Medical Research Archives

ESMED
European Society of Medicine

OPEN ACCESS

Published: May 31, 2022

Citation Scannell D, Desens L, et al., 2022. Combatting Mis/Disinformation: Combining Predictive Modeling and Machine Learning with Persuasion Science to Understand COVID-19 Vaccine Online Discourse, Medical Research Archives, [online] 10(5). https://doi.org/10.18103/mra.v10i5.2822

DOI
https://doi.org/10.18103/mra.v10i5.2822

ISSN: 2375-1924

RESEARCH ARTICLE

Combatting Mis/Disinformation: Combining Predictive Modeling and Machine Learning with Persuasion Science to Understand COVID-19 Vaccine Online Discourse

Denise Scannell, Ph.D.[1], Linda Desens, Ph.D.[1], David S.Day, Ph.D.[1], Yolande Tra, Ph.D.[1]

[1] MITRE Corporation

* ldesens@mitre.org

Conflicts of Interest Statement
The authors have no conflicts of interest to declare.

Funding Statement
This work was supported by The MITRE Corporation, 7515 Colshire Drive, McLean, VA 22102-7539

ABSTRACT

Health mis/disinformation can negatively impact health decisions and ultimately, health outcomes. Mis/disinformation related to COVID-19 vaccines has influenced vaccine hesitancy during a very critical time during the pandemic when globally, the vaccine was needed to attenuate the spread of the COVID-19 virus. This paper examines persuasive strategies used in Twitter posts, particular those with antivaccine sentiment. The authors developed a predictive model using variables based on the Elaboration Likelihood Model, Social Judgement Theory and the Extended Parallel Process Model to determine which persuasive tactics resulted in antivaccine, provaccine and neutral sentiment. The study also used machine learning to validate the persuasion variable algorithm to detect persuasion tactics in COVID-19 vaccine online discourse on Twitter. Understanding persuasive tactics used in antivaccine messaging can inform the development of a data-driven counter-response strategy.

Keywords: Misinformation, Disinformation, Persuasion, Algorithm, Sentiment

Introduction

Misinformation is not a new phenomenon in health. Throughout history, health innovations and discoveries have been misrepresented and surrounded by myth and conspiracy. In 1928, the discovery of penicillin by Alexander Fleming was rife with conspiracy and myth, some of which is still being refuted today.[1] In more recent history, a study conducted in 1998 linking the measles, mumps, and rubella vaccine to autism is still circulating in anti-vaccination circles, despite being debunked in 2010.[2] Vaccine hesitancy continues to have a negative influence on vaccine attitudes and behaviors due ever-growing exposure to vaccine misinformation.[3]

Health misinformation is not only a nuisance to public health due to its ability to steer individuals towards non-optimal health decisions, but it is also an issue of public safety.[4] Over the last few decades, the spread of inaccurate and dangerous information has led to public uprisings in protest of health interventions, and in some cases death.[4] Health communication scholars have struggled to address the growing trend of misinformation, searching for the one silver bullet – the one answer to address this mounting problem and ease some of the burden on public health communicators. However, misinformation is not a black and white issue. Rather, it is one that is as complex as humans themselves. The search for one answer ignores the many nuances of populations from culture, history, sex, community and creed. In addition, according to researchers from the Massachusetts Institute of Technology, misinformation also spreads faster and more broadly than factual information.[5] As evident during the COVID-19 pandemic, misinformation is also ever-changing and mutating to the rise and fall of rapidly changing science and public discourse Therefore, this is not a dragon to be slayed but one to be overcome or outsmarted through careful understanding of the patterns of persuasion and their intricate sway on individuals.

In 2020, a study was conducted to understand persuasion tactics used in COVID-19 vaccine messaging.[3] The study focused on persuasion tactics used in messaging in the three types of COVID-19 vaccine sentiments—Pro-Vaccine, Anti-Vaccine, and Neutral—with an additional focus on persuasion tactics around COVID-19 vaccines from what are likely bots. From that research, a framework was developed—the Health Information Persuasion Exploration (HIPE™) Framework—to identify mis/disinformation and persuasion tactics used in anti-vaccine messages and provide a path forward in the development of rapid response counter strategies and interventions. This research paper further builds on that study to explore the algorithm created to classify these patterns of persuasion in mis/disinformation narratives within COVID-19 social media discourse. Ineffective health communication during the pandemic has revealed a need to design more effective communication strategies that can be applied locally, nationally, and internationally to help combat the COVID-19 pandemic.[6]

Literature Review

This section explores research related to COVID-19 mis/disinformation on social media, evidence-based persuasive messaging, predictive models for pro-vaccine and anti-vaccine sentiment and machine learning models that facilitate the identification of not only mis/disinformation but also major topics in the COVID-19 vaccine online discourse. These models allow for the rapid analysis of the constantly evolving COVID-19 mis/disinformation that can have a negative effect on health decisions such as vaccine hesitancy.

Amplification of Mis/disinformation

In 2016, when there was a measles outbreak in Disneyland, Broniatowski et al stressed the critical importance of using social media to understand vaccine refusal before the next disease outbreak.[7] Current technology and social media have allowed the amplification of information at an unprecedented rate. These online social media platforms can accelerate the distribution of life-saving information to help people make informed health decisions that will protect themselves and their families. However, these same platforms can have the opposite effect. During the pandemic, social media has become a communication vehicle for mis- and disinformation. So much so that the World Health Organization (WHO) has referred to it as an "Infodemic."[8] An infodemic is an excessive overflow of information to include mis- and disinformation that can adversely influence people to make a health decision that could lead to severe illness or even death.

