

Original Paper

Natural Language Processing–Powered Real-Time Monitoring Solution for Vaccine Sentiments and Hesitancy on Social Media: System Development and Validation

Liang-Chin Huang¹, PhD; Amanda L Eiden², PhD; Long He¹, MS; Augustine Annan¹, PhD; Siwei Wang¹, MS; Jingqi Wang¹, PhD; Frank J Manion¹, PhD; Xiaoyan Wang¹, PhD; Jingcheng Du¹, PhD; Lixia Yao², PhD

¹Melax Tech, Houston, TX, United States

²Merck & Co, Inc, Rahway, NJ, United States

Corresponding Author:

Amanda L Eiden, PhD

Merck & Co, Inc

2025 E Scott Ave

Rahway, NJ, 07065

United States

Phone: 1 7325944000

Email: amanda.eiden@merck.com

Abstract

Background: Vaccines serve as a crucial public health tool, although vaccine hesitancy continues to pose a significant threat to full vaccine uptake and, consequently, community health. Understanding and tracking vaccine hesitancy is essential for effective public health interventions; however, traditional survey methods present various limitations.

Objective: This study aimed to create a real-time, natural language processing (NLP)–based tool to assess vaccine sentiment and hesitancy across 3 prominent social media platforms.

Methods: We mined and curated discussions in English from Twitter (subsequently rebranded as X), Reddit, and YouTube social media platforms posted between January 1, 2011, and October 31, 2021, concerning human papillomavirus; measles, mumps, and rubella; and unspecified vaccines. We tested multiple NLP algorithms to classify vaccine sentiment into positive, neutral, or negative and to classify vaccine hesitancy using the World Health Organization's (WHO) 3Cs (confidence, complacency, and convenience) hesitancy model, conceptualizing an online dashboard to illustrate and contextualize trends.

Results: We compiled over 86 million discussions. Our top-performing NLP models displayed accuracies ranging from 0.51 to 0.78 for sentiment classification and from 0.69 to 0.91 for hesitancy classification. Explorative analysis on our platform highlighted variations in online activity about vaccine sentiment and hesitancy, suggesting unique patterns for different vaccines.

Conclusions: Our innovative system performs real-time analysis of sentiment and hesitancy on 3 vaccine topics across major social networks, providing crucial trend insights to assist campaigns aimed at enhancing vaccine uptake and public health.

(*JMIR Med Inform* 2024;12:e57164) doi: [10.2196/57164](https://doi.org/10.2196/57164)

KEYWORDS

vaccine sentiment; vaccine hesitancy; natural language processing; NLP; social media; social media platforms; real-time tracking; vaccine; vaccines; sentiment; sentiments; vaccination; vaccinations; hesitancy; attitude; attitudes; opinion; perception; perceptions; perspective; perspectives; machine learning; uptake; willing; willingness; classification

Introduction

Vaccine is an essential public health intervention that has saved millions of lives and achieved a substantial global reduction in cases, hospitalizations, and health care costs associated with vaccine-preventable diseases (VPDs) [1-3]. Yet, despite their value, vaccine hesitancy persists as a barrier to full vaccine

uptake. The World Health Organization (WHO) defines vaccine hesitancy as the delay or refusal of vaccination, even when vaccination services are accessible [4]. Additionally, the WHO identifies vaccine hesitancy as one of the top 10 global health threats [5]. Delay or refusal of vaccines due to vaccine hesitancy can have broad-reaching implications; unvaccinated individuals not only put themselves at risk of VPDs, such as COVID-19,

but also pose a threat to the broader community or even global health [6]. This phenomenon has been documented since the advent of vaccines in over 90% of the countries [7]. Considering the case of measles, mumps, and rubella (MMR), it is crucial to uphold community protection or herd immunity, necessitating widespread vaccination to protect those unable to receive the vaccine [8]. A former London study successfully raised MMR vaccination rates from 80% to 94% in under 2 years through incentivized care packages and innovative technology use, approaching the desired herd immunity target [9].

There is a myriad of reasons for vaccine hesitancy, including personal or familial beliefs, concerns about adverse reactions or efficacy, and skepticism toward government and vaccine manufacturers [6,10-17]. This intricate web of motivations makes vaccine hesitancy a complex public health challenge [18].

Understanding vaccine hesitancy is crucial for developing effective interventions, public health education, and vaccination promotion strategies [19-22]. While surveys have traditionally served as a valuable tool for gathering public opinions on vaccination, they possess inherent limitations such as static data collection, resource intensiveness, and potential time lag [23-29]. To address these limitations, real-time tracking of vaccine hesitancy activities and trends offers public health professionals' valuable insights. This approach helps identify critical intervention points before the vaccination uptake wanes, allowing for more targeted and timely communication efforts.

The emergence of social media platforms has enabled billions of users to engage in discussions, information sharing, and opinion expression on various subjects, including health-related topics [30]. While this presents an unprecedented opportunity for public health improvement, it also poses a significant risk linked to the dissemination of vaccine-related misinformation and disinformation [31]. Previous research has used semiautomatic methods such as manual coding and hashtag or keyword analysis to study social media vaccine discussions [32-35]. Nevertheless, these approaches may sometimes encounter potential challenges with scalability and precision. Natural language processing (NLP) is an automated method designed to effectively and accurately decipher the wealth of information in natural language text, addressing challenges such as ambiguities and probabilistic parsing, and enabling applications such as information extraction and discourse analysis [36]. This technique has emerged as a promising solution, holding the potential to mitigate these challenges and improve the precision of vaccine-related public sentiment analysis [37,38].

To address these challenges, this study's principal aim was to create an NLP system for real-time monitoring of vaccine sentiment and hesitancy across English-language social media platforms targeting the US market. Our 3-fold contributions are (1) developing one of the first real-time monitoring systems for social media vaccine discussions that covers 3 major social media platforms and 3 vaccine topic groups [39]; (2) comprehensively evaluating multiple machine learning-based

NLP models for social media post classification tasks, thus establishing a benchmark for future research; and (3) analyzing decade-long trends of sentiment and hesitancy and linked real-world events to corresponding points on the trends for multiple vaccine targets.

Methods

Overview

We followed a systematic approach to monitor vaccine sentiment and hesitancy posts on Twitter (subsequently rebranded as X), Reddit, and YouTube. We selected Twitter, Reddit, and YouTube as they are the primary social media platforms offering substantial volumes of posts through application programming interface (API) access [40-43]. We focused exclusively on English language posts given the widespread use of English in the largest market countries for our target vaccines and with regard to the accessibility of English language social media. Other platforms and languages, such as Facebook and Spanish [44], may be of interest for future studies; however, these served as a first approach to research. Figure 1 illustrates our workflow, including data annotation, NLP algorithms, and an online dashboard.

First, we categorized vaccine sentiment into positive, negative, and neutral, which were the labels also used in other sentiment analyses using social media data [45,46]. Then, we aligned vaccine hesitancy with the WHO's 3Cs (confidence, complacency, and convenience) vaccine hesitancy model, described in further detail in the *3Cs Vaccine Hesitancy Annotation* section [4]. The definitions of post sentiment and vaccine hesitancy are comprehensively presented in Table 1. We collected data using vaccine-specific search queries (see Table S1 in Multimedia Appendix 1) for relevant posts from the 3 social media platforms. To ensure the quality and reliability of the data, we collaborated with medical experts to create annotated corpora aligned with the information model. These corpora were then used to train NLP algorithms to automatically extract vaccine sentiment and hesitancy content. Finally, we developed an online dashboard to provide real-time insights into vaccine sentiment and hesitancy trends. Our study focuses on evaluating the vaccine sentiment and hesitancy of human papillomavirus (HPV), MMR, and general or unspecified vaccines. The critical role of the vaccines is exemplified by the HPV vaccine, which has effectively reduced prevalent HPV infections and precancerous lesions, underlining the importance of global implementation [47], and the MMR vaccine is renowned for its safety and efficacy, which has greatly mitigated endemic diseases in the United States [48]. Despite these successes, challenges such as insufficient vaccination coverage, increasing hesitancy, and the resurgence of mumps, attributed to waning immunity and antigenic variation, persist worldwide. Throughout the COVID-19 pandemic up to 2022, HPV and MMR were the vaccines that maintained the greatest negative impact on routine vaccinations in the United States, suggesting a need for proactive efforts to increase vaccination coverage to prevent associated health complications and costs [49].

Figure 1. The overview of study design and classifications used to evaluate vaccine-related posts. 3Cs: confidence, complacency, and convenience; ML: machine learning; WHO: World Health Organization.

