Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 6967–6975
Marseille, 20-25 June 2022

VaccineLies: A Natural Language Resource for Learning to Recognize Misinformation about the COVID-19 and HPV Vaccines

Maxwell A. Weinzierl, Sanda M. Harabagiu
Human Language Technology Research Institute, Department of Computer Science, The University of Texas at Dallas
800 W. Campbell Rd., Richardson, TX 75080
{maxwell.weinzierl, sanda}@utdallas.edu

Abstract

Billions of COVID-19 vaccines have been administered, but many remain hesitant. Misinformation about the COVID-19 vaccines and other vaccines, propagating on social media, is believed to drive hesitancy towards vaccination. The ability to automatically recognize misinformation targeting vaccines on Twitter depends on the availability of data resources. In this paper we present VACCINELIES, a large collection of tweets propagating misinformation about two vaccines: the COVID-19 vaccines and the Human Papillomavirus (HPV) vaccines. Misinformation targets are organized in vaccine-specific taxonomies, which reveal the misinformation themes and concerns. The ontological commitments of the misinformation taxonomies provide an understanding of which misinformation themes and concerns dominate the discourse about the two vaccines covered in VACCINELIES. The organization into training, testing and development sets of VACCINELIES invites the development of novel supervised methods for detecting misinformation on Twitter and identifying the stance towards it. Furthermore, VACCINELIES can be a stepping stone for the development of datasets focusing on misinformation targeting additional vaccines.

Keywords: COVID-19, HPV, vaccine, misinformation, Twitter, social media, stance.

1. Introduction

Misinformation spreading, especially on social media, is believed to be responsible for vaccine hesitancy (Kouzy et al., 2020). It is imperative for public health practitioners to know *what* misinformation is spreading as well as who is *adopting* or *rejecting* it, such that interventions can be tailored appropriately. Public health messaging approaches could not only inoculate against misinformation, but also effectively reach social media users with the aim of shifting or bolstering vaccine attitudes. However, developing natural language processing methods targeting the identification of misinformation about vaccines in social media postings suffers from the lack of language resources where vaccine misinformation annotations are available.

When misinformation was cast as rumor detection, several well-known benchmark datasets for misinformation detection on Twitter became available. For example, the Twitter15 (Ma et al., 2016) and Twitter16 (Ma et al., 2017) datasets consist of a collection of tweets annotated as true rumors, false rumors, unverified rumors or non-rumors. Unfortunately, they did not cover any vaccine misinformation. Similarly, the PHEME dataset (Zubiaga et al., 2016) consists of Twitter conversation threads making a true and false claim, and a series of replies, but none of the conversations focused on vaccination. COVIDLIES (Hossain et al., 2020) is a dataset generated from 86 known misconceptions about COVID-19, for which tweets that evoke the misconceptions were retrieved and annotated with their *stance* towards the misconceptions. Inspired by COVIDLIES, we have created VACCINELIES[1], a dataset which addresses misinformation about two different vaccines: the COVID-19 vaccines and the vaccines protecting against the Human Papillomavirus (HPV).

Misinformation Target: *COVID-19 vaccine alters DNA.*
STANCE: **Accept** *Tweet:* @USER Good girl. The COVID-19 vaccination is an mRNA vaccine, this means it alters your body's DNA. This is incredibly dangerous, as no one knows the long term effects of this.
STANCE: **Reject** *Tweet:* @USER This is absolutely false. The mRNA from a COVID-19 vaccine never enters the nucleus of the cell, which is where our DNA is kept. The mRNA does not affect or interact with our DNA in any way. If you have other concerns, then fine, but don't fall for unfounded nonsense.
Misinformation Target: *HPV vaccine was banned.*
STANCE: **Accept** *Tweet:* @USER Don't worry most people I know don't take the vaccination. We've seen the countries that have banned the hpv vaccine because Gates drug was maiming and killing young girls!
STANCE: **Reject** *Tweet:* @USER Excuse me, actual person from Europe here you interfering trollop. WHAT vaccines are banned in the EU? A UK company released its own HPV vaccine that got outcompeted by Gardasil as it targeted a larger array of strains. Don't involve us to push your primitive agenda

Table 1: Misinformation Targets for the COVID-19 and HPV vaccines with tweets evoking them.

We present VACCINELIES, which consists of:

[1] github.com/Supermaxman/vaccine-lies

1. *Misinformation Targets* (MisTs) similar to those illustrated in Table 1, addressing misinformation towards COVID-19 or HPV vaccines;

2. The *tweet IDs* for those tweets that were judged as evoking any of the MisTs available in VACCINELIES;

3. Annotation of the *stance* of each tweet author that evoked a MisT, indicating if they *Accept* the Mist; *Reject* it, or they have no stance towards it.

4. A *taxonomy* of the MisTs, which enables the interpretation of the themes and concerns characterizing the vaccine misinformation available in VACCINELIES. The taxonomical organization into themes and concerns of the misinformation targets for each vaccine will illuminate the discovery of which targets of misinformation dominate when the vaccines are discussed in social media and, in addition, will lead to the discovery of which kinds of vaccine misinformation are most adopted or most rejected in VACCINELIES. Separate misinformation taxonomies were discerned for the COVID-19 vaccine and the HPV vaccine.

