CAVES: A Dataset to facilitate Explainable Classification and Summarization of Concerns towards COVID Vaccines

Soham Poddar
Indian Institute of Technology
Kharagpur, India

Azlaan Mustafa Samad
Indian Institute of Technology
Kharagpur, India

Rajdeep Mukherjee
Indian Institute of Technology
Kharagpur, India

Niloy Ganguly
Indian Institute of Technology
Kharagpur, India
Leibniz University of Hannover
Hannover, Germany

Saptarshi Ghosh
Indian Institute of Technology
Kharagpur, India

ABSTRACT

Convincing people to get vaccinated against COVID-19 is a key societal challenge in the present times. As a first step towards this goal, many prior works have relied on social media analysis to understand the specific concerns that people have towards these vaccines, such as potential side-effects, ineffectiveness, political factors, and so on. Though there are datasets that broadly classify social media posts into Anti-vax and Pro-Vax labels, there is no dataset (to our knowledge) that labels social media posts according to the specific anti-vaccine concerns mentioned in the posts. In this paper, we have curated CAVES, the first large-scale dataset containing about 10k COVID-19 anti-vaccine tweets labelled into various specific anti-vaccine concerns in a multi-label setting. This is also the first multi-label classification dataset that provides explanations for each of the labels. Additionally, the dataset also provides class-wise summaries of all the tweets. We also perform preliminary experiments on the dataset and show that this is a very challenging dataset for multi-label explainable classification and tweet summarization, as is evident by the moderate scores achieved by some state-of-the-art models.

CCS CONCEPTS

• **Human-centered computing** → **Collaborative and social computing**; • **Information systems** → **Clustering and classification**; *Summarization.*

KEYWORDS

COVID-19; anti-vaccine concerns; tweets; dataset; multi-label classification; explainable classification; summarization.

ACM Reference Format:
Soham Poddar, Azlaan Mustafa Samad, Rajdeep Mukherjee, Niloy Ganguly, and Saptarshi Ghosh. 2022. CAVES: A Dataset to facilitate Explainable Classification and Summarization of Concerns towards COVID Vaccines. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '22), July 11–15, 2022, Madrid, Spain.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3477495.3531745

SIGIR '22, July 11–15, 2022, Madrid, Spain

ACM ISBN 978-1-4503-8732-3/22/07…$15.00
https://doi.org/10.1145/3477495.3531745

1 INTRODUCTION

The COVID-19 pandemic has been ongoing since 2020 and has affected hundreds of millions of people till date. Medical professionals believe that vaccination at a near-societal scale is the best way to achieve herd immunity and eradicate the virus [1]. However, unfortunately a significant fraction of population are skeptical about taking COVID-19 vaccines due to different reasons [1, 29], and understanding their concerns is very important for convincing such people about the benefits of the vaccines.

Concerns about vaccination is not new. Even in pre-Covid times, there have been many people who are known to be *anti-vax* or *vaccine-hesitant*. Researchers have studied the concerns of such people towards vaccines, both through surveys as well as through the lens of social media like Twitter [1, 27]. Especially, understanding vaccine-related opinions through the lens of social media (e.g., Twitter) is a very popular method due to the large scale at which such studies can easily be conducted (as opposed to human surveys that are very difficult to conduct at large scale). Several datasets have also been developed to aid the study of vaccine-related opinions on social media. For instance, there exist datasets that label tweets (microblogs) into three broad categories of pro-vaccine (i.e., supporting vaccines), anti-vaccine (i.e., opposing vaccines) and neutral [3, 23]. However, when it comes to *understanding the reasons behind vaccine hesitancy* (i.e., a finer analysis of anti-vaccine concerns), researchers have only relied on human surveys / manual analysis at a small scale or unsupervised topic models which do not perform too well and require manual intervention. To the best of our knowledge, there does *not* exist any large-scale dataset of social media posts *labeled with various reasons of vaccine-hesitancy*.

In this paper, we provide the first such large-scale dataset by annotating tweets into these fine-grained concerns about vaccines. We first extensively collect tweets (via the Twitter API) using more than 200 keywords related to vaccines, and then employ a 3-class classifier to identify tweets that are very likely to be anti-vaccine. Subsequently we get such tweets labeled (by three human annotators) according to the specific anti-vaccine concern(s) stated in

[1] https://tinyurl.com/WHO-COVID-immunity

the tweet-text. The final dataset – which we call **C**oncerns **A**bout **V**accines with **E**xplanations and **S**ummaries (CAVES) – contains about 10K anti-vaccine tweets related to COVID-19 vaccination, which are manually labeled with 11 established anti-vaccine concerns, such as concerns about the vaccine ingredients, side-effects of vaccines, and so on (see Table 3 for a list of these concerns). Since a particular tweet often reflects multiple anti-vaccine concerns, the dataset is *multi-labeled*. Additionally, the dataset is curated with human-annotated *explanations* for the labels, indicating which exact parts of the tweet-text indicates a particular anti-vaccine concern.

Once these classes are identified, it also makes sense to summarize the tweets in a class, so that the authorities can quickly glance through the various concerns being raised by people and act accordingly to promote the use of vaccines. To this end, the CAVES dataset contains human-compiled summaries of the tweets pertaining to each anti-vaccine concern (i.e., class-wise summaries).

The CAVES dataset is very timely in the present pandemic situation, and can help social media researchers / authorities gain a fine-grained understanding of anti-vaccine concerns among masses. There are three major use-cases of the CAVES dataset – (i) the dataset can be used to develop models for automated fine-grained retrieval and analysis of concerns people have towards vaccines; (ii) the dataset can act as an excellent resource for performing the dual tasks of multi-label classification and explanation identification (where separate explanations for each of the different labels are provided), and (iii) the summaries provided for each class would act as an excellent resource for performing the task of multi-document or tweet-stream summarization.

We benchmark the CAVES dataset with respect to the three tasks stated above (multi-label classification, explanation generation, summarization). For each task, we apply several state-of-the-art methods. We demonstrate that the CAVES dataset is quite challenging for these tasks. For instance, in the multi-label classification task, the best performance achieved is a moderate Macro-F1 of 0.6, while the performances in the explanation generation task is even lower. Hence, there is a need for better models for these tasks. We have provided the dataset along with these benchmarking codes on Github (see Section 3.7 for details).

2 LITERATURE REVIEW

In this section we shall briefly touch on some previous works that have tried to understand the concerns about vaccines, highlighting the importance of our dataset. Additionally we briefly discuss some of the other datasets available for the tasks of multi-label classification, explanation generation and summarization.

2.1 Concerns about Vaccines

Several prior works have been trying to understand people's stance towards vaccines from social media like Twitter. One of the largest datasets about vaccine stances is that by Müller and Salathé [23] which contains about 28K tweets labelled into three different categories – i) Anti-Vax (tweets that are against vaccines), ii) Pro-Vax (tweets that support vaccines) and iii) Neutral (other tweets that talk about vaccine without a clear stance). Yuan et al. [44] provides another dataset that labels tweets in a similar fashion. After the onset of the COVID-19 pandemic the discourse around vaccines have risen a lot, especially hesitancy towards vaccines [1, 13, 29]. Naturally, researchers have tried to build labelled datasets for vaccine-stance detection using COVID-19 vaccine related tweets [3, 29]. However, these datasets only provide the broad-level classification of tweets (pro-vax, anti-vax and neutral) and do not deal with the particular reasons for vaccine hesitancy.

Several researchers have tried to understand the concerns of people causing the rise of vaccine hesitancy during COVID, usually with the help of surveys [5, 25, 27, 36, 42]. Some works have also examined Twitter posts to understand these concerns, through manual analysis of a small-sample [1], or by applying topic models such as LDA [30] or through a combination of both [29]. However, there exists no large-scale labelled dataset that Machine Learning models can leverage to perform automated detection of these fine-grained concerns towards vaccines.