There are many sources of mis/disinformation. For example, social media and online foreign disinformation campaigns have been shown to impact vaccination rates and attitudes towards vaccine safety.[9] The use of social media is

predictive of the belief that vaccines are unsafe, while online foreign disinformation campaigns are associated with negative discourse on social media about vaccines as well as a decrease in mean vaccination coverage.[9]

Social media bots, which are automated programs, can also be used to amplify mis/disinformation.[10] Bots can retweet content at a high degree of frequency within users in the same opinion group making it more difficult for factual health information to reach these groups.[11] Bots are assigned scores (1-99), which indicate how likely an account is a bot account. Intermediate scoring bot accounts posted more tweets overall and were more likely to post tweets that were more polarized and neutral while accounts with high bot scores posted more neutral tweets and less polarizing ones.[7] Additionally, Russian trolls promoted discord while bots containing malware and unsolicited content were more likely to spread antivaccine messages. Fake accounts also contributed negatively to public opinion on vaccination.[7]

Persuasion Messaging and Vaccine Sentiment

Understanding that mis/disinformation can impact both health decisions and health outcomes, it is critical to examine the persuasion strategies used in health messaging for vaccines, particularly the antivaccine messages, to inform public health communication efforts. For example, antivaccine tweets are retweeted more than provaccine and neutral tweets.[12] Public health and health communication experts must examine persuasion drivers that promote amplification of these messages and consider persuasion strategies for increasing the amplification of their messages. A first step is to look at persuasion elements of the messages.

Understanding the reasons that people share content can aid public health and health communication scientists design targeted messages with a greater chance of amplification. As an example, the emotion elicited by a message can impact the virality of a message. Physiological arousal can make content more or less viral. For example, Berger and Milkman[13] found that content that evokes low-arousal emotions such as sadness is less viral; whereas, high-arousal emotions, both positive and negative (e.g., anger, anxiety, awe) is more viral.

Other persuasion techniques that affect amplification include the type of content and the way it is presented. For example, Twitter content with pictures and celebrity endorsement were most amplified while text only content that contained information, promotion and participation were most amplified.[14] On the social media platform, Pinterest, antivaccine messages used narrative vaccination information more frequently when compared to provaccine messages, which used statistical information.[15] Storytelling can be a powerful tool for embedding facts versus communicating facts alone.[16]

Values and lifestyle norms are often used as part of persuasion tactics. Messages that focus on values that are important to the receiver of the message are more likely to be processed critically resulting in an increase in personal involvement and resistance to future attacks.[17] Anti-vaccine messages have been found to center around values such as freedom, choice and individuality and spread misinformation and fear that vaccines can cause adverse health outcomes.[18] Vaccination policies requiring vaccinations without an option for non-medical refusal challenges an individual's values of choice and freedom, and can result in increased antivaccine sentiment.[19] Healthcare providers play a significant role in vaccine acceptance. Creating a trusting relationship with the parent can increase vaccine adoption for the children.[19]

Recent analysis of social media discourse on COVID-19 have provided valuable insights into the most popular topics and themes that are also reflective of people's values. This provides an opportunity to see how the discourse changes as the pandemic evolves and to adapt messaging strategies appropriately. In addition to values, the following top themes were identified in an analysis of Tweets on COVID-19 - global nature; healthcare, illness, virus, government/government response, and individual concerns and strategies.[20] Conspiracy theories and loss of civil liberties were themes identified in vaccine content on Pinterest.[15] with conspiracy-focused misinformation gaining more support when compared to medical misinformation.[21]

Machine Learning and Misinformation

Machine learning offers the unique opportunity to combat mis/disinformation using algorithms to detect mis/disinformation before they are amplified and can impact health outcomes.

Machine-learning models have been developed to detect misinformation related to COVID-19.[22,23] Using short-term memory (LSTM) networks, a multichannel convolutional neural network and k-nearest neighbors showed excellent results in identifying COVID-19 misinformation.[22] Additionally, three different models have been used to identify misinformation in a dataset of COVID19 vaccine tweets – LSTM, XGBoost and the bidirectional encoder representations from transformers (BERT)-based model.[24] The highest F1, precision and recall scores were achieved using the BERT model. Natural Language Processing deep learning techniques and machine learning have also been used to classify datasets for a fake news detector.[25] Other traditional machine learning models that have been used to classify online misinformation with a high accuracy include Support Vector Machine (SVM), Decision Tree (DT), Random Forest (FR) and Stochastic Gradient Descent (SGD).[26]

With the proliferation and rapid amplification of online information, this automation allows the identification of misinformation as early as possible so that counter messaging strategies can be deployed. Machine learning has also enabled the identification of topics related to vaccine discourse on Twitter.[27,28] As COVID-19 vaccine sentiments and topics shift as the pandemic progresses, a rapid method to identify these changes in the trends and discourse is critical to public health in responding appropriately and in a timely fashion to these dynamic changes.[29]

Health Information Persuasion Exploration (HIPE™) Framework

The HIPE™ framework addresses mis/disinformation in online discourse on social media. The HIPE framework includes detection, analysis, design and evaluation. The detection phase includes the use of social listening tools to identify sentiment, key themes and trends as well as persuasion tactics. The algorithm developed to identify the persuasion variable are based on a combination of three theoretical frameworks: 1) Elaboration Likelihood Model (ELM);[30] 2) Social Judgement Theory (SJT);[31] and 3) the Extended Parallel Process Model (EPPM).[32] ELM provides a deeper understanding into the type of processing – central or peripheral - or the amount of effort used to evaluate a message.[30] Messages that require central processing include health information and statistics, questions and participation – elements which require the receiver of the message to think critically about the issue. Peripheral processing is more superficial. These types of messages include celebrities or other influencers, humor/sarcasms, inspiration, and stories. Social Judgement Theory focuses on people's values or issues within their latitude of acceptance.[31] EPPM includes messages that include fear appeal, perceived severity of a health threat (e.g., COVID-19 virus) or treatment (e.g. COVID-19 vaccine), perceived self- efficacy or perceived response efficacy.[32] Based on the results of the analysis, counterstrategies and messaging are designed that also take into consideration barriers to people and place as well as cultural and community-related nuances. Evaluation is the last part of the framework that includes formative evaluation of the messaging and the effectiveness of counter messaging strategies.

This current study aims to determine whether a model could be developed to predict vaccine sentiment based on the types of persuasion used int the messaging. Based on this, the authors proposed the first research question:

> RQ1. Can we build a model that predicts vaccine sentiment based on the types of persuasion used in the messaging?

When the HIPE™ framework was initially developed, persuasion variables and vaccine sentiment were manually coded. The challenge with this approach is the thousands of online posts on a health topic, presenting constraints and limitations related to amount of coding that can be done manually. The analysis of a greater amount of data would provide a more robust analysis of persuasion tactics across the spectrum of a disease. This led to the second research question:

> RQ2. Can machine learning be used to develop an algorithm to detect persuasion tactics in COVID-19 vaccine online discourse?

