Original Paper

Identifying False Human Papillomavirus (HPV) Vaccine Information and Corresponding Risk Perceptions From Twitter: Advanced Predictive Models

Tre Tomaszewski[1], MS; Alex Morales[2], PhD; Ismini Lourentzou[3], PhD; Rachel Caskey[4], MD; Bing Liu[5], PhD; Alan Schwartz[6], PhD; Jessie Chin[1,7], PhD

[1]School of Information Sciences, University of Illinois at Urbana-Champaign, Champaign, IL, United States
[2]Department of Computer Science, University of Illinois at Urbana-Champaign, Urbana, IL, United States
[3]Department of Computer Science, Virginia Polytechnic Institute and State University, Blacksburg, VA, United States
[4]College of Medicine, University of Illinois at Chicago, Chicago, IL, United States
[5]Department of Computer Science, University of Illinois at Chicago, Chicago, IL, United States
[6]Department of Medical Education, University of Illinois at Chicago, Chicago, IL, United States
[7]Cancer Center at Illinois, University of Illinois at Urbana-Champaign, Urbana, IL, United States

Corresponding Author:
Jessie Chin, PhD
School of Information Sciences
University of Illinois at Urbana-Champaign
501 E Daniel St
Champaign, IL, 61820
United States
Phone: 1 217 333 0125
Email: chin5@illinois.edu

Abstract

Background: The vaccination uptake rates of the human papillomavirus (HPV) vaccine remain low despite the fact that the effectiveness of HPV vaccines has been established for more than a decade. Vaccine hesitancy is in part due to false information about HPV vaccines on social media. Combating false HPV vaccine information is a reasonable step to addressing vaccine hesitancy.

Objective: Given the substantial harm of false HPV vaccine information, there is an urgent need to identify false social media messages before it goes viral. The goal of the study is to develop a systematic and generalizable approach to identifying false HPV vaccine information on social media.

Methods: This study used machine learning and natural language processing to develop a series of classification models and causality mining methods to identify and examine true and false HPV vaccine–related information on Twitter.

Results: We found that the convolutional neural network model outperformed all other models in identifying tweets containing false HPV vaccine–related information (*F* score=91.95). We also developed completely unsupervised causality mining models to identify HPV vaccine candidate effects for capturing risk perceptions of HPV vaccines. Furthermore, we found that false information contained mostly loss-framed messages focusing on the potential risk of vaccines covering a variety of topics using more diverse vocabulary, while true information contained both gain- and loss-framed messages focusing on the effectiveness of vaccines covering fewer topics using relatively limited vocabulary.

Conclusions: Our research demonstrated the feasibility and effectiveness of using predictive models to identify false HPV vaccine information and its risk perceptions on social media.

(J Med Internet Res 2021;23(9):e30451) doi: 10.2196/30451

KEYWORDS

misinformation; disinformation; social media; HPV; human papillomavirus vaccination; vaccination; causality mining; cause; effect; risk perceptions; vaccine; perception; risk; Twitter; machine learning; natural language processing; cervical cancer

Introduction

About 13,000 women are newly diagnosed with invasive cervical cancer and over 4000 women die from it every year [1]. Cervical cancer is caused by certain types of human papillomavirus (HPV) [2,3]. HPV is the most common sexually transmitted infection in the United States with an estimated 6.2 million new infections every year among persons 14 to 44 years of age [4-6]. In addition to cervical cancer, HPV is the causal mediator in multiple head and neck cancers, genital cancers, and anal cancers [7-9]. The overall burden of HPV-associated cancers has been increasing in the United States [9]. Prevention of HPV is more challenging than most sexually transmitted infections as condoms do not provide complete protection against infection [10]. Hence, prevention through vaccination is critical in decreasing the burden of cancer due to this ubiquitous infection.

The HPV vaccine is universally recommended for all adolescents [10]. Despite the exceptional efficacy (up to 90% protection) in preventing precancerous lesions caused by the targeted HPV types [11-13], only 56.8% of 13 to 17-year-old females and 51.8% of 13 to 17-year-old males in the United States have completed the HPV vaccine series [14]. There are many known barriers to HPV vaccination, including misconceptions about the side effects and adverse events from HPV vaccines, misbeliefs around the need for vaccines, inconsistent advice received from health care givers, costs to complete the vaccination, limited access to clinics, and violations to cultural beliefs [15-21]. Among these barriers, the bias in risk perceptions has not only been associated with low intention of vaccination [22-25] but also with the actual vaccination behavior [16,22,26-30]. The National Immunization Survey revealed the top 3 parental concerns of HPV vaccines to be a lack of knowledge, low perceived usefulness of vaccine (low perceived risk of HPV infection), and high perceived risks of side effects and safety concerns [31], underscoring the importance of risk perceptions in HPV vaccination decisions.

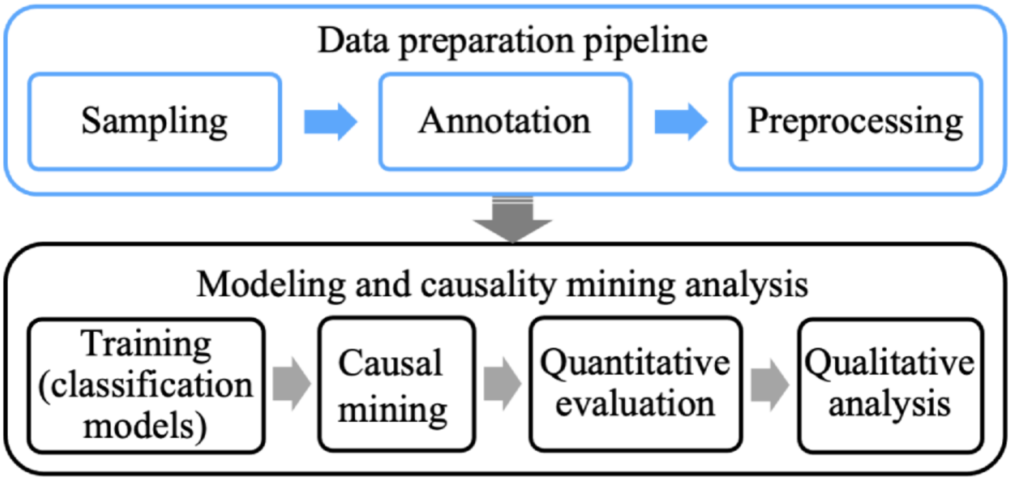
Social media has become an important information source for people to exchange vaccine-related opinions and form their attitudes toward vaccines [32-38]. Its impact is striking, especially for Twitter vaccine information, as HPV vaccine–related opinions on Twitter have been associated with actual vaccine acceptance and coverage [39]. Existing research has investigated the emerging themes and public attitudes toward the pro- and antivaccine online discussions about HPV vaccines [19,24,40-44]. Although multiple false conspiracies and myths around HPV vaccines have been identified, no research has used an automatic computational approach to extract the causal cues of the main vaccination arguments used in circulation of HPV vaccine misinformation. Research has shown that none of content quality, scientific robustness, or the veracity of the information has been found to be indicative of the spread of information, while false or unverified information sometimes becomes more viral than true information [45,46]. As attention to the propagation of false information on social media has surged [45,47-50], an automatic, systematic, and generalizable approach to detect socially endorsed false health information remains understudied. The threats of false information are critical because the reliance (ie, perceived accuracy) on false information can be amplified with each exposure to it and further magnified through social networks [51-53]. People can be especially victimized by the proliferation of user-generated false health information given their lack of health literacy, incompetence in credibility judgments, and the mixed quality of health news despite the sources cited [54-56]. Hence, detecting false health information before it propagates is an important step toward minimizing the threats of false information [57].

Several works have targeted health misinformation [58,59], with most studies using descriptive approaches to study known health misinformation and performing analysis to uncover the common misbeliefs, demographic and geographic patterns, and social media user behaviors [25,60,61]. A few studies have implemented computational models to identify health misinformation from other social medial platforms (such as YouTube and Instagram); however, none of them have attempted to identify health misinformation from short and sometimes incomplete text information, such as tweets [62,63]. In contrast with other related work, we combined a classification model for identifying false HPV vaccine information with unsupervised causality mining to extract the risk perceptions considered to be the attributable causes of HPV antivaccine health concerns based on the content expressed in Twitter messages. To this end, we conducted an infodemiology study to use natural language processing and machine learning methods, such as classification, clustering, dependency parsing, and phrase mining, to identify those false HPV vaccination arguments that frequently appear in social media. Our methodological analysis can be applied to other domains, such as COVID-19 vaccination, food safety, and politics, to extract insightful information regarding the differences and similarities between truthful and misleading claims shared online.