VACCINELIES was inspired by COVIDLIES (Hossain et al., 2020), a dataset of Twitter annotations focusing on misinformation about COVID-19. Like in COVIDLIES, we use the notion of Misinformation Target (MisT) to refer to misconceptions that are employed for propagating misinformation. In addition to misconceptions, we considered misinformation any reference to conspiracy theories or any flawed reasoning. Moreover, we extended the methodology for identifying MisTs, relying not only on misinformation that is readily available on Wikipedia web pages, but also on misinformation that is widely discussed in Twitter conversations. In addition to providing a set of MisTs focusing on two different vaccines, VACCINELIES provides a large set of IDs for tweets that evoke at least one of the MisTs, which were judged by language experts as being relevant to the misinformation expressed in MisTs. Furthermore, the stance of the author of each tweet that evokes a MisT was judged, indicating whether the author *Accepts* the MisT, because they agree with it; *Rejects* the MisT, as they disagree with it; or the author has *No Stance* towards the MisT, although it is evoked.

The annotations that enabled the creation of VACCINELIES were performed jointly by language experts from The University of Texas at Dallas and public health experts from The University of California, Irvine. In previous work (Weinzierl and Harabagiu, 2021; Weinzierl and Harabagiu, 2022), we used an earlier version of VACCINELIES to perform automatic detection of COVID-19 misinformation evocation and stance identification on Twitter. We also used VACCINELIES to identify vaccine hesitancy profiles of users on Twitter (Weinzierl et al., 2021). To our knowledge, VACCINELIES is the only publicly available resource tackling misinformation about the COVID-19 and HPV vaccines on Twitter. We believe our annotation efforts in constructing VACCINELIES fills a gap in vaccine misinformation research, which could greatly benefit both public health experts and natural language processing researchers.

VACCINELIES can be also seen as consisting of two vaccine-specific datasets, namely COVAXLIES and HPVAXLIES, corresponding to their focus on misinformation concerning the COVID-19 or the HPV vaccine, respectively. This organization of VACCINELIES presents the advantage that it allows language researchers and public health experts to contemplate efforts of bootstrapping the discovery of misinformation targeting other vaccines on social media.

The remainder of the paper is organized as follows. Section 2 introduces the process for identifying Misinformation Targets (MisTs), while Section 3 details the organization of MisTs into misinformation taxonomies targeting the COVID-19 and HPV vaccines. Section 4 describes the methodology used for recognizing tweets which evoke any of the MisTs, while Section 5 presents the *stance* annotation process used in VACCINELIES. Section 6 describes a cross-vaccine transfer learning approach and Section 7 presents and discusses the experimental results for cross-vaccine transfer learning. Section 8 summarizes the conclusions.

2. Vaccine Misinformation Targets

In VACCINELIES the identification of misinformation targeting the COVID-19 and HPV vaccines on Twitter was performed in two different ways. First, we have considered several trusted sources, such as the Mayo Clinic, University of California (UC) Davis Health, as well as the Wikipedia page[2], as illustrated in Figure 1 (A). These trusted sources have been actively collecting and debunking misinformation about COVID-19 since the beginning of the pandemic, and much of this misinformation is about the COVID-19 vaccine. MisTs from these trusted sources were merged into a final collection of 17 MisTs targeting the COVID-19 vaccines. However, the identification of trusted sources that debunk misinformation was more challenging for the HPV vaccine. For example, there is no Wikipedia page dedicated to misinformation about the HPV vaccine, although there is a page that is dedicated to the vaccine. There is a Wikipedia page dedicated to vaccine misinformation in general listing several misinformation themes, however, it did not provide specific MisTs for the HPV vaccine.

A second approach, illustrated in Figure 1 (B), was considered, which utilized questions from the Vaccine Confidence Repository (Rossen et al., 2019) to find answers from a vaccine-specific index of unique tweets

[2]en.wikipedia.org/wiki/COVID-19_misinformation#Vaccines

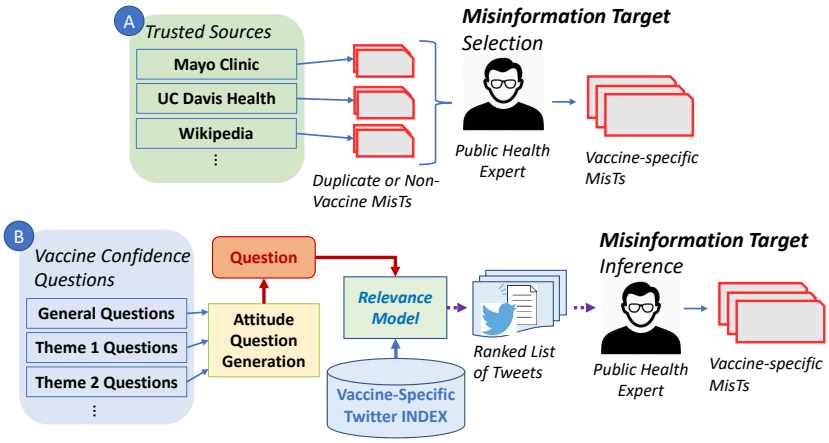

Figure 1: Misinformation Target (MisT) discovery utilizing (A): Trusted sources of vaccine misinformation, and (B): Our Question/Answering framework for vaccine misinformation discovery.

obtained from the Twitter API. Whenever these answers contain misinformation about either the COVID-19 or HPV vaccines, they were considered MisTs. Before using the second approach, approval from the Institutional Review Board at the University of Texas at Dallas was obtained in order to use the Twitter API to collect tweets discussing either the HPV or the COVID-19 vaccine: IRB-21-515 stipulated that our research met the criteria for exemption #8(iii) of the Chapter 45 of Federal Regulations Part 46.101.(b).