2.2 Multi-Label Classification, Explanation Generation and Summarization

Multi-Label text classification is a classical machine learning problem with several datasets available for it in different domains. Reuters dataset [14] is one of the oldest multi-label datasets containing news articles which are labelled into different categories. The MIMIC-III dataset [12] is a large dataset of medical records labelled into several ICD-9 codes. There also exist a few datasets performing multi-label classification on tweets, e.g. SemEval 2018 Task 1 dataset [20] containing several sub-classes of emotions, and TREC-IS dataset [18] which contains different label categories of disaster related tweets.

In the past few years, explainable machine learning models have been an important research area and there have been several models trying to provide explanations [32]. There also exist quite a few datasets that provide explanations/rationales for classification. A popular collection of datasets is the ERASER benchmark [4] which comprise of several datasets for providing explanations for different tasks. Another dataset from the domain of Hate-speech detection is HateXplain [17] which provide explanations for different tweets being hateful/offensive. Though there exist some methods that can generate explanations for multi-label classification [16, 21], to the best of our knowledge, *there exists no dataset that provides explanations in a multi-label setting, with separate explanations for different labels*. The CAVES dataset developed in this paper is the first dataset containing such distinct explanations for each label assigned to a text (tweet).

Multi-document summarization, as the name suggests deals with summarization of information contained in multiple documents. One such popular dataset is the Multi-News dataset [9]. Summarization of tweets also falls under this category, and there exist a few different datasets containing summaries of tweets from different categories/domains of tweets. For example, TGSum [2] contains summaries of news-related tweets, while TSix [24] deals with summaries tweets about some real-world events. He et al. [11] provides another dataset of summaries of events from Twitter, while Rudra et al. [33] and Dutta et al. [6] provide datasets containing summaries of disaster-related tweets. However, to our knowledge, there is no existing dataset focusing on summarization of general public opinions about COVID-19 vaccines.

Table 1: Some of the keywords used for collection of tweets.

Generic keywords: vaccine, vaxxer, vaxxed, vaccinated, vaccination, covid vaccine, corona vaccine
COVID vaccine-specific keywords: astra zeneca, novavax, pfizer, BioNTech, comirnaty, moderna, gamaleya, NIAID, bharat biotech, covaxin, covishield, sanofi, curevac, sinovac, sinopharm, janssen, oxford vaccine, johnson vaccine, russian vaccine, chinese vaccine

3 DATASET PREPARATION

In this section, we describe the CAVES dataset and its preparation in detail.

3.1 Selection of tweets

Fetching tweets: We fetched tweets using the official Twitter API with various keywords related to vaccines in general (e.g., 'vaccine', 'vaxxer') and COVID-19 vaccines in particular (e.g., names of COVID vaccines and their manufacturers such as 'comirnaty', 'covishield', 'moderna', 'gamaleya'). Some sample keywords are stated in Table 1. We also added the set of 126 Anti-Vaccine and 154 Pro-Vaccine hashtags provided by Gunaratne et al. [10] (in the supplemental material of their work). Using all of these keywords, we collected about 100M distinct vaccine-related tweets (excluding retweets) posted in between January 2020 and October 2021.

Extracting Anti-Vax tweets: In this work, we specifically wanted to deal with anti-vaccine / anti-vax tweets that exhibit hesitancy towards vaccines. Such tweets make up only a small fraction of all vaccine-related tweets, and selecting a complete random sample would lead to only a small amount of anti-vax tweets. Hence, we decided to specifically identify anti-vax tweets using a COVID-vaccine stance classifier developed in our prior work [29]. This BERT-based classifier gives the probability of a given tweet being Anti-Vax, Pro-Vax and Neutral. The classifier was tested over multiple datasets containing vaccine-related tweets (details in [29]) and achieved macro-F1 scores in the range of [0.78, 0.825]. This classifier was first run over all the tweets that we had collected to obtain the probability of the tweets being Anti-Vax. To be certain that we are left with a set of mostly anti-vax tweets, after testing several thresholds, we selected the tweets which the classifier predicted Anti-Vax with high confidence, i.e. with probability ≥ 0.8 (similar thresholds have been used in prior works [19] to filter out Anti-vax tweets with high precision). Finally we were left with about 7.5M tweets which are highly likely to be vaccine-hesitant (according to the classifier). Through the annotation study described later in Section 3.3, we observe that 98% of the tweets that are predicted to be anti-vaccine with probability ≥ 0.8 are actually anti-vaccine.

Removing duplicates for annotation: Even after removing the retweets, there were quite a lot of similar tweets remaining, e.g. tweets that contain the same text but a few different mentions or hashtags. For the human annotation (described in Section 3.3), we removed such near-duplicate tweets using some of the methods described in [41]. We measured the similarity between two tweets by their token overlap after removing the mentions and hashtags.

Table 2: Mapping the different anti-vaccine concerns given by previous works to our set of concerns (classes).

Our classes	Praveen [30]	Nuzhath [25]	Bonnevie [1]
Conspiracy	-	Unusual theories	-
Country	Skepticism over the nationality of the vaccine	-	-
Ineffective	-	Vaccine will be Ineffective	-
Ingredients	-	-	Vaccine Ingredients
Mandatory	-	Freedom of choice	-
Pharma	Negative feeling towards pharma companies	Profit from developing a COVID-19 vaccine; Mistrust of Scientists and vaccine advocates	Pharmaceutical Industry; Federal Health Authorities
Political	-	Mistrust in the government	Policies & Politics
Religious	-	Religious beliefs	Religion
Rushed	Skepticism over vaccine trails; Doubts regarding data; Rush in providing the vaccine	Vaccines are untested; Fast paced Vaccine Development	Research & Clinical Trials
Side-effect	Fear over health; Allergic reactions; Fear of death	Vaccine will have side effects; Vaccines cause illnesses of unknown origin	Negative health impacts; Vaccine Safety
Unnecessary	COVID-19 being exaggerated	Vaccine is Unnecessary	Disease Prevalence

We retain only one copy of tweets that had more than 80% token overlap [41]. After this step, we are left with around 6.5M anti-vax tweets.

3.2 Selection of anti-vaccine concern classes

Defining the classes: An important decision was the selection of the classes in our dataset, i.e., the specific anti-vaccine concerns. We initially consulted various prior works that curate lists of anti-vaccine concerns [1, 5, 25, 27, 29, 30] (see Section 2.1). The classes provided by a few such prior works are listed in the second, third and fourth columns of Table 2. We observed that different prior works have considered different sets/classes of anti-vaccine concerns. The classes also vary in granularity – multiple classes considered in one study can be combined into a single class considered in another study. Hence we decided to curate our own set of classes to appropriately capture the different concerns expressed by the tweets towards vaccines.

To this end, two of the authors examined the concern-classes listed in several prior works that attempted to categorize anti-vaccine concerns. They also manually examined a random sample of 500 tweets in 3 iterations to understand the different types of

Table 3: Description of different concerns towards vaccines. along with an example tweet (excerpts). These concerns define the classes in our dataset. The keywords marked in bold in the tweets are the explanations identified by annotators. The percentage of tweets in each class (last column) is calculated with respect to the total number of tweets in the final dataset.

Class Description	Example Tweet Excerpt	#Tweets
Unnecessary - The tweet indicates COVID is not dangerous, vaccines are unnecessary, or that alternate cures (such as hydroxychloroquine) are better.	I wouldn't get a vaccine for **a virus w/a 95% recovery rate**	1,560 (15.7%)
Mandatory - The tweet is against mandatory vaccination and talks about their freedom.	No one **should be forced** to get a vaccine! **#medicalfreedom**	694 (7.0%)
Pharma - The tweet indicates that the Big Pharmaceutical companies are just trying to earn money, or is against such companies in general because of their history.	**Pfizer** is about to unleash hell in the name of **profit** & reckless endangerment.	967 (9.7%)
Conspiracy - The tweet suggests some deeper conspiracy, e.g., vaccines are being used to track people via microchips, the entire COVID is a hoax, vaccines are used for behaviour and mind control, vaccines alter DNA, vaccines are bio-weapons, etc. (list of conspiracy theories compiled from [25, 35])	it won't be a covid vaccine though! Remember **depopulation, smart dust microchips**? It's what Bill Gates had planned all along!!	332 (3.3%)
Political - The tweet expresses concerns that the governments / politicians are pushing their own agenda though the vaccines.	It took **Donald Trump** to turn me into an anti-vaxxer.	483 (4.9%)
Country - The tweet is against some vaccine because of the country where it was developed / manufactured	I do not want a **EU crappy** mRNA vaccine!	158 (1.6%)
Rushed - The tweet expresses concerns that the vaccines have not been tested properly, have been rushed or that the published data is not accurate.	**no phase 3 trials**, Brazil rejected Covaxin	1,192 (12.0%)
Ingredients - The tweet expresses concerns about the ingredients present in the vaccines (eg. fetal cells, chemicals) or the technology used (e.g., mRNA)	**nanoparticles** in Pfizer's vaccine trigger rare allergic reactions	322 (3.2%)
Side-effect - The tweet expresses concerns about the side effects of the vaccines, including deaths caused.	did u hear about the Johnson and Johnson vaccine **blood clots**?	3,559 (35.9%)
Ineffective - The tweet expresses concerns that the vaccines are not effective enough and are useless.	They rushed into AZ, despite its **effectivity being 20% less**	1,489 (15.0%)
Religious - The tweet is against vaccines because of religious reasons	**Bishops discourage Catholics** from receiving Johnson & Johnson vaccine	46 (0.5%)
None - The tweet states no clear reason for vaccine hesitancy, or any of the other reasons.	They should be offered a free bullet with the vaccine	224 (2.3%)

concerns voiced in the tweets. Based on extensive discussions after each iteration, a list of *12 classes of anti-vaccine concerns* was finalized, that encompass the classes defined by most prior works as well. The 12 classes are described in Table 3. Note that the 'None' class is meant for tweets that oppose vaccines but do *not* give any reason / concern for such opposition. Table 2 gives a mapping of how these 12 classes correspond to the classes defined by a few of the prior works.

Multi-label setting with explanations: While manually examining the tweets, we noticed that several tweets actually express multiple concerns about vaccines; for instance, a tweet may say that the vaccines are insufficiently tested ("**Rushed**") and thus can cause unknown complications ("**Side-effects**"). Some example tweets which talk about multiple concerns are given in Table 4. Hence, we decided to label the tweets in a multi-label setting, assigning one or more classes to each tweet. This multi-label nature of the dataset can also be seen from the joint label distribution in Figure 1 which is explained later in Section 3.4.

Another consequence of the multi-label nature of the tweets is that different parts of the tweet-text explain the different labels/classes assigned to a tweet (see the examples given in Table 4). Since we want the dataset to support explainable classification, it is imperative for classifiers to get additional information to correctly identify the parts of the tweet that explain every label that is assigned to the tweet. Hence we also decided to get the explanations marked for each of the labels separately (and not just one explanation for the entire tweet).

3.3 Annotation of the tweets

Annotation setup: We took the help of a data annotation firm named Cogitotech (https://www.cogitotech.com/) to get the tweets annotated. We chose to employ this company instead of a crowdsourcing platform (such as Amazon Mechanical Turk), since we wanted to have discussions with the annotators to properly describe the different classes.

We first provided an instruction manual to the firm containing the description of the different class labels (anti-vaccine concerns) along with a couple of examples (as given in Table 3). Specifically, we provided the annotators the following instructions. **(I1):** For each tweet, select the labels based on the concern(s) that the tweet indicates towards COVID-19 vaccines. You can select more than one label if the tweet states multiple concerns towards vaccines. **(I2):** For each of the selected labels for a particular tweet, mark the specific part of the tweet that made you select that label. That is, mark a few keywords or small phrases from the tweet that explain your selection. **(I3):** If a tweet contains a URL, you can visit the content of the URL for better understanding of the tweet. **(I4):** Although we have tried to automatically select only tweets that are hesitant towards COVID vaccines, there may be some tweets that do not show any hesitancy towards the vaccines. Indicate

Table 4: Examples of Multi-Label tweets. The explanations for the three labels are highlighted in *blue*, *red* and brown. The overlaps between explanations are highlighted in cyan.

Tweet excerpt	Labels
No claims, no trials, no compensation: *Pfizer given protection* from legal action by **UK government**	*Untested*, *Pharma*, **Political**
Good only to implant their chips, of course that is the entire purpose... study says **jab ineffective** against local variant	*Conspiracy*, *Ingredients*, **Ineffective**
Johnson & Johnson, another company at the top of the broken *Big Pharma Ecosystem*, has created a vaccine that was *rapidly approved*.	*Pharma*, *Untested*
It's just another annual new corona virus. It's *on its way out*. Talk of a vaccine being pushed by a certain *billionaire computer nerd*	*Unnecessary*, *Pharma*

such tweets separately by selecting the option "The tweet is NOT hesitant towards any vaccines", and do not assign any label to such tweets. Keywords need not be provided for such tweets. **(I5):** If a tweet opposes vaccines but does *not* give any reason for such opposition, or gives some reason other than the concerns we have listed, then select the 'None' label. Keywords need not be provided for such tweets.

Annotation Process: We asked the firm to get each tweet annotated independently by three annotators. The annotators were a set of university graduates of the age group 20-30, who are well-versed in English and are conversant with Twitter. Each tweet was labeled by 3 annotators from this set.

We first provided a set of 1,000 randomly sampled tweets (from the set of tweets obtained after duplicate removal, as described in Section 3.1) to the firm to be annotated independently by three annotators. After a week, the firm completed the annotations and we cross-checked the annotations for 100 of these tweets for correctness. Other than a few minor corrections to the labels, the annotations looked good. We had a discussion with the annotators and clarified some of their doubts regarding the class descriptions. We also clarified that explanations (for a particular label) within a particular tweet can consist of non-contiguous words as well.

Once we were satisfied with the annotations, we then provided them another set of randomly sampled 10K tweets to be annotated. The firm agreed to annotate the total of 11k tweets for about INR 110,000 (~USD 1500), and completed the annotations in about 2 months. The tweets were marked by different annotators in their firm, with three annotators marking each tweet.

3.4 Finalizing the labels

Out of the 11k tweets we got annotated, 226 tweets were deemed to *not* express any hesitancy towards vaccines (i.e., these tweets were misclassified by the pro-vaccine / neutral / anti-vaccine classifier described in Section 3.1), and were thus removed from the dataset.