Methods

We collected a corpus related to HPV vaccines with tweets published from December 2013 until December 2017. We used the formerly known Crimson Hexagon's (now Brandwatch) social media analytics application programming interface and a list of HPV-related search terms, including, but not limited to, “HPV vaccine,” “papillomavirus vaccine,” “cervical cancer vaccine,” “HPV shot,” “cervical cancer shot,” and “Gardasil.” Our modeling pipeline consists of several steps: sampling, annotation and data preprocessing, training, and analysis (see Figure 1). The data preprocessing stage includes rule-based lexical normalization and unsupervised pretraining of word embeddings.

Figure 1. Causality mining data collection and modeling pipeline.


Data preparation pipeline
Sampling
Annotation
Preprocessing
Modeling and causality mining analysis
Training (classification models)
Causal mining
Quantitative evaluation
Qualitative analysis

First, we randomly sampled 1000 tweets per year and passed them to 2 annotators in 2 rounds. Both annotators received basic training about HPV vaccines (including extensive reading of the verified HPV vaccine-related materials from the National Cancer Institute, Center for Disease Control and Prevention, and American Cancer Society), and 1 had formal educational training in medical sciences. Similar to related work in misinformation detection [64-66], we framed the task as a binary classification, in which each tweet is categorized as true or false information (in which false information includes partial-false or partial-true information). We thus not only asked the annotators to judge the veracity of the content for each tweet but also allowed them to select an additional option as "not applicable" for tweets that did not fall under any of the 2 categories (eg, opinionated text and other nonfactual or irrelevant posts). Tweets labeled as not applicable were filtered out from the annotation pipeline. Any discrepancies of the ratings from the 2 annotators were reconciled through discussion. For the interrater reliability, a Cohen's kappa coefficient (κ) of 0.75, was considered to indicate good agreement on the task [67]. The resulting data set consisted of 5000 labeled and 702,858 unlabeled tweets. Character lengths of the tweets, including all mentions, retweets, and hashtags, ranged from 21 to 826 characters.

To reduce vocabulary size for the lexical normalization steps, words were formatted in lower case and URLs were removed; numerals, and Twitter-specific items, such as user mentions (usernames prefixed by "@") or retweets, were tagged and mapped to a common special token per category (ie, NUMBER, MENTION, RT, respectively). Selected contractions were then replaced with their canonical forms: for example, "Can't" was replaced with "Cannot," "You'll" was replaced with "You will," "&" was replaced with "and," etc. Additionally, hyphens and forward slashes were replaced with spaces, alphanumeric pairings were processed, instances of 2 or more user mentions were reduced to 2 "MENTION" tokens, hashtag quotes and other types of punctuation were removed, and multiple leading or trailing white spaces were replaced with a single one. This process reduced the length of each tweet, which could range between 18 and 295 characters.

The final vocabulary size based on the training set was 4098 terms (including 1 vocabulary term representing a blank space). Analysis of terms weighted by their frequency odds ratio (ie, the ratio of occurrence in each category) showed certain terms were overrepresented in the true category but appeared infrequently in the false category, for example, words that strongly indicated the effectiveness of HPV vaccines on cancer prevention spread online, such as "prevent," "protect," and "effective." On the other hand, false messages contain terms such as "danger," "adverse," and "deadly," and focus more on the negative causal effects that are used as arguments for vaccination.

Results

Classification Model

Word embeddings map discrete word tokens to real-valued vector representations, where semantically similar words have similar vectors and are therefore closer in the embedding space. In general, pretraining of such word embeddings has been found to be beneficial for several natural language processing tasks, allowing for faster model convergence and task performance improvements. Therefore, we trained an unsupervised embedding model, FastText [68], with our full Twitter collection as training data and with the aforementioned preprocessing. Compared to other word representation models, FastText can produce word vectors for out-of-vocabulary words and has been proven to be a strong baseline for short text similarity, with its open-sourced implementation allowing for faster training [69]. More specifically, FastText produced 300-dimensional vector representations for each term in our vocabulary, which was used as the initialization for our model's embedding layer. We also experimented with Wikipedia-pretrained embeddings and without any pretrained embeddings: our experiments showed that the model performed better in terms of accuracy when initialized with HPV-related pretrained word embeddings.

Finally, we divided the annotated data into 60% training, 20% validation, and 20% testing, keeping the same splits across all models for a fair comparison. Deduplication of tweets with exact matches within each set left 3661 tweets in total (2142 for training, 758 for validation, and 761 for the test set). We experimented with several model architectures, including convolutional neural network (CNN) [70], bidirectional long short-term memory (BiLSTM), and traditional models, including support vector machine and Naive Bayes. We trained with cross-entropy, Adaptive Moment Estimation with a 10^{-4} learning rate, 0.01 decay, and a 32 batch size for the neural models. Hyperparameter tuning was performed using the Tune library [71]. In Table 1, we report the mean and SD of the top-5 performing model variations. Our experimental evaluation showed that CNNs performed better than did the other models (see Figures 2 and 3 for respective confusion matrices and the area under the receiver operating characteristic curve comparisons between neural networks). Of the top-5 best performing models for either of the neural networks, the CNN required less training time than did the BiLSTM. The mean training time per epoch for the CNN was 11.5 ms (SD 1.09, minimum 16, maximum 16, median 12), whereas the mean training time per epoch for the BiLSTM was 51.3 ms (SD 34.07, minimum 14, maximum 88, median 81). Our best-performing CNN model had 256 convolutional filters, including –3 kernels of width (3,4, and 5) and rectified linear unit nonlinearities; a max pooling layer, a fully connected layer of 128 units with rectified linear unit activations and 0.1 dropout, and a final softmax output layer that produced the classification prediction.

Table 1. Identifying false human papillomavirus vaccine information: classification model comparison.

Model	Accuracy	Precision	Recall	*F* score
SVM[a], mean	57.424	57.806	56.721	55.532
Naive Bayes, mean	51.774	52.485	52.301	51.090
CNN[b], mean (SD)	91.958 (0.269)	91.953 (0.272)	91.946 (0.271)	*91.946 (0.270)* [c]
BiLSTM[d], mean (SD)	91.643 (0.432)	91.710 (0.396)	91.574 (0.453)	91.618 (0.438)

[a] SVM: support vector machine.

[b] CNN: convolutional neural network.

[c] Italics indicate the highest *F* score in the table.

[c] BiLSTM: bidirectional long short-term memory.

Figure 2. Confusion Matrix for best-performing CNN model. BiLSTM: bidirectional long short-term memory; CNN: convolutional neural network.