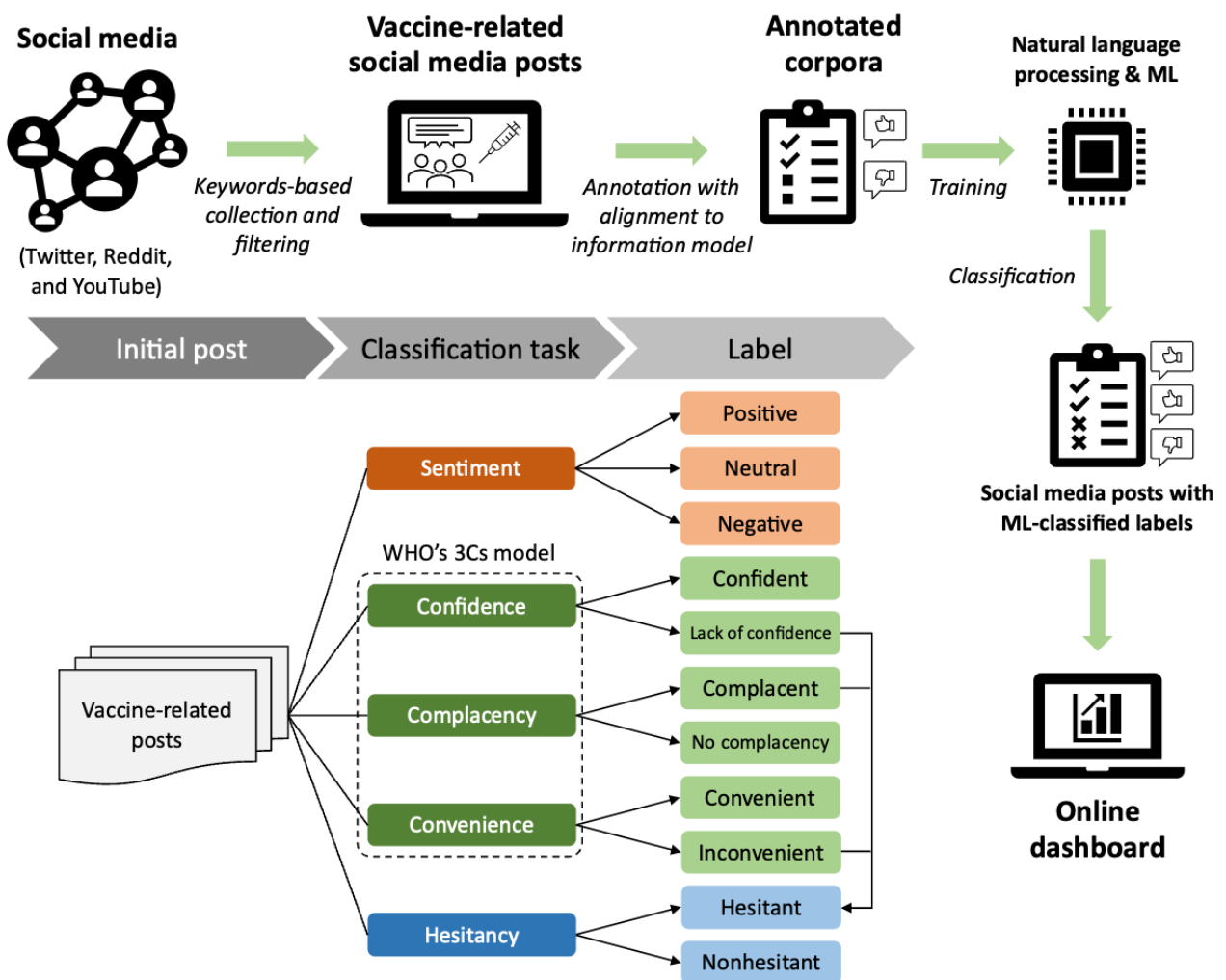


Table 1. Definitions of post sentiment and hesitancy.

Classification task and label	Definition
Sentiment	
Positive	Posts that mention, report, or share positive news, opinions, or stories about vaccines or vaccination.
Neutral	Posts that are related to vaccines or vaccination topics but contain no sentiment, the sentiment is unclear, or they contain both negative and positive sentiments.
Negative	Posts that mention, report, or share negative news, opinions, or stories about vaccines or vaccination, which may discourage vaccination.
Confidence	
Confident	Posts reflecting a trust in the effectiveness and safety of vaccines, the vaccine delivery system, or policy makers' motivations.
Lack of confidence	Posts reflecting a lack of trust in the effectiveness and safety of vaccines, the vaccine delivery system, or policy makers' motivations.
Complacency	
Complacent	Posts where the perceived risks of VPDs ^a are low and vaccination is deemed as an unnecessary preventive action.
No complacency	Posts where the perceived risks of VPDs are high and vaccination is deemed as a necessary preventive action.
Convenience	
Convenient	Posts where physical availability, affordability and willingness to pay, geographical accessibility, ability to understand (language and health literacy), and appeal of immunization services do not affect uptake.
Inconvenient	Posts where physical availability, affordability and willingness to pay, geographical accessibility, ability to understand (language and health literacy), and appeal of immunization services affect uptake.
Hesitancy	
Hesitant	The post is labeled as lack of confidence, complacent, or inconvenient.
Nonhesitant	The post is not labeled as lack of confidence, complacent, or inconvenient.

^aVPD: vaccine-preventable disease.

Social Media Data Collection

The systematic collection of social media data spanned from January 1, 2011, to October 31, 2021, across 3 platforms—Twitter, Reddit, and YouTube. During initial exploratory analysis, we recognized variations in text nature and query logic across these platforms, leading us to tailor our search queries for each platform to collect relevant posts while excluding irrelevant ones. Table S1 in [Multimedia Appendix 1](#) lists the customized queries on each platform for each vaccine topic group, which include both inclusion and exclusion keywords. We retrieved the results (relevant posts) using the APIs provided by the 3 platforms. Details about the software versions are described in the [Multimedia Appendix 1](#). To clarify ethical considerations and data privacy issues, when gathering data from Twitter, YouTube, and Reddit, we adhered to their API's data privacy policies and ensured the deidentification of all posts and videos by assigning them a unique random ID.

Ethical Considerations

Ethics board review was not required, as all modelling data came from public sources and there were no ethical issues. The data privacy policies of the application program interfaces (APIs) of Twitter, YouTube, and Reddit were followed when gathering data. We ensured the deidentification of all posts and videos by assigning them a unique random ID.

Data Annotation

From the retrieved results, approximately 90 million posts, we randomly selected 60,000 social media discussions. These posts were manually annotated to build both training and evaluation data sets, which were used for building the text classifiers. We selected 20,000 posts for annotation, including 10,000 tweets, 5000 Reddit posts, and 5000 YouTube comments for each vaccine topic group, including HPV vaccine, MMR vaccine, and general or unspecified vaccines. During annotator training, 4 annotators with a medical training background were recruited for the annotation. An annotation guideline was developed. All annotators first annotated the same 1000 tweets, 1000 Reddit posts, and 1000 YouTube posts independently, and then discussed collectively for any discrepancies. After all discrepancies were resolved through discussions, these annotators began to annotate the rest of the social media posts. A 2-fold annotation strategy was used, where first, we annotated the sentiment of the post as positive, neutral, or negative, assigning only 1 category to each post; and second, we annotated vaccine hesitancy based on the constructs of the WHO 3Cs model, which include confidence, complacency, and convenience ([Figure 1](#)). These annotation categories also define each classification task.

Sentiment Annotation

The annotation task involved assigning 1 of 3 sentiment labels

to each post, which constituted a multiple-class classification problem. The labels and corresponding illustrative examples are defined in [Textbox 1](#).

Textbox 1. Definitions and examples of sentiment labels.

- Positive: posts that mention, report, or share positive news, opinions, or stories about vaccines or vaccination.
 - Example: “HPV vaccine, prevents against the two HPV types, 16 and 18, which cause 70% of cervical cancers”
 - Example: “Get vaccinated against HPV to protect you in the future for now!”
- Neutral: posts that are related to vaccines or vaccination topics but contain no sentiment, the sentiment is unclear, or they contain both negative and positive sentiments.
 - Example: “The following report is specifically for the MMR vaccine, but you can browse around for others”
 - Example: “I just learned that there are more than 50 strains of HPV...I always thought the vaccine prevented all strains.”
- Negative: posts that mention, report, or share negative news, opinions, or stories about vaccines or vaccination, which may discourage vaccination.
 - Example: “According to a report, thousands of kids suffer permanent injury or death by getting vaccines”
 - Example: “Believe it? Vaccines have killed 1000 more kids than any measles!”

3Cs Vaccine Hesitancy Annotation

The annotation task involved assigning multiple labels to each post according to the 3Cs model constructs. Annotators checked each construct to determine whether the post was related to it separately. If any of the constructs were labeled as “lack of confidence,” “complacent,” or “inconvenient,” we considered the post as vaccine hesitant; otherwise, it was considered vaccine

nonhesitant. Definitions and examples for each 3Cs model construct are provided in [Textbox 2](#).

Table S2 in [Multimedia Appendix 1](#) provides examples of specific social media posts with annotations for the different categories. The distribution of annotated posts in each sentiment and 3Cs construct for each platform and vaccine topic group is shown in Table S3 in [Multimedia Appendix 1](#).

Textbox 2. Definitions and examples of World Health Organization’s 3Cs (confidence, complacency, and convenience) model.

- Lack of confidence: posts reflecting a lack of trust in the effectiveness and safety of vaccines, the vaccine delivery system, or policy makers’ motivations.
 - Example: “Fully vaccinated are 30 times more likely to get COVID-19, and 10 times more likely to require hospitalization.”
 - Example: “The vaccine label includes all these events. Concerns have been raised about reports of deaths occurring in individuals after receiving that vaccine.”
- Complacency: posts where the perceived risks of vaccine-preventable diseases are low, and vaccination is deemed as an unnecessary preventive action.
 - Example: “Why do adults need to know about the measles vaccine? The measles is a benign disease and there is no need for vaccines.”
 - Example: “I wasn’t vaccinated against a preventable disease. It’s not always just a life-or-death dichotomy - I recovered.”
- Inconvenience or convenience: posts where physical availability, affordability and willingness to pay, geographical accessibility, ability to understand (language and health literacy), and appeal of immunization services affect uptake.
 - Example: “I am 30-year-old man and am looking for an HPV vaccine. Unfortunately, my insurance only covers it for women. I am particularly at risk for certain cancers. I really don’t understand how insurance companies are allowed to make the gender distinction when the FDA approved it for both.”

Text Classification Algorithms

Overview

To classify the sentiment and hesitancy of social media posts, we compared the performance of 5 text classification algorithms—logistic regression (LR) [50], support vector machine (SVM) [51], random forest [52], extreme gradient boosting (XGBoost) [53], and Snorkel [54]. Each of these models has unique characteristics, which are summarized below.

LR Algorithm

LR is a classic statistical methodology that models a binary dependent variable using a logistic function. It is favored in medical research due to its ability to determine the odds ratio, indicating the potential change in outcome probabilities [55].

SVM Algorithm

SVM is one of the most robust classification methods based on statistical learning frameworks. It finds a hyperplane in an N-dimensional space that distinctly classifies data points. In

medical text mining, SVM combined with other algorithms has demonstrated effective performance in extracting and recognizing entities in clinical text, contributing notably to improved patient care [56].

Random Forest Algorithm

Random forest is a classifier that uses ensemble learning to combine decision tree classifiers through bagging or bootstrap aggregating. It has been applied to highly ranked features obtained through suitable ranker algorithms and has shown promising results in medical data classification tasks, enhancing the prediction accuracy for various diseases [57].

XGBoost Algorithm

XGBoost is an ensemble of algorithms that turn weak learners into strong learners by focusing on where the individual models went wrong. In gradient boosting, individual weak models train upon the difference between the classification and the actual results. It has been effective in mining and classifying suggestive sentences from online customer reviews by combining them with a word-embedding approach [58].

Snorkel Algorithm

Snorkel is a system that enables users to train models without hand labeling all training data by writing their labeling functions. Using Snorkel enables the extraction of chemical reaction relationships from biomedical literature abstracts, supporting the understanding of biological processes without requiring a large, labeled training data set [59].

We extracted the term frequency–inverse document frequency vector for each word in all text classification algorithms using *scikit-learn*'s *TfidfTransformer* function with default parameter settings. Term frequency–inverse document frequency evaluates how relevant a word is to a text in a collection of texts [60]. If the model encounters a new post with words or symbols not included in its original bag of words, it will effectively ignore those words during the transformation process. To ensure a balanced training set, the 3 class-balancing methods implemented by Python *imblearn* package applied were (1) random oversampling, (2) synthetic minority over-sampling technique (SMOTE) [61], and (3) SVM-based SMOTE [62] (with the default parameter settings, specifically $k_neighbors=5$, as they exhibited the optimal performance within the developer's data set [61]). SMOTE randomly selects a minority class instance, finds one of its nearest minority class neighbors, and then synthesizes an instance between these 2 instances in the feature space. SVM-based SMOTE uses support vectors to determine the decision boundaries and then synthesizes a minority class instance along the decision boundary.

NLP Evaluation

The evaluation data sets were created from the annotated corpora and randomly divided into training, validation, and test sets in a 6:2:2 ratio to assess the performance of the 5 text classification algorithms. The models were trained on the training sets,

optimized on the validation sets, and then evaluated on the test sets. The following key metrics were calculated to evaluate the models:

$$\text{Precision} = \frac{\text{True positive}}{\text{True positive} + \text{False positive}}$$

$$\text{Recall} = \frac{\text{True positive}}{\text{True positive} + \text{False negative}}$$

$$\text{F1 score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

$$\text{Accuracy} = \frac{\text{True positive} + \text{True negative}}{\text{True positive} + \text{True negative} + \text{False positive} + \text{False negative}}$$

A true positive occurs when the model accurately classifies the positive class (positive, negative, or neutral for sentiment; true for 3Cs model constructs). A true negative occurs when the model accurately classifies the negative class (nonpositive, nonnegative, or nonneutral for sentiment; false for 3Cs model constructs). A false positive is an incorrect positive classification, while a false negative is an incorrect negative classification. As the sentiment and hesitancy labels in Tweets, Reddit posts, and YouTube comments are imbalanced, we optimized our models based on F_1 -scores, which balance precision and recall, rather than accuracy. The purpose of optimizing a model based on F_1 -scores when dealing with imbalanced labels is to achieve a better balance between precision and recall, thereby improving the overall performance of the model. This is especially important in imbalanced data sets where the cost of misclassification can be high.

Dashboard Development

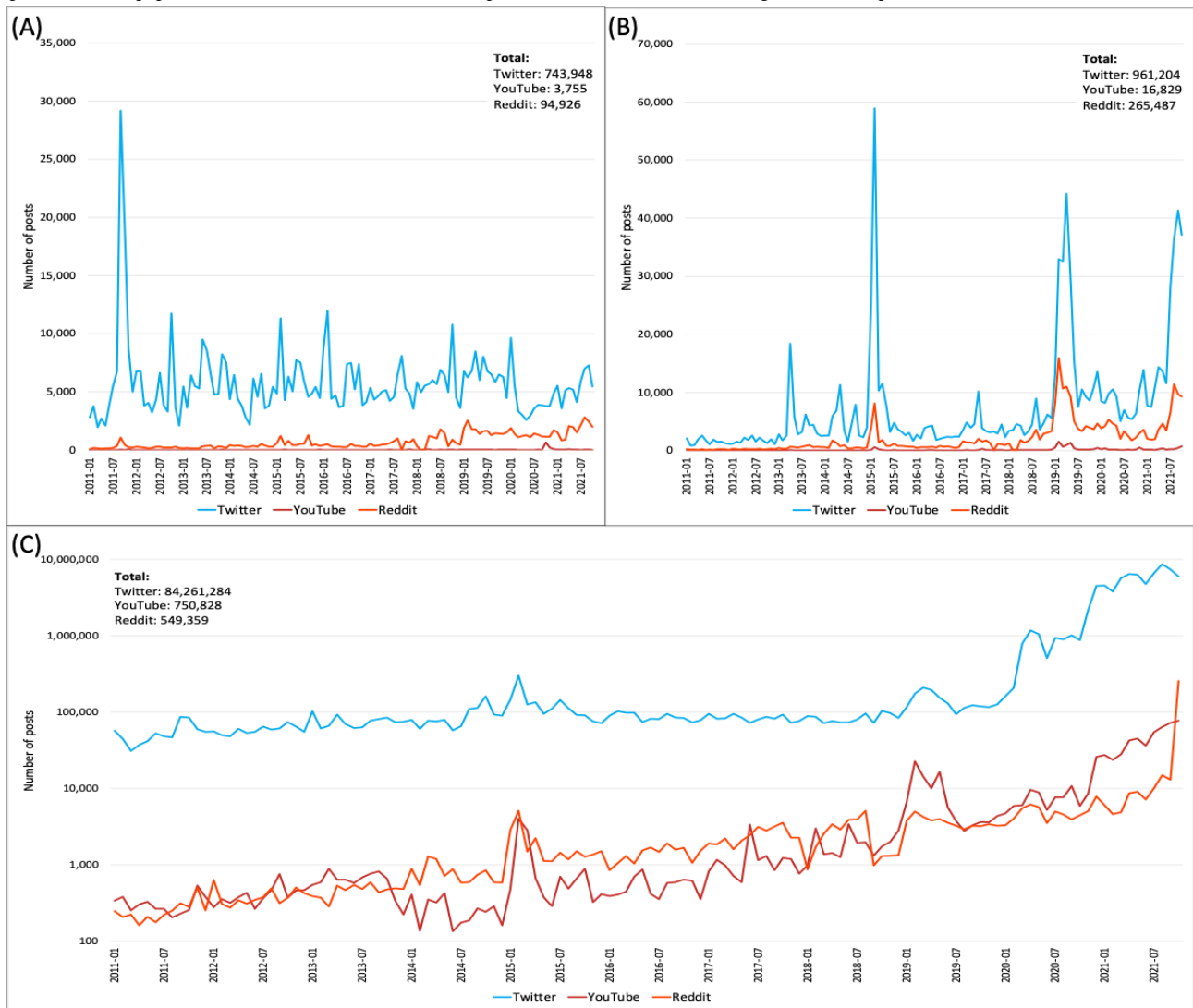
We designed a user-friendly, web-based visualization dashboard [39] for real-time analysis of trends in vaccine sentiment and hesitancy over time and geography (Figure S1A-C in [Multimedia Appendix 1](#)). The dashboard also allows for comparisons of sentiment and hesitancy across different social media platforms and vaccine topic groups (Figure S1D in [Multimedia Appendix 1](#)). The NLP models were optimized based on their F_1 -scores to address the imbalanced labels of sentiment and hesitancy in tweets, Reddit posts, and YouTube comments. The selected models are applied to all unlabeled data collected from 2011 to 2021. Technical details are described and represented in Figure S2 in [Multimedia Appendix 1](#).

Results

Social Media Data Collection Summary

From January 1, 2011, to October 31, 2021, we collected 86 million posts from Twitter, 0.9 million from Reddit, and 76,000 from YouTube, which were related to vaccines. The most widely discussed topic across all 3 platforms was the general or unspecified vaccine, followed by the MMR and then HPV vaccines. We observed a substantial increase in the general vaccine-related discussions on Twitter and Reddit starting in early 2020, coinciding with the onset of the COVID-19 pandemic. The collected social media data and growth trends are plotted in [Figure 2](#).

Figure 2. The long and short-term trends of collected vaccine-related social media post data across 3 different platforms for different vaccine topic groups: (A) human papillomavirus (HPV); (B) measles, mumps, and rubella (MMR); and (C) general or unspecified vaccine.



NLP Performance on Vaccine Sentiment and Hesitancy

We tested all combinations of the 5 NLP algorithms. The performances in sentiment classification, hesitancy classification, and 3Cs classifications are presented in Table 2. The best-performing algorithms (according to F_1 -scores) and detailed performance scores for different classification tasks are shown in Tables S4-S8 in Multimedia Appendix 1. In sentiment classification, LR outperformed other algorithms in 7 out of 9 platform–vaccine topic group combinations, with overall accuracies ranging from 0.51 to 0.78 (Table S4 in Multimedia Appendix 1). The macroaveraged F_1 -scores of negative, neutral, and positive sentiment classifications across different platforms and vaccine topic groups were 0.43, 0.67, and 0.53, respectively. In hesitancy classification, LR outperformed other algorithms in 6 platform–vaccine topic group combinations, with overall accuracies ranging from 0.69 to 0.91 (Table S5 in Multimedia Appendix 1). The macroaveraged F_1 -scores of nonhesitancy and hesitancy classifications were 0.86 and 0.40, respectively. Notably, Reddit users had fewer negative sentiment posts, resulting in lower performance in classifying negative sentiment. In addition, as

Reddit had fewer hesitancy posts, classifying hesitancy was more challenging than on Twitter and YouTube.

Our evaluation of various algorithms and class-balancing methods for each platform and vaccine topic group revealed that Snorkel performed best in 3 platform–vaccine topic group combinations in vaccine hesitancy classifications, with overall accuracies ranging from 0.69 to 0.98 (Table S6 in Multimedia Appendix 1). The macroaveraged F_1 -scores for lack of confidence and nonlack of confidence classifications were 0.88 and 0.45, respectively. Similarly, for complacency classifications, Snorkel outperformed other algorithms in 4 platform–vaccine topic group combinations, with overall accuracies ranging from 0.64 to 0.99 (Table S7 in Multimedia Appendix 1). The macroaveraged F_1 -scores for noncomplacency and complacency classifications were 0.89 and 0.49, respectively. Inconvenience classifications were significantly improved with Snorkel in 8 platform–vaccine topic group combinations, with overall accuracies ranging from 0.89 to 0.99 (Table S8 in Multimedia Appendix 1). However, the results are biased as there were limited posts with convenience information on all 3 social media platforms, which may impact generalizability. The macroaveraged F_1 -scores for

noninconvenience and inconvenience classifications were 0.98 and 0.38, respectively. Our findings demonstrate that advanced text classification algorithms such as XGBoost and Snorkel outperformed other algorithms in highly class-imbalanced situations, even when different class-balancing methods were applied.

We have created a web-based dashboard building upon those best-performing NLP algorithms to extract vaccine sentiment and hesitancy from social media posts. The dashboard summarizes posts from the 3 social media platforms and allows users to analyze temporal trends and geographic clustering easily. It offers different views, including 3 social media platform-centric views and a comparison view that enables

users to compare selected vaccine topic groups and sentiment or hesitancy (Figure S1 in [Multimedia Appendix 1](#)).

When analyzing the sentiment of HPV vaccine posts across 3 social media platforms from January 2011 to October 2021 (Figure 3A), we observed that the ratio of positive sentiment was generally higher than that of neutral and negative sentiment. We also compared vaccine sentiment across 3 social media platforms for MMR vaccines from January 2011 to October 2021 (Figure 3B). Overall, posts expressed positive sentiment toward MMR, with most being neutral. Taking the hesitancy of MMR vaccine as an example, the overall trend shows that the social media posts across 3 social media platforms have a higher ratio of nonhesitancy than hesitancy (Figure 3C).

Table 2. NLP^a performance (measured by F1-scores and accuracy) on vaccine sentiment and hesitancy.

Performance	Twitter			Reddit			YouTube		
	HPV ^b	MMR ^c	General ^d	HPV	MMR	General	HPV	MMR	General
Sentiment									
Positive F_1 -score	0.87	0.57	0.47	0.67	0.50	0.35	0.58	0.53	0.19
Neutral F_1 -score	0.71	0.67	0.83	0.67	0.65	0.86	0.51	0.59	0.51
Negative F_1 -score	0.41	0.53	0.43	0.32	0.26	0.21	0.60	0.49	0.59
Accuracy	0.78	0.61	0.73	0.63	0.55	0.75	0.56	0.55	0.51
Confidence									
Confident F_1 -score	0.35	0.31	0.52	0.35	0.62	0.56	0.44	0.29	0.63
Lack of confidence F_1 -score	0.88	0.95	0.79	0.86	0.74	0.84	0.89	0.99	0.98
Accuracy	0.80	0.90	0.71	0.77	0.69	0.77	0.82	0.98	0.95
Complacency									
Complacent F_1 -score	0.47	0.36	0.41	0.43	0.68	0.60	0.33	0.50	0.60
No complacency F_1 -score	0.94	0.91	0.81	0.93	0.59	0.96	0.91	1.00	0.97
Accuracy	0.89	0.84	0.71	0.88	0.64	0.93	0.84	0.99	0.95
Convenience									
Convenient F_1 -score	0.96	0.99	0.99	0.94	0.95	0.98	0.98	1.00	0.99
Inconvenient F_1 -score	0.48	0.18	0.55	0.67	0.17	0.50	0.17	0.50	0.20
Accuracy	0.92	0.98	0.98	0.89	0.91	0.97	0.96	0.99	0.98
Hesitancy									
Hesitant F_1 -score	0.40	0.44	0.38	0.19	0.23	0.20	0.58	0.53	0.61
Nonhesitant F_1 -score	0.94	0.90	0.89	0.87	0.81	0.95	0.81	0.83	0.76
Accuracy	0.90	0.83	0.82	0.78	0.69	0.91	0.73	0.75	0.70

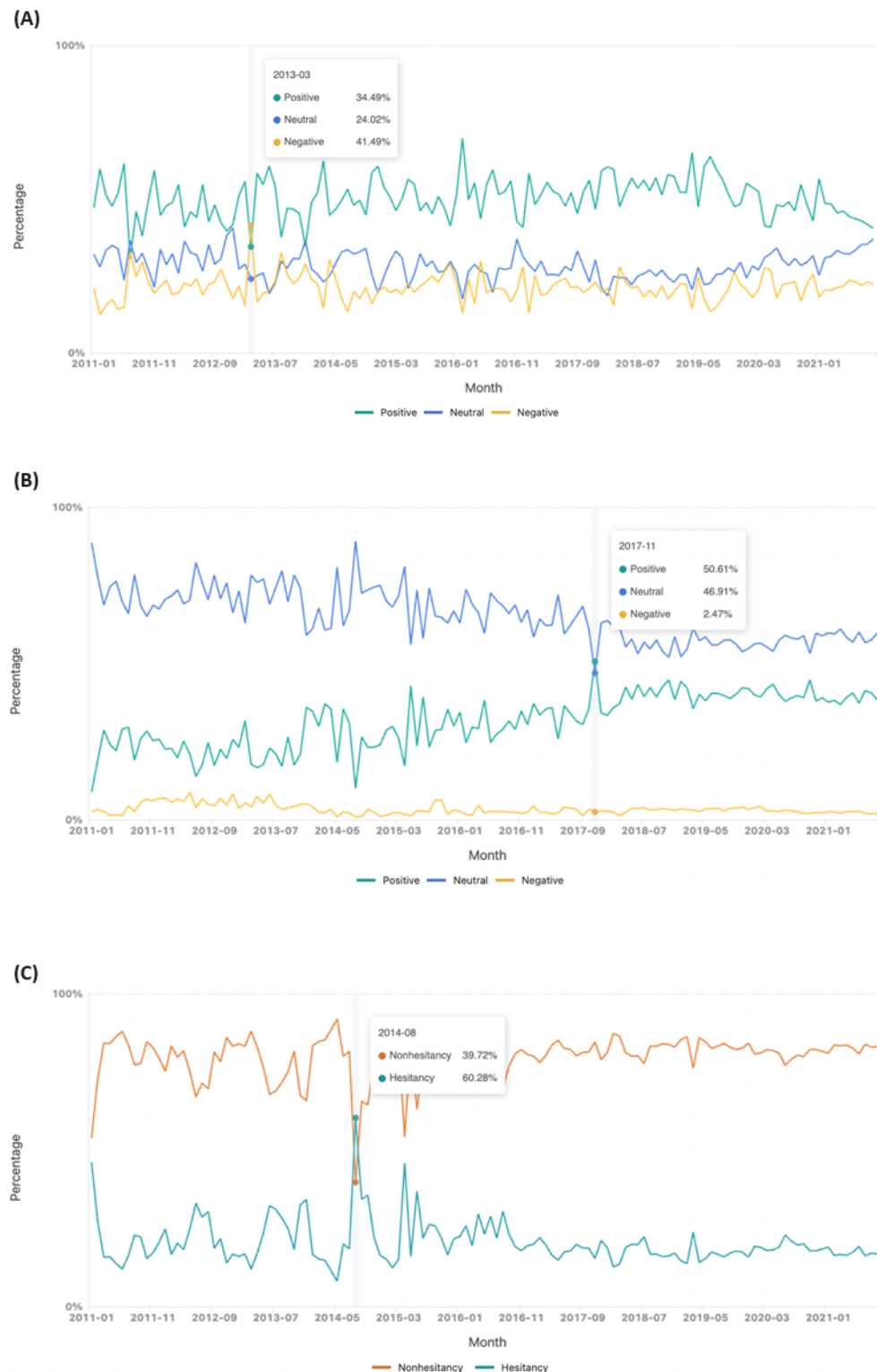
^aNLP: natural language processing.

^bHPV: human papillomavirus.

^cMMR: measles, mumps, and rubella.

^dGeneral: general or unspecified vaccines.

Figure 3. Temporal trends of vaccine sentiment and hesitancy. (A) Aggregation of 3 social media platform data sources to evaluate vaccine sentiment for HPV vaccine-related posts. (B) Comparison of vaccine sentiment for MMR vaccines. (C) Comparison of vaccine hesitancy for MMR vaccine. HPV: human papillomavirus; MMR: measles, mumps, and rubella.



Discussion

Principal Findings

Our analysis of temporal trends in vaccine-related sentiment on social media platforms yielded valuable insights into the dynamics of public perception. A total of 5 different classification algorithms were subjected to tests for performance

in sentiment and hesitancy classifications, revealing that advanced text classification algorithms such as XGBoost and Snorkel outperformed others in classifying hesitancy, complacency, and other factors, while LR had a superior performance for sentiment classification. The superior performance of LR could potentially be attributed to its enhanced ability to effectively handle binary classification

challenges and manage noise variables [63]. As the use of artificial intelligence platforms is increasingly becoming accessible for public use, it is crucial to gain an understanding of their accuracy and limitations. Traditional machine learning algorithms have the ability to predict outcomes but often lack transparency. Hence, enhancing public understanding and advancing toward explainable artificial intelligence is vital for error rectification and improved model efficacy for social media research [64].

When evaluating trends for the HPV vaccine, overall positive sentiment outweighed neutral and negative sentiment (Figure 3A), a notable exception occurred in March 2013. During this period, posts with negative sentiment on all 3 platforms surpassed those with 34% (2270/6582) positive and 24% (1581/6582) neutral sentiment, constituting 41% (2731/6582) of the total. This spike in negative sentiment can be attributed to news articles published in March 2013; for example, “Worried Parents Balk At HPV Vaccine For Daughters” by National Public Radio [65] and “Side Effect Fears Stop Parents from Getting HPV Vaccine for Daughters” by CBS News [66]. These articles highlighted concerns and fears about the HPV vaccine. Afterward, specific studies were conducted and published to further investigate these concerns and fears [67,68]. Notably, the HPV vaccines have been found to be safe in several studies and strongly recommended by the Centers for Disease Control and Prevention (CDC), etc [69,70].

Conversely, overall, posts expressed more neutral sentiment toward MMR than positive sentiment (Figure 3B), with an exception in November 2017. During this month, 51% (2844/5619) of posts expressed positive sentiment and 47% (2636/5619) were neutral. We found that a mumps outbreak was observed right before November 2017, which may have encouraged people to discuss the importance of MMR vaccination. News articles highlighted this outbreak, for example, “Third dose of mumps vaccine could help stop outbreaks, researchers say” by PBS News Hour [71] and “CDC recommends booster shot of MMR vaccine during mumps outbreaks” by CNN [72] mentioned the outbreak and recommended the booster shot of MMR vaccine.

When tracking vaccine hesitancy, we found that the social media posts with a higher ratio of hesitancy were only observed in August 2014 (Figure 3C). During this month, some examples of articles could be associated with vaccine hesitancy: “Journal questions validity of autism and vaccine study” by CNN [73] and “Whistleblower Claims CDC Covered Up Data Showing Vaccine-Autism Link” by TIME [74]. While speculation, particularly among antivaccination subpopulations, continues to surround the discredited study linking MMR vaccines with autism, it is crucial to emphasize that this link has been unequivocally debunked by subsequent research, and organizations such as the CDC and WHO have clarified that no such association exists [75-77]. Nonetheless, these news articles, considered by some as antivaccine propaganda, may partially explain the observed trends in MMR vaccine hesitancy during August 2014.

Strengths and Limitations

In this study, we introduced an NLP-powered online monitoring tool for tracking vaccine-related discussions on multiple social media platforms, covering 3 vaccine topic groups. Our system provides several features that distinguish it from existing tools. It uses NLP algorithms to perform sentiment analysis on social media posts and facilitates the tracking of temporal trends and geographic clustering of vaccine sentiment and hesitancy through visualization. In addition, our system enables users to compare vaccine sentiment and hesitancy across different social media platforms. We have publicly shared our annotated social media vaccine corpora, and we have evaluated several text classification algorithms, providing a benchmark for future research. One of the hypothetical use cases is that our NLP-based tool’s application spans from gauging vaccine sentiment during disease outbreaks to when a new vaccine is introduced. During an outbreak, the tool effectively analyzed sentiments toward measles vaccination, facilitating adjustments in public health campaigns.

While our proposed method uses the coarse-grained sentiment model (ie, represents the sentiment as a positive or negative class), fine-grained sentiment models, unlike traditional independent dimensional approaches, beneficially incorporate relations between dimensions, such as valence and arousal, into deep neural networks, thereby providing more nuanced, real-valued sentiment analysis and enhancing prediction accuracy [78-81]. These models prove particularly valuable in language-specific applications and are capable of classifying emotion categories and simultaneously predicting valence, arousal, and dominance scores for specific sentences, providing more nuanced sentiment analysis compared with simple positive or negative classifications.

Beyond the limitations inherent in the sentiment model, our approach also encounters constraints due to the use of traditional machine learning algorithms. Deep learning methods for word or sentiment embedding offer enhanced performance in sentiment analysis tasks by integrating external knowledge such as sentiment polarity and emotional semantics into word vectors [82-87]. They leverage neural networks and multitask learning to create task-specific embeddings, improving the accuracy of tasks such as sentiment and emotion analysis and sarcasm and stress detection [82-84,86]. Furthermore, these methods can adapt to the dynamic nature of language, handling out-of-vocabulary words and context-specific word meanings, proving more accurate and comprehensive than traditional word embeddings [86,87]. In future iterations, we plan to enrich our tool by integrating cutting-edge methods, alongside a more robust evaluation method such as time series cross-validation [88].

While previous studies have used NLP for sentiment analysis on COVID-19 vaccination and information exposure analysis regarding the HPV vaccine using Twitter data sets [40,89], and have investigated the temporal and geographic variations in public perceptions of the HPV vaccine [90], our tool extends its functionality to include a broader spectrum of platforms for tracking different vaccine sentiment and hesitancy on social media. Despite the scientific evidence supporting the safety and

efficacy of vaccines, vaccine hesitancy sentiments on social media can impact public confidence regarding vaccination [91]. Our tool is designed to quickly identify surges in vaccine hesitancy and thereby could be a tool to assist public health professionals in responding promptly with accurate information and effective vaccine promotion strategies.

However, it is essential to acknowledge the inherent limitations of using social media as a public health surveillance tool. These limitations include geography and language restrictions, as well as potential population, age, and gender biases, given that social media users may not represent the general population [92-94]. The user diversity across various social media platforms might partly account for the variation in sentiment and hesitancy label distributions. For example, YouTube has a high volume of users, but Twitter had the most activity in our study because people may view YouTube videos without leaving comments [93]. Moreover, owners of YouTube channels also have the option to disable comments on their uploaded videos. In addition, YouTube comments are highly tied to the content of the videos that the model might not have access to, leading to misinterpretations of sentiment and hesitancy. These biases and variabilities could partly account for the lower prediction accuracy observed for YouTube. Therefore, caution should be exercised when interpreting findings based on social media data, particularly considering the varying distributions of sentiment

and hesitancy across different social media platforms in our study. Another limitation pertains to the absence of a weighting system in the dashboard. Currently, the impact of each post, considering variables such as the number of views or reposts, is not considered. In addition, private interactions, specifically on sites such as Facebook, might go unnoticed and this lack of access to private dialogues could limit the comprehensiveness of the responses we capture. Finally, there is the possibility of shifts in user behavior to emerging social media platforms, such as TikTok, introducing additional population bias if such platforms are not included in further analyses.

Conclusions

This study successfully developed an innovative real-time monitoring system for analyzing vaccine sentiment and hesitancy across 3 major social media platforms. This system uses NLP and machine learning to mine and classify social media discussions on vaccines, providing valuable insights into public sentiment and hesitancy trends. The application of this tool presents significant implications for public health strategies, aiding in promptly identifying and mitigating vaccine misinformation, enhancing vaccine uptake, and assisting in the execution of targeted health campaigns. Moreover, it encourages health care professionals to foster an evidence-based discourse around vaccines, thus counteracting misinformation and improving public health outcomes.

Acknowledgments

This work was funded by Merck Sharp & Dohme Corp, a subsidiary of Merck & Co, Inc (Rahway, New Jersey). The content is the sole responsibility of the authors and does not necessarily represent the official views of Merck & Co, Inc or Melax Tech.

Authors' Contributions

JD, ALE, and LY conceptualized and designed the study. LCH and JD performed the experiments. LCH, ALE, JD, and LY drafted the paper. LCH, LH, and JD performed the acquisition, analysis, or interpretation of data. All authors performed critical revision of the paper for important intellectual content. JD, ALE, and LY performed study supervision.

Conflicts of Interest

ALE is a current employee of Merck Sharp & Dohme LLC, a subsidiary of Merck & Co, Inc, Rahway, New Jersey, United States, who may own stock and stock options in the Company. LY was an employee of Merck Sharp & Dohme LLC, a subsidiary of Merck & Co, Inc, Rahway, New Jersey, United States during the time of the study. Melax Tech, including JW, JD, and LCH, was compensated for activities related to the execution of the study. FJM was employed by Melax Tech and IMO Health during the research described. IMO Health retains interests in certain software described in this article.

Multimedia Appendix 1

The online dashboard's user interface, architecture, and performances.

[\[DOCX File , 2065 KB-Multimedia Appendix 1\]](#)

References

1. Watson OJ, Barnsley G, Toor J, Hogan AB, Winskill P, Ghani AC. Global impact of the first year of COVID-19 vaccination: a mathematical modelling study. *Lancet Infect Dis*. 2022;22(9):1293-1302. [\[FREE Full text\]](#) [doi: [10.1016/S1473-3099\(22\)00320-6](https://doi.org/10.1016/S1473-3099(22)00320-6)] [Medline: [35753318](https://pubmed.ncbi.nlm.nih.gov/35753318/)]
2. Lindmeier C. Measles vaccination has saved an estimated 17.1 million lives since 2000. World Health Organization. 2015. URL: <https://www.who.int/news/item/12-11-2015-measles-vaccination-has-saved-an-estimated-17-1-million-lives-since-2000> [accessed 2024-05-08]
3. Ehreth J. The global value of vaccination. *Vaccine*. 2003;21(7-8):596-600. [doi: [10.1016/s0264-410x\(02\)00623-0](https://doi.org/10.1016/s0264-410x(02)00623-0)] [Medline: [12531324](https://pubmed.ncbi.nlm.nih.gov/12531324/)]

4. MacDonald NE, SAGE Working Group on Vaccine Hesitancy. Vaccine hesitancy: definition, scope and determinants. *Vaccine*. 2015;33(34):4161-4164. [FREE Full text] [doi: [10.1016/j.vaccine.2015.04.036](https://doi.org/10.1016/j.vaccine.2015.04.036)] [Medline: [25896383](https://pubmed.ncbi.nlm.nih.gov/25896383/)]
5. Ten threats to global health in 2019. World Health Organization. URL: <https://www.who.int/news-room/spotlight/ten-threats-to-global-health-in-2019> [accessed 2024-05-08]
6. Dubé E, Vivion M, MacDonald NE. Vaccine hesitancy, vaccine refusal and the anti-vaccine movement: influence, impact and implications. *Expert Rev Vaccines*. 2015;14(1):99-117. [doi: [10.1586/14760584.2015.964212](https://doi.org/10.1586/14760584.2015.964212)] [Medline: [25373435](https://pubmed.ncbi.nlm.nih.gov/25373435/)]
7. Lane S, MacDonald NE, Marti M, Dumolard L. Vaccine hesitancy around the globe: analysis of three years of WHO/UNICEF joint reporting form data-2015-2017. *Vaccine*. 2018;36(26):3861-3867. [FREE Full text] [doi: [10.1016/j.vaccine.2018.03.063](https://doi.org/10.1016/j.vaccine.2018.03.063)] [Medline: [29605516](https://pubmed.ncbi.nlm.nih.gov/29605516/)]
8. Black FL. The role of herd immunity in control of measles. *Yale J Biol Med*. 1982;55(3-4):351-360. [FREE Full text] [Medline: [7180027](https://pubmed.ncbi.nlm.nih.gov/7180027/)]
9. Cockman P, Dawson L, Mathur R, Hull S. Improving MMR vaccination rates: herd immunity is a realistic goal. *BMJ*. 2011;343:d5703. [doi: [10.1136/bmj.d5703](https://doi.org/10.1136/bmj.d5703)] [Medline: [21971162](https://pubmed.ncbi.nlm.nih.gov/21971162/)]
10. Lieu TA, Ray GT, Klein NP, Chung C, Kulldorff M. Geographic clusters in underimmunization and vaccine refusal. *Pediatrics*. 2015;135(2):280-289. [doi: [10.1542/peds.2014-2715](https://doi.org/10.1542/peds.2014-2715)] [Medline: [25601971](https://pubmed.ncbi.nlm.nih.gov/25601971/)]
11. Omer SB, Pan WKY, Halsey NA, Stokley S, Moulton LH, Navar AM, et al. Nonmedical exemptions to school immunization requirements: secular trends and association of state policies with pertussis incidence. *JAMA*. 2006;296(14):1757-1763. [FREE Full text] [doi: [10.1001/jama.296.14.1757](https://doi.org/10.1001/jama.296.14.1757)] [Medline: [17032989](https://pubmed.ncbi.nlm.nih.gov/17032989/)]
12. Dempsey AF, Schaffer S, Singer D, Butchart A, Davis M, Freed GL. Alternative vaccination schedule preferences among parents of young children. *Pediatrics*. 2011;128(5):848-856. [doi: [10.1542/peds.2011-0400](https://doi.org/10.1542/peds.2011-0400)] [Medline: [21969290](https://pubmed.ncbi.nlm.nih.gov/21969290/)]
13. Sadaf A, Richards JL, Glanz J, Salmon DA, Omer SB. A systematic review of interventions for reducing parental vaccine refusal and vaccine hesitancy. *Vaccine*. 2013;31(40):4293-4304. [doi: [10.1016/j.vaccine.2013.07.013](https://doi.org/10.1016/j.vaccine.2013.07.013)] [Medline: [23859839](https://pubmed.ncbi.nlm.nih.gov/23859839/)]
14. Zhao Z, Luman ET. Progress toward eliminating disparities in vaccination coverage among U.S. children, 2000-2008. *Am J Prev Med*. 2010;38(2):127-137. [doi: [10.1016/j.amepre.2009.10.035](https://doi.org/10.1016/j.amepre.2009.10.035)] [Medline: [20117568](https://pubmed.ncbi.nlm.nih.gov/20117568/)]
15. Zimet GD, Weiss TW, Rosenthal SL, Good MB, Vichnin MD. Reasons for non-vaccination against HPV and future vaccination intentions among 19-26 year-old women. *BMC Womens Health*. 2010;10:27. [FREE Full text] [doi: [10.1186/1472-6874-10-27](https://doi.org/10.1186/1472-6874-10-27)] [Medline: [20809965](https://pubmed.ncbi.nlm.nih.gov/20809965/)]
16. Dredze M, Broniatowski DA, Smith MC, Hilyard KM. Understanding vaccine refusal: why we need social media now. *Am J Prev Med*. 2016;50(4):550-552. [FREE Full text] [doi: [10.1016/j.amepre.2015.10.002](https://doi.org/10.1016/j.amepre.2015.10.002)] [Medline: [26655067](https://pubmed.ncbi.nlm.nih.gov/26655067/)]
17. Peretti-Watel P, Larson HJ, Ward JK, Schulz WS, Verger P. Vaccine hesitancy: clarifying a theoretical framework for an ambiguous notion. *PLoS Curr*. Feb 25, 2015;7:ecurrents.outbreaks.6844c80ff9f5b273f34c91f71b7fc289. [FREE Full text] [doi: [10.1371/currents.outbreaks.6844c80ff9f5b273f34c91f71b7fc289](https://doi.org/10.1371/currents.outbreaks.6844c80ff9f5b273f34c91f71b7fc289)] [Medline: [25789201](https://pubmed.ncbi.nlm.nih.gov/25789201/)]
18. Galagali PM, Kinikar AA, Kumar VS. Vaccine hesitancy: obstacles and challenges. *Curr Pediatr Rep*. 2022;10(4):241-248. [FREE Full text] [doi: [10.1007/s40124-022-00278-9](https://doi.org/10.1007/s40124-022-00278-9)] [Medline: [36245801](https://pubmed.ncbi.nlm.nih.gov/36245801/)]
19. Larson HJ, Smith DMD, Paterson P, Cumming M, Eckersberger E, Freifeld CC, et al. Measuring vaccine confidence: analysis of data obtained by a media surveillance system used to analyse public concerns about vaccines. *Lancet Infect Dis*. 2013;13(7):606-613. [doi: [10.1016/S1473-3099\(13\)70108-7](https://doi.org/10.1016/S1473-3099(13)70108-7)] [Medline: [23676442](https://pubmed.ncbi.nlm.nih.gov/23676442/)]
20. Lawrence HY, Hausman BL, Dannenberg CJ. Reframing medicine's publics: the local as a public of vaccine refusal. *J Med Humanit*. 2014;35(2):111-129. [doi: [10.1007/s10912-014-9278-4](https://doi.org/10.1007/s10912-014-9278-4)] [Medline: [24682632](https://pubmed.ncbi.nlm.nih.gov/24682632/)]
21. WHO T. The guide to tailoring immunization programmes. WHO Regional Office for Europe. 2013. URL: <https://iris.who.int/handle/10665/351166> [accessed 2024-05-08]
22. Yaqub O, Castle-Clarke S, Sevdalis N, Chataway J. Attitudes to vaccination: a critical review. *Soc Sci Med*. 2014;112:1-11. [FREE Full text] [doi: [10.1016/j.socscimed.2014.04.018](https://doi.org/10.1016/j.socscimed.2014.04.018)] [Medline: [24788111](https://pubmed.ncbi.nlm.nih.gov/24788111/)]
23. Cox DS, Cox AD, Sturm L, Zimet G. Behavioral interventions to increase HPV vaccination acceptability among mothers of young girls. *Health Psychol*. 2010;29(1):29-39. [doi: [10.1037/a0016942](https://doi.org/10.1037/a0016942)] [Medline: [20063933](https://pubmed.ncbi.nlm.nih.gov/20063933/)]
24. Cates JR, Ortiz R, Shafer A, Romocki LS, Coyne-Beasley T. Designing messages to motivate parents to get their preteenage sons vaccinated against human papillomavirus. *Perspect Sex Reprod Health*. 2012;44(1):39-47. [FREE Full text] [doi: [10.1363/4403912](https://doi.org/10.1363/4403912)] [Medline: [22405151](https://pubmed.ncbi.nlm.nih.gov/22405151/)]
25. Clayton EW, Hickson GB, Miller CS. Parents' responses to vaccine information pamphlets. *Pediatrics*. 1994;93(3):369-372. [Medline: [8115193](https://pubmed.ncbi.nlm.nih.gov/8115193/)]
26. Kagashe I, Yan Z, Suheryani I. Enhancing seasonal influenza surveillance: topic analysis of widely used medicinal drugs using Twitter data. *J Med Internet Res*. 2017;19(9):e315. [FREE Full text] [doi: [10.2196/jmir.7393](https://doi.org/10.2196/jmir.7393)] [Medline: [28899847](https://pubmed.ncbi.nlm.nih.gov/28899847/)]
27. Eichstaedt JC, Schwartz HA, Kern ML, Park G, Labarthe DR, Merchant RM, et al. Psychological language on Twitter predicts county-level heart disease mortality. *Psychol Sci*. 2015;26(2):159-169. [FREE Full text] [doi: [10.1177/0956797614557867](https://doi.org/10.1177/0956797614557867)] [Medline: [25605707](https://pubmed.ncbi.nlm.nih.gov/25605707/)]
28. Chan B, Lopez A, Sarkar U. The canary in the coal mine tweets: social media reveals public perceptions of non-medical use of opioids. *PLoS One*. 2015;10(8):e0135072. [FREE Full text] [doi: [10.1371/journal.pone.0135072](https://doi.org/10.1371/journal.pone.0135072)] [Medline: [26252774](https://pubmed.ncbi.nlm.nih.gov/26252774/)]

29. Mitra T, Counts S, Pennebaker J. Understanding anti-vaccination attitudes in social media. 2016. Presented at: Tenth International AAAI Conference on Web and Social Media; May 17-20, 2016:269-278; Cologne, Germany. URL: <https://ojs.aaai.org/index.php/ICWSM/issue/view/272> [doi: [10.1609/icwsm.v10i1.14729](https://doi.org/10.1609/icwsm.v10i1.14729)]
30. McDonald L, Malcolm B, Ramagopalan S, Syrad H. Real-world data and the patient perspective: the promise of social media? *BMC Med*. 2019;17(1):11. [FREE Full text] [doi: [10.1186/s12916-018-1247-8](https://doi.org/10.1186/s12916-018-1247-8)] [Medline: [30646913](https://pubmed.ncbi.nlm.nih.gov/30646913/)]
31. 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;15(4):e85. [FREE Full text] [doi: [10.2196/jmir.1933](https://doi.org/10.2196/jmir.1933)] [Medline: [23615206](https://pubmed.ncbi.nlm.nih.gov/23615206/)]
32. Becker BFH, Larson HJ, Bonhoeffer J, van Mulligen EM, Kors JA, Sturkenboom MCJM. Evaluation of a multinational, multilingual vaccine debate on Twitter. *Vaccine*. 2016;34(50):6166-6171. [FREE Full text] [doi: [10.1016/j.vaccine.2016.11.007](https://doi.org/10.1016/j.vaccine.2016.11.007)] [Medline: [27840012](https://pubmed.ncbi.nlm.nih.gov/27840012/)]
33. Radzikowski J, Stefanidis A, Jacobsen KH, Croitoru A, Crooks A, Delamater PL. The measles vaccination narrative in Twitter: a quantitative analysis. *JMIR Public Health Surveill*. 2016;2(1):e1. [FREE Full text] [doi: [10.2196/publichealth.5059](https://doi.org/10.2196/publichealth.5059)] [Medline: [27227144](https://pubmed.ncbi.nlm.nih.gov/27227144/)]
34. Love B, Himelboim I, Holton A, Stewart K. Twitter as a source of vaccination information: content drivers and what they are saying. *Am J Infect Control*. 2013;41(6):568-570. [doi: [10.1016/j.ajic.2012.10.016](https://doi.org/10.1016/j.ajic.2012.10.016)] [Medline: [23726548](https://pubmed.ncbi.nlm.nih.gov/23726548/)]
35. Keelan J, Pavri V, Balakrishnan R, Wilson K. An analysis of the human papilloma virus vaccine debate on MySpace blogs. *Vaccine*. 2010;28(6):1535-1540. [FREE Full text] [doi: [10.1016/j.vaccine.2009.11.060](https://doi.org/10.1016/j.vaccine.2009.11.060)] [Medline: [20003922](https://pubmed.ncbi.nlm.nih.gov/20003922/)]
36. Chowdhary KR. Natural language processing. In: *Fundamentals of Artificial Intelligence*. New York City. Springer; 2020:603-649.
37. Vinet L, Zhedanov A. A 'missing' family of classical orthogonal polynomials. *J Phys A Math Theor*. 2011;44(8):085201. [doi: [10.1088/1751-8113/44/8/085201](https://doi.org/10.1088/1751-8113/44/8/085201)]
38. Salathé M, Khandelwal S. Assessing vaccination sentiments with online social media: implications for infectious disease dynamics and control. *PLoS Comput Biol*. 2011;7(10):e1002199. [FREE Full text] [doi: [10.1371/journal.pcbi.1002199](https://doi.org/10.1371/journal.pcbi.1002199)] [Medline: [22022249](https://pubmed.ncbi.nlm.nih.gov/22022249/)]
39. Vaccine Sentiments on Social Media. URL: <https://vaccine.social/> [accessed 2024-05-08]
40. Qorib M, Oladunni T, Denis M, Ososanya E, Cotaé P. Covid-19 vaccine hesitancy: text mining, sentiment analysis and machine learning on COVID-19 vaccination Twitter dataset. *Expert Syst Appl*. 2023;212:118715. [FREE Full text] [doi: [10.1016/j.eswa.2022.118715](https://doi.org/10.1016/j.eswa.2022.118715)] [Medline: [36092862](https://pubmed.ncbi.nlm.nih.gov/36092862/)]
41. Kumar N, Corpus I, Hans M, Harle N, Yang N, McDonald C, et al. COVID-19 vaccine perceptions in the initial phases of US vaccine roll-out: an observational study on reddit. *BMC Public Health*. 2022;22(1):446. [FREE Full text] [doi: [10.1186/s12889-022-12824-7](https://doi.org/10.1186/s12889-022-12824-7)] [Medline: [35255881](https://pubmed.ncbi.nlm.nih.gov/35255881/)]
42. Li HOY, Pastukhova E, Brandts-Longtin O, Tan MG, Kirchhof MG. YouTube as a source of misinformation on COVID-19 vaccination: a systematic analysis. *BMJ Glob Health*. 2022;7(3):e008334. [FREE Full text] [doi: [10.1136/bmjgh-2021-008334](https://doi.org/10.1136/bmjgh-2021-008334)] [Medline: [35264318](https://pubmed.ncbi.nlm.nih.gov/35264318/)]
43. Kwon S, Park A. Examining thematic and emotional differences across Twitter, Reddit, and YouTube: the case of COVID-19 vaccine side effects. *Comput Human Behav*. 2023;144:107734. [FREE Full text] [doi: [10.1016/j.chb.2023.107734](https://doi.org/10.1016/j.chb.2023.107734)] [Medline: [36942128](https://pubmed.ncbi.nlm.nih.gov/36942128/)]
44. Aleksandric A, Anderson HI, Melcher S, Nilizadeh S, Wilson GM. Spanish Facebook posts as an indicator of COVID-19 vaccine hesitancy in Texas. *Vaccines (Basel)*. 2022;10(10):1713. [FREE Full text] [doi: [10.3390/vaccines10101713](https://doi.org/10.3390/vaccines10101713)] [Medline: [36298580](https://pubmed.ncbi.nlm.nih.gov/36298580/)]
45. Yue L, Chen W, Li X, Zuo W, Yin M. A survey of sentiment analysis in social media. *Knowl Inf Syst*. 2018;60(2):617-663. [doi: [10.1007/s10115-018-1236-4](https://doi.org/10.1007/s10115-018-1236-4)]
46. Nandwani P, Verma R. A review on sentiment analysis and emotion detection from text. *Soc Netw Anal Min*. 2021;11(1):81. [FREE Full text] [doi: [10.1007/s13278-021-00776-6](https://doi.org/10.1007/s13278-021-00776-6)] [Medline: [34484462](https://pubmed.ncbi.nlm.nih.gov/34484462/)]
47. Bonanni P, Bechini A, Donato R, Capei R, Sacco C, Levi M, et al. Human papilloma virus vaccination: impact and recommendations across the world. *Ther Adv Vaccines*. 2015;3(1):3-12. [FREE Full text] [doi: [10.1177/2051013614557476](https://doi.org/10.1177/2051013614557476)] [Medline: [25553242](https://pubmed.ncbi.nlm.nih.gov/25553242/)]
48. Bankamp B, Hickman C, Icenogle JP, Rota PA. Successes and challenges for preventing measles, mumps and rubella by vaccination. *Curr Opin Virol*. 2019;34:110-116. [doi: [10.1016/j.coviro.2019.01.002](https://doi.org/10.1016/j.coviro.2019.01.002)] [Medline: [30852425](https://pubmed.ncbi.nlm.nih.gov/30852425/)]
49. Eiden AL, DiFranzo A, Bhatti A, Wang HE, Bencina G, Yao L, et al. Changes in vaccine administration trends across the life-course during the COVID-19 pandemic in the United States: a claims database study. *Expert Rev Vaccines*. 2023;22(1):481-494. [FREE Full text] [doi: [10.1080/14760584.2023.2217257](https://doi.org/10.1080/14760584.2023.2217257)] [Medline: [37218717](https://pubmed.ncbi.nlm.nih.gov/37218717/)]
50. Kleinbaum DG, Klein M, Pryor ER. *Logistic Regression: A Self-Learning Text*. Berlin, Heidelberg, Dordrecht, New York City. Springer; 2002.
51. Noble WS. What is a support vector machine? *Nat Biotechnol*. 2006;24(12):1565-1567. [doi: [10.1038/nbt1206-1565](https://doi.org/10.1038/nbt1206-1565)]
52. Pal M. Random forest classifier for remote sensing classification. *Int J Remote Sens*. 2007;26(1):217-222. [doi: [10.1080/01431160412331269698](https://doi.org/10.1080/01431160412331269698)]

53. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. New York, NY, United States. Association for Computing Machinery; 2016. Presented at: KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; August 13-17, 2016:785-794; San Francisco, California, USA. URL: <https://dl.acm.org/doi/proceedings/10.1145/2939672> [doi: [10.1145/2939672.2939785](https://doi.org/10.1145/2939672.2939785)]
54. Ratner A, Bach SH, Ehrenberg H, Fries J, Wu S, Ré C. Snorkel: rapid training data creation with weak supervision. Proceedings VLDB Endowment. 2017;11(3):269-282. [FREE Full text] [doi: [10.14778/3157794.3157797](https://doi.org/10.14778/3157794.3157797)] [Medline: [29770249](https://pubmed.ncbi.nlm.nih.gov/29770249/)]
55. Schober P, Vetter TR. Logistic regression in medical research. Anesth Analg. 2021;132(2):365-366. [FREE Full text] [doi: [10.1213/ANE.0000000000005247](https://doi.org/10.1213/ANE.0000000000005247)] [Medline: [33449558](https://pubmed.ncbi.nlm.nih.gov/33449558/)]
56. Sun W, Cai Z, Li Y, Liu F, Fang S, Wang G. Data processing and text mining technologies on electronic medical records: a review. J Healthc Eng. 2018;2018:4302425. [FREE Full text] [doi: [10.1155/2018/4302425](https://doi.org/10.1155/2018/4302425)] [Medline: [29849998](https://pubmed.ncbi.nlm.nih.gov/29849998/)]
57. Alam MZ, Rahman MS, Rahman MS. A random forest based predictor for medical data classification using feature ranking. Inform Med Unlocked. 2019;15:100180. [FREE Full text] [doi: [10.1016/j.imu.2019.100180](https://doi.org/10.1016/j.imu.2019.100180)]
58. Alotaibi Y, Malik MN, Khan HH, Batool A, Alsufyani A, Alghamdi S, et al. Suggestion mining from opinionated text of big social media data. CMC-Comput Mater Con. 2021;68(3):3323-3338. [FREE Full text] [doi: [10.32604/cmc.2021.016727](https://doi.org/10.32604/cmc.2021.016727)]
59. Mallory EK, de Rochemonteix M, Ratner A, Acharya A, Re C, Bright RA, et al. Extracting chemical reactions from text using Snorkel. BMC Bioinformatics. 2020;21(1):217. [FREE Full text] [doi: [10.1186/s12859-020-03542-1](https://doi.org/10.1186/s12859-020-03542-1)] [Medline: [32460703](https://pubmed.ncbi.nlm.nih.gov/32460703/)]
60. Ramos J. Using TF-IDF to determine word relevance in document queries. Proceedings of the first instructional conference on machine learning. 2003. URL: <https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=b3bf6373ff41a115197cb5b30e57830c16130c2c> [accessed 2024-05-13]
61. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. J Artif Intell Res. 2002;16:321-357. [FREE Full text] [doi: [10.1613/jair.953](https://doi.org/10.1613/jair.953)]
62. Tang Y, Zhang YQ, Chawla NV, Krasser S. SVMs modeling for highly imbalanced classification. IEEE Trans Syst Man Cybern B Cybern. 2009;39(1):281-288. [doi: [10.1109/TSMCB.2008.2002909](https://doi.org/10.1109/TSMCB.2008.2002909)] [Medline: [19068445](https://pubmed.ncbi.nlm.nih.gov/19068445/)]
63. Kirasich K, Smith T, Sadler B. Random forest vs logistic regression: binary classification for heterogeneous datasets. SMU Data Science Review. 2018;1(3):9.
64. Mehta H, Passi K. Social media hate speech detection using Explainable Artificial Intelligence (XAI). Algorithms. 2022;15(8):291. [FREE Full text] [doi: [10.3390/a15080291](https://doi.org/10.3390/a15080291)]
65. Hensley S. Worried parents balk at HPV vaccine for daughters. NPR. 2013. URL: <https://www.npr.org/sections/health-shots/2013/03/18/174617709/worried-parents-balk-at-hpv-vaccine-for-daughters> [accessed 2024-05-08]
66. Castillo M. Side effect fears stop parents from getting HPV vaccine for daughters. CBS News. 2013. URL: <https://www.cbsnews.com/news/side-effect-fears-stop-parents-from-getting-hpv-vaccine-for-daughters/> [accessed 2024-05-08]
67. Zimet GD, Rosberger Z, Fisher WA, Perez S, Stupiansky NW. Beliefs, behaviors and HPV vaccine: correcting the myths and the misinformation. Prev Med. 2013;57(5):414-418. [FREE Full text] [doi: [10.1016/j.ypmed.2013.05.013](https://doi.org/10.1016/j.ypmed.2013.05.013)] [Medline: [23732252](https://pubmed.ncbi.nlm.nih.gov/23732252/)]
68. Karafillakis E, Simas C, Jarrett C, Verger P, Peretti-Watel P, Dib F, et al. HPV vaccination in a context of public mistrust and uncertainty: a systematic literature review of determinants of HPV vaccine hesitancy in Europe. Hum Vaccin Immunother. 2019;15(7-8):1615-1627. [FREE Full text] [doi: [10.1080/21645515.2018.1564436](https://doi.org/10.1080/21645515.2018.1564436)] [Medline: [30633623](https://pubmed.ncbi.nlm.nih.gov/30633623/)]
69. HPV, the vaccine for HPV, and cancers caused by HPV. Centers for Disease Control and Prevention. 2022. URL: <https://tinyurl.com/2d45j3jz> [accessed 2024-05-08]
70. Meites E, Szilagyi PG, Chesson HW, Unger ER, Romero JR, Markowitz LE. Human papillomavirus vaccination for adults: updated recommendations of the advisory committee on immunization practices. MMWR Morb Mortal Wkly Rep. 2019;68(32):698-702. [FREE Full text] [doi: [10.15585/mmwr.mm6832a3](https://doi.org/10.15585/mmwr.mm6832a3)] [Medline: [31415491](https://pubmed.ncbi.nlm.nih.gov/31415491/)]
71. Branswell H. Third dose of mumps vaccine could help stop outbreaks, researchers say. STAT. 2017. URL: <https://www.statnews.com/2017/09/06/mumps-vaccine-study/> [accessed 2024-05-08]
72. Scutti S. CDC recommends booster shot of MMR vaccine during mumps outbreaks. CNN Health. 2017. URL: <https://www.cnn.com/2017/10/25/health/cdc-mumps-outbreak-syracuse-university/index.html> [accessed 2024-05-08]
73. Goldschmidt D. Journal questions validity of autism and vaccine study. CNN Health. 2014. URL: <https://www.cnn.com/2014/08/27/health/irpt-cdc-autism-vaccine-study/index.html> [accessed 2024-05-08]
74. Park A. Whistleblower claims CDC covered up data showing vaccine-autism link. TIME. 2014. URL: <https://time.com/3208886/whistleblower-claims-cdc-covered-up-data-showing-vaccine-autism-link/> [accessed 2024-05-08]
75. Autism and vaccines. Centers for Disease Control and Prevention. 2021. URL: <https://www.cdc.gov/vaccinesafety/concerns/autism.html> [accessed 2024-05-08]
76. Epidemiological WW. MMR and autism. World Health Organization. 2003. URL: <https://www.who.int/groups/global-advisory-committee-on-vaccine-safety/topics/mmr-vaccines-and-autism> [accessed 2024-05-08]
77. The Editors of The Lancet, Caplan AL. Retraction—ileal-lymphoid-nodular hyperplasia, non-specific colitis, and pervasive developmental disorder in children. Lancet. 2010;375(9713):445. [doi: [10.1016/s0140-6736\(10\)60175-4](https://doi.org/10.1016/s0140-6736(10)60175-4)]

78. Xie H, Lin W, Lin S, Wang J, Yu LC. A multi-dimensional relation model for dimensional sentiment analysis. *Inf Sci.* 2021;579:832-844. [FREE Full text] [doi: [10.1016/j.ins.2021.08.052](https://doi.org/10.1016/j.ins.2021.08.052)]
79. Park S, Kim J, Ye S, Jeon J, Park HY, Oh A. Dimensional emotion detection from categorical emotion. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Online and Punta Cana, Dominican Republic. Association for Computational Linguistics; Nov 2021:4367-4380.
80. Lee LH, Li JH, Yu LC. Chinese EmoBank: building valence-arousal resources for dimensional sentiment analysis. *ACM Trans Asian Low-Resour Lang Inf Process.* 2022;21(4):1-18. [FREE Full text] [doi: [10.1145/3489141](https://doi.org/10.1145/3489141)]
81. Wang J, Yu LC, Lai KR, Zhang X. Tree-structured regional CNN-LSTM model for dimensional sentiment analysis. *IEEE/ACM Trans Audio Speech Lang Process.* 2020;28:581-591. [FREE Full text] [doi: [10.1109/taslp.2019.2959251](https://doi.org/10.1109/taslp.2019.2959251)]
82. Tang D, Wei F, Qin B, Yang N, Liu T, Zhou M. Sentiment embeddings with applications to sentiment analysis. *IEEE Trans Knowl Data Eng.* 2016;28(2):496-509. [doi: [10.1109/tkde.2015.2489653](https://doi.org/10.1109/tkde.2015.2489653)]
83. Xu P, Madotto A, Wu CS, Park JH, Fung P. Emo2Vec: learning generalized emotion representation by multi-task training. In: Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis. Brussels, Belgium. Association for Computational Linguistics; Oct 2018:292-298.
84. Ye Z, Li F, Baldwin T. Encoding sentiment information into word vectors for sentiment analysis. Association for Computational Linguistics; 2018. Presented at: Proceedings of the 27th International Conference on Computational Linguistics; August, 2018; Santa Fe, New Mexico, USA. URL: <https://aclanthology.org/C18-1085/>
85. Yu LC, Wang J, Lai KR, Zhang X. Refining word embeddings using intensity scores for sentiment analysis. *IEEE/ACM Trans Audio Speech Lang Process.* 2018;26(3):671-681. [doi: [10.1109/taslp.2017.2788182](https://doi.org/10.1109/taslp.2017.2788182)]
86. Wang J, Zhang Y, Yu LC, Zhang X. Contextual sentiment embeddings via bi-directional GRU language model. *Knowl-Based Syst.* 2022;235:107663. [FREE Full text] [doi: [10.1016/j.knosys.2021.107663](https://doi.org/10.1016/j.knosys.2021.107663)]
87. Zhu L, Li W, Shi Y, Guo K. SentiVec: learning sentiment-context vector via kernel optimization function for sentiment analysis. *IEEE Trans Neural Netw Learn Syst.* 2021;32(6):2561-2572. [doi: [10.1109/TNNLS.2020.3006531](https://doi.org/10.1109/TNNLS.2020.3006531)] [Medline: [32673198](https://pubmed.ncbi.nlm.nih.gov/32673198/)]
88. Bergmeir C, Hyndman RJ, Koo B. A note on the validity of cross-validation for evaluating autoregressive time series prediction. *Comput Stat Data An.* 2018;120:70-83. [doi: [10.1016/j.csda.2017.11.003](https://doi.org/10.1016/j.csda.2017.11.003)]
89. Dunn AG, Surian D, Leask J, Dey A, Mandl KD, Coiera E. Mapping information exposure on social media to explain differences in HPV vaccine coverage in the United States. *Vaccine.* 2017;35(23):3033-3040. [FREE Full text] [doi: [10.1016/j.vaccine.2017.04.060](https://doi.org/10.1016/j.vaccine.2017.04.060)] [Medline: [28461067](https://pubmed.ncbi.nlm.nih.gov/28461067/)]
90. Du J, Luo C, Shegog R, Bian J, Cunningham RM, Boom JA, et al. Use of deep learning to analyze social media discussions about the human papillomavirus vaccine. *JAMA Netw Open.* 2020;3(11):e2022025. [FREE Full text] [doi: [10.1001/jamanetworkopen.2020.22025](https://doi.org/10.1001/jamanetworkopen.2020.22025)] [Medline: [33185676](https://pubmed.ncbi.nlm.nih.gov/33185676/)]
91. Zhang Q, Zhang R, Wu W, Liu Y, Zhou Y. Impact of social media news on COVID-19 vaccine hesitancy and vaccination behavior. *Telemat Inform.* 2023;80:101983. [FREE Full text] [doi: [10.1016/j.tele.2023.101983](https://doi.org/10.1016/j.tele.2023.101983)] [Medline: [37122766](https://pubmed.ncbi.nlm.nih.gov/37122766/)]
92. Zhao Y, He X, Feng Z, Bost S, Prospero M, Wu Y, et al. Biases in using social media data for public health surveillance: a scoping review. *Int J Med Inform.* 2022;164:104804. [doi: [10.1016/j.ijmedinf.2022.104804](https://doi.org/10.1016/j.ijmedinf.2022.104804)] [Medline: [35644051](https://pubmed.ncbi.nlm.nih.gov/35644051/)]
93. Auxier B, Anderson M. Social media use in 2021. Pew Research Center. 2021;1:1-4. [FREE Full text] [doi: [10.4135/9781412963947.n376](https://doi.org/10.4135/9781412963947.n376)]
94. Shor E, van de Rijdt A, Fotouhi B. A large-scale test of gender bias in the media. *SocScience.* 2019;6:526-550. [FREE Full text] [doi: [10.15195/v6.a20](https://doi.org/10.15195/v6.a20)]

Abbreviations

- 3Cs:** confidence, complacency, and convenience
- API:** application programming interface
- CDC:** Centers for Disease Control and Prevention
- HPV:** human papillomavirus
- LR:** logistic regression
- MMR:** measles, mumps, and rubella
- NLP:** natural language processing
- SMOTE:** synthetic minority over-sampling technique
- SVM:** support vector machine
- VPD:** vaccine-preventable disease
- WHO:** World Health Organization
- XGBoost:** extreme gradient boosting

Edited by C Lovis; submitted 07.02.24; peer-reviewed by M Chatzimina, S Lee, Liang-Chih Yu, X Vargas Meza; comments to author 25.03.24; revised version received 08.04.24; accepted 11.04.24; published 21.06.24

Please cite as:

Huang Liang-Chin, Eiden AL, He L, Annan A, Wang S, Wang J, Manion FJ, Wang X, Du J, Yao L

Natural Language Processing–Powered Real-Time Monitoring Solution for Vaccine Sentiments and Hesitancy on Social Media: System Development and Validation

JMIR Med Inform 2024;12:e57164

URL: <https://medinform.jmir.org/2024/1/e57164>

doi: [10.2196/57164](https://doi.org/10.2196/57164)

PMID: [38904984](https://pubmed.ncbi.nlm.nih.gov/38904984/)

©Liang-Chin Huang, Amanda L Eiden, Long He, Augustine Annan, Siwei Wang, Jingqi Wang, Frank J Manion, Xiaoyan Wang, Jingcheng Du, Lixia Yao. Originally published in JMIR Medical Informatics (<https://medinform.jmir.org>), 21.06.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on <https://medinform.jmir.org/>, as well as this copyright and license information must be included.