Recall that 3 annotators independently assigned one or more labels to each tweet. Taking a union of all the labels for each tweet, we were left with 5,680 tweets with a single label, 4,208 tweets with two distinct labels, 821 tweets with three distinct labels, and 65

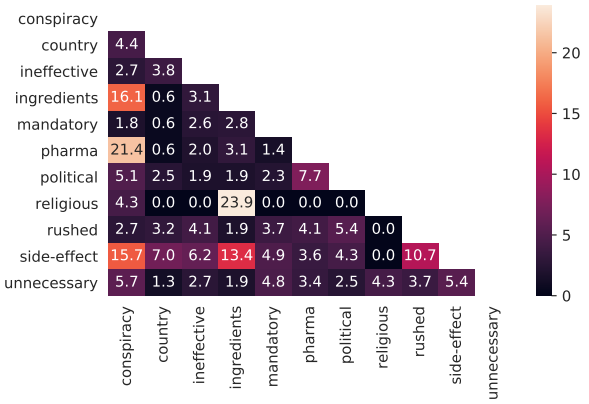


Figure 1: Joint distribution of different classes. For a pair of classes c_i and c_j, the value in a cell is $n_{ij} \times 100/\min\{|c_i|, |c_j|\}$, where n_{ij} is the number of tweets that have been labeled with both c_i and c_j.

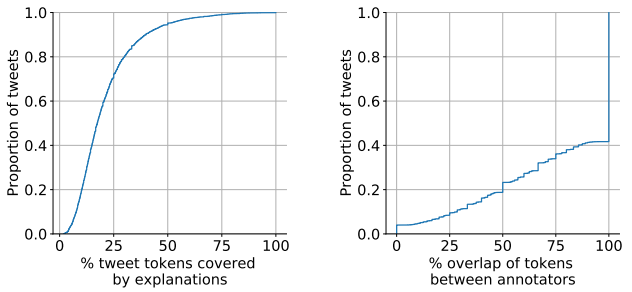
tweets with more than three distinct labels. However, some of these tweets were marked with a particular label by only one annotator out of the three. On manually examining some of these tweets and labels, we found a few cases where the label did *not* accurately represent the concerns in the tweets, and probably represented personal opinions. Thus, for each tweet, *we consider only those labels that are given by at least two annotators*, since such labels are more likely to represent the concerns expressed in the tweet. For example, suppose a tweet has been labelled by the three annotators ($A1$–$A3$) as follows – $A1 : \{C1, C2, C3\}$; $A2 : \{C2, C3, C4\}$; $A3 : \{C3\}$ (where $C1 - C4$ represent some classes), then the final ground truth labels for the tweet will be $L : \{C2, C3\}$, since these labels have been selected by at least two annotators. We then removed 853 tweets that did not have any label assigned by at least two annotators.

Using this strategy we were left with 9,921 tweets, out of which 957 have exactly two distinct labels, 74 have three distinct labels, and the rest have a single label. No tweet has more than 3 labels. The distribution of labels class-wise is given in Table 3 (last column). To check the *Inter Annotator Agreement* (IAA), we calculated the Krippendorff's alpha agreement score (using MASI distance [26]). The agreement score came out as 0.9557 on the final dataset (considering only the labels assigned by at least two annotators) which shows very high agreement.

To get an idea of which classes frequently occur with some other classes, we have also calculated the joint-distribution of classes (shown in Figure 1). Let c_i and c_j be two classes. We check the number n_{ij} of tweets that have been labeled with both c_i and c_j, and divide this number by the maximum number of possible tweets where both classes could have been present (i.e. the number of tweets in the smaller class out of c_i and c_j). This distribution of $n_{ij} \times 100/\min\{|c_i|, |c_j|\}$ is given in Figure 1. It can be seen that some particular classes co-exist frequently with other classes. For example, the 'Conspiracy' class often co-occurs with the 'Ingredients', 'Pharma', and 'Side-effect' classes. Similarly, the 'Ingredients' and 'Rushed' classes also co-exist often with the 'Side-effect' class.

Table 5: Sample explanations given by different annotators. The tweets marked by the three annotators are highlighted in *blue*, *red* and brown. The overlaps are highlighted in cyan.

Class	Tweet
Side-effect	*53 dead in Gibraltar in 10 days* after experimental Pfizer mRNA COVID injections started. Could the new vaccines be *causing all those "COVID deaths"*? Tiny Gibraltar Shines Huge Light on **Vaccine Deaths**.
Ineffective	Get vaxxed with *Pfizer* with 39% efficacy or with J&J with even much lower efficacy so you can spread virus without testing.
Pharma	Somebody is *going to* make money *from all of this*, why not let it be you? Find the **Big Pharma companies that are front runners** for creating the vaccine, which would = huge $$$.
Ineffective	@USER *You can't actually say it* will save *others or even those vaccinated.*



(a) CDF of % of tweet tokens covered by explanations, after taking union of keywords selected by individual annotators

(b) CDF of % overlap between keyword-sets marked by individual annotators (intersection over union).

Figure 2: Statistics of explanations selected by the annotators

3.5 Finalizing the explanations

Recall that for a particular tweet t, we consider only those labels (classes) that have been marked by at least 2 annotators. Each annotator also marks some keywords as an explanation of each label that he/she assigns to the tweet t. Now, even if two annotators label t with the same label l, there can be differences in the explanations they give for the same label. Hence, for each selected label l for t, we need to take a combination of all the keywords marked by the 2 or 3 annotators (as explanations) who labeled t with the label l. A few examples of such explanations have been given in Table 5.

To combine the explanations provided by multiple annotators for the same label for t, we can either take the *union* or the *intersection* of the keywords that have been marked by at least two annotators. After observing the explanations of about 100 tweets, we felt that taking intersection reduces the amount of information available about the explanations, making it insufficient for training classifiers. On the other hand, taking the union of the explanations (i.e., considering all words that have been marked as explanation by any annotator) seemed to give an appropriate amount of information.

Table 6: Some of the most frequent keywords from the explanations for each class (after removing '*covid*', '*corona*', '*virus*', '*vaccine*' and generic stopwords).

Class	Keywords
Conspiracy	population, depopulation, control, world, chip, agenda, plan
Country	russian, chinese, china, russia, want, develop, accept
Ineffective	effective, efficacy, work, pfizer, get, stop, prevent
Ingredients	cell, aborted, chip, ingredient, fetal, tissue, contain
Mandatory	force, passport, mandatory, push, people, mandate, passports
Pharma	pfizer, pharma, money, company, gates, moderna, billion, profit
Political	government, trump, election, political, pfizer, borisjohnson, politician
Religious	catholic, religion, catholics, avoid, leader, bishop, morally
Untested	rush, trial, test, experimental, untested, testing, datum
Side-effect	die, effect, death, reaction, pfizer, clot, cause, adverse
Unnecessary	need, don't, take, want, people, rate, vaccinate

Hence, for label l for tweet t, we consider as explanation the union of the keywords selected by all annotators who labeled t with l.

We now check the quality of the explanations (that are generated by taking union of the keywords selected by multiple annotators) in various ways. First, we wanted to check if too many tweets have a large portion covered up with explanations; if so, this could imply that the explanations do not contain enough distinct meaningful information to guide models about the important parts of the tweet. To this end, we checked the distribution of the % of tokens in the tweet-text that is covered by the explanations, i.e., the number of tokens in the explanations expressed as a percentage of number of tokens in the entire tweet. The distribution is plotted in Figure 2a. We see that, only in very few cases (<1%) does the explanations cover more than 80% of the tweet-text; whereas, for more than 80% of the tweets, the explanations (considering union of keywords selected by multiple annotators) cover less than 30% of the tweet-text. This implies that even after taking unions, the explanations are terse enough to produce distinctive information. The <1% cases where the explanations cover most of the tweet, seem to be cases where the tweets are small and most of it is actually relevant to the corresponding class label. Such an example is given in the last row of Table 5.

Second, to understand the inter-annotator agreement about explanations, we computed the distribution of the % overlap of tokens between the keywords marked by different annotators (for the same tweet and label). We adopt an approach based on (intersection over union). For each tweet we calculated the number of keywords that were marked by at least 2 annotators (intersection), and divided it by the number of keywords that were marked by at least 1 annotator (union). This distribution is given in Figure 2b. As can be seen from this figure, in more than 60% of the tweets, the annotators have complete overlap of the keywords they marked, thus implying good agreement between the annotators as well as the distinct nature of the explanations found in the tweet.

Table 7: Evaluation scores of the class-wise summaries (averaged over all 3 summaries in each class).

Class	Consistency	Fluency	Relevance
Conspiracy	4.445	4.223	4.000
Country	4.000	3.556	3.776
Ineffective	3.999	4.332	3.778
Ingredients	4.556	4.334	4.223
Mandatory	3.777	4.334	4.111
Pharma	3.776	3.889	4.111
Political	3.889	3.667	4.445
Rushed	4.445	4.000	4.333
Side-Effect	3.778	3.889	4.112
Unnecessary	4.112	4.112	4.112
Overall	4.077	4.033	4.111

Finally, Table 6 shows some of the most frequent words in the final explanations associated with the different labels/classes. It is evident that the explanations are of good quality, containing keywords that are very relevant to the labels.

3.6 Generating class-wise summaries

To enable the use of the CAVES dataset for summarization, we wanted to provide summaries of the tweets from each class. Since the 'Religious' class has only 46 tweets (as seen in Table 3), we decided to exclude this class from the summary generation process. The 'None' class is also excluded, since it does not represent any particular theme of discussion. The labelled tweets are then segregated into 10 groups based on the remaining 10 classes – we put a tweet into a class if the class is present in any one of its labels. Thus a particular tweet can be included in multiple groups if it has multiple labels.[2]

Writing summaries: For writing the summaries of the tweets, we employed workers on the crowdsourcing platform Prolific (https://prolific.co/). We selected workers who are fluent in English and are conversant with Twitter. Additionally, to ensure quality annotations we also added filters selecting workers who had completed at least 1000 tasks on the platform with a 100% acceptance rate. Each group (class) of tweets were summarized by 3 different Prolific workers. We asked the annotators to write *abstractive* summaries. Specifically, the annotators were given the following instructions – *Write a coherent summary of 200 - 250 words of all the tweets you read. Write this summary in your own words. Try to cover as much of the content as you can in the summary, and also avoid redundancy as much as possible.*

We asked 3 annotators to write summaries for each class of tweets. Thus we finally compiled a list of 30 summaries (10 classes × 3 summaries).

Evaluation of summaries: We next evaluate the quality of the summaries written by the Prolific workers. To this end, we tried three different quality metrics for summaries [8]:

[2]Due to this reason, the summaries for a particular class can contain some information about other classes, due to the multi-label nature of the tweets.

Table 8: Overview of the CAVES dataset.

# Tweets classified as Anti-Vax with high confidence	7.5M
# Distinct Anti-vax tweets after duplicate removal	6.5M
# Tweets annotated by human workers	11,000
# Labeled tweets in final dataset	9,921
Total # label-explanation tuples	10,802
Total # of summaries	30

• **Consistency** – A consistent summary contains only statements / information that are present in the original set of tweets, and not any additional information.
• **Fluency** - The quality of individual sentences in the summary with regards to how easy they are to read and understand, whether the sentences in the summary are grammatically correct, etc.
• **Relevance** - The summary should include only important information from the original set of tweets. Summaries which do not contain enough important information, or redundancies or excess information are given a lower score.

We again employed some Prolific workers for evaluating the summaries, taking care to choose different workers for the evaluation, and *not* the same workers who wrote the summaries. Three workers were assigned to rate each of the 3 summaries individually for each of the 10 classes. They were also given the original set of tweets shown to the summarizing workers, as reference. For each of the metrics defined above, the workers were asked to rate a summary on a scale of 1–5, with 1 being the worst/lowest score and 5 being the best/highest score.

We found 4 particular summaries to be of low quality (evaluation metrics < 3.5). We removed these summaries and again floated the summary-writing task on Prolific, followed by evaluation of the summaries. All of these were done by different workers than the ones who were previously employed (which is possible to set on Prolific). The scores on the final set of summaries were averaged class-wise and have been stated in Table 7. We see that the final class-wise summaries are of sufficiently good quality, achieving scores higher than 3.5 (out of 5) for all three metrics.

3.7 CAVES dataset: overview and availability

The CAVES dataset contains 9,921 anti-vax tweets labeled with their anti-vaccine concerns (class), explanations for the labels, and class-wise summaries. Table 8 gives a summary of the dataset. For the classification and explanation generation tasks, we divide the set of tweets into train, validation, and test sets (split in the ratio 70-10-20 as defined later in Section 4.1).

Note that, according to the Twitter sharing policy[3], it is only permitted to share the tweet IDs, whereas the tweet text (or hydrated tweet objects) should not be shared publicly. In compliance with the Twitter policy, we have publicly released the tweet-IDs, the labels and the abstractive summaries on Github at https://github.com/sohampoddar26/caves-data. We will be providing the full-text and the explanations on request to any researcher who agrees to use the dataset for research purposes. Detailed instructions are given in the Github repository.

[3]https://developer.twitter.com/en/developer-terms/more-on-restricted-use-cases

Table 9: Comparison of models for multi-label classification.

Model	Macro-F1	Weighted-F1	Jaccard	Accuracy
Only predicts labels				
LLDA	0.1188	0.2134	0.1543	0.0675
RoBERTa	0.5837	0.6993	0.6713	0.6010
CT-BERT	0.5860	0.7144	0.6884	0.6087
HateXplain	**0.6007**	**0.7153**	**0.6928**	**0.6294**
Predicts labels and explanations				
CAML	0.3449	0.5039	0.4365	0.3675
ExPred	0.5870	0.6924	0.6652	0.5715
Multi-Task	0.5775	0.7029	0.6761	0.5937
Multi-Task (w GRU)	**0.5991**	**0.7104**	**0.6855**	**0.6037**

We have also made publicly available the implementations of the benchmark methods tried on the dataset (as described in the next section) as well as some helper scripts (such as scripts for loading the dataset into appropriate Numpy arrays).

4 TASKS ON THE DATASET

In this section we shall discuss three tasks that can be directly applied to the dataset – (1) multi-label classification, (2) explainable classification, and (3) tweet summarization. We also apply several state-of-the-art methods for benchmarking each of the three tasks.

4.1 Multi-label classification

In the standard multi-label classification task, each data point (tweet in our case) has to be assigned to one or more classes (anti-vaccine concerns in our case). This is in contrast with the multi-class (or single-label) classification in which a data point has to be assigned with exactly one (out of many) classes.

Data splitting: For creating a benchmark, we split the tweet-set into a train, a validation and a test set in the ratio of 70-10-20. Since the distribution of classes is skewed (as seen in Table 3), a random split would cause a class imbalance. Thus we decided to split the data using the iterative stratification method [34, 39] that aims to balance the classes in a multi-label setting. We used the sk-multilearn library [40] to perform this stratification.

Metrics for Evaluation: Given the set of predicted and gold standard labels, we have used 4 different metrics to estimate the performance of the models. First we calculate the F1-score for all the 12 classes separately and find the (i) Macro-average, and (ii) Weighted-average (where the weights of the classes are proportional to the class frequencies). We also calculate the Jaccard similarity between the predicted label-set and the gold standard label-set over each tweet, and average the Jaccard similarity values over all tweets. Finally we also calculate the subset accuracy – for a particular tweet, a predicted set of labels is considered a match only if it *exactly* matches with the set of gold standard labels. All metrics were calculated using standard functions from the Scikit-Learn package [28].

Benchmarking methods: We experimented with representative methods from different families of multi-label classifiers. We try out a topic modelling method (Labeled LDA) since these are what have been used by prior works for understanding anti-vaccine opinions (as described in Section 2.1). We tried some basic transformer based classifiers, namely RoBERTa-Large (known to perform very well for several NLP tasks) and CT-BERT-v2 [22] which is a BERT-Large model pretrained on millions of COVID-19 tweets (known to give good classification performance on COVID-related tweets [29]). Next we modified the HateXplain [17] model to work for multi-label scenario by joining the explanations for different labels.

We also tried a few models which incorporate the explanations, either to predict only the labels or to predict the labels and generate explanations jointly. In this family of models, we tried the CAML [21] model which uses CNN and Attention mechanism and can be used to get explanations for all the labels. We also use a simplified version of the recently-developed ExPred [46] model, and modify it for a multi-label setting. This model generates explanations along with a label prediction auxiliary task and then predicts the labels again. The "*Multi-task*" model contains a shared CT-BERT encoder and two decoders which separately predict the labels and generate the explanations. We also try a variation of the Multi-Task model where we pass the token embeddings through a GRU first before generating the explanations.

For each of the above-stated methods, we used a batch size of 16 and a maximum sequence length of 128. For HateXplain in particular we obtained best results with a batch size of 2. The models were trained and validated on corresponding datasets for 20 epochs. The model that yielded the best Macro-F1 score on the validation dataset was saved, and this model was used to calculate the performance metrics on the test dataset.

Performances: The results of different methods are given in Table 9. It is seen that CT-BERT outperforms RoBERTa slightly due to the domain pretraining. In contrast, the LLDA model vastly underperforms, highlighting the limitations of the previous works which tried to use topic modelling for extracting concerns about vaccines. CAML does not perform too well because it uses word2vec embeddings which are known to not be as good as BERT embeddings. The HateXplain model seems to perform the best achieving a Macro-F1 score of 0.6007, while the Multi-Task model with GRU performs second best with a Macro-F1 score of 0.5991, while also generating explanations. In terms of class-wise F1 score, all the models seem to perform badly on the 'None' and 'Conspiracy' classes which contain some confusing arguments. The 'Country' and 'Religious' classes are also very sparse, which lead to models under-performing on these classes. Details are omitted for brevity.

4.2 Explainable classification

The next task is that of generating explanations along with the predicted labels. The explanation generation part is a sequence-labelling task, where for a given tweet text as input, each token in the text is to be marked if it is part of the explanation or not. This is to be repeated for each of the predicted labels for the tweet. Thus, for a given tweet, the output of the task will be (i) the predicted labels, and (ii) a list of binary vectors, one for each predicted label (where each vector is of the same length as the number of tokens in the tweet). In these binary vectors, an element being 1 implies that the corresponding token in the tweet-text is part of the explanation for the corresponding label.

Table 10: Comparison of models for explanation generation.

	Explanation metrics			*Tuple metrics*	
Model	**IOU-F1**	**Jaccard**	**Token F1**	**M-F1**	**Acc**
CAML	0.0774	0.1899	0.2886	0.0351	0.0498
ExPred	0.0560	0.1582	0.2487	0.0249	0.0156
Multi-Task	0.3476	0.3336	0.4154	0.2704	0.2854
Multi-Task (with GRU)	0.3015	0.2920	0.3451	0.2355	0.2724

Table 11: Comparison of models for Summarization task.

Model	**ROUGE-1 F1**	**ROUGE-2 F1**	**ROUGE-L F1**
Extractive models			
LexRank [7]	21.48	2.29	15.73
PacSum [47]	19.14	2.09	12.69
COWTS [33]	31.54	4.52	21.07
Abstractive models			
BART [15]	28.60	5.04	19.19
Pegasus [45]	25.20	5.16	20.00
T5 [31]	27.73	5.34	19.38

Metrics for evaluation: For evaluating the explanations generated by different methods, we use some of the hard selection metrics defined by DeYoung et al. [4]. Among these, the first metric IOU-F1 (Intersection-Over-Union F1) is useful for detecting partial matches. The IOU is the size of the overlap between two sets of tokens divided by the size of the union of the two sets of tokens (also known as the Jaccard coefficient). For a particular tweet and label, a predicted explanation is counted as a match if and only if the IOU with the ground truth explanation is ≥ 0.5 [4]. These matches are then used to calculate the multi-label F1 score as described previously in Section 4.1. Additionally, the token-level F1 score and Jaccard coefficient is also calculated and averaged over all data points. We collectively refer to these metrics as *Explanation metrics*.

Separately, we also *evaluate the tuples of predicted labels and explanations together*. We consider a (label, explanation) tuple to be a match if and only if the predicted label is present in the gold standard set of labels and the predicted explanation overlaps at least 50% with the corresponding gold standard explanation (IOU ≥ 0.5). The Macro-F1 and the Accuracy is then calculated over these matches as before. We refer to these metrics as *Tuple metrics*.

Benchmarking methods: For this task, we use the same models which generate explanations (CAML, ExPred, Multi-Task and Multi-Task w GRU), as described in the previous subsection on multi-label classification. We also use the same data splits along with the experimental settings. Hence we are omitting these details here for brevity. It must however be noted that for the 'CAML' model, the explanations were extracted from the Attention layer weights, by taking the top-8 tokens with the highest weights. For the rest of the models, the explanations were predicted as a sequence-labelling task, with those tokens being generated that have the sigmoid of the logits ≥ 0.5 (similar to a multi-label prediction).

Performances: The metric scores on the test set for the explanation generation task are given in Table 10. The two variations of the Multi-task models seem to perform the best on these tasks, with the standard model achieving an IOU-F1 score of 0.3476 and the one with a GRU achieving 0.3015. Especially for the tuple metrics, these models comprehensively outperform the CAML and ExPred models. Similar to the multi-class classification scores, the IOU-F1 scores are low for the 'Country' and 'Religious' (sparse) classes.

4.3 Summarization

Multi-document summarization (a special case of which is tweet summarization) is a classical IR task that aims to produce a short summary of a large set of documents, that contains as much information content as possible while reducing redundant information. Summarization algorithms come in two flavours – i) Extractive, which select a few tweets out of all the tweets, and ii) Abstractive, which generates words to create a coherent summary like humans do. We have tried a few methods of each type to generate class-wise summaries and get some benchmark scores.

Metrics for Evaluation: For evaluation of the summaries, we use the popular ROUGE metrics. Specifically, we report the ROUGE-1 F1-score (that considers unigram matches between the gold standard summaries and an algorithm-generated summary), ROUGE-2 F1-score (that considers bigram matches) and the ROUGE-L F1-score (that considers longest sequence matches).

Benchmarking methods: We have employed a few popular summarization algorithms to benchmark our summarization dataset. Among extractive methods, we have used the graph-based LexRank summarizer [7], PacSum [47] that defines its own centrality measures, and COWTS [33] which is an Integer Linear Programming based summarizer designed for disaster-related tweets. Among abstractive methods, we have employed different *pretrained* transformer based encoder-decoder models such as T5 [31], BART [15], and Pegasus [45], that differ in their pre-training strategies.

For each class, we used the extractive models to generate summaries of around 20 tweets each, from the document consisting of all labeled tweets belonging to that class. Similarly, the abstractive models were used to generate class-wise summaries of around 250 words each. It is to be noted here that the pre-trained abstractive models have a limitation on the size of input documents that they can summarize. Hence, for each class, we first split the corresponding tweet dataset into almost equal-sized chunks such that the length of each chunk is less than or equal to the maximum permissible length. We obtain smaller summaries from each chunk and concatenate these to form a longer summary. The final summary of length around 250 words was obtained by greedily selecting the top-ranked sentences (from the longer summary) based on their TF-IDF scores w.r.t. all tweets in the corresponding class.

Performances: The ROUGE scores on the summarization task are given in Table 11. Among the extractive models, COWTS performs the best achieving a ROUGE-2 F1 score of 4.52 and a ROUGE-L F1 score of 21.07. The performances of the different abstractive models are similar. All models achieve especially low ROUGE-2 F1 scores, and it is a potential research direction to improve summarization performance over this dataset.

4.4 Summary of the section

We tried different state-of-the-art models on three tasks to establish benchmark results on the CAVES dataset. For the classification task the highest Macro-F1 score achieved was 0.6007, which is a moderate score considering it is a multi-label setting. Explanations generated were of decent quality with the best IOU-F1 of 0.3476, though the performance for some other models seem to be quite low. For the summarization task, the ROUGE-2 scores achieved were low. This may be due to various reasons such as the specific vocabulary used in the anti-vax text may be unique, repetition of similar concepts in multiple tweets, etc. These results suggest that the CAVES dataset and associated tasks pose interesting challenges, and more specific models need to be developed to tackle the tasks.

5 OTHER POTENTIAL USES OF THE DATASET

In this section, we highlight some potential applications of our proposed dataset, other than the ones we have discussed so far.

Distribution of concerns over time: Our dataset can be used to study the change in the distribution of different anti-vaccine concerns over time (and for training models to track such changes when applied on large scale Twitter data). In Figure 3, we show the month-wise frequency distribution of the largest 6 classes in our labeled set of tweets (in terms of the number of tweets in a class). We observe some interesting peaks in the figure which can be mapped to certain real-world events (e.g., as reported in the AJMC articles on COVID vaccine developments throughout 2020 [37] and 2021 [38])–

- There is a spike in the *Pharma* class around September 2020 which could be explained by two events – Pfizer expanding phase 3 trials of its vaccine, and Astrazeneca trial halting due to complications faced by a patient.
- There is a spike in the *Unnecessary* class in October 2020, which may be due to the FDA's approval of Remdesivir as a COVID-19 drug (which made many people feel that vaccines are unnecessary).
- The *Rushed* class has a spike in November 2020 likely due to Pfizer and AstraZeneca reporting completion of their trials.
- The *Side-effects* class has the peak in April-May of 2021, likely due to some adverse reactions to the Johnson & Johnson vaccine being reported at the end of April.
- The spike in the *Ineffective* class in July 2021 is likely owing to the reports of the Pfizer Vaccine not being effective against the delta variant of COVID-19.

It is to be noted here that the labeled set of tweets in the CAVES dataset might not be fully representative of the actual temporal distribution of tweets, due to inadvertent biases that might have crept in as part of the various steps we took for selecting the tweets. Hence, one should be careful in drawing strong trends or temporal conclusions just by analyzing the labeled dataset.

Generating highlights for explainable summarization: The benchmark methods we have used in Section 4.3 are standard summarization models that work only with the tweet texts. Our dataset can facilitate designing of models that utilize the explanations to improve the summarization task. Similar to Wang et al. [43], a select-then-generate framework could be designed that first highlights the reasoning spans as explanations and then generates a summary while focusing on the highlighted spans. Such as approach can not only improve the interpretability of extractive summarization models, but can also provide suitable explanations behind generation of particular phrases in case of abstractive models.

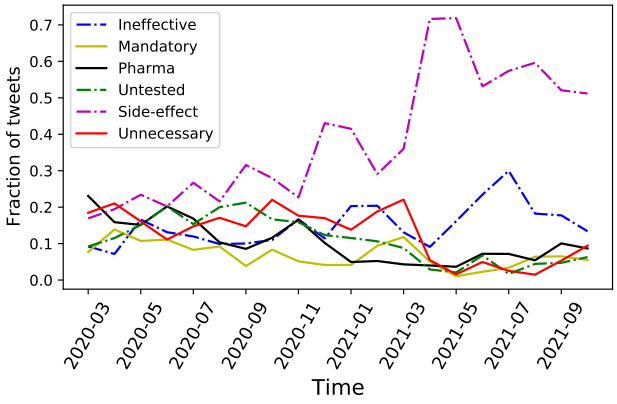


Figure 3: Frequency distribution of tweets corresponding to different concerns over time.

Conspiracy detection: Our dataset contains a set of tweets related to conspiracies around COVID-19 vaccines, as reported in Table 3. The dataset can thus be used to benchmark automatic COVID-19 conspiracy theory detection models such as [35]. Our dataset can further facilitate the design and evaluation of multi-task models that not only detect conspiracy-related tweets but also generate explanations for the same.

6 CONCLUSION

We have built a dataset of tweets that is important from a societal standpoint as it identifies concerns that people have towards vaccines, as well as facilitates explainable classification in a multi-label setting. The dataset also contains summaries of different classes and hence can be used to develop or test summarization algorithms. We have provided some benchmark results on the three different primary tasks, and discussed some other potential retrieval tasks.

The benchmark results point towards the need for improved, customized models for addressing the tasks. For example, apart from the tweet texts, the models can potentially incorporate additional (meta) information from the tweets or the users who posted the tweets to improve scores. Given the timely importance of the CAVES dataset, we believe it will instill enough interest within the community in the near future, to develop better methods for the proposed tasks.

ACKNOWLEDGMENTS

The project is partially supported by research grants from Accenture Corporation and DRDO, Government of India (through the research project titled "Claim Detection and Verification using Deep NLP: an Indian Perspective"). S. Poddar is also supported by the Prime Minister's Research Fellowship (PMRF) from the Ministry of Education, Government of India.

REFERENCES

[1] Erika Bonnevie, Allison Gallegos-Jeffrey, Jaclyn Goldbarg, Brian Byrd, and Joseph Smyser. 2021. Quantifying the rise of vaccine opposition on Twitter during the COVID-19 pandemic. *Journal of communication in healthcare* 14, 1 (2021), 12–19.

[2] Ziqiang Cao, Chengyao Chen, Wenjie Li, Sujian Li, Furu Wei, and Ming Zhou. 2016. Tgsum: Build tweet guided multi-document summarization dataset. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 30.

[3] Liviu-Adrian Cotfas, Camelia Delcea, Ioan Roxin, Corina Ioanăş, Dana Simona Gherai, and Federico Tajariol. 2021. The longest month: Analyzing covid-19 vaccination opinions dynamics from tweets in the month following the first vaccine announcement. *IEEE Access* 9 (2021), 33203–33223.

[4] Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C Wallace. 2020. ERASER: A Benchmark to Evaluate Rationalized NLP Models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 4443–4458.

[5] Kuldeep Dhama, Khan Sharun, Ruchi Tiwari, Manish Dhawan, Talha Bin Emran, Ali A Rabaan, and Saad Alhumaid. 2021. COVID-19 vaccine hesitancy–reasons and solutions to achieve a successful global vaccination campaign to tackle the ongoing pandemic. *Human Vaccines & Immunotherapeutics* 17, 10 (2021), 3495–3499.

[6] Soumi Dutta, Vibhash Chandra, Kanav Mehra, Asit Kumar Das, Tanmoy Chakraborty, and Saptarshi Ghosh. 2018. Ensemble Algorithms for Microblog Summarization. *IEEE Intelligent Systems* 33, 3 (2018), 4–14.

[7] Günes Erkan and Dragomir R Radev. 2004. Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of artificial intelligence research* 22 (2004), 457–479.

[8] Alexander R Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. Summeval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics* 9 (2021), 391–409.

[9] Alexander R Fabbri, Irene Li, Tianwei She, Suyi Li, and Dragomir R Radev. 2019. Multi-news: A large-scale multi-document summarization dataset and abstractive hierarchical model. *arXiv preprint arXiv:1906.01749* (2019).