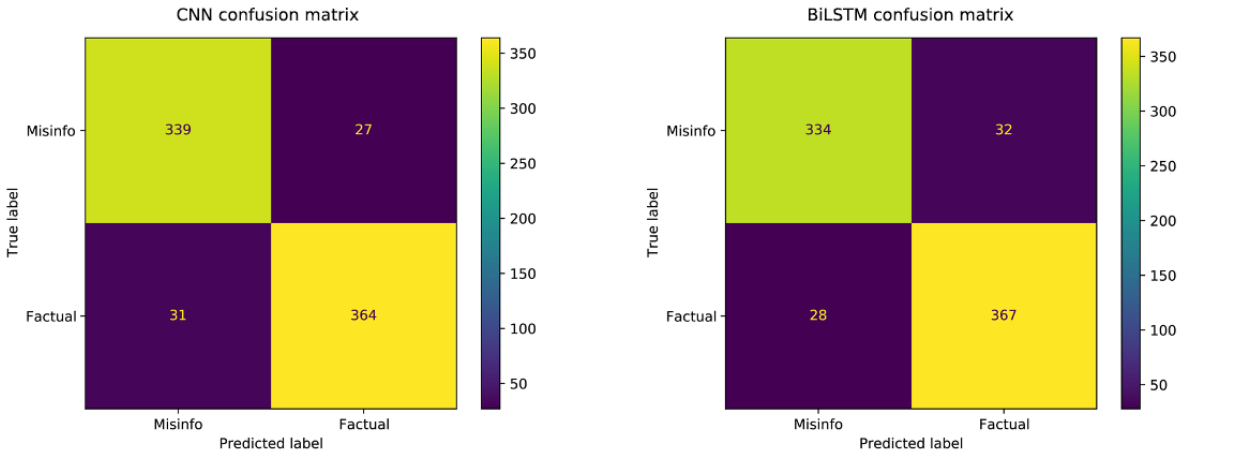
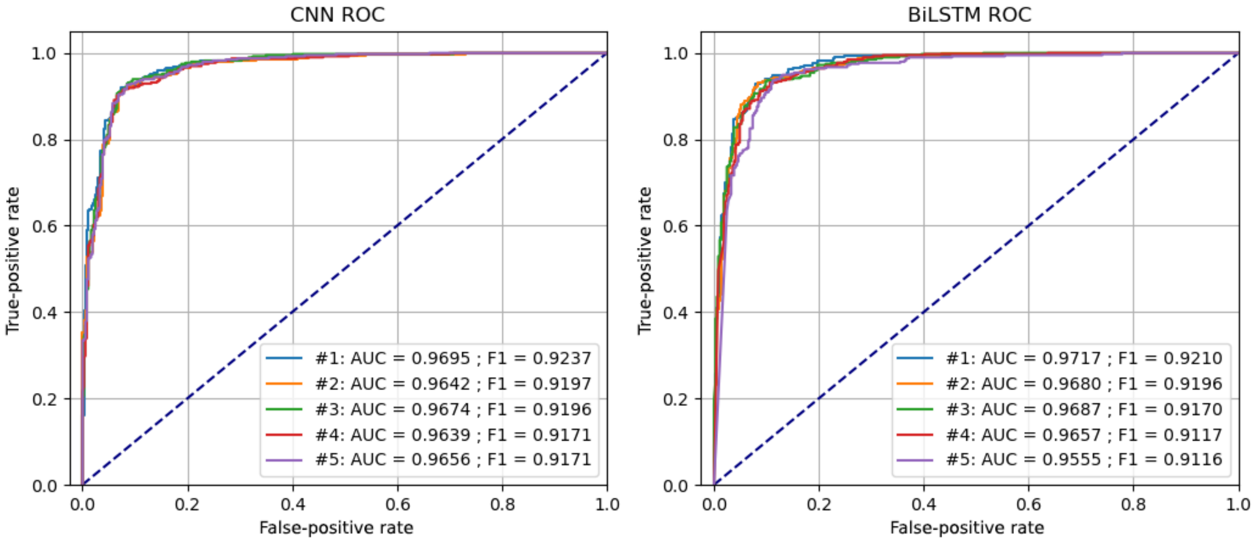


Figure 3. ROC for the best-performing convolutional neural network and bidirectional long short-term memory models. AUC: area under the curve; ROC: receiver operating characteristic.



Causality Mining

To identify the risk perceptions attributed to the HPV vaccines, we first applied our classifier to a set of 291,037 tweets from which we are able to tag 124,031 as false tweets and 167,006 as true tweets. Using a dictionary of causal terms derived by Kayesh et al [72] for Twitter causality detection, we screened for tweets that contained at least 1 of these terms and kept tweets classified as false information if the classification confidence was at least 0.998, as this maintained high fidelity with our classifier. Thus, a total of 9352 tweets were used for the causal relationship mining process (Table 2). We then used a dependency parser for tweets to tag and merge multi-word expressions [73]. As tweets can have multiple utterances (ie, independent sentences or fragments), we kept the noun phrases that appeared with the causal cue regardless of whether they had a dependency related to the causal cue, which is in contrast to the work by Kayesh et al [72]. A candidate causal phrase is a set of terms pertaining to a tweet that contains a causal cue and precipitates the candidate effect phrase.

Table 2. Number of messages after applying several filters.

Model	False	True	Total
No filter, n	124,031	167,006	291,037
+ Confidence threshold, n (%)	72,172 (58.19)	105,166 (62.97)	177,338 (60.93)
+ Contains causal cue, n (%)	3667 (2.96)	5685 (3.40)	9352 (3.21)

We could then compute the pointwise mutual information (PMI) for the causal set $C = \{c_1,...,c_m\}$ and the effect set $E = \{e_1,...,e_m\}$ where the candidate causal phrase, c_i and effect phrase, e_j, are sets that contain terms $w_c \in \boldsymbol{V}_c$ and $w_e \in V_e$, respectively. Here, $\boldsymbol{V}_c$ is the set of terms, noun phrases, and multi-word expressions derived from candidate causal phrases in the tweets (excluding terms with a minimum frequency of 1 and removing stopwords) and V_e is the vocabulary derived from the candidate effect phrases.

To compute the PMI for terms $w_e \in e_j$ and $w_c \in c_i$ we have,

$$\text{pmi}(w_e, w_c) = \log\left(\frac{P(w_e, w_c)}{P(w_c)P(w_e)}\right)$$

We can apply Laplace smoothing to ensure the probability distributions are nonzero [74] and can compute the normalized pointwise mutual information (NPMI) [75] as follows:

$$\text{npmi}(w_e, w_c) = \frac{\text{pmi}(w_e, w_c)}{-\log P(w_e, w_c)} = \frac{\log(P(w_c)P(w_e))}{\log P(w_e, w_c)} - 1$$

The range of values of NPMI are from –1 to 1, where –1 means the terms never occur together, 0 means they are independent, and 1 is complete co-occurrence.

Collapsing Candidate Effect Phrases and Ranking Effects

As our model was completely unsupervised and included retweets, tweet messages could become very redundant, but our method could detect many near-duplicate candidate effect phrases. To collapse these phrases, we clustered the terms using semantic similarity derived from embedding representations of the candidate effect phrases. In particular, we used the package HuggingFace [76] to acquire the sum of the last 4 layers of the bidirectional encoder representations from transformers model [77]. To compute the word embedding, we then averaged these word embeddings in the candidate effect phrase to produce the embedding vectors.

Density-based spatial clustering of applications with noise (DBSCAN) [78] was then used to cluster the candidate effect phrases. There were 2 parameters of importance: (1)

reachability, which is the max distance between the 2 “points” to be considered in the same cluster; and (2) the minimum number of “points” to be considered clusters. By points here we mean an embedded real-valued vector representation of the effect sets. We set the reachability to 0.1 and the minimum number of points to 1, as we wanted to retrieve only the closest semantically similar words while maintaining meaningful clusters. Note that there are other alternatives to DBSCAN, such as ordering points to identify the clustering structure (OPTICS) [79] and hierarchical density-based spatial clustering of applications with noise (HDBSCAN) [80] for spectral clustering; however, we only needed to reduce the number of effects compared at query time, meaning that DBSCAN was sufficient. We then selected the cluster cores as representatives for each cluster to collapse the effects.

To identify the perceptions associated with different causal words, we formulated this as a retrieval problem. Given a causality-related query q, we ranked the associated effects by using the NPMI. To compute the scoring function, we used the following:

$$\mathrm{rank}(e_i, q) = \sum\nolimits_{w_e \in e_i, w \in q} \frac{\mathrm{npmi}(w_e, w)}{|e_i| \cdot |q|}$$

This scoring function computes the average NPMI for all pairs of terms in the query and candidate effect phrase. We can then compute the cumulative NPMI score for a category C_a of effect phrases as follows:

$$\mathrm{cnpmi}(C_a, q) = \sum\nolimits_{e \in C_a} \mathrm{rank}(e_i, q)$$

Causality Mining Results

To validate the candidate causal mining approach, we took a lexicon pertaining to risk perceptions (ie, perceived effects) concerning HPV vaccines. The HPV-Vaccine Risk Lexicon (HPVVR) is a consumer-facing lexicon to capture how laymen describe their risk perceptions about HPV vaccines (including their perceived harms and benefits about HPV vaccines) [81]. The HPVVR was developed in 2 stages. The first stage involved adopting the risk expressions and HPV-vaccine–related consumer-facing vocabulary from the Department of Homeland Security Risk Lexicon, MedlinePlus Consumer Health Topic Vocabulary, and Consumer Health Vocabulary (Unified Medical Language System) [82,83]. The second stage was to extract layman language about the descriptions of risk perceptions based on the user-generated content (from randomly sampled user-generated content from 2013 to 2018, including from Twitter and Facebook) by 2 trained annotators (interrater reliability: Cohen's kappa coefficient (κ)=0.80). The HPVVR covers more than 200 terms or phrases across 29 categories of risk perception-related vocabulary.