Methods

Two data sets were used for this study. The first data set consisted of 1000 tweets retrieved by the Social Integrity Platform between July 14-23, 2020 [3], which was a period of time when the COVID-19 vaccine discussion became the main theme of online discourse for COVID-19. A test for inter-rater reliability was conducted using Gwet's AC1 agreement coefficient metric[33] that resulted in an

inter-rater reliability score that was statistically significant (*p*-values were less than 0.05). A second set of Twitter data consisting of 1000 tweets were manually annotated by two of the same researchers, who annotated the first data set. This second data set was obtained using the same search strategy as the first data set. The Tweets were coded by three annotators as to the type of persuasion tactic that was used in the messaging and the type of vaccine sentiment. Some tweets were excluded in the course of the coding process because they were not in English or the link to the Tweet no longer worked. The final set of data consisted of 1,845 annotated tweets.

Predicting vaccine sentiment from persuasion type

There were two issues that we wanted to explore using machine learning on the manually annotated data created in this study. One question was which of the different types of persuasion messaging was most likely to predict an author's sentiment towards vaccination.

Models were built with machine learning algorithms to identify which combination of persuasion messaging will predict the sentiments towards the vaccine. The persuasion variables were used as independent variables, also known as features, in a classification approach leading to a prediction of the sentiments (the outcome). Sentiments are categorized into Provax, Antivax, and Neutral. Two algorithms were selected: Decision Tree and Random Forest, which are two types of machine learning appropriate for classification.[34] The R software was used for the analysis.

Decision Tree

Decision tree (DT is an algorithm that learns from the data, builds, and validates a model. It breaks down the data into smaller subsets resulting in a tree with decision nodes and leaf nodes. In this paper, a Recursive PARTitioning (RPART) library was used in the analysis of the data.[35]

Random Forest

Random forest (RF) is an algorithm that produces a collection of decision tress (DT's).[36] It is also known as an ensemble learning method. Prediction by one decision tree may not be accurate. Thus, combining many DT's improves accuracy of the prediction/classification, on the average. A specific number of DT's are trained based on Bootstrap samples, in a parallel fashion. The final classification is obtained through a majority vote fashion. The selected model extracts the most important features that influence the classification. Using a testing set, performance across all classes were assessed with precision and recall metrics. Furthermore, a macro F1 score is calculated. This score is derived from the macro-averaged precision and recall. It is suitable for data with imbalanced class distribution.

A DT was built on the entire dataset, using all the variables, whereas a RF was built on a training set (70%) and validated on a test set (30%). A RF randomly selected observations and specific variables to build multiple DT's and averages the results.

Predicting persuasion types automatically at scale

Another question that this research addressed was whether we could develop automated models that could detect persuasion types from the content of Tweets alone, whether such models could scale to the volume of material that is generated in social media, and whether the models would be sufficiently reliable that they could support the development of counter messaging in a timely fashion. As described above there were eighteen different types of persuasion that were identified in the data (Table 1).

Table 1. Persuasion Types

Theoretical Framework	Persuasion Constructs
Elaboration Likelihood Model (ELM)	High Elaboration (Deep processing) 1. Information/Statistics 2. Question 3. Participation Low Elaboration (Superficial Processing) 4. Celebrity 5. Inspiration 6. Humor/Sarcasm 7. Story
Social Judgment Theory (Values)	8. Health Evidence (Health/evidence-based information) 9. Safety 10. Religion 11. Choice 12. Political 13. Social Equity 14. Altruism
Extended Parallel Process Model (EPPM)	15. Fear Appeal (Perceived Severity) 16. Perceived susceptibility 17. Self-efficacy 18. Response Efficacy

We decided to approach the modeling problem using eighteen independent binary text classification tasks, with the assumption that this would allow each of the models to focus their discriminative capability on a single task, which could lead to better overall performance. In the analytic tool suite for which these models were developed, we selected the text classification modeling tool FastText [37] for its combination of accuracy, speed and modest computing requirements. Later, we conducted some additional experimentation using Huggingface Transformers, tuning on top of a pre-trained text classification model developed for tweet sentiment analysis. [38,39] The summary results are reported in the next section.

As can be seen in Table 1, many of the persuasion categories were detected at very low rates during manual annotation, which meant that the training data was sometimes strongly skewed towards the absence of a given persuasion category. If left unaddressed, the resulting models would strongly prefer predicting negative labels, with an overall accuracy rate that would be superficially high, but where the precision and recall measured relative to the positive categories would be very low. We took two approaches to address this problem of skewed training data. First, we trained the FastText prediction models on a subsampled population of negative exemplars that was no more than two times the number of positive exemplars. Second, to get a clear picture of the model behavior, we report precision, recall and F-measure in two different ways, treating separately the presence and absence of a persuasion strategy as the target class to be predicted, and then also report the macro average of the respective F-measures. This serves as a useful way of understanding the quality of the individual binary classification models beyond the simple accuracy metric. In a similar vein, we include the precision recall curve – area under the curve (PRC AUC) to allow for additional insight into the models' performance characteristics across the full range of precision-recall tradeoffs.

Another approach to addressing the issue of low positive exemplars was to introduce a more coarse-grained set of categories that merged some of the original categories and would therefore include a larger number of positive exemplars in those categories. Of course, for this to make sense the categories would have to be closely related conceptually. For this reason, we introduced two new persuasion categories that represented the merging of two sets of fine-grained distinctions: persuasion variables 1 (information & statistics), 2 (questions) and 3 (participation) were merged into

"Elaboration Likelihood Model – High", or what we refer to here as "Deep Processing" (via information and questions) and "Elaboration Likelihood Model – Low" or "Superficial Processing" (via celebrities and stories).

The performance numbers reported in the next section are the result of conducting 10-fold cross validation, with 15% of the data being held out for testing on each fold. For each of the binary classifiers we specified that FastText should use n-grams of up to length 3, a learning rate of 1.0, and to run for 25 epochs.

