this performance boost comes with zero additional manual annotation.

Our study raises the following points to consider.

Topical Diversity in Anticontent. We annotate 200 random samples from $\mathcal{D}^{ALIL}_{1000}$ into one of the 10 topics. We do not intend to be formal or exhaustive, but rather to be illustrative of the broad range of topics present in anticontents that are misclassified as misinformation. Figure 1 shows that while politics is the dominating topic, the selected anticontent possesses a healthy diversity of a broad range of topics.

Misinformation about Previous Outbreaks. Although there cannot exist COVID-19-specific misinformation in the pre-COVID-19 era, conspiracies, anti-vaccine sentiment and fake cure/treatment, topics that appear frequently in our combined dataset, existed well before COVID-19 from the previous outbreaks (e.g. SARS, Zika Virus, MERS). In fact, Figure 1 shows that two comments (1%) are marked as misinformation related to non-COVID-19 diseases/viruses, where one comment discusses biohazard weapons related to tuberculosis and the other discusses how flu shot is related to Bill Gates as a medium of population control. Both topics appear as COVID-19 misinformation in our combined dataset, and adding these samples as non-misinformation to the train set for COVID-19 misinformation detection can either i) improve the classifier's performance by letting it learn

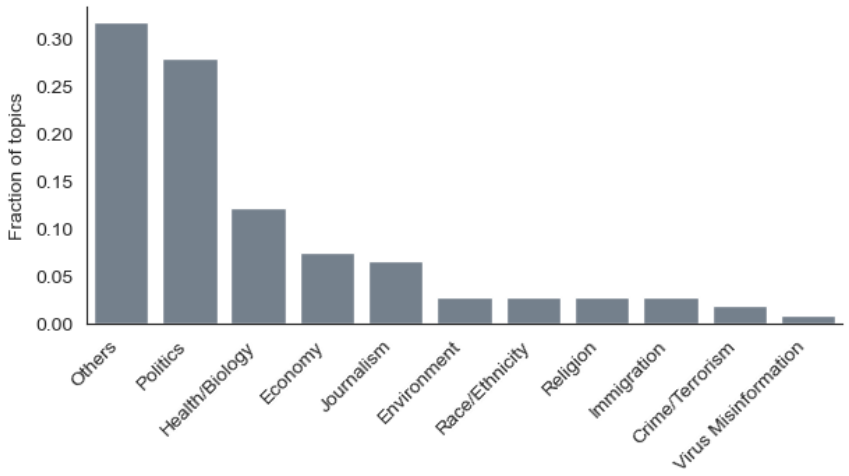

Figure 1: Distribution of annotated topics present in 200 randomly selected YouTube comments from $\mathcal{D}_{1000}^{ALIL}$.

how to distinguish between COVID-19 and other diseases (tuberculosis and flu in this case) or ii) confuse the classifier by providing similar comments with opposing labels. Although the impact of adding these samples may be small due to their size (1%), figuring out precisely how they impact performance merits a deeper future exploration.

Generalizability to Tasks without Temporal Structure. COVID-19 is an exogenous shock with a crisp temporal structure, making it straightforward to construct a set of anticontent predating COVID-19. ***How can we use a similar approach to improve classifier performance in tasks without crisp temporal structures, such as detecting anti-Semitism, misogyny, or racism?*** Defining implicitly labeled anticontent for such problems is not as straightforward as COVID-19 misinformation detection since there are no well-established points of the time these phenomena started.

We present a path forward in incorporating our approach to combat these types of inappropriate content. For instance, a strictly moderated community, $\mathcal{C}$, designated as a safe space for say, the LGBTQ+ community, is highly unlikely to have a large number of anti-LGBTQ+ posts. A hate speech classifier designed to protect the LGBTQ+ community can adopt ALIL. Even if hate speech exists in $\mathcal{C}$, augmenting the train set may still lead to performance gain so long as the overall proportion of hate speech is low.

To test our hypothesis, we assume that there exists no temporal structure in COVID-19 and that the overall misinformation rate is low, possibly close to 1.7% as computed from $\mathcal{D}_{youtube}^{test}$. To augment $\mathcal{D}_{combined}^{train}$, we randomly sample 1,000 post-COVID-19 comments and implicitly label them as non-misinformation ($\mathcal{D}_{1000}^{random,post}$). Even though there may exist misinformation among 1,000 samples, we observe that augmentation leads to (1) an improvement in weighted F-1 score from 87.7% to 93.7%, (2) a reduction of pre-COVID-19 misinformation rate from 14.3% to 3.4%, and (3) bringing post-COVID-19 misinformation rate closer to 1.7% from 13.6% to 4.0% (Table 4). Although not as effective as using the temporal structure and ALIL, this shows that our approach of finding implicitly labeled anticontent in a low probability (of a positive label) setting can be extended to solving more general problems than a special case like detecting COVID-19 misinformation.

Value in the Active Learning Framework. Since the entire comment set predating COVID-19 could serve as anticontent, it is possible to augment the train data using randomly sampled pre-COVID-19 comments. One may then wonder if the active learning pipeline is adding any value to performance gain. We construct $\mathcal{D}_{1000}^{random,pre}$ by randomly sampling 1,000 pre-COVID-19 comments from $\mathcal{D}_{pre}$, ensuring that there is no overlap with comments in $\mathcal{D}_{youtube}^{test}$ and $\mathcal{D}_{pre}^{1m}$. We observe that adding 1,000 random samples has similar performance effects in (1) improving F-1 score on $\mathcal{D}_{combined}^{test}$ from 87.7% to 93.3%, (2) reducing pre-COVID-19 misinformation rate from 14.3% to 3.7%, and (3) bringing the post-COVID-19 misinformation rate closer to 1.7% from 5.0% (Table 4). However, the improvement is less pronounced than that of using ALIL.

Rehydration of Tweets and its Potential Impact. Our experiments are conducted on the currently available snapshots of the datasets. To preserve a user's *right to be forgotten*, most of these datasets only provide Tweet IDs that need to be re-hydrated. Some of the samples present in the released datasets are no longer available due to reasons like deletion of posts (e.g. Twitter's COVID-19 misleading information policy) and user suspension.

Ethics Statement

Although our focus is to build a robust COVID-19 misinformation classifier, any content filter can be inverted for malicious purposes. For example, a real-world application can utilize our misinformation classifier to effectively detect misinformation but maliciously choose to filter out non-misinformation and filter in misinformation.

Acknowledgements

We thank Anirban Chowdhury, Rupak Sarkar, and Ritam Dutta for their input.

References

Alam, F.; Shaar, S.; Nikolov, A.; Mubarak, H.; Martino, G. D. S.; Abdelali, A.; Dalvi, F.; Durrani, N.; Sajjad, H.; Darwish, K.; and Nakov, P. 2020. Fighting the COVID-19 Infodemic: Modeling the Perspective of Journalists, Fact-Checkers, Social Media Platforms, Policy Makers, and the Society. *ArXiv*, abs/2005.00033.

Arango, A.; Pérez, J.; and Poblete, B. 2019. Hate speech detection is not as easy as you may think: A closer look at model validation. In *Proceedings of the 42nd international acm sigir conference on research and development in information retrieval*, 45–54.

Attenberg, J.; Melville, P.; and Provost, F. 2010. A unified approach to active dual supervision for labeling features and examples. In *ECML/PKDD*, 40–55. Springer.

Baram, Y.; Yaniv, R. E.; and Luz, K. 2004. Online choice of active learning algorithms. *Journal of Machine Learning Research*, 5(Mar): 255–291.

Blanchard, G.; Lee, G.; and Scott, C. 2011. Generalizing from several related classification tasks to a new unlabeled sample. *Advances in neural information processing systems*, 24: 2178–2186.

Blitzer, J.; McDonald, R.; and Pereira, F. 2006. Domain adaptation with structural correspondence learning. In *Proceedings of the 2006 conference on empirical methods in natural language processing*, 120–128.

Cheng, M.; Wang, S.; Yan, X.; Yang, T.; Wang, W.; Huang, Z.; Xiao, X.; Nazarian, S.; and Bogdan, P. 2021. A COVID-19 Rumor Dataset. *Frontiers in Psychology*, 12.

Cinelli, M.; Quattrociocchi, W.; Galeazzi, A.; Valensise, C. M.; Brugnoli, E.; Schmidt, A. L.; Zola, P.; Zollo, F.; and Scala, A. 2020. The COVID-19 social media infodemic. *Scientific reports*, 10(1): 1–10.

Cohen, J. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1): 37–46.

Cui, L.; and Lee, D. 2020. CoAID: COVID-19 Healthcare Misinformation Dataset. *ArXiv*, abs/2006.00885.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL*.

Donmez, P.; Carbonell, J. G.; and Bennett, P. N. 2007. Dual strategy active learning. In *European Conference on Machine Learning*, 116–127. Springer.

Dutta, S.; Li, B.; Nagin, D. S.; and KhudaBukhsh, A. R. 2022. A Murder and Protests, the Capitol Riot, and the Chauvin Trial: Estimating Disparate News Media Stance. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022*, 5059–5065.

Gebru, T.; Morgenstern, J.; Vecchione, B.; Vaughan, J. W.; Wallach, H.; Iii, H. D.; and Crawford, K. 2021a. Datasheets for datasets. *Communications of the ACM*, 64(12): 86–92.

Gebru, T.; Morgenstern, J.; Vecchione, B.; Vaughan, J. W.; Wallach, H. M.; III, H. D.; and Crawford, K. 2021b. Datasheets for datasets. *Commun. ACM*, 64(12): 86–92.

Geirhos, R.; Jacobsen, J.; Michaelis, C.; Zemel, R. S.; Brendel, W.; Bethge, M.; and Wichmann, F. A. 2020. Shortcut learning in deep neural networks. *Nat. Mach. Intell.*, 2(11): 665–673.