This gold standard list of effects, G, was then matched with the effect set E. In particular, we defined a partial match to be present if some terms in the ground truth effect phrase were matched with some of the candidate effect phrase (ie, $g \cap e \neq \varnothing$). For example, we mined “prevent throat cancer” from the data, which is a partial match to “prevent cervical cancer” in the HPVVR. There were 2 other kinds of partial match. We defined a match to be proper if the candidate effect phrase was a more specific example of the ground truth effect phrase, $g \in G$. For example, we mined “early onset menopause” from the data, which is a proper match to “menopause” in the HPVVR. We considered a reverse match to be present if the candidate effect phrase was a more general form of the ground truth effect phrase, $e \subseteq g$. For example, we mined “fatigue” from the data, which is a reverse match to “extreme fatigue” in the HPVVR. Out of a total of 136 ground-truth effect phrases, we found 55 (40.4%) matches, 78 (57.4%) partial matches, 48 (35.3%) reverse matches, and 103 (75.7%) either partial or proper matches or both. Meanwhile, there were also some candidate effect phrases which were newly discovered effects.

As the causality mining method is a completely bottom-up, unsupervised method, we could automatically mine candidate effects for any set of tweets. In particular, for the predicted false tweets, one of the largest candidate effect clusters contained terms relating to the reactions of different entities, such as “Japan,” “Denmark,” and “college,” on the potential issues with the HPV vaccine, such as “recall,” “lose support,” and “banned.” Another such cluster contained terms relating to infertility misconceptions of the HPV vaccine, such as “premature ovarian failure” and “early menopause on young girls.” Another large candidate effect cluster was about the misconceptions of severe adverse events and complications, such as “sudden death,” “paralysis,” and “stroke.” Note that it is possible that some candidate effect phrases may not be directly related to health effects. Thus, to alleviate this limitation in further analysis, we limited the effects to the terms in the ground truth (ie, $V_e = G$).

Discussion

Principal Results

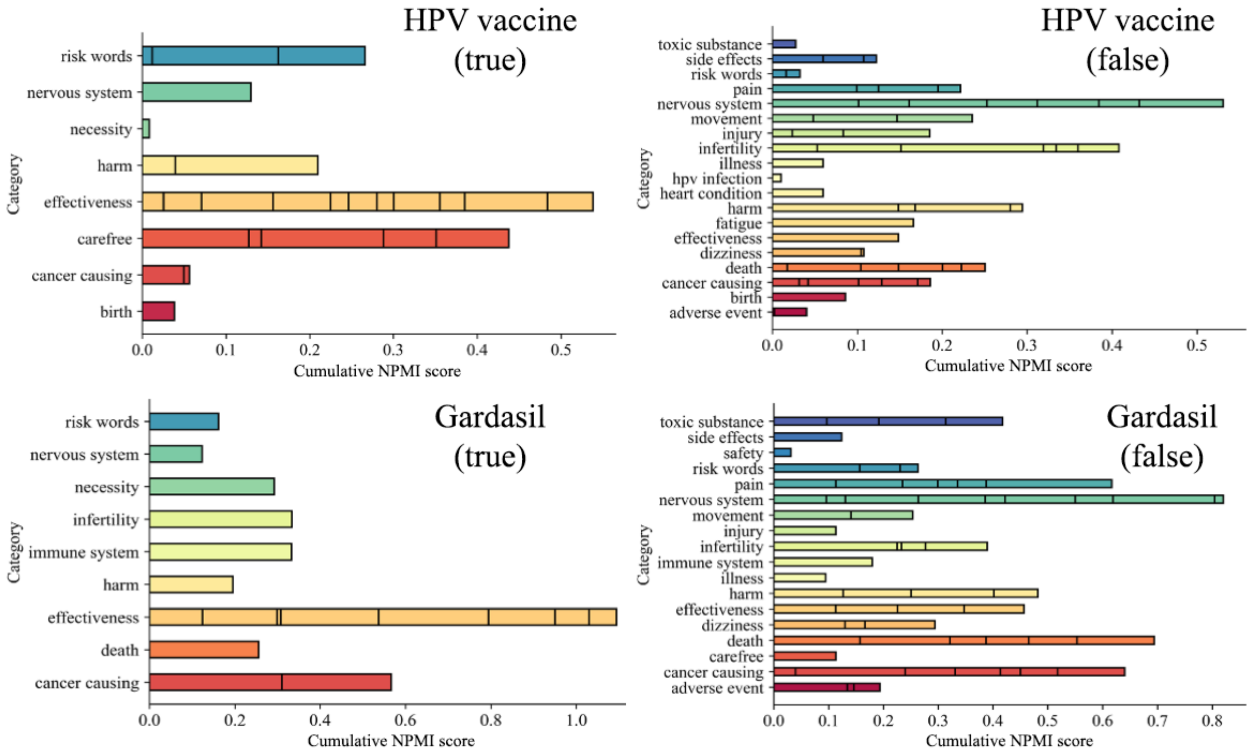
The performance of the CNN and BiLSTM models used in this study showed the feasibility of discerning misinformation from factual information regarding HPV vaccines using the text of tweets. On average, both models predicted either class with high confidence. Although both models performed almost identically in terms of accuracy (and confidence) during testing, the CNN trained much more expediently than did the BiLSTM model, leading to its choice as the preferred model.

To examine the risk perceptions pertaining to HPV vaccines, we leveraged the false information classifier and the effect ranker. Figure 4 shows the cumulative NPMI scores after our effect ranker querying for both “HPV Vaccine” and “Gardasil” was applied. We could categorize these perceptions around the costs and benefits of HPV vaccines. In general, people discussed the benefits or low risk of harms in the true HPV vaccine tweets and various adverse events in the false HPV vaccine tweets. The main effects associated with the HPV vaccines in the true HPV vaccine tweets were about the prevention of HPV infection–related cancer and the denial of risk of increased unprotected sexual behavior of the vaccinated teens. The main effects associated with the HPV vaccines in the false HPV vaccine tweets were regarding infertility-related conditions (such as ovarian injury), child developmental disorder, death, and toxic ingredients in the HPV vaccines. Following our previous work on the patient-driven HPVVR, the findings from causality mining aided us in identifying the major concerns related to HPV vaccines, whose solutions could then be prioritized.

The results show that false HPV vaccine messages not only span a wide variety of topics in risk perceptions but also involve a more diverse vocabulary to describe these topics compared to the fewer topics and relatively limited terminology found in true messages. This phenomenon of medical-based term frequency and topic diversity within false or misleading messages has also been noted in similar work regarding anti- and provaccine literature [84]. A possible explanation for this discrepancy is that true information requires an evidentiary consensus, thereby restricting terminology and outcomes to a specific selection of topics or phrases used to describe these topics. Misinformation lacks such restrictions to terms or outcomes and tends to use narrative language or mention novel topics to gauge attention [46,85].

We also observed differences in message framing in true and false HPV vaccine messages through causality mining (see Figure 4). True information contained both gain-framed and loss-framed messages, especially those highlighting the effectiveness of the vaccine at preventing HPV-related cancers, the link between HPV infection and cancer, and negating the potential harms of vaccines such as carefree or unprotected sexual behavior (Figure 4). Conversely, false messages were largely loss-framed, focusing on negative outcomes purportedly caused by the vaccine, such as those causing HPV-related cancer or other serious adverse events (infertility, neurological disorder, or death; Figure 4). The use of the risk-indicating causative verb (eg, vaccines “prevent” versus vaccines “harm” or “cause”, etc) might be diagnostic for differentiating the true and false information. Future studies should leverage previous findings on the effectiveness of message framing to examine the impact of misinformation with different framing [86,87].

Figure 4. The cumulative NMPI scores when querying for “HPV Vaccine” and “Gardasil”. The sections in the bar width correspond to the NPMI contribution of effect terms for each category. HPV: human papillomavirus; NPMI: normalized pointwise mutual information.



Comparison With Prior Work

Health-related misinformation research spans a broad range of disciplines [58,59], with several studies focusing on different medical domains, such as cancer, sexually transmitted disease and infections, influenza, and more recently, COVID-19 [25,60,61]. In vaccine-related domains, several papers have examined vaccine behavior as well as geographic and demographic patterns on the dissemination of antivaccine and misinformation tweets in social media with respect to autism spectrum disorder [60], influenza (flu) vaccines [88], and cancer treatments [89]. Several research endeavors tackle key issues, such as mitigating label scarcity with additional weak social supervision signals, improving intractability with attention mechanisms, and leveraging network and group or user information [65,90-92]. In general, the distinction between vaccine hesitancy identification and vaccination behavior detection is that the former involves an attitude or stance, while the latter is concerned with detecting the action of getting vaccinated [93]. Our study is more similar to the research in vaccine hesitancy but differs in that we focused on extracting causality from tweets through examining risk perceptions; attributable causes of HPV vaccine–related health concerns or expected gain; and using natural language processing, machine learning, and unsupervised causal mining techniques.