Results

Predicting vaccine sentiment from persuasion type

After removing rows with missing values and "0" value for sentiment, 1474 tweets were used in this analysis. Sentiments among the tweets are broken down as follows: Provax (52.71%), Antivax (19.74%), and Neutral (27.54%). The persuasion type Elaboration Likelihood Model (ELM) with values 1 to 7 was dichotomized for better representation of each group. For the single decision tree, at the root node (the topmost node) EPPM1 = 0 (No Fear Appeal) was displayed. It is selected to be the best/useful predictor of the classification. A set of *if-then-else* decision rules was then obtained from applying the algorithm (Figure 1). In the figure, the green node represents pro-vaccine sentiment; the red node represents anti-vaccine sentiment; and the yellow node represents neutral sentiment. The leaf nodes are displayed at the bottom of the tree. In each leaf node, the top number is the sentiment category, the next three numbers below the category are the percentage of each sentiment category in the data for that node. The last number is the predicted percent for each category. When the tree flow is read from top to bottom, the following decision rules from terminal nodes are extracted:

- If there is no Fear Appeal and there is Response Efficacy, then 27% are classified as Provax.
- If there is no Fear Appeal and no Response Efficacy, and there is Self-Efficacy, then 14% are classified as Provax.
- If there is no Fear Appeal, no Response Efficacy and no Self-Efficacy and there is Equity/Access or Altruism, then 5% are classified as Provax.
- If there is no Fear Appeal, no Response Efficacy and no Self-Efficacy, and there is Health Evidence or no Value identified and there is Superficial Processing (Story), then 3% are classified as Provax.
- If there is Fear Appeal, then 14% are classified as Antivax
- If there is no Fear Appeal, no Response Efficacy, and no Self-efficacy, and Value is Safety, Religion or Choice, and there is Superficial Processing (Story), then 2% are classified as Antivax
- If there is no Fear Appeal, no Response Efficacy and no Self-efficacy, and Value is Safety or Religion or Choice and there is no Superficial Processing (Story), and there is Superficial Processing (Humor/Sarcasm), then 1% are classified as Antivax
- If there is no Fear Appeal, no Response Efficacy and no Self-Efficacy, and Value is Political, then 16% are classified as Neutral
- If there is no Fear Appeal, no Response Efficacy and no Self-Efficacy, and there is Health Evidence or no Value identified, and there is no Superficial Processing (Story), then 15% are classified as Neutral
- If there is no Fear Appeal, no Response Efficacy and no Self-Efficacy, and Value is Safety, Religion or Choice, and there is no Superficial Processing (Story) and no Superficial Processing (Humor/Sarcasm), then 4% are classified as Neutral

For the collection of DTs, the RF, the data was split into training set (70%) and testing set (30%). The model achieved an overall accuracy of 78.51% (95% CI: 0.7588, 0.8098) using the training set. The top important feature was Self-Efficacy, a different best feature derived from a single decision tree (No Fear Appeal). One of the metrics produced by the algorithm is the Mean Decrease in Accuracy (MDA).[40] This measure, when ranked in descending order displays the top important features and assesses the usefulness of the persuasion variables techniques in the prediction of the sentiment (Figure 2). The MDA for Self-Efficacy was 71.8, which means that if Self-Efficacy is excluded from the features, the classification accuracy decreases by 71.8%. The larger the MDA value, the more important and useful is the feature in predicting/classifying the outcome. The next most important feature was Fear appeal with an MDA value of 68.9, showing a drop from 71.8. The least important feature was Story.