Tweets discussing either the COVID-19 or the HPV vaccines were obtained by querying the Twitter API. A collection of 9,133,471 tweets was obtained from the Twitter streaming API as a result of the query *"(covid OR coronavirus) vaccine lang:en -is:retweet"*. We perform Locality Sensitive Hashing (LSH) (Das et al., 2007) with term trigrams, 100 permutations, and a Jaccard threshold of 50%, to produce 5,865,046 unique tweets discussing COVID-19 vaccines. These tweets were authored in the time frame from December 18th, 2019, to July 21st, 2021. Similarly, a collection of 864,008 tweets was obtained from the Twitter historical API as a result of the query *"(human papillomavirus vaccination) OR (human papillomavirus vaccine) OR gardasil OR cervarix OR (hpv vaccine) OR (hpv vaccination) OR (cervical vaccine) OR (cervical vaccination) lang:en -is:retweet"*. After using LSH for detecting near-duplication, we obtained 422,078 unique tweets discussing HPV vaccines. Both tweet collections were organized in vaccine-specific indexes, obtained using Lucene (Foundation, 1999) with the BM25 vector relevance model (Beaulieu et al., 1997), which informed the Q/A framework illustrated in Figure 1 (B).

The questions that were asked originate in the Vaccine Confidence Repository (VCR) (Rossen et al., 2019). For each of the 19 questions available in VCR, we generated attitude-evoking questions using simple regular expressions, such that the expected answers would evoke various attitude responses, on a scale from 1 (no confidence) to 5 (complete confidence) in vaccines. For example, the vaccine confidence question *"Have COVID-19 vaccines been adequately tested for safety?"* was modified to evoke low-confidence attitudes by asking *"Why are you completely sure that the COVID-19 vaccine has not been adequately tested?"*, and was modified to evoke high-confidence attitudes by asking *"What makes you think that the COVID-19 vaccine has certainly been tested adequately?"* We therefore produced $19 \times 5 = 95$ attitude-evoking questions, which retrieved ranked lists of tweets. Public health experts have analyzed the relevance of the top 300 ranked tweets while language experts have selected the discourse units that are shared by sets of tweets, that have the same attitude towards the predication of the VCR question that was originally asked. Using the Pyramid method (Nenkova and Passonneau, 2004), the framing of the vaccine hesitancy was inferred. MisTs were discovered from framings that contained misinformation. The decision of whether a framing contained misinformation was based on finding evidence on the Web, as retrieved by search engines, that the framing expressed known misconceptions, or conspiracy theories. In addition, whenever flawed reasoning was observed, the framing was categorized as misinformation. One researcher with expertise in Web search and an expert on Public Health independently judged the framings that contain misinformation. The two researchers adjudicated their differences and decided that (33%) expressed misinformation. In this way, we identified an additional set of 38 MisTs targeting the COVID-19 vaccine, out of which 7 MisTs were already known to us from the first approach, illustrated in Figure 1 (A). Similarly, 21 MisTs were identified targeting the HPV vaccines. Therefore, VACCINELIES contains 69 vaccine-specific MisTs, with COVAXLIES containing 48 COVID-19 vaccine MisTs and HPVAXLIES containing 21 HPV vaccine MisTs.

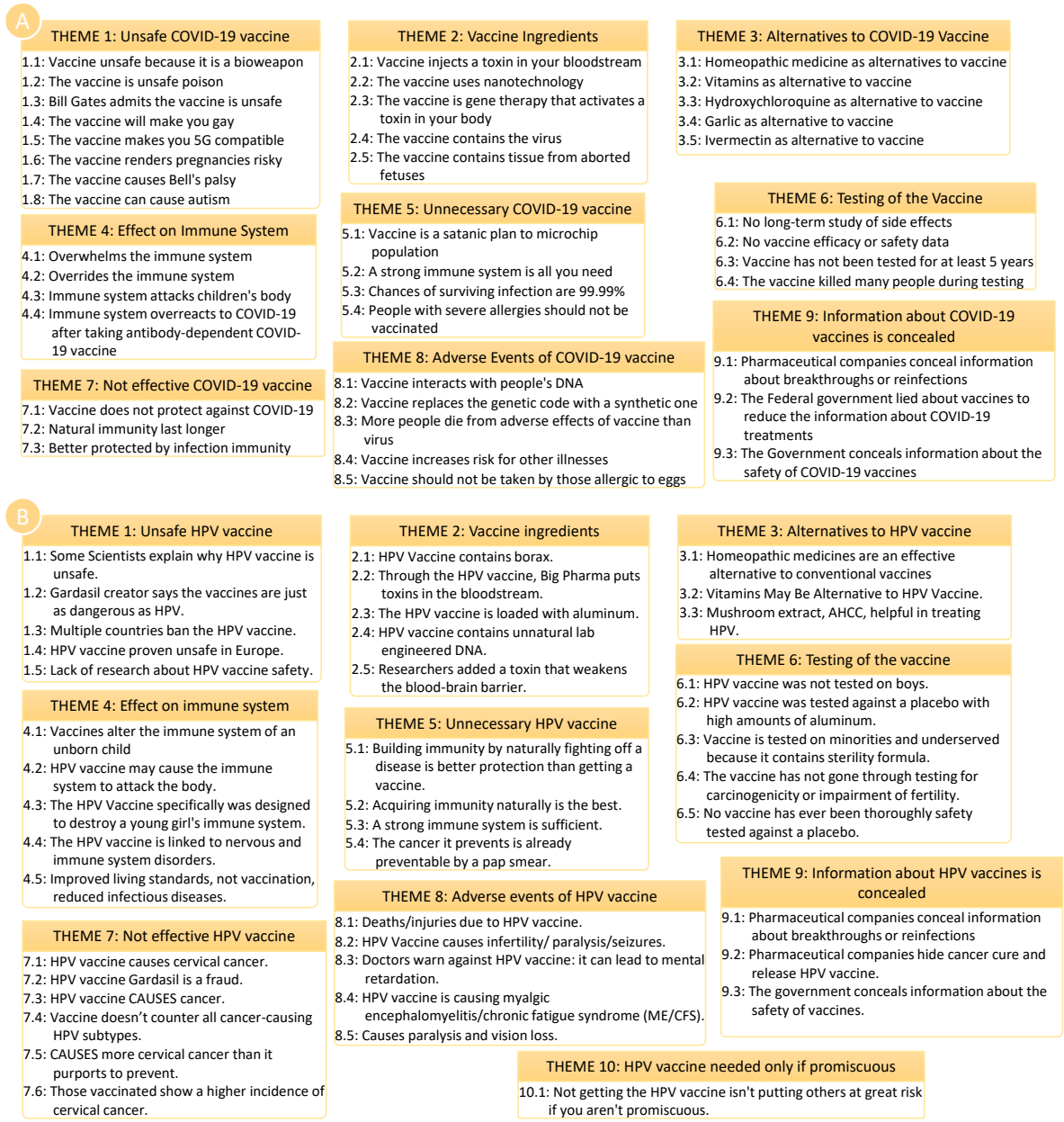

Figure 2: Vaccine misinformation taxonomy for (A) the COVID-19 vaccine and (B) the HPV vaccine.

3. Taxonomy of Vaccine Misinformation

The MisTs were ontologically examined with the goal of discovering common themes and concerns. As in any taxonomy, all MisTs that shared the same theme were further categorized to uncover the concerns that distinguish MisTs within the theme. In this way, the vaccine-specific taxonomy of misinformation has three layers: (1) themes; (2) concerns within each MisT; and (3) MisTs. Misinformation themes represent the highest level of abstraction, while misinformation concerns differentiate the various MisTs from VACCINELIES. Each of the 69 MisTs from VACCINELIES were included in the two vaccine-specific taxonomies. Nine misinformation themes were revealed for COVAXLIES, illustrated in Figure 2 (A), and ten misinformation themes were revealed for HPVAXLIES, illustrated in Figure 2 (B). For each COVAXLIES misinformation theme, a different number of concerns was revealed: the largest number of concerns pertain to the Theme 1, predicating the fact that the COVID-19 vaccines are unsafe (8 concerns) while the smallest number of concerns pertain to the Theme 7 (3 concerns) claiming that the vaccines are not effective (3 concerns) or Theme 9 (3 concerns) that predicates that information about the vaccines is concealed. Although the misinformation taxonomies for the COVID-19 and HPV vaccine share nine themes, it is interesting to note that the concerns are vastly different, as illustrated in Figure 2. For each HPVAXLIES misinformation theme, a different number of concerns were revealed: the largest

number of concerns relating to the Theme 7, predicating that the HPV vaccine is not effective (6 concerns) while the smallest number of concerns involved the Theme 10, claiming that the HPV vaccine is only needed if promiscuous. Nine misinformation themes are shared between the CoVaxLies and HpVaxLies taxonomies, all characterizing aspects that impact confidence in their respective vaccines, while one unique theme was identified for the HPV vaccine. While all the other themes generally focus on the factor of confidence in their respective vaccines, HpVaxLies theme 10, that the HPV vaccine is needed only if promiscuous, touches on the factor of complacency. Confidence, along with convenience and complacency, are well known universal factors contributing to vaccine hesitancy, according to the 3C model (Macdonald, 2015).

4. Recognizing Tweets that Propagate Misinformation

As in COVIDLies (Hossain et al., 2020), which inspired our work, the recognition of tweets that evoked any of the MisTs from VaccineLies relies on (a) the identification of the tweets deemed to evoke a MisT; and (b) the recognition of the stance of the tweet author towards the evoked MisT. This process of recognizing tweets that evoke MisTs was detailed in (Weinzierl and Harabagiu, 2021), presenting the challenges in using BM25 (Beaulieu et al., 1997) and BERTScore (Zhang et al., 2020) as retrieval models. To recognize tweets evoking any MisT we reused the vaccine-specific tweet indexes for COVID-19 and HPV, presented in Section 2, and performed retrieval using each MisT as a query. While we had identified that the BM25 model outperformed BERTScore in retrieving truly evoking misinformation, we recognized the value in moving beyond term-based retrieval systems and considering the advantages of BERTScore.

The tweets retrieved when using BERTScore are characterized by less term overlap with the textual content of the MisTs, and, thus, BERTScore emphasizes more semantic relevancy. We relied on this observation by combining the benefits of the retrieval model provided by BM25 scoring with the semantic relevancy provided by BERTScore. Our tweet retrieval system used the BM25 (Beaulieu et al., 1997) scoring function to select the top 1,000 initial candidate tweets for each MisT, which were then re-ranked against each MisT using a BERT-RERANK (Nogueira and Cho, 2020) system. We initialized the BERT weights to a re-ranking model which was trained on MSMARCO (Nguyen et al., 2016), a large-scale question answering collection, using BioBERT (Lee et al., 2019), a biomedical domain-specific BERT (Devlin et al., 2019) language model. This system had found success in a recent COVID-19 question answering shared task (Weinzierl and Harabagiu, 2020), and produced 80% MisT-evoking tweets from initial experiments on COVID-19 vaccine MisTs, nearly doubling the 42%

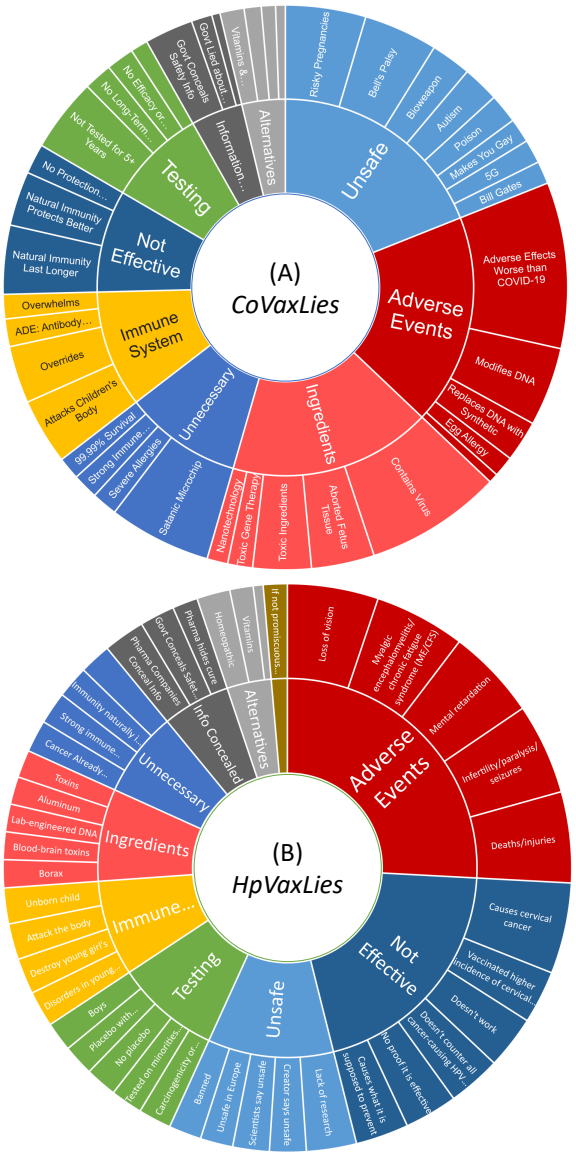

Figure 3: Distribution of (A) COVID-19 and (B) HPV vaccine misinformation themes and concerns in the tweets available from VaccineLies.

MisT-evoking tweets found in prior work (Weinzierl and Harabagiu, 2021). The final top 200 tweets for each MisT were judged by language experts for relevance. We selected the 200 best scored tweets because (1) the same number of tweets was considered in the most similar prior work (Hossain et al., 2020); and (2) it was a number of tweets that did not overwhelm our human judges.

In addition, as shown in Figure 3, the misinformation taxonomies, outlined in Section 3, enabled us to identify the most common misinformation themes and concerns across both the COVID-19 and HPV vaccines. Figure 3 (A) illustrates the most commonly evoked misinformation themes and concerns for the COVID-19 vaccine, as was judged in VaccineLies, while Figure 3 (B) illustrates the most commonly evoked misinformation themes and concerns for the HPV vaccine.

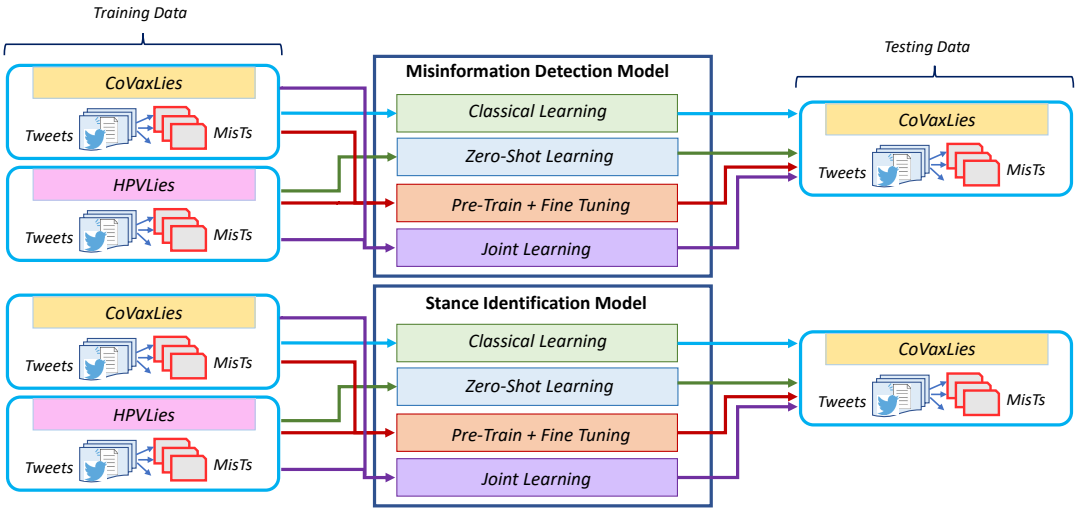

Figure 4: Scenarios for transfer learning on VACCINELIES for (a) Misinformation Detection, or (b) Misinformation Stance Identification.

	COVAXLIES	HPVAXLIES	Total
MisTs	48	21	69
Evoke	7,152	2,230	9,382
Accept	3,720	1,365	5,085
Reject	2,194	617	2,811
No Stance	1,238	248	1,486
Tweets	12,118	2,524	14,642

Table 2: Distribution of MisTs, tweets evoking them and their stance in VACCINELIES.

Misinformation themes most often evoked from COVAXLIES are the belief that the COVID-19 vaccines are unsafe, that they cause adverse events, and that the ingredients of the vaccines should be a major concern. Moreover, the primary concerns regarding the lack of safety of the COVID-19 vaccines involves risky pregnancies or Bell's palsy. Misinformation themes most often evoked from HPVAXLIES are the belief that the HPV vaccines cause adverse events, that they are not effective, and that the vaccines are unsafe. Moreover, the primary concerns regarding the adverse events of the HPV vaccines involves loss of vision, myalgic encephalomyelitis/chronic fatigue syndrome (ME/CFS), and mental retardation.

5. Annotation of Stance towards Misinformation

Researchers from the Human Language Technology Research Institute (HLTRI) at the University of Texas at Dallas judged (a) whether a tweet evokes any of the MisTs from VACCINELIES; and (b) the stance of the tweet author towards the MisT. 14,642 tweets were judged, with 9,382 tweets evoking one or more MisTs from VACCINELIES. They were organized in [tweet, MisT] *pairs*, annotated with a stance value that could be *Accept*, *Reject* or *No Stance*. The retrieval of tweets produced 84% of tweets evoking Mists towards COVID-19 vaccines, and 88% of tweets evoking MisTs towards the HPV vaccine, which is a significant improvement from the prior best of 42% (Weinzierl and Harabagiu, 2021). Statistics for the number of tweets evoking a MisT, as well as of the stance their authors have towards the MisT, are provided in Table 2. To evaluate the quality of judgements, we randomly selected a subset of 1,000 tweets (along with the MisT against which they have been judged a *stance* value), which have been judged by at least two different language experts. Inter-judge agreement was computed using the Cohen Kappa score, yielding a score of 0.63 for the *stance* of tweets for COVID-19 vaccine MisTs and 0.67 for the *stance* of tweets for the HPV vaccine MisTs, which indicates moderate agreement between annotators (0.60-0.79) (McHugh, 2012).

To enable the usage of VACCINELIES in supervised learning frameworks targeting misinformation detection on Twitter, we provide: (a) a training collection; (b) a development collection; and (c) a test collection. The VACCINELIES training collection, which consists of 10,637 [tweet, MisT] pairs (8,777 for COVAXLIES and 1,860 for HPVAXLIES), was utilized to train our MisT-evoking detection and stance identification systems, described in Section 6. The VACCINELIES development collection, which consists of 1,109 [tweet, MisT] pairs (920 for COVAXLIES and 189 for HPVAXLIES), was used to select model hyperparameters, such as threshold values. The VACCINELIES test collection, which consists of 2,896 [tweet, MisT] pairs (2,421 for COVAXLIES and 475 for HPVAXLIES), was used to evaluate the detection of tweets which evoke MisTs along with stance identification approaches, enabling us to report the results in Section 7.

6. Transfer Learning for Recognizing Misinformation and Stance

The goal of transfer learning is to build a learner from one domain by transferring information from a related domain, and VACCINELIES provides two vaccine misinformation domains, namely COVAXLIES and HPVAXLIES. As it can be expensive and time-consuming to acquire and annotate sufficient examples of vaccine-specific misinformation for a new vaccine, it is of great interest to natural language processing researchers and public health practitioners alike to best utilize existing vaccine misinformation collections.

Transfer learning has found success over a wide variety of tasks and modalities (Weiss et al., 2016), therefore we examine four different learning scenarios, illustrated in Figure 4, characterized by the training data that is available for learning (a) COVID-19 vaccine *misinformation detection* and (b) COVID-19 vaccine misinformation *stance* identification when:

(Scenario 1) training on COVAXLIES;
(Scenario 2) training on HPVAXLIES;
(Scenario 3) training on the entire VACCINELIES; or
(Scenario 4) pre-training on HPVAXLIES and fine-tuning on COVAXLIES.

Scenario 1 represents the *Classical* non-transfer learning approach of training and evaluating on the same domain. Scenario 2 utilizes *Zero-Shot* learning to rely solely on a different domain during training by training on the HPV vaccine misinformation collection. This scenario provides significant value to public health practitioners, as it represents the most rapid approach possible when there is interest in the detection of misinformation for a new vaccine, as it requires zero examples of tweets evoking misinformation targeting the new vaccine. Scenario 3 performs *Joint* multi-domain training on both COVID-19 and HPV vaccine misinformation, and represents the benefit of including additional vaccines in the VACCINELIES collection. Scenario 4 is similar to how pre-trained language models, like BERT (Devlin et al., 2019), are pre-trained on one domain, such as online English text, and then fine-tuned on a different domain. We *Pre-Train* our model on HPV vaccine misinformation and then *Fine-Tune* the pre-trained model on COVID-19 vaccine misinformation. This scenario highlights the value of discovering misinformation targeting a new vaccine when misinformation targeting a different vaccine is available, thus avoiding learning from scratch. All four scenarios also apply for HPV vaccine *misinformation detection* and *stance* identification.

Misinformation detection involves determining whether a tweet evokes a specific MisT, given a [tweet, MisT] pair. We cast misinformation detection as a binary classification problem, and therefore we design a neural architecture to perform binary classification. Misinformation stance identification involves identifying which *stance value* the author of a tweet holds towards a specific MisT, given a [tweet, MisT] pair. We cast misinformation stance identification as a three-way classification problem between *stance* values of *"Accept"*, *"Reject"*, and *"No Stance"*. For both tasks, we utilize COVID-Twitter-BERT-v2 (Müller et al., 2020), a pre-trained domain-specific language model which started with neural weights equal to those of BERT (Devlin et al., 2019) but was additionally pre-trained on the masked language modeling task for 97 million COVID-19 tweets. Joint Word-Piece Tokenization is performed for both a MisT m_j and a tweet t_i, which produces a single sequence of word-piece tokens for both the misinformation target and the tweet separated by a special $[SEP]$ token. The beginning $[CLS]$ token and end $[SEP]$ token are placed at the beginning and end of the joint sequence respectively. COVID-Twitter-BERT-v2 produces contextualized embeddings for each word-piece token, and we select the first contextualized embedding to represent the entire joint sequence, representing the initial $[CLS]$ token embedding. This embedding is provided to a fully-connected layer with a softmax activation function, which outputs a task-dependent probability distribution. The vaccine misinformation detection model, which we call the BERT Vaccine Misinformation Evocation Detector (BERT-VMED), outputs a probability distribution over $P(Evoke|t_i, m_j)$, where $Evoke$ can take the value of *"True"* or *"False"*. The vaccine misinformation *stance* identification model, which we call the BERT Vaccine Misinformation Stance Identifier (BERT-VMSI), outputs a probability distribution over $P(Stance|t_i, m_j)$, where $Stance$ can take the value of *"Accept"*, *"Reject"*, and *"No Stance"*. Misinformation is detected for BERT-VMED when the probability is larger than a predefined threshold T, and *stance* is identified based for BERT-VMSI by the maximum *stance* value probability. In our experiments, the value of the threshold T was determined by maximizing the F_1 score of each model on the development collection. Both BERT-VMED and BERT-VMSI are trained end-to-end using the cross-entropy loss function minimized with ADAM (Kingma and Ba, 2014), a variant of gradient descent.

Testing	Scenario	F_1	P	R
COVAXLIES	Classical	90.7	84.6	97.7
	Zero-Shot	73.5	58.1	**100.0**
	Joint	91.2	87.3	95.5
	Pre-Train	**91.7**	**87.7**	96.1
HPVAXLIES	Classical	93.6	88.3	99.5
	Zero-Shot	92.9	86.7	**100.0**
	Joint	93.8	89.1	99.0
	Pre-Train	**94.5**	**90.6**	98.8

Table 3: Vaccine misinformation detection results for the BERT Vaccine Misinformation Evocation Detector (BERT-VMED) utilizing vaccine transfer learning scenarios.

Testing	Scenario	Macro			*Accept*			*Reject*		
		F_1	P	R	F_1	P	R	F_1	P	R
CoVaxLies	Classical	83.4	81.6	**85.5**	85.9	81.5	**90.8**	80.9	81.8	**80.1**
	Zero-Shot	72.5	84.1	64.2	79.0	85.4	73.5	65.9	82.9	54.8
	Joint	83.3	82.8	84.1	85.2	81.9	88.8	**81.4**	**83.6**	79.4
	Pre-Train	**83.6**	**85.7**	81.5	**86.8**	**88.7**	85.0	80.3	82.7	78.1
HpVaxLies	Classical	79.6	79.9	79.4	83.1	82.2	84.0	76.2	77.7	74.8
	Zero-Shot	74.6	71.8	79.7	79.2	68.4	**93.9**	70.0	75.3	65.4
	Joint	80.5	80.5	80.6	85.9	83.8	88.2	75.0	77.2	72.9
	Pre-Train	**84.0**	**85.1**	**83.2**	**88.1**	**86.4**	89.7	**80.0**	**83.7**	**76.6**

Table 4: Vaccine misinformation *stance* identification results for the BERT Vaccine Misinformation Stance Identifier (BERT-VMSI) utilizing several vaccine transfer learning scenarios.

7. Experimental Results

7.1. Misinformation Detection

Table 3 lists the experimental results we obtained for misinformation detection, where bolded numbers are the best results obtained. We show in Table 3 the training scenarios and the testing collections used for BERT-VMED. To evaluate the quality of vaccine transfer learning on misinformation identification on the test collections from CoVaxLies and HpVaxLies we used the Precision (P), Recall (R), and F_1 metrics when detecting whether a tweet evoked a MisT for each [tweet, MisT] pair in the test collection. Evaluation of the four vaccine transfer learning scenarios discussed in Section 6 was performed using BERT-VMED across two different evaluations, the CoVaxLies test collection and the HpVaxLies test collection. The *Classical* scenario involves training BERT-VMED on the same domain as it was evaluated, and provides a baseline comparison when no cross-vaccine transfer learning is utilized, achieving F_1 scores of 90.7 on CoVaxLies and 93.6 on HpVaxLies. The *Zero-Shot* scenario involves training BERT-VMED on a different domain than the evaluation, and demonstrates zero-shot vaccine transfer learning, achieving F_1 scores of 73.5 on CoVaxLies and 92.9 on HpVaxLies. This zero-shot approach performs worse than the baseline, but still achieves competitive performance with zero vaccine-specific training data, indicating that this approach could provide significant value as new or less-studied vaccines are discussed on social media. *Joint* training of BERT-VMED on both vaccine domains demonstrates the value of training on multi-vaccine collections, achieving F_1 scores of 91.2 on CoVaxLies and 93.8 on HpVaxLies. The *Pre-Train* scenario of BERT-VMED was pre-trained on one domain and fine-tuned on a different domain, enabling quick adaptation of the BERT-VMED model to new vaccines, achieving the best F_1 scores of 91.7 on CoVaxLies and 94.5 on HpVaxLies.

7.2. Misinformation Stance Identification

Table 4 lists the experimental results we obtained when recognizing the stance of tweet authors towards the evoked MisT. We show in Table 4 the training scenarios and the testing collections used for BERT-VMSI. The bolded numbers represent the best results we obtained. To evaluate the quality of vaccine transfer learning on misinformation *stance* identification on the test collections from CoVaxLies and HpVaxLies we used the Precision (P), Recall (R), and F_1 metrics for identifying the *Accept* and *Reject* values of stance. We also compute a Macro averaged Precision, Recall, and F_1 score. Evaluation of the four vaccine transfer learning scenarios discussed in Section 6 was performed using BERT-VMSI across two different evaluations, the CoVaxLies test collection and the HpVaxLies test collection. We see similar transfer learning results for misinformation *stance* identification when compared to misinformation detection. Training BERT-VMSI on a different domain than the evaluation continues to perform worse than training BERT-VMSI on the same domain as it was evaluated, but this zero-shot approach still produces competitive results, achieving Macro F_1 scores of 72.5 on CoVaxLies and 74.6 on HpVaxLies. Jointly training BERT-VMSI only results in performance improvements for HpVaxLies over training BERT-VMSI on the same domain, while pre-training BERT-VMSI on one domain and fine-tuning on a different domain continues to perform best, achieving Macro F_1 scores of 83.6 on CoVaxLies and 84.0 on HpVaxLies.

8. Conclusion

We have described the annotation effort that made possible the creation of the VaccineLies dataset, which consists of tweets propagating misinformation about two types of vaccines, namely the COVID-19 and the HPV vaccines. Misinformation targeting these vaccines was represented as Misinformation Targets (MisTs), which were discovered by two different methods. Moreover, the MisTs were organized in vaccine-specific taxonomies, revealing the misinformation themes and concerns. A large set of tweets evoking any of the MisTs were identified and are provided as part of VaccineLies, along with annotations of the stance of the tweet authors towards the evoked MisT.

Because VACCINELIES provides misinformation targeting two different vaccines, we also presented several scenarios of transfer learning, highlighting the advantages of having a resource such as VACCINELIES for the case when misinformation about yet another new vaccine shall be needed to be discovered.

9. Bibliographical References

Beaulieu, M. M., Gatford, M., Huang, X., Robertson, S., Walker, S., and Williams, P. (1997). Okapi at trec-5. In *The Fifth Text REtrieval Conference (TREC-5)*, pages 143–165, January.

Das, A. S., Datar, M., Garg, A., and Rajaram, S. (2007). Google news personalization: Scalable online collaborative filtering. In *Proceedings of the 16th International Conference on World Wide Web*, WWW '07, page 271–280, New York, NY, USA. Association for Computing Machinery.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Foundation, A. S. (1999). Apache lucene. `https://github.com/apache/lucene`.

Hossain, T., Logan IV, R. L., Ugarte, A., Matsubara, Y., Young, S., and Singh, S. (2020). COVIDLies: Detecting COVID-19 misinformation on social media. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*. Association for Computational Linguistics.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kouzy, R., Abi Jaoude, J., Kraitem, A., El Alam, M. B., Karam, B., Adib, E., Zarka, J., Traboulsi, C., Akl, E. W., and Baddour, K. (2020). Coronavirus goes viral: Quantifying the covid-19 misinformation epidemic on twitter. *Cureus*, 12(3):e7255–e7255, Mar. 32292669[pmid].

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., and Kang, J. (2019). Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*.

Ma, J., Gao, W., Mitra, P., Kwon, S., Jansen, B. J., Wong, K.-F., and Cha, M. (2016). Detecting rumors from microblogs with recurrent neural networks. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, IJCAI'16, page 3818–3824. AAAI Press.

Ma, J., Gao, W., and Wong, K.-F. (2017). Detect rumors in microblog posts using propagation structure via kernel learning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 708–717, Vancouver, Canada, July. Association for Computational Linguistics.

Macdonald, N. (2015). Vaccine hesitancy: Definition, scope and determinants. *Vaccine*, 32, 04.

McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3):276–282.

Müller, M., Salathé, M., and Kummervold, P. E. (2020). Covid-twitter-bert: A natural language processing model to analyse covid-19 content on twitter. *https://arxiv.org/abs/2005.07503*.

Nenkova, A. and Passonneau, R. (2004). Evaluating content selection in summarization: The pyramid method. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 145–152, Boston, Massachusetts, USA, May 2 - May 7. Association for Computational Linguistics.

Nguyen, T., Rosenberg, M., Song, X., Gao, J., Tiwary, S., Majumder, R., and Deng, L. (2016). MS MARCO: A human generated machine reading comprehension dataset. *CoRR*, abs/1611.09268.

Nogueira, R. and Cho, K. (2020). Passage re-ranking with bert.

Rossen, I., Hurlstone Angeli, M. J., Dunlop, P. D., and Lawrence, C. (2019). Accepters, fence sitters, or rejecters: Moral profiles of vaccination attitudes. *Social Science Medicine*, pages 23–27.

Weinzierl, M. and Harabagiu, S. M. (2020). The university of texas at dallas hltri's participation in epic-qa: Searching for entailed questions revealing novel answer nuggets. In *Thirteenth Text Analysis Conference*, volume 13. Text Analysis Conference.

Weinzierl, M. A. and Harabagiu, S. M. (2021). Automatic detection of covid-19 vaccine misinformation with graph link prediction. *Journal of Biomedical Informatics*, 124:103955.

Weinzierl, M. and Harabagiu, S. (2022). Identifying the adoption or rejection of misinformation targeting covid-19 vaccines in twitter discourse. In *Proceedings of the ACM Web Conference 2022*, WWW '22, page 3196–3205, New York, NY, USA. Association for Computing Machinery.

Weinzierl, M. A., Hopfer, S., and Harabagiu, S. M. (2021). Scaling up the discovery of hesitancy profiles by identifying the framing of beliefs towards vaccine confidence in twitter discourse. *medRxiv*.

Weiss, K., Khoshgoftaar, T. M., and Wang, D. (2016). A survey of transfer learning. *Journal of Big Data*, 3(1):9, May.

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., and Artzi, Y. (2020). Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Zubiaga, A., Hoi, G. W. S., Liakata, M., Procter, R., and Tolmie, P. (2016). Analysing how people orient to and spread rumours in social media by looking at conversational threads. *PLoS ONE*, 11.