We observed that the convolutional models with multiple filter sizes [70,94] worked better than did BiLSTM models for domains with short text, such as tweets. Intuitively, the CNN architecture captures the most common n-grams (of lengths 3, 4, and 5) and therefore is better at discovering discriminative text patterns in short text. Although we tested more sophisticated BiLSTM architectures, overall, the CNN model performed better than did the other model variations and was faster to train. These findings can be useful for social media health–related analysis, in particular with regards to the set of models that practitioners in this domain should explore for social media text classification.

With respect to causality mining, early works use hand-coded, domain-specific knowledge bases [95,96]. A challenge in identifying causal relations is the variety of ways in which it can be observed via various linguistic constructions. A previous study [97] showed that a classifier can determine whether a causal-verb expression, automatically extracted from predefined linguistic patterns of the form <noun phrase–verb–noun phrase> is a causal relationship or not. However, supervised methods require large amounts of manually annotated causes and effects and are thus resource demanding. Recent work has compared unsupervised methods for causal relationship mining including co-occurrence methods, such as point-wise mutual information, and discourse cue-based methods, which are based on information retrieval techniques, to count the number of matches in a cause-effect query [98]. Such comparisons were performed on large-scale document collections, and thus their insights are not applicable to our tasks that, in contrast, have limited amounts of data. Finally, event causality detection in tweets restricts the causal relationship mining to certain events of interest. In "Event causality detection in tweets by context word extension and neural net" [72], the authors propose an approach to encode both the candidate causal phrase and the candidate effect phrase for developing a feed-forward network classifier. Our method is not restricted to certain events. Most importantly, we focus on health-related messages pertaining to HPV vaccines, an approach which can be generalized to other health topics.

Limitations

One common bottleneck when applying supervised learning methods is the requirement of large amounts of high-quality annotated data for training. Due to the complex nature of the task-at-hand, and the need for extensive manual effort, our data set size might be restrictive in providing insights that can generalize across other domains and data sources. Additionally, due to frequent linguistic variations found in informal user-generated language, closely worded instances might have evaded deduplication. In the future, we hope to address the shortage of available labeled data by incorporating weak supervision methods and denoising mechanisms. Nevertheless, we chose to continue with supervised learning for higher precision, as weak supervision may result in label noise being injected into the false information detection models and thus affecting the subsequent causality mining steps.

Another limitation stems from the misalignment of model confidence and accuracy. In other words, model confidence might not be indicative of model correctness, a problem that is well-known in the machine learning research community [99]. In our experiments, we observed that the BiLSTM model produced high confidence estimates for most false negatives (ie, it misplaced more confidence when predicting factual text), while the CNN model had an equal number of false positives and false negatives for high-confidence examples. Approximately 20% of CNN's incorrect predictions had low confidence. Overall, the BiLSTM model seems to be overconfident in one direction and could be potentially calibrated better. Further analyses on these high-confidence inaccurate predictions are required to discover interpretable patterns that can identify misinformation subtopics and statements that share strong similarities to factual counterparts.

Finally, we should note that any use of additional metadata requires caution, especially for information that is added by the user, such as user profile characteristics, as well as reported timestamps and social network links, as recent studies show that misinformation spreaders tend to manipulate not only social network structure by forming groups to increase influence [100] but also several types of metadata [101]. In this study, we did not use these types of additional data sources, and thus we can only interpret content-based results and not along any other dimension other than the relationships found in the text.

Conclusions

The study has demonstrated a systematic, automatic approach to developing computational models for identifying false HPV vaccine–related information and its associated effects on social media. This approach could be generalized to other social media health information and provide insights into estimating the potential effects of a given health topic.

Acknowledgments

We thank Chieh-Li (Julian) Chin and Wen-Yuh Su for their assistance in the data collection and preparation for this study. We also thank the Technology Services and the Cancer Center at Illinois at the University of Illinois at Urbana-Champaign for their support of our study. This study was supported by a grant from the National Institutes of Health (#1R21CA237483). Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the National Institutes of Health.

Conflicts of Interest

None declared.

References

1. Siegel R, Miller K, Jemal A. Cancer statistics, 2019. CA A Cancer J Clin 2019 Jan 08;69(1):7-34 [FREE Full text] [doi: 10.3322/caac.21551]

2. Schiffman MH, Bauer HM, Hoover RN, Glass AG, Cadell DM, Rush BB, et al. Epidemiologic evidence showing that human papillomavirus infection causes most cervical intraepithelial neoplasia. J Natl Cancer Inst 1993 Jun 16;85(12):958-964. [doi: 10.1093/jnci/85.12.958] [Medline: 8388478]
3. Bosch FX, Manos MM, Muñoz N, Sherman M, Jansen AM, Peto J, et al. Prevalence of human papillomavirus in cervical cancer: a worldwide perspective. International biological study on cervical cancer (IBSCC) Study Group. J Natl Cancer Inst 1995 Jun 07;87(11):796-802. [Medline: 7791229]
4. Dunne EF, Unger ER, Sternberg M, McQuillan G, Swan DC, Patel SS, et al. Prevalence of HPV infection among females in the United States. JAMA 2007 Feb 28;297(8):813-819. [doi: 10.1001/jama.297.8.813] [Medline: 17327523]
5. Myers ER, McCrory DC, Nanda K, Bastian L, Matchar DB. Mathematical model for the natural history of human papillomavirus infection and cervical carcinogenesis. Am J Epidemiol 2000 Jun 15;151(12):1158-1171. [doi: 10.1093/oxfordjournals.aje.a010166] [Medline: 10905528]
6. Weinstock H, Berman S, Cates W. Sexually transmitted diseases among American youth: incidence and prevalence estimates, 2000. Perspect Sex Reprod Health 2004;36(1):6-10. [doi: 10.1363/psrh.36.6.04] [Medline: 14982671]
7. Mork J, Lie AK, Glattre E, Hallmans G, Jellum E, Koskela P, et al. Human papillomavirus infection as a risk factor for squamous-cell carcinoma of the head and neck. N Engl J Med 2001 Apr 12;344(15):1125-1131. [doi: 10.1056/NEJM200104123441503] [Medline: 11297703]
8. Watson M, Saraiya M, Ahmed F, Cardinez CJ, Reichman ME, Weir HK, et al. Using population-based cancer registry data to assess the burden of human papillomavirus-associated cancers in the United States: overview of methods. Cancer 2008 Nov 15;113(10 Suppl):2841-2854 [FREE Full text] [doi: 10.1002/cncr.23758] [Medline: 18980203]
9. Viens LJ, Henley SJ, Watson M, Markowitz LE, Thomas CC, Thompson TD, et al. Human Papillomavirus-Associated Cancers - United States, 2008-2012. MMWR Morb Mortal Wkly Rep 2016 Jul 08;65(26):661-666 [FREE Full text] [doi: 10.15585/mmwr.mm6526a1] [Medline: 27387669]
10. HPV vaccine schedule and dosing. Center for Disease Control and Prevention. URL: https://www.cdc.gov/hpv/hcp/schedules-recommendations.html [accessed 2021-06-17]
11. Garland SM, Kjaer SK, Muñoz N, Block SL, Brown DR, DiNubile MJ, et al. Impact and effectiveness of the quadrivalent human papillomavirus vaccine: a systematic review of 10 years of real-world experience. Clin Infect Dis 2016 Aug 15;63(4):519-527 [FREE Full text] [doi: 10.1093/cid/ciw354] [Medline: 27230391]
12. Villa LL, Costa RLR, Petta CA, Andrade RP, Ault KA, Giuliano AR, et al. Prophylactic quadrivalent human papillomavirus (types 6, 11, 16, and 18) L1 virus-like particle vaccine in young women: a randomised double-blind placebo-controlled multicentre phase II efficacy trial. Lancet Oncol 2005 May;6(5):271-278. [doi: 10.1016/S1470-2045(05)70101-7] [Medline: 15863374]
13. FUTURE II Study Group. Quadrivalent vaccine against human papillomavirus to prevent high-grade cervical lesions. N Engl J Med 2007 May 10;356(19):1915-1927. [doi: 10.1056/NEJMoa061741] [Medline: 17494925]
14. Elam-Evans LD, Yankey D, Singleton JA, Sterrett N, Markowitz LE, Williams CL, et al. National, regional, state, and selected local area vaccination coverage among adolescents aged 13-17 years - United States, 2019. MMWR Morb Mortal Wkly Rep 2020 Aug 21;69(33):1109-1116 [FREE Full text] [doi: 10.15585/mmwr.mm6933a1] [Medline: 32817598]
15. CDC recommends only two HPV shots for younger adolescents. Center for Disease Control and Prevention. 2016. URL: http://www.cdc.gov/media/releases/2016/p1020-hpv-shots.html [accessed 2021-06-17]
16. Muhwezi WW, Banura C, Turiho AK, Mirembe F. Parents' knowledge, risk perception and willingness to allow young males to receive human papillomavirus (HPV) vaccines in Uganda. PLoS One 2014;9(9):e106686 [FREE Full text] [doi: 10.1371/journal.pone.0106686] [Medline: 25203053]
17. Brewer NT, Fazekas KI. Predictors of HPV vaccine acceptability: a theory-informed, systematic review. Prev Med 2007;45(2-3):107-114. [doi: 10.1016/j.ypmed.2007.05.013] [Medline: 17628649]
18. Newman PA, Logie CH, Doukas N, Asakura K. HPV vaccine acceptability among men: a systematic review and meta-analysis. Sex Transm Infect 2013 Nov;89(7):568-574 [FREE Full text] [doi: 10.1136/sextrans-2012-050980] [Medline: 23828943]
19. Larson HJ, Wilson R, Hanley S, Parys A, Paterson P. Tracking the global spread of vaccine sentiments: the global response to Japan's suspension of its HPV vaccine recommendation. Hum Vaccin Immunother 2014;10(9):2543-2550 [FREE Full text] [doi: 10.4161/21645515.2014.969618] [Medline: 25483472]
20. Dempsey AF, Zimet GD, Davis RL, Koutsky L. Factors that are associated with parental acceptance of human papillomavirus vaccines: a randomized intervention study of written information about HPV. Pediatrics 2006 May;117(5):1486-1493. [doi: 10.1542/peds.2005-1381] [Medline: 16651301]
21. Jain N, Euler GL, Shefer A, Lu P, Yankey D, Markowitz L. Human papillomavirus (HPV) awareness and vaccination initiation among women in the United States, National Immunization Survey-Adult 2007. Prev Med 2009 May;48(5):426-431. [doi: 10.1016/j.ypmed.2008.11.010] [Medline: 19100762]
22. Pask EB, Rawlins ST. men's intentions to engage in behaviors to protect against human papillomavirus (HPV): testing the risk perception attitude framework. Health Commun 2016;31(2):139-149. [doi: 10.1080/10410236.2014.940670] [Medline: 26098812]