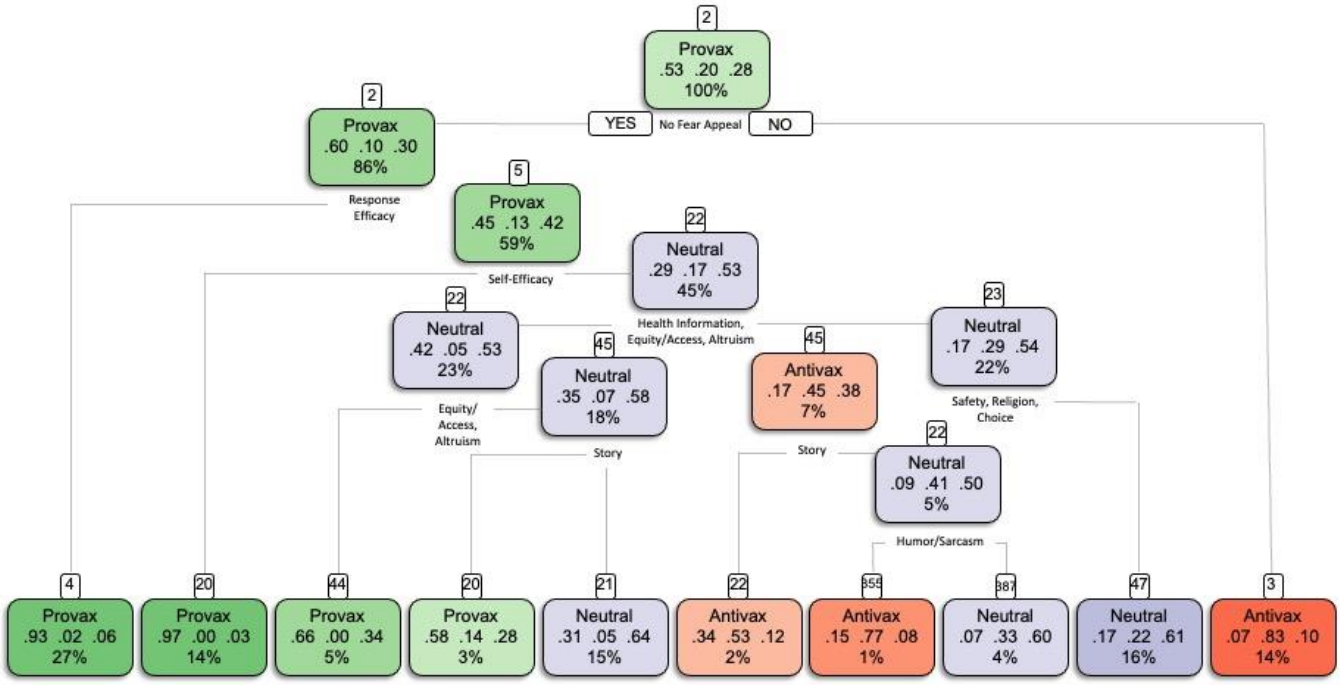

Figure 1: Classification Tree of Sentiments

The overall accuracy using the testing set was 74.6% (95% CI: 0.7027, 0.786). Note that the accuracy for a multiclass classifier is calculated as the average accuracy per class. The sensitivity metric showed that Provax sentiment is 78.5% accurate, Antivax sentiment is 55.2% accurate and Neutral Sentiment is 81% accurate. Precision (P) and Recall (R) were obtained for each class: Provax (P=0.88, R=0.79); Antivax (P=0.81, R=0.55); Neutral (P=0.56, R=0.81). Overall, the performance is high. However, the classifier underperforms for Antivax (Recall) and Neutral (Precision). Based on these Precision and Recall metrics, the macro F1 score was 0.73. This value indicates that the classifier performs reasonably well for each individual class. Results from DT and RF are slightly different. No Fear appeal was the top important feature for the single DT while it was the second most important feature for the RF.

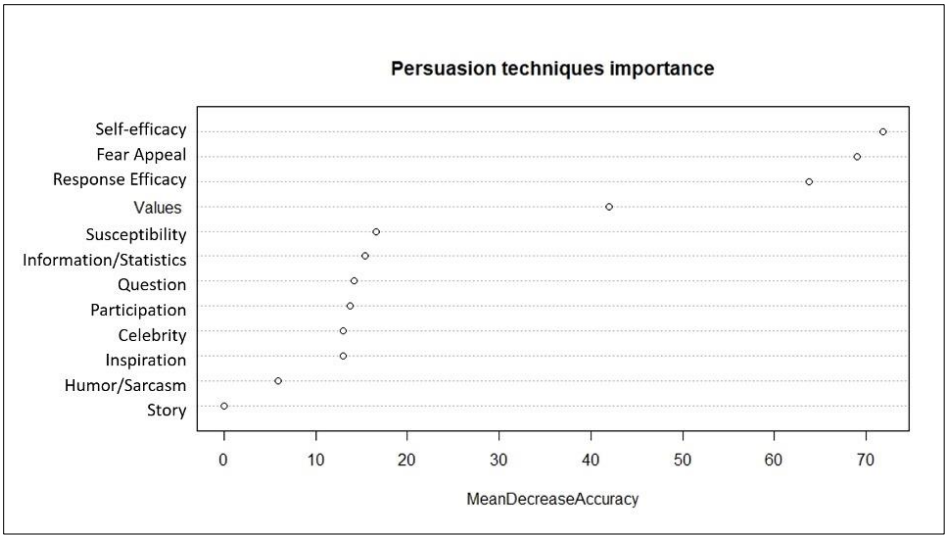

Figure 1. Persuasion Techniques Importance

Predicting persuasion types automatically at scale

Table 2 provides a summary of the performance of the FastText-based binary classifiers for each of the eighteen persuasion categories, as well as the merged persuasion categories of "Deep Processing" and "Superficial Processing." The table presents a range of metrics to support a careful consideration of how these models could be used in practice. The results are sorted in order of decreasing macro averaged F-Measures.

Table 2. Evaluation results of persuasion detection models.

	FastText-based Models									Huggingface-based Models			
persuasion category	positive precision	positive recall	positive f-measure	negative precision	negative recall	negative f-measure	macro avg fmeasure	accuracy	positive precision recall AUC	macro avg fmeasure	accuracy	category present count	category absent count
1. Information, Statistics	0.804	0.573	0.667	0.819	0.933	0.872	0.781	81.6%	0.919	0.801	82.4%	601	1,244
1-3. Deep Processing	0.769	0.626	0.689	0.749	0.856	0.798	0.750	75.6%	0.862	0.755	75.7%	787	1,057
18. Response Efficacy	0.756	0.384	0.504	0.847	0.964	0.902	0.732	83.6%	0.935	0.749	83.8%	399	1,446
8. Health Evidence	0.744	0.467	0.572	0.791	0.925	0.853	0.730	78.1%	0.896	0.779	80.9%	566	1,279
17. Self Efficacy	0.754	0.177	0.282	0.891	0.990	0.938	0.682	88.6%	0.952	0.777	90.0%	246	1,599
9. Safety	0.797	0.111	0.192	0.923	0.998	0.959	0.670	92.2%	0.947	0.579	92.6%	151	1,694
4-7. Superficial Processing	0.618	0.442	0.515	0.706	0.832	0.763	0.649	68.3%	0.796	0.687	70.2%	719	1,125
15. Fear Appeal	0.733	0.088	0.154	0.897	0.995	0.943	0.648	89.3%	0.951	0.708	91.3%	210	1,634
4. Celebrity	0.642	0.088	0.152	0.878	0.993	0.932	0.630	87.4%	0.926	0.760	89.4%	244	1,601
13. Social Equity	0.600	0.072	0.123	0.940	1.000	0.969	0.619	94.0%	0.959	0.515	93.3%	124	1,721
12 Political	0.582	0.095	0.163	0.841	0.986	0.908	0.614	83.4%	0.921	0.700	83.0%	351	1,494
7. Story	0.489	0.070	0.120	0.848	0.985	0.911	0.588	83.9%	0.904	0.585	81.3%	284	1,561
2. Question	0.350	0.036	0.065	0.941	0.997	0.968	0.565	93.9%	0.975	0.648	95.2%	106	1,739
16. Perceived Susceptability	0.283	0.041	0.070	0.958	0.998	0.977	0.559	95.6%	0.975	0.493	93.9%	84	1,761
3. Participation	0.200	0.020	0.036	0.963	1.000	0.981	0.533	96.2%	0.971	0.545	94.6%	80	1,765
11. Choice	0.200	0.013	0.025	0.925	0.996	0.959	0.524	92.2%	0.942	0.574	92.3%	137	1,708
5. Inspiration	0.000	0.000	0.000	0.981	1.000	0.990	0.495	98.1%	0.983	0.496	98.6%	28	1,817
14. Altruism	0.000	0.000	0.000	0.973	0.998	0.986	0.493	97.2%	0.985	0.491	96.3%	53	1,792
6. Humor, Sarcasm	0.000	0.000	0.000	0.910	0.997	0.951	0.476	90.7%	0.946	0.491	91.0%	164	1,681

A first observation is that, not surprisingly, the performance of the models is strongly correlated with the amount of positive training data available. Some of the persuasion categories have so few positive exemplars, such as "5. Inspiration" with 28, and "6. Altruism" with 53, that it is understandable that these paltry training amounts would strain the ability of the machine learning algorithm to identify the common semantics that could then be used to recognize unseen tweets that also make use of these very general types of persuasion. The correlation of positive training size and prediction performance is not absolute, however, which reflects the differing levels of complexity and variability by which a particular type of persuasion is employed in natural language. For example, the merged category of "4-7. Superficial Processing" has the second largest number of positive exemplars (719), but its inclusion of different types of persuasion, one of which is the challenging task of detecting humor and sarcasm, results in the prediction model performing well below other models with far fewer positive exemplars, such as "17. Self -Efficacy" and "9. Safety".