[10] Keith Gunaratne, Eric A Coomes, and Hourmazd Haghbayan. 2019. Temporal trends in anti-vaccine discourse on Twitter. *Vaccine* 37, 35 (2019), 4867–4871.

[11] Ruifang He, Liangliang Zhao, and Huanyu Liu. 2020. TWEETSUM: Event oriented social summarization dataset. In *Proceedings of the 28th International Conference on Computational Linguistics*. 5731–5736.

[12] Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. MIMIC-III, a freely accessible critical care database. *Scientific data* 3, 1 (2016), 1–9.

[13] Neil F Johnson, Nicolas Velásquez, Nicholas Johnson Restrepo, Rhys Leahy, Nicholas Gabriel, Sara El Oud, Minzhang Zheng, Pedro Manrique, Stefan Wuchty, and Yonatan Lupu. 2020. The online competition between pro-and anti-vaccination views. *Nature* 582, 7811 (2020), 230–233.

[14] David D Lewis, Yiming Yang, Tony Russell-Rose, and Fan Li. 2004. Rcv1: A new benchmark collection for text categorization research. *Journal of machine learning research* 5, Apr (2004), 361–397.

[15] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 7871–7880.

[16] Irene Li, Tianxiao Li, Yixin Li, Ruihai Dong, and Toyotaro Suzumura. 2021. Heterogeneous Graph Neural Networks for Multi-label Text Classification. *arXiv preprint arXiv:2103.14620* (2021).

[17] Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. HateXplain: A Benchmark Dataset for Explainable Hate Speech Detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 14867–14875.

[18] Richard McCreadie, Cody Buntain, and Ian Soboroff. 2019. Trec incident streams: Finding actionable information on social media. (2019).

[19] Tanushree Mitra, Scott Counts, and James W Pennebaker. 2016. Understanding anti-vaccination attitudes in social media. In *Tenth International AAAI Conference on Web and Social Media*.

[20] Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. Semeval-2018 task 1: Affect in tweets. In *Proceedings of the 12th international workshop on semantic evaluation*. 1–17.

[21] James Mullenbach, Sarah Wiegreffe, Jon Duke, Jimeng Sun, and Jacob Eisenstein. 2018. Explainable Prediction of Medical Codes from Clinical Text. In *Proceedings of NAACL-HLT*. 1101–1111.

[22] Martin Müller, Marcel Salathé, and Per E Kummervold. 2020. Covid-twitter-bert: A natural language processing model to analyse covid-19 content on twitter. *arXiv preprint arXiv:2005.07503* (2020).

[23] Martin M Müller and Marcel Salathé. 2019. Crowdbreaks: tracking health trends using public social media data and crowdsourcing. *Frontiers in public health* 7 (2019), 81.

[24] Minh-Tien Nguyen, Dac Viet Lai, Huy Tien Nguyen, and Minh Le Nguyen. 2018. Tsix: a human-involved-creation dataset for tweet summarization. In *Proc. International Conference on Language Resources and Evaluation (LREC)*.

[25] Tasmiah Nuzhath, Samia Tasnim, Rahul Kumar Sanjwal, Nusrat Fahmida Trisha, Mariya Rahman, SM Farabi Mahmud, Arif Arman, Susmita Chakraborty, and Md Mahbub Hossain. 2020. COVID-19 vaccination hesitancy, misinformation and conspiracy theories on social media: A content analysis of Twitter data. (2020).

[26] Rebecca Passonneau. 2006. Measuring agreement on set-valued items (MASI) for semantic and pragmatic annotation. (2006).

[27] Elise Paul, Andrew Steptoe, and Daisy Fancourt. 2021. Attitudes towards vaccines and intention to vaccinate against COVID-19: Implications for public health communications. *The Lancet Regional Health-Europe* 1 (2021), 100012.

[28] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[29] Soham Poddar, Mainack Mondal, Janardan Misra, Niloy Ganguly, and Saptarshi Ghosh. 2022. Winds of Change: Impact of COVID-19 on Vaccine-related Opinions of Twitter users. In *Proceedings of the Sixteenth International AAAI Conference on Web and Social Media (ICWSM'22)*.

[30] SV Praveen, Rajesh Ittamalla, and Gerard Deepak. 2021. Analyzing the attitude of Indian citizens towards COVID-19 vaccine–A text analytics study. *Diabetes & Metabolic Syndrome: Clinical Research & Reviews* 15, 2 (2021), 595–599.

[31] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683* (2019).

[32] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why should i trust you?" Explaining the predictions of any classifier. In *Proc. ACM SIGKDD Conference on Knowledge Discovery and Data Mining*.

[33] Koustav Rudra, Subham Ghosh, Niloy Ganguly, Pawan Goyal, and Saptarshi Ghosh. 2015. Extracting situational information from microblogs during disaster events: a classification-summarization approach. In *Proceedings of the 24th ACM international on conference on information and knowledge management*. 583–592.

[34] Konstantinos Sechidis, Grigorios Tsoumakas, and Ioannis Vlahavas. 2011. On the stratification of multi-label data. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 145–158.

[35] Shadi Shahsavari, Pavan Holur, Tianyi Wang, Timothy R Tangherlini, and Vwani Roychowdhury. 2020. Conspiracy in the time of corona: Automatic detection of emerging COVID-19 conspiracy theories in social media and the news. *Journal of computational social science* 3, 2 (2020), 279–317.

[36] Kalyani Sonawane, Catherine L Troisi, and Ashish A Deshmukh. 2021. COVID-19 vaccination in the UK: Addressing vaccine hesitancy. *The Lancet Regional Health–Europe* 1 (2021).

[37] AJMC Staff. 2021. A Timeline of COVID-19 Developments in 2020. *AJMC* (2021). https://www.ajmc.com/view/a-timeline-of-covid19-developments-in-2020

[38] AJMC Staff. 2021. A Timeline of COVID-19 Vaccine Developments in 2021. *AJMC* (2021). https://www.ajmc.com/view/a-timeline-of-covid-19-vaccine-developments-in-2021

[39] Piotr Szymański and Tomasz Kajdanowicz. 2017. A network perspective on stratification of multi-label data. In *First International Workshop on Learning with Imbalanced Domains: Theory and Applications*. PMLR, 22–35.

[40] Piotr Szymański and Tomasz Kajdanowicz. 2017. A scikit-based Python environment for performing multi-label classification. *arXiv preprint arXiv:1702.01460* (2017).

[41] Ke Tao, Fabian Abel, Claudia Hauff, Geert-Jan Houben, and Ujwal Gadiraju. 2013. Groundhog day: near-duplicate detection on twitter. In *Proceedings of the 22nd international conference on World Wide Web*. 1273–1284.

[42] Gianmarco Troiano and Alessandra Nardi. 2021. Vaccine hesitancy in the era of COVID-19. *Public health* 194 (2021), 245–251.

[43] Haonan Wang, Yang Gao, Yu Bai, Mirella Lapata, and Heyan Huang. 2021. Exploring Explainable Selection to Control Abstractive Summarization. In *Proc. AAAI Conference on Artificial Intelligence*. 13933–13941.

[44] Xiaoyi Yuan, Ross J Schuchard, and Andrew T Crooks. 2019. Examining emergent communities and social bots within the polarized online vaccination debate in Twitter. *Social media+ society* 5, 3 (2019), 2056305119865465.

[45] Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning*. PMLR, 11328–11339.

[46] Zijian Zhang, Koustav Rudra, and Avishek Anand. 2021. Explain and predict, and then predict again. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*. 418–426.

[47] Hao Zheng and Mirella Lapata. 2019. Sentence Centrality Revisited for Unsupervised Summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 6236–6247.