23. Jolley D, Douglas KM. The effects of anti-vaccine conspiracy theories on vaccination intentions. PLoS One 2014;9(2):e89177 [FREE Full text] [doi: 10.1371/journal.pone.0089177] [Medline: 24586574]
24. Nan X, Madden K. HPV vaccine information in the blogosphere: how positive and negative blogs influence vaccine-related risk perceptions, attitudes, and behavioral intentions. Health Commun 2012 Nov;27(8):829-836. [doi: 10.1080/10410236.2012.661348] [Medline: 22452582]
25. Zimet GD, Rosberger Z, Fisher WA, Perez S, Stupiansky NW. Beliefs, behaviors and HPV vaccine: correcting the myths and the misinformation. Prev Med 2013 Nov;57(5):414-418 [FREE Full text] [doi: 10.1016/j.ypmed.2013.05.013] [Medline: 23732252]
26. Brewer NT, Chapman GB, Gibbons FX, Gerrard M, McCaul KD, Weinstein ND. Meta-analysis of the relationship between risk perception and health behavior: the example of vaccination. Health Psychol 2007 Mar;26(2):136-145. [doi: 10.1037/0278-6133.26.2.136] [Medline: 17385964]
27. Mayhew A, Mullins TLK, Ding L, Rosenthal SL, Zimet GD, Morrow C, et al. Risk perceptions and subsequent sexual behaviors after HPV vaccination in adolescents. Pediatrics 2014 Mar;133(3):404-411 [FREE Full text] [doi: 10.1542/peds.2013-2822] [Medline: 24488747]
28. van der Pligt J. Risk perception and self-protective behavior. European Psychologist 1996 Jan;1(1):34-43. [doi: 10.1027/1016-9040.1.1.34]
29. Weinstein ND, Kwitel A, McCaul KD, Magnan RE, Gerrard M, Gibbons FX. Risk perceptions: Assessment and relationship to influenza vaccination. Health Psychology 2007;26(2):146-151. [doi: 10.1037/0278-6133.26.2.146]
30. Betsch C, Renkewitz F, Betsch T, Ulshöfer C. The influence of vaccine-critical websites on perceiving vaccination risks. J Health Psychol 2010 Apr;15(3):446-455. [doi: 10.1177/1359105309353647] [Medline: 20348365]
31. Reagan-Steiner S, Yankey D, Jeyarajah J, Elam-Evans LD, Singleton JA, Curtis CR, et al. National, regional, state, and selected local area vaccination coverage among adolescents aged 13-17 years--United States, 2014. MMWR Morb Mortal Wkly Rep 2015 Jul 31;64(29):784-792 [FREE Full text] [doi: 10.15585/mmwr.mm6429a3] [Medline: 26225476]
32. Stein RA. The golden age of anti-vaccine conspiracies. Germs 2017 Dec;7(4):168-170 [FREE Full text] [doi: 10.18683/germs.2017.1122] [Medline: 29264353]
33. Covolo L, Ceretti E, Passeri C, Boletti M, Gelatti U. What arguments on vaccinations run through YouTube videos in Italy? A content analysis. Hum Vaccin Immunother 2017 Jul 03;13(7):1693-1699 [FREE Full text] [doi: 10.1080/21645515.2017.1306159] [Medline: 28362544]
34. Zhang C, Gotsis M, Jordan-Marsh M. Social media microblogs as an HPV vaccination forum. Hum Vaccin Immunother 2013 Nov;9(11):2483-2489 [FREE Full text] [doi: 10.4161/hv.25599] [Medline: 23842072]
35. Kang GJ, Ewing-Nelson SR, Mackey L, Schlitt JT, Marathe A, Abbas KM, et al. Semantic network analysis of vaccine sentiment in online social media. Vaccine 2017 Jun 22;35(29):3621-3638 [FREE Full text] [doi: 10.1016/j.vaccine.2017.05.052] [Medline: 28554500]
36. Orr D, Baram-Tsabari A, Landsman K. Social media as a platform for health-related public debates and discussions: the Polio vaccine on Facebook. Isr J Health Policy Res 2016;5:34 [FREE Full text] [doi: 10.1186/s13584-016-0093-4] [Medline: 27843544]
37. Salathé M, Khandelwal S. Assessing vaccination sentiments with online social media: implications for infectious disease dynamics and control. PLoS Comput Biol 2011 Oct;7(10):e1002199 [FREE Full text] [doi: 10.1371/journal.pcbi.1002199] [Medline: 22022249]
38. Aquino F, Donzelli G, De Franco E, Privitera G, Lopalco PL, Carducci A. The web and public confidence in MMR vaccination in Italy. Vaccine 2017 Aug 16;35(35 Pt B):4494-4498. [doi: 10.1016/j.vaccine.2017.07.029] [Medline: 28736200]
39. Dunn AG, Leask J, Zhou X, Mandl KD, Coiera E. Associations between exposure to and expression of negative opinions about human papillomavirus vaccines on social media: an observational study. J Med Internet Res 2015 Jun 10;17(6):e144 [FREE Full text] [doi: 10.2196/jmir.4343] [Medline: 26063290]
40. Keim-Malpass J, Mitchell EM, Sun E, Kennedy C. Using twitter to understand public perceptions regarding the #HPV vaccine: opportunities for public health nurses to engage in social marketing. Public Health Nurs 2017 Mar 06;34(4):316-323. [doi: 10.1111/phn.12318]
41. Surian D, Nguyen DQ, Kennedy G, Johnson M, Coiera E, Dunn AG. Characterizing twitter discussions about HPV vaccines using topic modeling and community detection. J Med Internet Res 2016;18(8):e232 [FREE Full text] [doi: 10.2196/jmir.6045] [Medline: 27573910]
42. Massey PM, Leader A, Yom-Tov E, Budenz A, Fisher K, Klassen AC. Applying multiple data collection tools to quantify human papillomavirus vaccine communication on twitter. J Med Internet Res 2016 Dec 05;18(12):e318 [FREE Full text] [doi: 10.2196/jmir.6670] [Medline: 27919863]
43. Bahk CY, Cumming M, Paushter L, Madoff LC, Thomson A, Brownstein JS. Publicly available online tool facilitates real-time monitoring of vaccine conversations and sentiments. Health Aff (Millwood) 2016 Feb;35(2):341-347. [doi: 10.1377/hlthaff.2015.1092] [Medline: 26858390]
44. Du J, Xu J, Song H, Tao C. Leveraging machine learning-based approaches to assess human papillomavirus vaccination sentiment trends with Twitter data. BMC Med Inform Decis Mak 2017 Jul 05;17(Suppl 2):69 [FREE Full text] [doi: 10.1186/s12911-017-0469-6] [Medline: 28699569]