He, B.; Ahamad, M.; and Kumar, S. 2021. PETGEN: Personalized Text Generation Attack on Deep Sequence Embedding-based Classification Models. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 575–584.

Hearst, M. 1991. Noun homograph disambiguation using local context in large text corpora. *Using Corpora*, 185–188.

Hendrycks, D.; Basart, S.; Mu, N.; Kadavath, S.; Wang, F.; Dorundo, E.; Desai, R.; Zhu, T.; Parajuli, S.; Guo, M.; et al. 2021. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8340–8349.

III, H. D. 2007. Frustratingly Easy Domain Adaptation. In Carroll, J. A.; van den Bosch, A.; and Zaenen, A., eds., *ACL 2007, Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, June 23-30, 2007, Prague, Czech Republic*. The Association for Computational Linguistics.

KhudaBukhsh, A. R.; Sarkar, R.; Kamlet, M. S.; and Mitchell, T. M. 2021. We Don't Speak the Same Language: Interpreting Polarization through Machine Translation. In *The Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021*, 14893–14901. AAAI Press.

KhudaBukhsh, A. R.; Sarkar, R.; Kamlet, M. S.; and Mitchell, T. M. 2022. Fringe News Networks: Dynamics of US News Viewership following the 2020 Presidential Election. In *WebSci '22: 14th ACM Web Science Conference 2022*, 269–278. ACM.

Knight, W. 2021. Why a YouTube Chat About Chess Got Flagged for Hate Speech. https://www.wired.com/story/why-youtube-chat-chess-flagged-hate-speech/. Online; accessed 15-August-2022.

Krueger, D.; Caballero, E.; Jacobsen, J.-H.; Zhang, A.; Binas, J.; Zhang, D.; Le Priol, R.; and Courville, A. 2021. Out-of-distribution generalization via risk extrapolation (rex). In *International Conference on Machine Learning*, 5815–5826. PMLR.

Lewis, D. D.; and Gale, W. A. 1994. A sequential algorithm for training text classifiers. In *SIGIR'94*, 3–12. Springer.

Memon, S. A.; and Carley, K. M. 2020. Characterizing COVID-19 Misinformation Communities Using a Novel Twitter Dataset. *ArXiv*, abs/2008.00791.

Micallef, N.; He, B.; Kumar, S.; Ahamad, M.; and Memon, N. 2020. The Role of the Crowd in Countering Misinformation: A Case Study of the COVID-19 Infodemic. *2020 IEEE International Conference on Big Data (Big Data)*, 748–757.

Palakodety, S.; KhudaBukhsh, A. R.; and Carbonell, J. G. 2020. Voice for the Voiceless: Active Sampling to Detect Comments Supporting the Rohingyas. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020*, 454–462. AAAI Press.

Plank, B. 2011. *Domain adaptation for parsing*. Citeseer.

Ramponi, A.; and Plank, B. 2020. Neural Unsupervised Domain Adaptation in NLP—A Survey. *arXiv preprint arXiv:2006.00632*.

Roy, N.; and McCallum, A. 2001. Toward Optimal Active Learning through Sampling Estimation of Error Reduction. In *Proceedings of the Eighteenth International Conference on Machine Learning (ICML 2001), Williams College, Williamstown, MA, USA, June 28 - July 1, 2001*, 441–448.

Sarkar, R.; and KhudaBukhsh, A. R. 2021. Are Chess Discussions Racist? An Adversarial Hate Speech Data Set (Student Abstract). In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021*, 15881–15882. AAAI Press.

Settles, B. 2012. *Active Learning*. Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan & Claypool Publishers.

Shen, Z.; Liu, J.; He, Y.; Zhang, X.; Xu, R.; Yu, H.; and Cui, P. 2021. Towards out-of-distribution generalization: A survey. *arXiv preprint arXiv:2108.13624*.

Sindhwani, V.; Melville, P.; and Lawrence, R. D. 2009. Uncertainty sampling and transductive experimental design for active dual supervision. In *Proceedings of the 26th ICML*, 953–960. ACM.

Teney, D.; Abbasnejad, E.; and Hengel, A. v. d. 2020. Unshuffling data for improved generalization. *arXiv preprint arXiv:2002.11894*.

Verma, G.; Bhardwaj, A.; Aledavood, T.; Choudhury, M. D.; and Kumar, S. 2022. Examining the impact of sharing COVID-19 misinformation online on mental health. *Scientific reports*, 12 1: 8045.

Yang, L.; and Carbonell, J. 2013. Buy-in-bulk active learning. In *Advances in neural information processing systems*, 2229–2237.

Zhou, Y.; Kantarcioglu, M.; and Thuraisingham, B. 2012. Self-training with selection-by-rejection. In *2012 IEEE 12th international conference on data mining*, 795–803. IEEE.

Zhu, J.; Wang, H.; Tsou, B. K.; and Ma, M. 2009. Active learning with sampling by uncertainty and density for data annotations. *IEEE Transactions on audio, speech, and language processing*, 18(6): 1323–1331.