Construction of Large-Scale Misinformation Labeled Datasets from Social Media Discourse using Label Refinement

Karishma Sharma
krsharma@usc.edu
University of Southern California
USA

Emilio Ferrara
University of Southern California
USA
emiliofe@usc.edu

Yan Liu
University of Southern California
USA
yanliu.cs@usc.edu

ABSTRACT

Malicious accounts spreading misinformation has led to widespread false and misleading narratives in recent times, especially during the COVID-19 pandemic, and social media platforms struggle to eliminate these contents rapidly. This is because adapting to new domains requires human intensive fact-checking that is slow and difficult to scale. To address this challenge, we propose to leverage news-source credibility labels as weak labels for social media posts and propose model-guided refinement of labels to construct large-scale, diverse misinformation labeled datasets in new domains. The weak labels can be inaccurate at the article or social media post level where the stance of the user does not align with the news source or article credibility. We propose a framework to use a detection model self-trained on the initial weak labels with uncertainty sampling based on entropy in predictions of the model to identify potentially inaccurate labels and correct for them using self-supervision or relabeling. The framework will incorporate social context of the post in terms of the community of its associated user for surfacing inaccurate labels towards building a large-scale dataset with minimum human effort. To provide labeled datasets with distinction of misleading narratives where information might be missing significant context or has inaccurate ancillary details, the proposed framework will use the few labeled samples as class prototypes to separate high confidence samples into false, unproven, mixture, mostly false, mostly true, true, and debunk information. The approach is demonstrated for providing a large-scale misinformation dataset on COVID-19 vaccines. **Dataset:** https://github.com/USC-Melady/Constructing-Misinformation-Datasets-WWW-2022.

CCS CONCEPTS

• **Information systems → Social networks**.

KEYWORDS

misinformation, datasets, labeling, social media, covid-19, vaccines

ACM Reference Format:
Karishma Sharma, Emilio Ferrara, and Yan Liu. 2022. Construction of Large-Scale Misinformation Labeled Datasets from Social Media Discourse using Label Refinement. In *Proceedings of the ACM Web Conference 2022 (WWW '22), April 25–29, 2022, Virtual Event, Lyon, France.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3485447.3512271

WWW'22, April 25–29, 2022, Virtual Event, Lyon, France

ACM ISBN 978-1-4503-9096-5/22/04.
https://doi.org/10.1145/3485447.3512271

1 INTRODUCTION

In recent times, malicious and coordinated promotion of misinformation coupled with uncertainties in real-world events, have sparked a plethora of *false* and *misleading* information on social media platforms [24]. Social media platforms struggle to eliminate these contents effectively and in a timely manner, and are recently attempting to solve the problem through more crowdsourced approaches to misinformation identification [3]. Twitter introduced *'Birdwatch'* in 2021, which allows people to identify tweets they believe are misleading and provide notes with additional context. This is an effort to respond more quickly to diverse false and misleading claims present on the platform. However, this comes with the biggest challenge for Twitter in ensuring that Birdwatch itself does not become prey to malicious coordinated operations. Secondly, due to ideological biases about the truth it is a challenge to build reliable consensus from platforms like Birdwatch [3].

The central challenge in timely misinformation detection, mitigation, and analysis is the difficulty in obtaining labeled misinformation datasets at scale, especially in new domains. Moreover, diverse and evolving false and misleading information based on changing real-world events, constantly surface on social media [22]. In the literature, the two primary approaches to constructing misinformation datasets are based on either collecting available **fact-checked claims** from organizations like Snopes, PolitiFact, etc. [14, 27], or utilizing **news-source credibility** labels based on reliable or unreliable sources listed by fact-checking organizations [21, 34]. The former approach suffers from claim selection bias with a slow and not scalable, human intensive fact-checking process, unsuitable for timely identification in new and evolving domains. The latter approach allows for more diverse, large-scale misinformation labeled social media posts, from a handful of analyzed news sources but can contain inaccurate labels in the dataset.

Proposed Approach: We address the above shortcomings with an alternate approach to construct misinformation datasets. We propose to utilize news-source credibility labels as weak labels for social media posts, and use *model-guided* refinement of labels to construct large-scale, diverse misinformation datasets in new domains. The news-source credibility based labels can be inaccurate at the article or social media post level when the stance of the user does not align with the news-source or article credibility. Therefore, for label refinement, we propose to use self-supervision from *any* generic misinformation detection model, with social context modeling of the social media posts. In this framework, we use a misinformation detection model trained on the initial weak labels,

with uncertainty sampling based on entropy in predictions of the model to identify potentially inaccurate labels and correct for them using self-supervision or relabeling. In addition, we incorporate the social context of the post in terms of the community of its associated user to model user credibility and stance in the discourse. The model-guided refinement is used to surface inaccurate labels iteratively, and minimize human labeling efforts, enabling timely scaling to large misinformation datasets with greater coverage.

The model-guided confidence in the labels is used to filter out or correct inaccurate weak labels, and the resulting dataset of social media posts with its engagements are labeled as misinformation/reliable, and associated with a model confidence in its label. The misinformation can be further segregated at finer-grained labels (such as false, mixture, true [22]) which can be obtained after label refinement with a *semi-supervised classification* setup [10, 32] in the proposed approach. Specifically, to provide finer-grained labels, we use the few human labeled examples as class prototypes to separate high confidence examples into *false, unproven, mixture, mostly false, mostly true, true, and debunk* information. More details on the labels and annotation guidelines are discussed later. The approach is demonstrated and applied for constructing and providing the research community with a large-scale public misinformation dataset on COVID-19 vaccines.

Contributions: Our contributions developed in this work are:

- Model-guided label refinement approach for timely construction of large-scale misinformation datasets.
- Label annotation guidelines and flexible framework that can generalize to other misinformation domains.
- Evaluation and construction of public misinformation dataset on COVID-19 vaccine social media data from Twitter.

In the following sections, we discuss the challenges in misinformation dataset construction, limitations of existing methods, related works, and the proposed approach and experiments, and conclude with a discussion of limitations and future work.

2 RELATED WORK

Misinformation datasets. Misinformation referring to false and distorted facts on social media has been addressed in numerous studies. A review of misinformation detection, mitigation techniques, and related datasets and tasks is comprehensively surveyed in [22]. The construction of misinformation datasets is a central task to enable research on misinformation detection, mitigation and analysis. Existing misinformation datasets are either general, such as over a specific time period [14] and cover content, social media engagements, and temporal features [27], or topic-specific datasets such as on the Syrian war [19]. The label scheme of datasets and the type of information collected vary based on the specifics of the task. For instance, for claim verification with external knowledge, datasets include content and evidence collected from the web that support or refute claims in the content [16]. The general detection task requires learning discriminative classifiers for misinformation claims, and usually includes content and its social media engagements, and the labels depend on the distinction made during data collection e.g. fake/real news [5] unreliable/reliable [34] rumors/non-rumors [14]. A comprehensive summary of several popular datasets in terms of their label classes and features is available in [22, 27].

Misinformation detection. Misinformation detection relies on learning discriminative features from labeled datasets, often utilizing the propagation features, content features, and account features [14, 17]. Wang et al. [30] in addition use weak supervision from user's reports to augment labeled misinformation datasets with unlabeled examples for misinformation detection. Shu et al. [26] use weak social supervision to similarly improve misinformation detection, i.e., where social media engagements are abundant but labeled misinformation content is not, modeling the interactions between social media users and contents to improve discrimination of misinformation content. Both these works are similar in flavor, in terms of *augmenting* the misinformation labeled datasets with auxiliary information to improve detection. In our work, we address how to scale the construction of misinformation labeled datasets using news-source credibility as initial weak labels.

Label refinement. Label noise in real-world data is common, and there are many different approaches to detect, remove, or correct it, which are relevant to this work [1, 11, 20]. Some works use local label inconsistencies in the feature space for detection [20], others utilize the training loss of deep neural network classifiers on the dataset to filter examples with high training loss in early epochs as noisy [1], or utilize entropy or variance in classifier predictions [8]. Other works focus on making classifiers more robust to label noise in datasets [29]. Active learning works address selection of instances from unlabeled or labeled datasets that are most useful to get human labels for to learn better models from the data, but depend on the presence of an 'oracle' i.e., human labeler, and utilize the expected model change from human labeling to select which instance to pick [11]. Here, we propose to additionally incorporate social context in label refinement, since in social media applications, the structures and context of social media users, as we show, provides relevant, complementary signals.

3 DATA COLLECTION

We collected social media posts on COVID-19 vaccines using Twitter's streaming API from December 9, 2020 - Feb 24, 2021 with keywords related to the vaccines ("Vaccine", "Pfizer", "BioNTech", "Moderna", "Janssen", "AstraZeneca", "Sinopharm"). The stream fetches a $\sim$ 1% sample of all tweets containing at least one of the keywords from the platform in real time. The data collection period started just prior to the first Emergency use authorization of Pfizer-BioNTech COVID-19 vaccine in the U.S. The dataset contains **4,764,701** unique user accounts with **15,158,523** collected tweets.

3.1 News-source credibility labels

Previous works that use news-source credibility for misinformation labeling, include news sources analyzed by different fact-checking organizations [4]. Bozarth et al. 2020 found that differences in lists based on the fact-checker it is compiled from affect prevalence, but not the temporal trends or differences in narratives of misinformation vs. legitimate contents labeled by these methods. In this work, more than prevalence, we are interesting in curating news sources to provide weak labels covering a diverse set of possible misinformation found in social media posts.

Therefore, we compile news-source credibility labels from multiple fact-checking resources, to encompass a wide range of low-credibility news sources. Following [21], we collect lists of unreliable and conspiracy news sources from three fact-checking resources: Media Bias/Fact[1], NewsGuard[2], and Zimdars [35][3]. NewsGuard maintains a repository of news publishing sources that have actively published false information during the COVID-19 pandemic. The listed sources from NewsGuard, accessed on September 22, 2020 are included, along with low and very-low factual reporting listed as questionable from Media Bias/Fact Check, and sources tagged with unreliable or related labels and conspiracy/pseudoscience from Zimdar's list. List of reliable sources[4] [21], covering high factual sources is also collected for obtaining the weak labels. In total, we obtained **1380** unreliable (or conspiracy) and **124** reliable sources. This choice of lists provides informative weak labels (ref. Section 4 and 6) but can be replaced or updated with other resources on news-source credibility analysis as needed.

3.2 News-related tweet cascades

On social media, content propagates through the network when accounts engage with posts by re-sharing (retweet), replying (reply tweets), quoting (quote tweets are retweets without a comment). A reply tweet can also be retweeted or quoted, and likewise for quote tweets. Therefore, source posts receive direct and subsequent indirect engagements through propagation over the network. This flow of information is referred to as an 'information cascade' [33] or tweet cascade. We represent it as a sequence of tweets, ordered by their time-stamp. The source post is the first tweet in the cascade. Formally, a cascade can be represented as follows with the user (u), tweet (tw), and temporal (t) features of when the users posted the engagements [18],

$$C_j = [(u_1, tw_1, t_1), (u_2, tw_2, t_2), \cdots (u_n, tw_n, t_n)] \quad (1)$$

Extracting tweet cascades. To extract the content cascades from the collected data, we use the retweet/reply/quote links between the tweets available from its metadata, and construct a directed graph of the tweets. We find the weakly-connected components of this graph, and each corresponds to one tweet cascade [23]. **Weak labels using news-sources.** The cascade is weakly labeled based on news-source credibility lists if the source post references one of the news sources. The news-source label (unreliable, conspiracy, reliable) is assigned to the cascade as its weak label.

We extracted tweet cascades from the collected Twitter data stream sample, keeping 490,638 user accounts that have at least 5 collected tweets in the sampled stream. The total tweets for these accounts is 9M. We weakly labeled the tweets as mentioned, and obtained **10,377** reliable cascades, and **4,267** unreliable or conspiracy cascades. These 14.6k cascades with weak labels will be used to construct the misinformation dataset as described in later sections.

[1]https://mediabiasfactcheck.com/
[2]https://www.newsguardtech.com/covid-19-resources/
[3]https://21stcenturywire.com/wpcontent/ uploads/2017/02/2017-DR-ZIMDARS-False-Misleading- Clickbait-y-and-Satirical-%E2%80%9CNews%E2%80%9DSources- Google-Docs.pdf
[4]https://en.wikipedia.org/wiki/Wikipedia:Reliable_sources/Perennial_sources

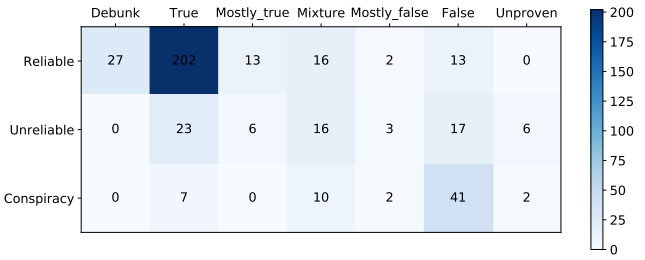


Figure 1: News-source credibility labels correlation with fact-checked claim based labels on validation and test tweets.

4 CHALLENGES IN MISINFORMATION LABELING

4.1 Existing approaches and limitations

Existing works apply two primary approaches to construct misinformation labeled datasets, either from claims verified by fact-checking organizations, or using credibility of the sources publishing the content. Both approaches suffer from drawbacks. We describe the approaches and summarize the drawbacks below. **Fact-checking based labeling.** One approach to collecting labeled misinformation contents, in the form of news articles, claims, or social media posts, is from contents verified by fact-checking websites (e.g., Snopes, PolitiFact). This approach is frequently used to construct datasets with few hundred or thousand labeled misinformation contents [12, 14, 27]. Then, for the fact-checked claims, social media engagements related to it are collected by search for content keywords using social media API's (e.g. Twitter search API). The matched social media posts containing these keywords are inspected to determine if they are relevant to the content [12], or the search keywords are refined until reasonably relevant matches of social media engagements are collected [14]. **News credibility-based labeling.** The other approach used for misinformation labeling is based on credibility of news sources [4]. Social media posts referencing content from any of these sources is labeled based on the source credibility to provide a dataset of unreliable and reliable contents. This is used frequently to identify misinformation posts from social media discourse for timely analysis in new domains [28, 34]. The drawbacks are summarized as follows.

- *Fact-checking based labeling.* It can have claim *selection bias*, since fact-checkers usually select claims to verify based on relevance or popularity (e.g., PolitiFact[5]), which can limit the diversity of collected claims for detection, and bias the analysis. Plus it is slow, human-intensive, and less scalable.
- *News source credibility-based labeling.* It scales to many social media posts using a handful of unreliable (questionable) and reliable sources, resulting in more diversity in claims, but has inherent label noise at the article or social media post level. Therefore, it can only provide weak labels.

[5]https://www.politifact.com/article/2018/feb/12/principles-truth-o-meter-politifacts-methodology-i/#How%20we%20choose%20claims

4.2 Correlation between news-source credibility and fact-checked claim based labels

First, we analyze the correlation between the labels from the two approaches. We collected fact-checked claims from Snopes.com[6] and NewsGuard[7] on COVID-19 vaccines. For each fact-checked claim, we find tweet cascades that discuss the claim by searching for text matches to words related to the claim. E.g. "Myth: The COVID-19 vaccine will use microchip surveillance technology created by Bill Gates-funded research." We search for source tweets with words "chip", "microchip", "surveillance" for matches. If nothing is found, we refine the search with"gates" and sample to check for matches.

NewsGuard provides only Myths (false claims), while Snopes provides varying factuality labels (true, mostly true, mixture, false etc.). From the Snopes collection tagged as COVID-19 vaccines, we obtained the claims labeled as one of these types (tagged as fact-checked) or labeled as news articles (AP news, The Conversation). Associated Press or AP news are rated as very high factual reporting and least biased by Media Bias/Fact Check. Therefore, we take claims from AP News as reliable. For more reliable claims, we also directly crawl the websites of AP news and NPR news (same factual and bias rating as AP news) for news articles (extracting the article heading, claim/short description, date) using python web-scraping. Snopes, AP news and NPR news together give us 400 claims, which we sample to find matching tweet cascades. We inspected the source tweet and labeled it based on the stance of the tweet to the fact-checked claim or reliable news article as 'true, mostly true, mixture, false, mostly false, unproven, debunk', similar to Snopes.

We found 256 tweet cascades to label based on the Snopes fact-checks and News articles. This forms our **evaluation test set of tweets with human expert ground-truth labels**. To additionally construct a **validation set** of human labeled tweets, we used stratified sampling of 150 additional tweets from the 14k tweet cascades and labeled them based on similar annotation guidelines as Snopes, described in the next subsection.

In Fig 1, we compare the news-source credibility based labels (unreliable/conspiracy/reliable) with the inspected fact-checked claim based labels for the human labeled tweets in the evaluation test and validation sets mentioned above. Overall, the news credibility labels appear to be well correlated with actual human labels. Individual inaccuracies can still exist, but with the large-scale weakly-labeled data smoothing out individual errors, we could learn to refine the weak labels to construct the misinformation dataset.

4.3 Annotation scheme and guidelines

Labeling misinformation is already challenging, more so because misinformation is not easy to specify [22]. It can lie on a spectrum of truth, including false, conspiracy [7], and misleading or *distorted* information such as missing or misleading contexts or mixture [13, 31]. We find that fact-checking organization Snopes uses a well-defined label schema that is **general enough to fit any domain**, and yet manages to **cover all types** and nuances of distortion we found upon examining tweets in the vaccines dataset, and generally in the literature [22]. Snopes includes several labels to cover the varying degree of truth and other deceptive tactics like miscaptioned, misattributed, scam. We work with the 6 most relevant Snopes categories, and add the 'Debunk' category based on what we observed in tweets. These cover even very specific types of anti-vaccine misinformation and science distortions [13].

The label scheme is proposed below, derived from Snopes and tweet inspection. We refine the label definitions to make the distinction between them and its coverage explicitly clear based on the inspected tweet data, for labeling social media posts based on their factualness. **Guideline:** Label the tweet based on what the tweet is trying to say or claim, and how factual its claim is. Choose one of the below labels for the tweet:

- True: Primary elements of the claim are demonstrably true.
- Debunk: Tweet calls out or debunks inaccurate information.
- Mostly true: Primary elements of a claim are demonstrably true, but some of the *ancillary details* surrounding the claim *may be inaccurate.*
- Mixture: Claim has *significant elements of both* truth and falsehood (including for e.g. *significant missing context* or *misleading* which might cause one to be misled about truth).
- Mostly false: Primary elements of a claim are false, but ancillary details may be accurate.
- False: Primary elements of a claim are false or conspiratorial.
- Unproven: Insufficient evidence that it is true, but for which declaring it false would require a difficult (if not impossible) task of proving a negative.

We evaluated the label scheme and guidelines on a random 200 sample subset from the collected tweet cascades. We compute the inter-annotator credibility for the tweets between two annotators, one graduate non-native English speaker familiar with misinformation research vs. one undergraduate native English speaker not familiar with the research. The agreement is moderate if we consider across the 7 label categories (0.61 Cohen's kappa), and substantial (0.77 Cohen's kappa) if binarized as (true, debunk, mostly true) vs. (mixture, mostly false, false, unproven) as high-level abstractions. Both annotators followed the same guideline and instructions with typical and difficult examples (noted in Appendix A).[8]

5 MODEL-GUIDED LABEL REFINEMENT

We propose an alternate approach to constructing misinformation datasets at scale, addressing the shortcomings of existing approaches. We propose to use news-source credibility as weak labels and leverage the large-scale weak labeled data with label refinement to construct misinformation datasets minimizing time-consuming, and not scalable human labeling efforts.

In the previous section, we observed that the news credibility labels are correlated overall with actual fact-checked claim based labels, and with the large scale of the weak labeled dataset smoothing out individual errors, we could learn to remove inaccuracies. The inaccuracies in these weak labels arise at two levels: 1) **Article**

[6]https://www.snopes.com/tag/covid-19-vaccine/

[7]https://www.newsguardtech.com/special-reports/special-report-top-covid-19-vaccine-myths/

[8]This label scheme and instruction set was finalized after two iterations. At first, we started with 11 labels with vaccine specific instructions e.g. potential to mislead (can be misinterpreted as unsafe), missing context, etc) but found difficulties in making clear distinctions and following the specified instructions. This label scheme was clear to follow and we found that sufficient examples of tricky and typical cases in each category was very helpful to the annotators, who were asked to review the instructions and examples before annotating, and had access to reference it when in doubt.

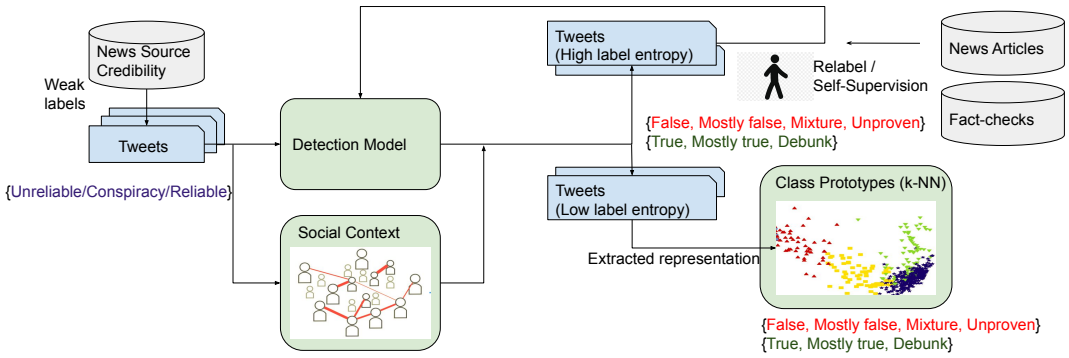


Figure 2: Proposed approach: Model-guided label refinement with self-supervision from a generic misinformation detection model and social context modeling to construct large-scale misinformation labeled social media datasets in a timely manner.

level. First, not all contents published by misinformation news sources might contain misinformation, although they tend to be unreliable or repeatedly violate journalistic reporting principles, enough for the source to be included as a low factual reporting source by experts. 2) **Tweet level.** Secondly, the weak label can be incorrect based on the stance of the social media post to the content from the news source. The post or tweet might reference content from the source with its supporting viewpoint or restating it as is. Or in other cases, oppose or distort the content from the source, which would result in a mislabeling at the level of a tweet.

In **Fig. 2**, we provide the proposed framework for misinformation dataset construction and labeling in new domains at scale. The weak labels **unreliable, conspiracy, reliable** on tweet cascades are from news-source credibility. It is utilized to construct the initial dataset with $y \in \{0, 1\}$ as weak labels with 1 as unreliable/conspiracy and 0 as reliable. Our goal is to remove or correct inaccurate weakly labeled instances in the dataset, and output high-level distinctions of misinformation label as 1 and reliable information label as 0, with the model-guided predictions of confidence in the labels.

5.1 Self-supervision from misinformation detection modeling

In the proposed framework, we make use of *any* generic misinformation detection model to guide the weak label refinement. Classifiers are often utilized to estimate uncertainty in the instance labels from the loss or model predictions in label noise methods [1, 6]. In this work, we use entropy of the misinformation detection model predictions to measure closeness from the decision boundary [6]. High entropy indicates greater model uncertainty about the label. The entropy S_i in the model predictions for an instance i is defined as follows, where $p(x_i)$ is the vector of predicted probabilities from the detection model, and k is the classes, here for $y \in \{0, 1\}$.

$$S_i = -\sum_{k=1}^{L} p_k(x_i) \log(p_k(x_i)) \tag{2}$$

We train the detection model on all weak labeled data, and then filter out high entropy instances. We also filter out tweets with low entropy model predictions if the initial weak label and predicted model label for the tweet are inconsistent with each other. This is for instances where the model is confident in its prediction, either has an incorrect weak label or predicted label.

With the filtered dataset, we retrain the detection model, and repeat until the model has marginal improvements on a held-out small human-labeled or fact-checked labeled validation set is marginal. This iterative self-training improves the detection model and its signals of the inaccuracies in weak labels. In each iteration, the retrained detection model is applied to all instances in the initial dataset to calculate the entropy scores, and filter for the next iteration as it can now make more informed filtering decisions than those in the previous iterations.

5.2 Social context modeling

In misinformation applications, social media engagements are known to provide useful signals for misinformation detection [27]. Here, we propose that the social context can also be useful in guiding the construction of misinformation labeled datasets. We describe how *social context can be modeled and leveraged* in this application.

We incorporate social context of the post using the community of its associated user to model a user account's credibility and stance in the discourse. Social media discourse tends to be segregated into *echo-chambers* of user accounts sharing similar opinions [9]. User accounts follow each other based on their interests, and become more exposed to contents that align with their interest and ideologies [22]. The retweet graph between user accounts that retweeted each other's tweets can be used to identify **user communities**. Retweets are seen as a form of endorsement of content and edges with at least two retweets are retained to capture links of similar interests in the user accounts [9]. We identify user account communities from the retweet graph using *Louvain* method [2].

To leverage the social context, we identify communities that dominantly post or share misinformation sources vs. reliable sources. Several works have found that *informed* and *misinformed* user accounts exhibit echo-chambers in their network structure [15, 25]. For the identified communities, if tweets of user accounts in the community dominantly contain references to misinformation news sources, the communities are likely less to be credible, or are more misinformed. Misinformation communities would involve either malicious groups promoting misinformation, or groups with beliefs

that support or are vulnerable to believing and sharing misinformation on the topic of the discourse [15]. We can thereby leverage this to encode a **user account's credibility and stance** with respect to misinformation on the topic of the discourse. Here, we denote social context of a tweet as the community of the user account that posted the tweet. Given the community structure, the tweet cascade is detected as possibly mislabeled if,

- the user account belongs to an identified dominant misinformation (unreliable/conspiracy) news-source sharing community, but the tweet is weakly-labeled as reliable.
- or, if the account is in a dominantly reliable information sharing community, but the weak label is unreliable/conspiracy.

For mixed communities with unclear dominant reliable or misinformation sharing patterns, we have no definitive social context for label refinement, and use only the detection entropy.

5.3 Iterative label refinement

We jointly use the social context signal and the detection model entropy to guide the identification of post-level mistakes in the weakly-labeled data. The proposed framework (Fig 2) is iterative and flexible. We can replace the misinformation detection model with any modeling choice, and use either **self-supervision and/or human/model based relabeling**. The detection model is first trained by itself with self-supervision from the model predictions. Then the improved detection model signals are jointly combined with social context for further label refinement. The process is iteratively repeated with evaluation of detection model on small held-out human labeled or fact-checked validation set as a proxy for label quality in the large-scale dataset.

The procedure for label refinement from detection model and social context is described in **Algorithm 1**. The subroutine assumes as input the instance (tweet cascade denoted as x_i, with its weak label y_i), the detection model M trained in the previous iteration, and the social context S. Given the model state, we generate three possible **actions:** (1) RETAIN weak label (2) FLIP weak label (3) QUERY label. Action retain keeps the instance with its weak label in the dataset, flip is model-guided relabeling (without human resource), and query is for active human relabeling of the model suggested instance. If the human resource is not available, then QUERY can be replaced by REMOVAL (discarding the instance due to low confidence in its label or due to contradictory confident signals from detection model M and social context S).

The **states** from the detection model M and social context S are defined as follows, for instance x_i,

- M-lc: If high-entropy S_i in detection model prediction (M-lc stands for low confidence, that is, high entropy)
- M-consistent and M-inconsistent: If low entropy S_i, and M predicted label equals weak label then it is consistent (opposite predicted label and weak label, then inconsistent)
- S-unk: no social context signal, either its user's community is not dominantly reliable or unreliable/conspiracy, but a mixture; or the user is not clustered in any main community.
- S-consistent and S-inconsistent: If the social context of a user account (its community label) is (in)consistent with the weak label of its tweet in x_i (as described earlier).

Algorithm 1 Label Refinement Procedure

Require: Dataset instance x_i, weak label y_i, and detection model M, and social context S
Ensure: Action: RETAIN, FLIP, QUERY label

```
if M-consistent and S-consistent then
    RETAIN with weak label // reinforced signals from M and S
else if M-inconsistent and S-inconsistent then
    FLIP weak label // reinforced signals from M and S
else if (M-consistent and S-inconsistent) or (M-inconsistent and S-consistent) then
    QUERY label // contrasting signals from M and S
else if M-consistent and S-unk then
    RETAIN with weak label // only signals from M
else if M-inconsistent and S-unk then
    FLIP weak label // only signals from M
else
    QUERY label // detection model high entropy filtering
end if
```

The **objective** of the procedure is to minimize human relabel queries, and incorporate high confidence signals from both detection model M and social context S to ultimately remove or correct as many inaccurate weak labels, keeping as many correctly weak labeled instances. If the signals reinforce each other, the procedure can more confidently take an action without human label querying (or removal/discarding of the instance). Given the state, the appropriate action is selected by the procedure Alg. 1.

Fine-grained semi-supervised classification. The dataset is refined based on retaining weak labels, model based relabeling, and human relabeling or removal of the instance. The retained and refined instances form the output constructed dataset with the associated model confidence in its label. The fine-grained labels are obtained by the human labeling but only on selected instances **false, unproven, mixture, mostly false, mostly true, true, and debunk**. For the remaining instances, we can use a *semi-supervised classification* setup [10, 32] to obtain fine-grained distinctions. The few obtained human labeled instances become class prototypes to separate the rest into the seven classes. The distinctions can be very nuanced with varying degrees of truth, and difficult for a model to distinguish very accurately, so we provide these as auxiliary outputs.

6 EXPERIMENTS

We study the proposed approach for constructing a **large-scale public misinformation dataset** on **COVID-19 vaccines**. We use iterative self-training of the misinformation detection model CSI [18] trained first on the initial weakly-labeled cascades. We use low-quality news sources for weak labels on the collected Twitter dataset, compiled from fact-checking resources as described earlier in Section 3. We have a total of 14.6k tweet cascades with roughly 10,377 weakly labeled as reliable and 4,267 weakly labeled as unreliable/conspiracy. With this setting, we experiment with the proposed framework for large-scale misinformation dataset construction from weak news-source labels.

Table 1: Results on classification performance on test set from detection model with label refinement proposed approach for misinformation dataset construction on COVID-19 vaccines. Metrics: AP (average precision), AUC (ROC), F1 and Macro F1.

Experiment	AP	AUC	F1	Macro F1
Weak labels	0.722 ± 0.03	0.876 ± 0.01	0.774 ± 0.02	0.812 ± 0.01
Self-training (iteration 1)	0.768 ± 0.01	0.888 ± 0.0	0.812 ± 0.01	0.842 ± 0.01
Self-training (iteration 2)	0.775 ± 0.02	0.891 ± 0.0	0.811 ± 0.01	0.842 ± 0.01
Social-context only	0.764 ± 0.02	0.891 ± 0.01	0.810 ± 0.01	0.837 ± 0.01
Social+Detection model	0.785 ± 0.02	0.895 ± 0.0	0.813 ± 0.01	0.842 ± 0.01
Social+Detection (+label correction)	**0.800 ± 0.01**	**0.895 ± 0.0**	**0.818 ± 0.01**	**0.845 ± 0.01**

Table 2: Results for noise detection in weak labels with label refinement proposed approach for misinformation dataset construction on COVID-19 vaccines. Evaluation metrics: Rec (noise recall), Prec (precision), Frac UQ (fraction of unwanted queries), F1 (F1 of detected noise in weak labels).

	Test set				Validation set			
Experiment	**Rec**	**Prec**	**Frac UQ**	**F1**	**Rec**	**Prec**	**Frac UQ**	**F1**
Naive	1.0	0.1719	1.0	0.2934	1.0	0.1533	1.0	0.2658
Self-training (iteration 2)	0.5682	0.3846	0.1887	0.4587	0.6522	0.4054	0.1732	0.5000
Social-context only	0.4545	0.4444	0.1179	0.4494	0.4348	0.3704	0.1339	0.4000
Social+Detection model	0.8409	0.3978	0.2642	0.5401	0.7826	0.3673	0.2441	0.5000
Social+Detection (+label flipping)	0.8409	0.3978	0.2406	0.5401	0.7826	0.3673	0.2126	0.5000

Table 3: Fine-grained classification from human labeled class prototypes to remaining examples in the dataset.

Expt (5-fold)	Macro F1	Weighted F1	F1 (debunk)	(true)	(mostly_true)	(mixture)	(mostly_false)	(false)	(unproven)	Acc
Random	0.113 +/- 0.037	0.201 +/- 0.067	0.071 +/- 0.089	0.254 +/- 0.111	0.042 +/- 0.084	0.16 +/- 0.101	0.125 +/- 0.125	0.097 +/- 0.058	0.044 +/- 0.089	0.164 +/- 0.043
Majority	0.105 +/- 0.007	0.428 +/- 0.078	0.0 +/- 0.0	0.733 +/- 0.049	0.0 +/- 0.0	0.0 +/- 0.0	0.0 +/- 0.0	0.0 +/- 0.0	0.0 +/- 0.0	0.581 +/- 0.063
Unweighted MLP	0.219 +/- 0.022	0.552 +/- 0.073	0.04 +/- 0.08	0.757 +/- 0.045	0.0 +/- 0.0	0.257 +/- 0.074	0.0 +/- 0.0	0.483 +/- 0.064	0.0 +/- 0.0	0.595 +/- 0.054
Weighted MLP	0.266 +/- 0.055	0.57 +/- 0.077	0.235 +/- 0.145	0.735 +/- 0.05	0.033 +/- 0.067	0.338 +/- 0.165	0.0 +/- 0.0	0.52 +/- 0.1	0.0 +/- 0.0	0.546 +/- 0.061

6.1 Evaluation

The evaluation **test set** of tweet cascades contains **256 tweets** with ground-truth fact-checked claim based labels obtained by searching for tweets related to Snopes fact-checks and AP news/NPR news on COVID-19 vaccines and labeling from the 7 fine-grained labels according to the annotation scheme and fact-checked claims. For experiments, a human-labeled **validation set** of **150 tweets**, based on the annotation scheme and guidelines, is also constructed and held-out from the **14.6k tweet cascades** (as described in Sec 4.2).

6.1.1 Evaluation tasks. We cannot directly measure the quality of the constructed misinformation dataset, since we cannot obtain ground-truth fact-checker (e.g. Snopes) labels on all 14.6k tweet cascades. We instead evaluate on the fact-checked claim based test subset of 256 tweet cascades using the following evaluation metrics (i) **Misinformation detection performance** on test set. Label quality in the dataset should be positively correlated with misinformation detection accuracy on ground-truth labeled data. (ii) **Label correction accuracy** on validation and test sets and (iii) # of wasted queries generated in the label refinement procedure, to measure human **resources that are inefficiently utilized**.

6.1.2 Evaluation metrics. (i) The baselines and proposed experiments are evaluated for **misinformation detection (classification) performance** on the ground-truth test set of 256 tweet cascades. The detection model performance is averaged over 5 random seeds. The evaluation includes the classification metrics namely, Area under the precision-recall curve (AP) and Area under the ROC curve (AUC) and F1 and Macro-F1 for the misinformation classification. It measures how well the refined/constructed dataset from baselines or the proposed approach work in separating the misinformation from reliable information tweet cascades.

(ii) The baselines and proposed experiments are evaluated for **label correction accuracy** on the ground-truth test set and validation set. We have the initial weak labels and correct misinformation labels for the test and validation set. Therefore, we can measure the recall (Rec), precision (Prec), and F1 of the noise in the weak labels (i.e., weak label and ground-truth fact-checked label are not aligned). The instances detected as noise by the methods are ones selected for FLIP or QUERY (REMOVE) actions (as it is predicted by the method as having a possibly mislabeled weak label). Recall is the fraction of actual noise in weak labels that are correctly detected by the methods, and Precision measures the correctly recalled noise in weak labels out of all instances detected as noise by the methods. F1 is the harmonic mean of precision and recall.

(iii) We additionally propose a metric to also measure how **efficiently the resources are utilized** by the baselines and proposed methods. We define Frac UQ (fraction of correctly weak-labeled instances that are assigned QUERY (REMOVE) action for human

relabeling (removal), i.e. unwanted or wasted queries) as follows,

$$\text{Frac UQ} = \frac{|\text{ (QUERY action assigned) \& (correct weak label) }|}{|\text{correct weak label}|} \tag{3}$$

The # of instances with correct weak labels assigned QUERY (REMOVE) action is the numerator, measuring human resource wastage. *Lower value of Frac UQ is better*, while maintaining high noise recall.

6.2 Results

Misinformation detection performance. In **Table 1** we provide results of the proposed framework to construct misinformation datasets from weak labels. We trained the CSI [18] misinformation detection model on weak labels from news source credibility to classify misinformation (unreliable/conspiracy) tweets from reliable information tweets, as a baseline. The held-out validation set is used by the detection model for early stopping in model optimization, and for calculating the threshold for detection based on AUC curve, trading off sensitivity and specificity on the validation set. The reported results are on the held-out ground-truth test set of fact-checked based labels. The same setting is used in all experiments.

The removal (entropy filtering) guided by the detection model (self-training iteration 1 and 2) improves the classification on the ground-truth test set, indicative of higher label quality in the retained tweets. After 2 iterations, the improvement was insignificant.

Further, incorporating social context modeling, we first evaluate Social-context only, wherein the tweets with labels opposite to their community label (dominantly reliable or dominantly misinformation) are surfaced as to be queried or removed. We find that combining the social context modeling and detection model guidance is more informative about possible mislabeling (tweets to be removed) in the weak labels (Social+Detection model). Finally, Social+Detection model (+label correction) is used to correct the labels that the two signals suggest should be oppositely labeled, and remove ones that the model is unsure about either from the detection model or social context (i.e., using the label refinement procedure in Alg 1). We find results in model-guided label refinement for construction of misinformation datasets is effective and significantly improves both recall in misinformation detection, (since the misinformation examples are fewer in the imbalanced data), and the precision of detected misinformation, and other metrics separating the two classes of misinformation and reliable information.

Label correction accuracy and resource efficiency of label refinement. In **Table 2**, we provide the results of performance on label correction using the signals from the detection model and/or social context. Naive baseline is trivially set to assume all weak labels are mislabeled, and QUERY all of them. Therefore, all correctly weak-labeled instances are Queried with worst resource utilization of 1, and low precision, F1 scores. For the removal (entropy filtering) guided by the detection model (self-training iteration 2) has roughly 56% recall in inaccurate weak labels, with reasonable precision and low Frac UQ. It is similarly the case for Social-context only method.

With the proposed approach (i.e., using the label refinement procedure in Alg 1) combining the social context modeling and detection model guidance is more informative about possible mislabeling in the weak labels (Social+Detection model) and we see a massive increase in recall on combining the two signals, resulting in

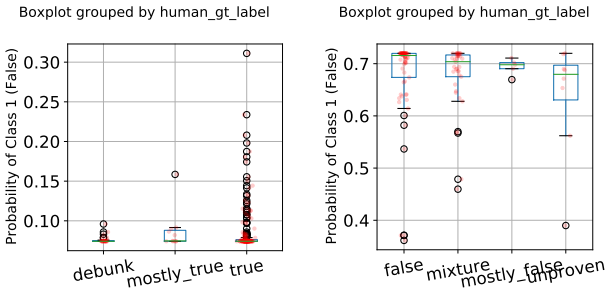


Figure 3: Predicted probabilities from detection model after label refinement: correlation with true labels on ground-truth fact-checked human-labeled test and validation set.

0.84 recall. This might suggest that the signals provided by detection model entropy filtering and social context are **complementary** to each other, and jointly inform label refinement in the proposed algorithm most effectively. With label FLIP action included (Social+Detection (+label flipping), the difference is only in Frac UQ, where now some of the detected noise will be directly selected for FLIP action instead of for removal (or query), minimizing the wasted queries, if the FLIP was assigned to actual noisy instances.

These evaluation metrics suggest how well the proposed method works at constructing high-quality misinformation labels, with the least cost incurred in terms of the human labeling resources, or mistakes in identification of possible incorrect weak labels.

Constructed misinformation dataset. In the constructed misinformation dataset derived from weak labels with the proposed method, in Fig. 3, we examine scatter plot of instances on the predicted probability of misinformation from the detection model, which as we see is correlated with the fine-grained human labels available on the validation and test set, capturing the varying degree of truth. In Table 3, we show the fine-grained classification from human labeled class prototypes on remaining examples in the dataset, using 5-fold cross validation on stratified splits of the validation plus test set for evaluation. For classification, we additionally labeled 400 instances to include as human-labeled class prototypes. We used extracted representations of tweet cascades from the CSI detection model used here, to train an MLP. With class-weighting, the fine-grained classifier has 0.57 weighted F1 distinguishing over the 7 nuanced label categories which is a difficult task.

7 DISCUSSION AND CONCLUSIONS

The proposed label refinement approach is effective at constructing large-scale datasets from weak labels with high recall of inaccurate weak labels when incorporating social context jointly with entropy filtering. We provide discussions of the proposed approach and three potential, concrete future research directions: (1) The weak labels are collected from news-source credibility, so posts with references to news sources form the basis of the dataset. The dataset might be more centered on news-worthy contents, which is one limitation of the approach. Other sources to augment weak labels could be considered in the future. Also, the annotation scheme

label categories are general, but effective ways to construct expert examples as annotator guidelines in new domains could be explored.

(2) The proposed approach models *instance credibility* and *user credibility* through entropy filtering and social context modeling. The label refinement procedure could be provided additional signals of news-source credibility by modeling each news source separately (since each might cause different noise rates, and unreliable sources are more mixed than conspiracy sources, as we observed). (3) The fine-grained classification is a difficult task and future research directions could explore model-guided selection of instances for human labeling, to act as the most effective class prototypes, with richer semi-supervised fine-grained classification techniques.

To conclude, the proposed label refinement with social context modeling is a useful, new approach for constructing misinformation datasets, in a timely and scalable way for new or evolving domains.

ACKNOWLEDGMENTS

This work is supported by NSF Research Grant (CCF-1837131) and DARPA (HR001121C0169 and W911NF-17-C-0094). Views and conclusions are of the authors and should not be interpreted as representing the social policies of the funding agency, or the U.S. Government. We thank Feng Pan and Cindy Lin for helping with annotation guidelines and labeling efforts.

REFERENCES

[1] Eric Arazo, Diego Ortego, Paul Albert, Noel O'Connor, and Kevin Mcguinness. 2019. Unsupervised Label Noise Modeling and Loss Correction. In *International Conference on Machine Learning*. 312–321.

[2] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 2008. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment* 2008, 10 (2008), P10008.

[3] Shannon Bond. 2021. *Twitter's 'Birdwatch' Aims to Crowdsource Fight Against Misinformation.* Retrieved 2021 from https://www.npr.org/2021/02/10/965839888/twitters-birdwatch-aims-to-crowdsource-fight-against-misinformation

[4] Lia Bozarth, Aparajita Saraf, and Ceren Budak. 2020. Higher Ground? How Groundtruth Labeling Impacts Our Understanding of Fake News about the 2016 U.S. Presidential Nominees. *Proceedings of the International AAAI Conference on Web and Social Media* 14, 1 (May 2020), 48–59. https://ojs.aaai.org/index.php/ICWSM/article/view/7278

[5] Enyan Dai, Yiwei Sun, and Suhang Wang. 2020. Ginger Cannot Cure Cancer: Battling Fake Health News with a Comprehensive Data Repository. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 14. 853–862.

[6] Sergiy Fefilatyev, Matthew Shreve, Kurt Kramer, Lawrence Hall, Dmitry Goldgof, Rangachar Kasturi, Kendra Daly, Andrew Remsen, and Horst Bunke. 2012. Label-noise reduction with support vector machines. In *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*. IEEE, 3504–3508.

[7] Emilio Ferrara, Herbert Chang, Emily Chen, Goran Muric, and Jaimin Patel. 2020. Characterizing social media manipulation in the 2020 US presidential election. *First Monday* (2020).

[8] Benoît Frénay and Michel Verleysen. 2013. Classification in the presence of label noise: a survey. *IEEE transactions on neural networks and learning systems* 25, 5 (2013), 845–869.