45. Bessi A, Coletto M, Davidescu GA, Scala A, Caldarelli G, Quattrociocchi W. Science vs conspiracy: collective narratives in the age of misinformation. PLoS One 2015 Feb 23;10(2):e0118093 [FREE Full text] [doi: 10.1371/journal.pone.0118093] [Medline: 25706981]
46. Vosoughi S, Roy D, Aral S. The spread of true and false news online. Science 2018 Mar 09;359(6380):1146-1151. [doi: 10.1126/science.aap9559] [Medline: 29590045]
47. Acemoglu D, Ozdaglar A, ParandehGheibi A. Spread of misinformation in social networks. arXiv. 2009. URL: http://arxiv.org/abs/0906.5007 [accessed 2021-06-17]
48. Lewandowsky S, Ecker UK, Cook J. Beyond misinformation: understanding and coping with the “post-truth” era. Journal of Applied Research in Memory and Cognition 2017 Dec;6(4):353-369. [doi: 10.1016/j.jarmac.2017.07.008]
49. Bessi A, Caldarelli G, Del Vicario M, Scala A, Quattrociocchi W. Social determinants of content selection in the age of (mis)information. 2014 Presented at: International Conference on Social Informatics; November, 2014; Barcelona, Spain p. 259-268 URL: http://arxiv.org/abs/1409.2651 [doi: 10.1007/978-3-319-13734-6_18]
50. Allcott H, Gentzkow M. Social media and fake news in the 2016 election. National Bureau of Economic Research. 2017. URL: http://www.nber.org/papers/w23089 [accessed 2021-06-17]
51. Dechêne A, Stahl C, Hansen J, Wänke M. The truth about the truth: a meta-analytic review of the truth effect. Pers Soc Psychol Rev 2010 May;14(2):238-257. [doi: 10.1177/1088868309352251] [Medline: 20023210]
52. Unkelbach C, Rom SC. A referential theory of the repetition-induced truth effect. Cognition 2017 Mar;160:110-126. [doi: 10.1016/j.cognition.2016.12.016]
53. Schwartz M. Repetition and rated truth value of statements. The American Journal of Psychology 1982;95(3):393-407. [doi: 10.2307/1422132]
54. Habel MA, Liddon N, Stryker JE. The HPV vaccine: a content analysis of online news stories. J Womens Health (Larchmt) 2009 Mar;18(3):401-407. [doi: 10.1089/jwh.2008.0920] [Medline: 19281323]
55. Eysenbach G, Powell J, Kuss O, Sa E. Empirical studies assessing the quality of health information for consumers on the world wide web: a systematic review. JAMA 2002;287(20):2691-2700. [Medline: 12020305]
56. Cline RJ, Haynes KM. Consumer health information seeking on the Internet: the state of the art. Health Educ Res 2001 Dec;16(6):671-692 [FREE Full text] [Medline: 11780707]
57. Moorhead SA, Hazlett DE, Harrison L, Carroll JK, Irwin A, Hoving C. A new dimension of health care: systematic review of the uses, benefits, and limitations of social media for health communication. J Med Internet Res 2013 Apr 23;15(4):e85 [FREE Full text] [doi: 10.2196/jmir.1933] [Medline: 23615206]
58. Dhoju S, Main URM, Ashad KM, Hassan N. Differences in health news from reliable and unreliable media. 2019 Presented at: Companion Proceedings of The 2019 World Wide Web Conference; May 13-17 2019; San Francisco p. 981-987. [doi: 10.1145/3308560.3316741]
59. Suarez-Lledo V, Alvarez-Galvez J. Prevalence of health misinformation on social media: systematic review. J Med Internet Res 2021 Jan 20;23(1):e17187 [FREE Full text] [doi: 10.2196/17187] [Medline: 33470931]
60. Tomeny TS, Vargo CJ, El-Toukhy S. Geographic and demographic correlates of autism-related anti-vaccine beliefs on Twitter, 2009-15. Soc Sci Med 2017 Oct;191:168-175 [FREE Full text] [doi: 10.1016/j.socscimed.2017.08.041] [Medline: 28926775]
61. Huang X, Smith M, Paul M, Ryzhkov D, Quinn S, Broniatowski D, et al. Examining patterns of influenza vaccination in social media. 2017 Presented at: AAAI Workshops; Feb 4-5, 2017; San Francisco.
62. Hou R, Pérez-Rosas V, Loeb S, Mihalcea R. Towards automatic detection of misinformation in online medical videos. 2019 Presented at: International Conference on Multimodal Interaction; October 2019; Suzhou, China p. 235-243 URL: http://arxiv.org/abs/1909.01543 [doi: 10.1145/3340555.3353763]
63. Wang Z, Yin Z, Argyris Y. Detecting medical misinformation on social media using multimodal deep learning. IEEE J. Biomed. Health Inform 2021 Jun;25(6):2193-2203 [FREE Full text] [doi: 10.1109/jbhi.2020.3037027]
64. Cheng L, Guo R, Shu K, Liu H. Towards causal understanding of fake news dissemination. arXiv. 2020. URL: https://arxiv.org/abs/2010.10580 [accessed 2021-06-17]
65. Ruchansky N, Seo S, Liu Y. CSI: A hybrid deep model for fake news detection. 2017 Presented at: ACM Conference on Information and Knowledge Management; Nov 6-10, 2017; New York p. 797-806. [doi: 10.1145/3132847.3132877]
66. Shu K, Cui L, Wang S, Lee D, Liu H. dEFEND: Explainable Fake News Detection. 2019 Presented at: 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining; Aug 4-8, 2019; New York p. 395-405. [doi: 10.1145/3292500.3330935]
67. Hunt R. Percent agreement, Pearson's correlation, and kappa as measures of inter-examiner reliability. J Dent Res 1986 Feb;65(2):128-130. [doi: 10.1177/00220345860650020701] [Medline: 3455967]
68. Bojanowski P, Grave E, Joulin A, Mikolov T. Enriching word vectors with subword information. TACL 2017 Dec;5:135-146. [doi: 10.1162/tacl_a_00051]
69. Joulin A, Grave E, Bojanowski P, Mikolov T. Bag of tricks for efficient text classification. 2017 Presented at: 15th Conference of the European Chapter of the Association for Computational Linguistics; Apr 3-7, 2017; Valencia, Spain p. 427-431. [doi: 10.18653/v1/e17-2068]