Overall, these performance results point to an optimistic takeaway: Those additional positive exemplars would likely contribute to additional improvement of the prediction models, especially those with the fewest numbers. To exploit this fact, we have developed an integrated analytic platform[3] in which analysts can easily identify additional positive and negative exemplars in the course of their work, which in turn supports an iterative retraining of the prediction models in a so-called "tag-a-little, learn-a-little" model improvement paradigm. [41]

The relatively high positive precision values for seven of these persuasion detection models, ranging from .733 to .804, indicates that these models could already play a useful role in the detection, tracking and mitigation of mis- and dis-information regarding vaccines as envisioned by this research. When applied against a large volume of on-topic tweets these models would be able to detect the ebb and flow of the associated types of persuasion, which in turn would allow for the types of intervention described elsewhere in this paper.

Subsequent to the development and experimentation with the analytic tools described above we conducted a small comparative study of the prediction modeling approach using Transformer technology. We tuned eighteen binary text classification models using this same data, tuning on top of a pre-trained Tweet sentiment prediction model.[39] As can be seen in the summary results included in Table 2, the Huggingface-based Transformer models showed modest improvement in 16 of the 19 categories, the macro averaged F-measures increasing by an average of 2 percentage points over all 19, indicating that this is a promising direction for adoption in future versions of our analytic environment. In most cases this improved performance was gained through an improvement in positive recall (not shown in the table).

Discussion

One of the aims of this study was to highlight the opportunities for a predictive model to determine persuasion patterns within vaccine online discourse. The application of machine learning technology advances critical insights to counter mis/disinformation and informs the development of data-driven counter-response strategies. Health misinformation is on the rise and will continue to pose a threat to public safety. It is imperative that public health communicators lean into technological advancements to provide critical insights to build effective strategies and behavioral interventions.

The persuasion variables were originally manually curated and shown to be credible. We developed two models for predicting sentiment based solely on the (manually annotated) type of persuasion being employed. The first model, using a decision tree approach, was effective in finding a combination of the persuasion messaging to classify an author's sentiment towards vaccination. The second model built with the random forest was able to extract a ranked list of persuasion variables ordered with mean decrease in accuracy. Based on precision, recall and F1 score, the latter model performs reasonably well. Model performance in both models was hampered by a combination of data sparsity and a significant skew in the distribution of categories.

In order to provide a fully-automated system we trained models to attempt to predict the persuasion variables themselves based on the content of the messages, as described in the previous section. As with the sentiment prediction models, it is clear that the performance of the sentiment models is limited first and foremost by the combination of limited amounts of training data for many categories, and the associated skew in the distribution of categories. Examining the performance metrics in Table 2, we see a strong correlation between positive category counts and the macro-averaged F1 scores.

These persuasion prediction models were created within the context of a larger analytic tool environment and one element of this system included the ability of analysts to view the results of model prediction on new text messages in the course of their work. The user interface allowed the analyst to either confirm or correct the predicted persuasion label, which could then be fed back into an iterative model re-training process. While not obligating the analyst to perform this annotation effort all the time, the integration of this confirmation/correction process into the working environment was intended to promote additional corpus creation. This so-called "tag-a-little, learn-a-little" iterative model building is one step towards increasing the amount of manually annotated data, which we hope will lead to still higher performance metrics going forward.

As noted elsewhere in this paper, there is an increasing amount of work developing predictive models to attempt to aid in the identification and amelioration of mis- and dis-information regarding community health. Some of the models being described cite higher performance metrics.[42,24] It is important to distinguish the different approach being pursued in our work, where we are attempting to detect the *types of persuasion* being employed by a communicator. These types of persuasion are available to anyone attempting to convey their point of view, whether or not they are actively spreading disinformation. In this way one can see that the natural language processing models needed for persuasion detection are more dissimilar to traditional sentiment analysis classification tasks, both in the fact that they extend a binary classification task to a multi-class task, but also in the types of natural language features they are likely to depend on.

The research community is encouraged to expand on these approaches to predict vaccine sentiments. Researchers can extend this study to a larger corpus of Twitter data and automate the annotation of the persuasion variables through Natural Language Processing (NLP). This work is a starting point to

further improve the identification of persuasion tactics, COVID-19 vaccine sentiment detection, and tracking and mitigation of mis- and dis-information in social media platforms.

Early Exposure of Persuasion Patterns

As this paper posits, advancements in the identification of persuasion patterns in social media posts could aid in limiting the spread of misinformation before widespread amplification. Misinformation, if not caught early, is resistant to counter response. Research indicates this is especially true if the information is challenging to an individual's world view.[43] Proactive counter strategies are a critical path to behavioral intervention and provide more successful inoculation of potentially harmful health misinformation.[3] Inoculation strategies prepare an individual to identify possibly deceptive or misleading messages and dismiss them, ultimately limiting the persuasive nature of a message, and decreasing the spread of misinformation.[3] Early exposure of persuasion patterns is an inherent benefit of the HIPE™ machine learning approach and will lessen the burden on public health communicators in their pursuit of improving health literacy and disseminating factual health information within their communities.

Prediction as a Means to Reduce Risk and Poor Health Outcomes

The analysis of persuasion classifications identified an opportunity to illustrate the prediction of sentiment and potential outcome. The HIPE™ framework proposes that the greater the number of persuasion tactics applied, the more complex the response must be. The analysis of classifications also revealed that some persuasion tactics, such as fear, self-efficacy, and response efficacy, have greater weight in terms of predicting sentiment. The combination of these tactics, therefore, would indicate a greater persuasive force. Early indicators of these persuasive tactics in messages could forewarn and forearm health communicators to focus energy on proactive response early on, reducing risk of serious harm and bolstering positive health outcomes. In parallel, health communicators could use this opportunity to design digital health education programs to focus on those particular topic areas.

Addressing Digital Health Literacy through Targeted Interventions

Bolstering digital health literacy plays a critical role in addressing health misinformation. Individuals with low digital health literacy are more susceptible to misinformation.[44] The World Health Organization states that the prevention of misinformation facilitates the delivery or access of reliable and comprehensible health information, driving optimal health outcomes.[45] The U.S. Department of Health and Human Services defines health literacy as "the degree to which individuals have the ability to find, understand, and use information and services to inform health-related decisions and actions for themselves and others."[46] Digital health literacy extends that definition to include the appraisal of health information from electronic sources. Digital health literacy includes ability to seek, find, understand, and appraise health information from electronic sources and apply the knowledge gained to addressing or solving a health problem.[47] If the persuasion algorithm can provide clarity to gaps in digital health literacy, one can create targeted education programs designed to fill the gaps in reliable health information, ultimately improving healthcare quality, reducing costs, and decreasing burden on the public health system. In addition, early identification of celebrity or influencer persuasive content, can serve as a critical strategy to drive individuals to more accurate health information, an important step to improving digital health literacy and health equity.[3,44] Individuals with low health literacy may not have sufficient capability to discern between accurate and false information and may engage in less critical processing of information. Therefore, celebrity or influencer connection with information provides a significant nudge or attractor towards certain viewpoints or types of information.

Conclusion

Combining predictive modeling and machine learning with persuasive messaging can serve as an innovative approach to understanding COVID-19 vaccine online discourse, and effectively be scaled to address mis/disinformation for health information of the future at the local, national and global level. The analysis of the data obtained from online discourse can be used to provide evidence-based counterstrategies for health mis/disinformation to achieve positive health behaviors and health outcomes especially during a public health crisis.

References

1. Diggins FW. The true history of the discovery of penicillin, with refutation of the misinformation in the literature. *Br J Biomed Sci*. 1999;56(2):83-93.
2. Koslap-Petraco M. Vaccine hesitancy: Not a new phenomenon, but a new threat. *J Am Assoc Nurse Pract*. Nov 2019;31(11):624-626. doi:10.1097/jxx.0000000000000342
3. Scannell D, Desens L, Guadagno M, et al. COVID-19 Vaccine Discourse on Twitter: A Content Analysis of Persuasion Techniques, Sentiment and Mis/Disinformation. *Journal of Health Communication*. 2021/07/03 2021;26(7):443-459. doi:10.1080/10810730.2021.1955050
4. Krishna AT, Teresa L. . Misinformation About Health: A Review of Health Communication and Misinformation Scholarship. *American Behavioral Scientist* 2021;65(2):316-332.
5. Vosoughi S, Roy D, Aral S. The spread of true and false news online. *Science*. 2018/03/09 2018;359(6380):1146-1151. doi:10.1126/science.aap9559
6. Kim DKD, Kreps GL. An Analysis of Government Communication in the United States During the COVID-19 Pandemic: Recommendations for Effective Government Health Risk Communication. *World Med Health Policy*. 2020:10.1002/wmh3.363. doi:10.1002/wmh3.363
7. Broniatowski D, Jamison A, Qi S, et al. Weaponized health communication: Twitter bots and Russian trolls amplify the vaccine debate. *American Journal of Public Health*. 2018;108(10):1378-84.
8. Organization WH. Infodemic. March 1, 2022, 2022. Accessed 03/01/2022, 2022. https://www.who.int/health-topics/infodemic#tab=tab_1
9. Wilson SL, Wiysonge C. Social media and vaccine hesitancy. *BMJ Global Health*. 2020;5(10):e004206. doi:10.1136/bmjgh-2020-004206
10. Security H. Social Media Bots Overview. May 2018 2018;
11. Yuan X, Schuchard R, Crooks A. Examining Emergent Communities and Social Bots Within the Polarized Online Vaccination Debate in Twitte. *Social Media + Society*. April 2019;
12. Blankenship E, Goff M, Yin J, et al. Sentiment, Contents, and Retweets: A Study of Two Vaccine-Related Twitter Datasets. *Permanente journal*. 2018;22:17-138.
13. Berger J, Milkman KL. What Makes Online Content Viral? *Journal of Marketing Research*. 2012/04/01 2012;49(2):192-205. doi:10.1509/jmr.10.0353
14. Veale HJ, Sacks-Davis R, Weaver ERN, Pedrana AE, Stoové MA, Hellard ME. The use of social networking platforms for sexual health promotion: identifying key strategies for successful user engagement. *BMC Public Health*. 2015/02/06 2015;15(1):85. doi:10.1186/s12889-015-1396-z
15. Guidry JP, Carlyle K, Messner M, Jin Y. On pins and needles: how vaccines are portrayed on Pinterest. *Vaccine*. Sep 22 2015;33(39):5051-6. doi:10.1016/j.vaccine.2015.08.064
16. Krause RJ, Rucker DD. Strategic Storytelling: When Narratives Help Versus Hurt the Persuasive Power of Facts. *Personality and Social Psychology Bulletin*. 2020/02/01 2019;46(2):216-227. doi:10.1177/0146167219853845
17. Blankenship K, Wegener D. Opening the Mind to Close It: Considering a Message in Light of Important Values Increases Message Processing and Later Resistance to Change. *Journal of Personality and Social Psychology*. 03/01 2008;94:196-213. doi:10.1037/0022-3514.94.2.94.2.196
18. Moran M, Lucas M, Everhart K, Morgan A, Prickett E. What makes anti-vaccine websites persuasive? A content analysis of techniques used by anti-vaccine websites to engender anti-vaccine sentiment. *Journal of Communication in Healthcare*. 10/03 2016;9:1-13. doi:10.1080/17538068.2016.1235531
19. Bester JC. Vaccine Refusal and Trust: The Trouble With Coercion and Education and Suggestions for a Cure. *Journal of Bioethical Inquiry*. 2015/12/01 2015;12(4):555-559. doi:10.1007/s11673-015-9673-1
20. Singh L, Bansal S, Bode L, et al. A first look at COVID-19 information and misinformation sharing on Twitter. *ArXiv*. 2020:arXiv:2003.13907v1.
21. School HK. Misinformation Review. *Creative Commons Attribution 40 International (CC BY 40)* 2020;1(8)
22. Alenezi MN, Alqenaei ZM. Machine Learning in Detecting COVID-19 Misinformation on Twitter. *Future Internet*. 2021;13(10):244.
23. Hayawi KS, S.; Serhani, M.A.; Ta;eb. O/,; Mathew, S. ANTi-Vax: a novel Twitter dataset for COVID-19 vaccine misinformation detection. *Public Health*. 2022;203:23-30.

24. Hayawi K, Shahriar S, Serhani MA, Taleb I, Mathew SS. ANTi-Vax: a novel Twitter dataset for COVID-19 vaccine misinformation detection. *Public Health*. 2022/02/01/ 2022;203:23-30. doi:https://doi.org/10.1016/j.puhe.2021.11.022
25. Krešňáková VM, Sarnovský M, Butka P. Deep learning methods for Fake News detection," 2019 IEEE 19th International Symposium on Computational Intelligence and Informatics and 7th IEEE International Conference on Recent Achievements in Mechatronics, Automation, Computer Sciences and Robotics (CINTI-MACRo. 2019;
26. Choudrie J, Banerjee S, Kotecha K, Walambe R, Karende H, Ameta J. Machine learning techniques and older adults processing of online information and misinformation: A covid 19 study. *Computers in Human Behavior*. 2021/06/01/ 2021;119:106716. doi:https://doi.org/10.1016/j.chb.2021.106716
27. Jamison A, Broniatowski DA, Smith MC, et al. Adapting and Extending a Typology to Identify Vaccine Misinformation on Twitter. *American Journal of Public Health*. 2020;110(S3):S331-S339. doi:10.2105/ajph.2020.305940
28. Sear R, Velaxquez N, Leahy R, et al. Quantifying COVID-19 Content in the Online Health Opinion War Using Machine Learning. *IEEE*. 2020;8:91886-91893.
29. Hu T, Wang S, Luo W, et al. Revealing Public Opinion Towards COVID-19 Vaccines With Twitter Data in the United States: Spatiotemporal Perspective. *JOURNAL OF MEDICAL INTERNET RESEARCH*. 2021;23(9):e30854.
30. Petty RE, Cacioppo JT. The elaboration likelihood model of persuasion. *Communication and persuasion*. Springer; 1986:1-24.
31. Smith SW, Atkin CK, Martell D, Allen R, Hembroff L. A social judgment theory approach to conducting formative research in a social norms campaign. *Communication Theory*. 2006;16(1):141-152.
32. Witte K, Allen M. A meta-analysis of fear appeals: Implications for effective public health campaigns. *Health education & behavior*. 2000;27(5):591-615.
33. Wongpakaran N, Wongpakaran T, Wedding D, Gwet KL. A comparison of Cohen's Kappa and Gwet's AC1 when calculating inter-rater reliability coefficients: a study conducted with personality disorder samples. *BMC Medical Research Methodology*. 2013/04/29 2013;13(1):61. doi:10.1186/1471-2288-13-61
34. Breiman L, Friedman JH, Olshen RA, Stone CJ. *Classification and regression trees*. Routledge; 2017.
35. Therneau T, Atkinson B, Ripley B. Rpart: Recursive Partitioning. R Package Version 4.1-3. 2022. http://CRAN.R-project.org/package=rpart
36. Liaw A, Wiener M. Classification and Regression by RandomForest. *Forest*. 11/30 2001;23
37. Joulin A, E. G, Bojanowski P, Mikolove T. Conference Proceedings. 2017:427-431.
38. Debut L, Sanh V, Chaumond J, et al. Transformers: State-of-the-Art Natural Language Processing. Association for Computational Linguistics; 2020:38-45.
39. Pérez JM, Giudici JC, Luque F. A Python Toolkit for Sentiment Analysis and SocialNLP tasks. arXiv:2106.09462v1 [cs.CL] https://arxiv.org/abs/2106.09462
40. Hong H, Xiaoling G, Hua Y. Variable selection using Mean Decrease Accuracy and Mean Decrease Gini based on Random Forest. 2016:219-224.
41. Aberdeen J, Bayer S, Yeniterzi R, et al. The MITRE Identification Scrubber Toolkit: design, training, and assessment. *Int J Med Inform*. Dec 2010;79(12):849-59. doi:10.1016/j.ijmedinf.2010.09.007
42. Abdelminaam DS, Ismail FH, Taha M, Taha A, Houssein EH, Nabil A. CoAID-DEEP: An Optimized Intelligent Framework for Automated Detecting COVID-19 Misleading Information on Twitter. *IEEE Access*. 2021;9:27840-27867. doi:10.1109/access.2021.3058066
43. Cook J, Lewandowsky S, Ecker UKH. Neutralizing misinformation through inoculation: Exposing misleading argumentation techniques reduces their influence. Article. *PLoS ONE*. 2017;12(5):1-21. doi:10.1371/journal.pone.0175799
44. Scherer LD, Pennycook G. Who Is Susceptible to Online Health Misinformation? *Am J Public Health*. Oct 2020;110(S3):S276-s277. doi:10.2105/ajph.2020.305908
45. Ilona Kickbusch I, Pelikan, J.M., Apfel, F., Tsouros, A.D. (Eds.) Health Literacy—The Solid Facts. WHO Regional Office for Europe: Copenhagen, Denmark; 2013. Accessed April 8, 2022. https://apps.who.int/iris/bitstream/handle/10665/128703/e96854.pdf
46. Promotion USDoHaHSOoDPaH. Health Literacy in Healthy People 2030.

https://health.gov/healthypeople/priority-areas/health-literacy-healthy-people-2030

47. Bin Naeem S, Kamel Boulos MN. COVID-19 Misinformation Online and Health Literacy: A Brief Overview. *International journal of environmental research and public health*. 2021;18(15):8091. doi:10.3390/ijerph18158091