[9] Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. 2018. Quantifying controversy on social media. *ACM Transactions on Social Computing* 1, 1 (2018), 1–27.

[10] Thomas N Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907* (2016).

[11] Jan Kremer, Fei Sha, and Christian Igel. 2018. Robust active label correction. In *International conference on artificial intelligence and statistics*. PMLR, 308–316.

[12] Sejeong Kwon, Meeyoung Cha, and Kyomin Jung. 2016. Rumor detection over varying time windows.. In *Harvard Dataverse*. Harvard Dataverse. https://doi.org/10.7910/DVN/BFGAVZ

[13] Stephan Lewandowsky, John Cook, Philipp Schmid, Dawn Liu Holford, Adam Finn, Julie Leask, Angus Thomson, Doug Lombardi, Ahmed K Al-Rawi, Michelle A Amazeen, et al. 2021. The COVID-19 Vaccine Communication Handbook. A practical guide for improving vaccine communication and fighting misinformation.

[14] Jing Ma, Wei Gao, Prasenjit Mitra, Sejeong Kwon, Bernard J Jansen, Kam-Fai Wong, and Meeyoung Cha. 2016. Detecting Rumors from Microblogs with Recurrent Neural Networks.. In *IJCAI*. 3818–3824.

[15] Shahan Ali Memon and Kathleen M Carley. 2020. Characterizing COVID-19 misinformation communities using a novel twitter dataset. In *CEUR Workshop Proceedings*, Vol. 2699.

[16] Kashyap Popat, Subhabrata Mukherjee, Jannik Strötgen, and Gerhard Weikum. 2017. Where the Truth Lies: Explaining the Credibility of Emerging Claims on the Web and Social Media. In *Proceedings of the 26th International Conference on World Wide Web Companion - WWW '17 Companion*. ACM Press, Perth, Australia, 1003–1012. https://doi.org/10.1145/3041021.3055133

[17] Feng Qian, Chengyue Gong, Karishma Sharma, and Yan Liu. 2018. Neural User Response Generator: Fake News Detection with Collective User Intelligence.. In *IJCAI*, Vol. 3834. 3840.

[18] Natali Ruchansky, Sungyong Seo, and Yan Liu. 2017. CSI: A Hybrid Deep Model for Fake News Detection. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. ACM, 797–806.

[19] Fatima K Abu Salem, Roaa Al Feel, Shady Elbassuoni, Mohamad Jaber, and May Farah. 2019. FA-KES: a fake news dataset around the Syrian war. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 13. 573–582.

[20] Karishma Sharma, Pinar Donmez, Enming Luo, Yan Liu, and I Zeki Yalniz. 2020. Noiserank: Unsupervised label noise reduction with dependence models. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVII 16*. Springer, 737–753.

[21] Karishma Sharma, Emilio Ferrara, and Yan Liu. 2022. Characterizing Online Engagement with Disinformation and Conspiracies in the 2020 U.S. Presidential Election. In *Proceedings of the International AAAI Conference on Web and Social Media*.

[22] Karishma Sharma, Feng Qian, He Jiang, Natali Ruchansky, Ming Zhang, and Yan Liu. 2019. Combating fake news: A survey on identification and mitigation techniques. *ACM Transactions on Intelligent Systems and Technology (TIST)* 10, 3 (2019), 1–42.

[23] Karishma Sharma, Sungyong Seo, Chuizheng Meng, Sirisha Rambhatla, and Yan Liu. 2020. Covid-19 on social media: Analyzing misinformation in twitter conversations. *arXiv e-prints* (2020), arXiv–2003.

[24] Karishma Sharma, Yizhou Zhang, Emilio Ferrara, and Yan Liu. 2021. Identifying Coordinated Accounts on Social Media through Hidden Influence and Group Behaviours. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining* (Virtual Event, Singapore) *(KDD '21)*. Association for Computing Machinery, 1441–1451. https://doi.org/10.1145/3447548.3467391

[25] Karishma Sharma, Yizhou Zhang, and Yan Liu. 2021. COVID-19 Vaccines: Characterizing Misinformation Campaigns and Vaccine Hesitancy on Twitter. arXiv:2106.08423 [cs.SI]

[26] Kai Shu, Susan Dumais, Ahmed Hassan Awadallah, and Huan Liu. 2020. Detecting fake news with weak social supervision. *IEEE Intelligent Systems* 36, 4 (2020), 96–103.

[27] Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. 2020. FakeNewsNet: A Data Repository with News Content, Social Context, and Spatiotemporal Information for Studying Fake News on Social Media. *Big Data* 8, 3 (2020), 171–188.

[28] Lisa Singh, Shweta Bansal, Leticia Bode, Ceren Budak, Guangqing Chi, Kornraphop Kawintiranon, Colton Padden, Rebecca Vanarsdall, Emily Vraga, and Yanchen Wang. 2020. A first look at COVID-19 information and misinformation sharing on Twitter. *arXiv preprint arXiv:2003.13907* (2020).

[29] Hwanjun Song, Minseok Kim, Dongmin Park, Yooju Shin, and Jae-Gil Lee. 2020. Learning from noisy labels with deep neural networks: A survey. *arXiv preprint arXiv:2007.08199* (2020).

[30] Yaqing Wang, Weifeng Yang, Fenglong Ma, Jin Xu, Bin Zhong, Qiang Deng, and Jing Gao. 2020. Weak supervision for fake news detection via reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 516–523.

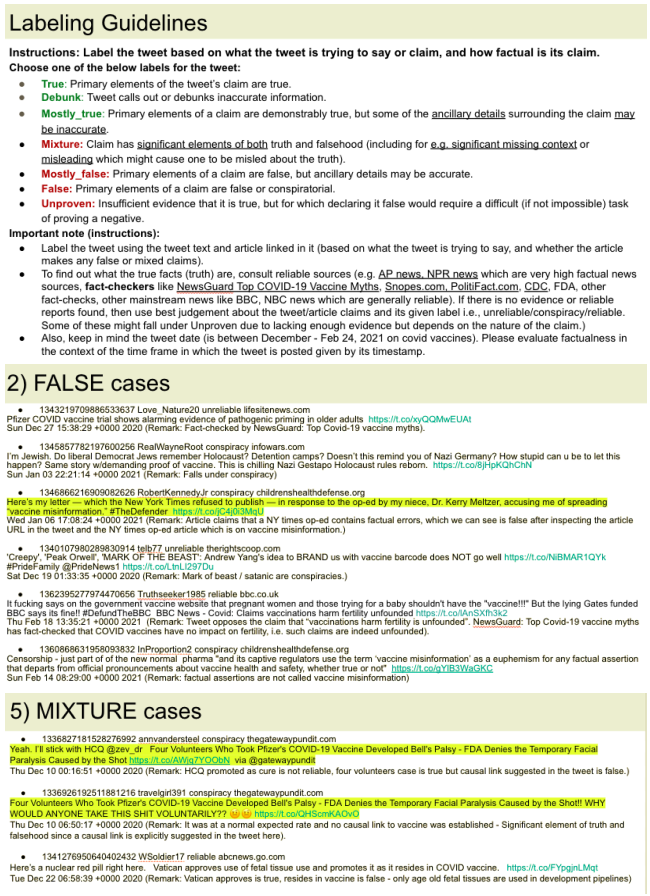
[31] Claire Wardle. 2017. *Fake news. It's complicated.* Retrieved 2019 from https://firstdraftnews.org/fake-news-complicated/

[32] Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le. 2020. Unsupervised Data Augmentation for Consistency Training. *Advances in Neural Information Processing Systems* 33 (2020).

[33] Jaewon Yang and Jure Leskovec. 2010. Modeling information diffusion in implicit networks. In *2010 IEEE International Conference on Data Mining*. IEEE, 599–608.

[34] Xinyi Zhou, Apurva Mulay, Emilio Ferrara, and Reza Zafarani. 2020. Recovery: A multimodal repository for covid-19 news credibility research. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. 3205–3212.

[35] Melissa Zimdars. 2016. *False, Misleading, Clickbait-Y, and Satirical 'News' Sources.* Retrieved 2019 from https://21stcenturywire.com/wp-content/uploads/2017/02/2017-DR-ZIMDARS-False-Misleading-Clickbait-y-and-Satirical-%E2%80%9CNews%E2%80%9D-Sources-Google-Docs.pdf



Labeling Guidelines

Instructions: Label the tweet based on what the tweet is trying to say or claim, and how factual is its claim.
Choose one of the below labels for the tweet:

- **True**: Primary elements of the tweet's claim are true.
- **Debunk**: Tweet calls out or debunks inaccurate information.
- **Mostly_true**: Primary elements of a claim are demonstrably true, but some of the ancillary details surrounding the claim may be inaccurate.
- **Mixture:** Claim has significant elements of both truth and falsehood (including for e.g. significant missing context or misleading which might cause one to be misled about the truth).
- **Mostly_false:** Primary elements of a claim are false, but ancillary details may be accurate.
- **False:** Primary elements of a claim are false or conspiratorial.
- **Unproven:** Insufficient evidence that it is true, but for which declaring it false would require a difficult (if not impossible) task of proving a negative.

Important note (instructions):

- Label the tweet using the tweet text and article linked in it (based on what the tweet is trying to say, and whether the article makes any false or mixed claims).
- To find out what the true facts (truth) are, consult reliable sources (e.g. AP news, NPR news which are very high factual news sources, **fact-checkers** like NewsGuard Top COVID-19 Vaccine Myths, Snopes.com, PolitiFact.com, CDC, FDA, other fact-checks, other mainstream news like BBC, NBC news which are generally reliable). If there is no evidence or reliable reports found, then use best judgement about the tweet/article claims and its given label i.e., unreliable/conspiracy/reliable. Some of these might fall under Unproven due to lacking enough evidence but depends on the nature of the claim.)
- Also, keep in mind the tweet date (is between December - Feb 24, 2021 on covid vaccines). Please evaluate factualness in the context of the time frame in which the tweet is posted given by its timestamp.

2) FALSE cases

- 1343219709886533637 Love_Nature20 unreliable lifesitenews.com
Pfizer COVID vaccine trial shows alarming evidence of pathogenic priming in older adults https://t.co/xyQQMwEUAt
Sun Dec 27 15:38:29 +0000 2020 (Remark: Fact-checked by NewsGuard: Top Covid-19 vaccine myths).

- 1345857782197600256 RealWayneRoot conspiracy infowars.com
I'm Jewish. Do liberal Democrat Jews remember Holocaust? Detention camps? Doesn't this remind you of Nazi Germany? How stupid can u be to let this happen? Same story w/demanding proof of vaccine. This is chilling Nazi Gestapo Holocaust rules reborn. https://t.co/8jHpKQhChN
Sun Jan 03 22:21:14 +0000 2021 (Remark: Falls under conspiracy)

- 1346866216909082626 RobertKennedyJr conspiracy childrenshealthdefense.org
Here's my letter — which the New York Times refused to publish — in response to the op-ed by my niece, Dr. Kerry Meltzer, accusing me of spreading "vaccine misinformation." #TheDefender https://t.co/jC4d93MsU
Wed Jan 06 17:08:24 +0000 2021 (Remark: Article claims that a NY times op-ed contains factual errors, which we can see is false after inspecting the article URL in the tweet and the NY times op-ed article which is on vaccine misinformation.)

- 1340107980289830914 telb77 unreliable therightscoop.com
'Creepy', 'Peak Orwell', 'MARK OF THE BEAST': Andrew Yang's idea to BRAND us with vaccine barcode does NOT go well https://t.co/NiBMAR1QYk
#PrideFamily @PrideNews1 https://t.co/LtnLI297Du
Sat Dec 19 01:33:35 +0000 2020 (Remark: Mark of beast / satanic are conspiracies.)

- 1362395277974470656 Truthseeker1985 reliable bbc.co.uk
It fucking says on the government vaccine website that pregnant women and those trying for a baby shouldn't have the "vaccine!!!" But the lying Gates funded BBC says its fine!! #DefundTheBBC BBC News - Covid: Claims vaccinations harm fertility unfounded https://t.co/lAnSXfh3k2
Thu Feb 18 13:35:21 +0000 2021 (Remark: Tweet opposes the claim that "vaccinations harm fertility is unfounded". NewsGuard: Top Covid-19 vaccine myths has fact-checked that COVID vaccines have no impact on fertility, i.e. such claims are indeed unfounded).

- 1360868631958093832 InProportion2 conspiracy childrenshealthdefense.org
Censorship - just part of of the new normal pharma "and its captive regulators use the term 'vaccine misinformation' as a euphemism for any factual assertion that departs from official pronouncements about vaccine health and safety, whether true or not" https://t.co/gYlB3WaGKC
Sun Feb 14 08:29:00 +0000 2021 (Remark: factual assertions are not called vaccine misinformation)

5) MIXTURE cases

- 1336827181528276992 annvandersteel conspiracy thegatewaypundit.com
Yeah. I'll stick with HCQ @zev_dr Four Volunteers Who Took Pfizer's COVID-19 Vaccine Developed Bell's Palsy - FDA Denies the Temporary Facial Paralysis Caused by the Shot https://t.co/AWaJYQObN via @gatewaypundit
Thu Dec 10 00:16:51 +0000 2020 (Remark: HCQ promoted as cure is not reliable, four volunteers case is true but causal link suggested in the tweet is false.)

- 1336926192511881216 travelgirl391 conspiracy thegatewaypundit.com
Four Volunteers Who Took Pfizer's COVID-19 Vaccine Developed Bell's Palsy - FDA Denies the Temporary Facial Paralysis Caused by the Shot!! WHY WOULD ANYONE TAKE THIS SHIT VOLUNTARILY?? lol lol https://t.co/QH5cmKAOv0
Thu Dec 10 06:50:17 +0000 2020 (Remark: It was at a normal expected rate and no causal link to vaccine was established - Significant element of truth and falsehood since a causal link is explicitly suggested in the tweet here).

- 1341276950640402432 WSoldier17 reliable abcnews.go.com
Here's a nuclear red pill right here. Vatican approves use of fetal tissue use and promotes it as it resides in COVID vaccine. https://t.co/FYpgjnLMqt
Tue Dec 22 06:58:39 +0000 2020 (Remark: Vatican approves is true, resides in vaccine is false - only age old fetal tissues are used in development pipelines)

Figure 4: Annotation guidelines.

A ANNOTATION GUIDELINES

In Fig 4, the instructions and guidelines specified for annotators is included. The annotators are asked to label in the context of when the tweet is posted, with examination of facts from high-factual, low bias news article sources, fact-checking resources, and official information sources. The annotators are provided tweets with the screen name, news source domain, news source label, full tweet text (including the news URL hyperlink), tweet timestamp are provided to aid the annotator. The article URL provides context to the tweet content, and is needed at times to understand the tweet's claim. Typical and trick examples with remarks in each category were provided to review and revisit while annotating, which is a useful guide to provide the distinctions between label types.