70. Kim Y. Convolutional neural networks for sentence classification. : Association for Computational Linguistics; 2014 Presented at: The 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP); Oct 24-29, 2014; Doha, Qatar p. 1746-1751. [doi: 10.3115/v1/d14-1181]
71. Liaw R, Liang E, Nishihara R, Moritz P, Gonzalez J, Stoica I. Tune: a research platform for distributed model selection and training. arXiv. 2018. URL: https://arxiv.org/abs/1807.05118 [accessed 2021-06-17]
72. Kayesh H, Islam M, Wang J. Event causality detection in tweets by context word extension and neural networks. 2019 Presented at: 20th International Conference on Parallel and Distributed Computing, Applications and Technologies (PDCAT); Dec 5-7, 2019; Gold Coast, Australia p. 352-357. [doi: 10.1109/pdcat46702.2019.00070]
73. Kong L, Schneider N, Swayamdipta S, Bhatia A, Dyer C, Smith N. A dependency parser for tweets. 2014 Presented at: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP); Oct 25-29, 2014; Doha, Qatar p. 1001-1012. [doi: 10.3115/v1/d14-1108]
74. Zhai C, Lafferty J. A study of smoothing methods for language models applied to ad hoc information retrieval. 2017 Presented at: ACM SIGIR Forum; July 2017; New York p. 268-276. [doi: 10.1145/3130348.3130377]
75. Bouma G. Normalized (pointwise) mutual information in collocation extraction. 2009 Presented at: Proceedings of GSCL; Sep 30, 2009; Potsdam, Germany p. 31-40.
76. Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi A, et al. HuggingFace's transformers: state-of-the-art natural language processing. arXiv. 2019. URL: https://arxiv.org/abs/1910.03771 [accessed 2021-06-17]
77. Devlin J, Chang M, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv. 2018. URL: https://arxiv.org/abs/1810.04805 [accessed 2021-06-17]
78. Schubert E, Sander J, Ester M, Kriegel H, Xu X. DBSCAN revisited, revisited: why and how you should (Still) use DBSCAN. ACM Trans. Database Syst 2017 Aug 24;42(3):1-21. [doi: 10.1145/3068335]
79. Ankerst M, Breunig M, Kriegel H, Sander J. OPTICS: ordering points to identify the clustering structure. SIGMOD Rec 1999 Jun;28(2):49-60. [doi: 10.1145/304182.304187]
80. McInnes L, Healy J, Astels S. hdbscan: Hierarchical density based clustering. JOSS 2017 Mar;2(11):205. [doi: 10.21105/joss.00205]
81. Chin J, Chin C, Caskey R, Liu B, Schwartz A. Building a patient-driven human papillomavirus (HPV) vaccine risk lexicon. 2018 Presented at: 40th Annual Meeting of the Society for Medical Decision Making; Oct 13-17, 2018; Montreal, Canada. [doi: 10.5005/jp/books/10027_10]
82. Smith C, Stavri P. Consumer health vocabulary. In: Lewis D, Eysenbach G, Kukafka R, Stavri P, Jimison H, editors. Consumer Health Informatics: Informing Consumers and Improving Health Care. New York, NY: Springer; 2005:122-128.
83. Zeng QT, Tse T. Exploring and developing consumer health vocabularies. J Am Med Inform Assoc 2006;13(1):24-29 [FREE Full text] [doi: 10.1197/jamia.M1761] [Medline: 16221948]
84. Xu Z. I don’t understand you but I trust you: using computer-aided text analysis to examine medical terminology use and engagement of vaccine online articles. Journal of Communication in Healthcare 2020 May 21;14(1):61-67. [doi: 10.1080/17538068.2020.1755137]
85. Meng J, Peng W, Tan P, Liu W, Cheng Y, Bae A. Diffusion size and structural virality: The effects of message and network features on spreading health information on twitter. Comput Human Behav 2018 Dec;89:111-120 [FREE Full text] [doi: 10.1016/j.chb.2018.07.039] [Medline: 32288177]
86. Gerend MA, Shepherd JE. Using message framing to promote acceptance of the human papillomavirus vaccine. Health Psychol 2007 Nov;26(6):745-752. [doi: 10.1037/0278-6133.26.6.745] [Medline: 18020847]
87. O'Keefe DJ, Nan X. The relative persuasiveness of gain- and loss-framed messages for promoting vaccination: a meta-analytic review. Health Commun 2012;27(8):776-783. [doi: 10.1080/10410236.2011.640974] [Medline: 22292904]
88. Weissenbacher D, Sarker A, Paul M, Gonzalez-Hernandez G. Overview of the third social media mining for health (SMM4H) shared tasks. In: Association for Computational Linguistics. 2018 Presented at: The 2018 EMNLP workshop SMM4H: the 3rd social media mining for health applications workshop & shared task; Oct 2018; Brussels, Belgium. [doi: 10.18653/v1/w18-5904]
89. Ghenai A, Mejova Y. Fake Cures. In: Proc. ACM Hum.-Comput. Interact. 2018 Nov Presented at: Proceedings of the ACM on Human-Computer Interaction; 2018; New York p. 1-20. [doi: 10.1145/3274327]
90. Jin Z, Cao J, Zhang Y, Luo J. News verification by exploiting conflicting social viewpoints in microblogs. 2016 Presented at: Proceedings of the AAAI Conference on Artificial Intelligence; Feb 12-17, 2016; Pheonix, Arizona.
91. Wang Y, Yang W, Ma F, Xu J, Zhong B, Deng Q, et al. Weak supervision for fake news detection via reinforcement learning. In: AAAI. 2020 Apr 03 Presented at: Proceedings of the AAAI Conference on Artificial Intelligence 2020; Feb 7-12; New York p. 516-523. [doi: 10.1609/aaai.v34i01.5389]
92. Lu Y, Li C. GCAN: Graph-aware Co-Attention Networks for Explainable Fake News Detection on Social Media. 2020 Presented at: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics; Jul 6-8, 2020; Online p. 505-514. [doi: 10.18653/v1/2020.acl-main.48]
93. Joshi A, Dai X, Karimi S, Sparks R, Paris C, MacIntyre C. Shot or not: Comparison of NLP approaches for vaccination behaviour detection. 2018 Presented at: Proceedings of the 2018 EMNLP Workshop SMM4H: The 3rd Social Media Mining for Health Applications Workshop & Shared Task; Oct 31, 2018; Brussels, Belgium. [doi: 10.18653/v1/w18-5911]

94. Yin W, Kann K, Yu M, Schütze H. Comparative study of CNN and RNN for natural language processing. arXiv. 2017. URL: https://arxiv.org/abs/1702.01923 [accessed 2021-06-17]
95. Joskowicz L, Ksiezyck T, Grishman R. Deep domain models for discourse analysis. 1989 Presented at: The Annual AI Systems in Government Conference; Mar 27-31, 1989; Washington, DC p. 195. [doi: 10.1109/aisig.1989.47325]
96. Kaplan RM, Berry-Rogghe G. Knowledge-based acquisition of causal relationships in text. Knowledge Acquisition 1991 Sep;3(3):317-337. [doi: 10.1016/1042-8143(91)90009-c]
97. Girju R. Automatic detection of causal relations for question answering. 2003 Presented at: The ACL Workshop on Multilingual Summarization and Question Answering; July 2003; Sapporo, Japan p. 76-83. [doi: 10.3115/1119312.1119322]
98. Hassanzadeh O, Bhattacharjya D, Feblowitz M, Srinivas K, Perrone M, Sohrabi S, et al. Answering binary causal questions through large-scale text mining: an evaluation using cause-effect pairs from human experts. Proceedings of the 28 International Joint Conference on Artificial Intelligence (IJCAI-19) 2019:5003-5009. [doi: 10.24963/ijcai.2019/695]
99. Guo C, Pleiss G, Sun Y, Weinberger K. On calibration of modern neural networks. 2017 Presented at: 34th International Conference on Machine Learning; Aug 6-11, 2017; Sydney, Australia p. 1321-1330.
100. Wu L, Morstatter F, Carley KM, Liu H. Misinformation in social media: Definition, manipulation, and detection. SIGKDD Explor. Newsl 2019 Nov 26;21(2):80-90. [doi: 10.1145/3373464.3373475]
101. Acker A. Data craft: the manipulation of social media metadata. Data & Society Research Institute. 2018. URL: https://datasociety.net/wp-content/uploads/2018/11/DS_Data_Craft_Manipulation_of_Social_Media_Metadata.pdf [accessed 2021-06-16]

Abbreviations

BiLSTM: bidirectional long short-term memory
CNN: convolutional neural network
DBSCAN: density-based spatial clustering of applications with noise
HDBSCAN: hierarchical density-based spatial clustering of applications with noise
HPV: human papillomavirus
HPVVR: The HPV-Vaccine Risk Lexicon
NPMI: normalized pointwise mutual information
OPTICS: ordering points to identify the clustering structure
PMI: pointwise mutual information

Edited by G Eysenbach; submitted 14.05.21; peer-reviewed by R R, D Huang; comments to author 08.06.21; accepted 04.08.21; published 09.09.21

Please cite as:
Tomaszewski T, Morales A, Lourentzou I, Caskey R, Liu B, Schwartz A, Chin J
Identifying False Human Papillomavirus (HPV) Vaccine Information and Corresponding Risk Perceptions From Twitter: Advanced Predictive Models
J Med Internet Res 2021;23(9):e30451
URL: https://www.jmir.org/2021/9/e30451
doi: 10.2196/30451
PMID: