INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY

RESEARCH ARTICLE

# Understanding Vaccine Hesitancy with Application of Latent Dirichlet Allocation to Reddit Corpora

**Samuel Duraivel[1]*, Lavanya[2], Aby Augustine[3]***

**1** Assistant Professor, Department of Media Studies, Kristu Jayanti College Autonomous, Bengaluru, India
**2** Assistant Professor, Department of Media Sciences, CEG Anna University, Chennai, India
**3** Kristu Jayanti College Autonomous, Department of Media Studies, Bengaluru

OPEN ACCESS

**Received:** 25-03-2022
**Accepted:** 04-09-2022
**Published:** 26-09-2022

**Citation:** Duraivel S, Lavanya , Augustine A (2022) Understanding Vaccine Hesitancy with Application of Latent Dirichlet Allocation to Reddit Corpora. Indian Journal of Science and Technology 15(37): 1868-1875. https://doi.org/10.17485/IJST/v15i37.687

* **Corresponding authors**.

duraivelsamuel@gmail.com

abyaugustine@kristujayanti.com

**Funding:** None

**Competing Interests:** None



Published By Indian Society for Education and Environment (iSee)

**ISSN**
Print: 0974-6846
Electronic: 0974-5645

## Abstract

**Objectives**: To understand the vaccine-hesitant behavior among social media user cohorts; to identify the underlying factors that contribute to vaccine hesitancy; to help policymakers make informed decisions to improve the success rate of vaccination campaigns. **Methods:** Latent Dirichlet Allocation (LDA)—a popular topic modeling technique—was used to extract topics from the Reddit corpus on vaccine hesitancy discussion. The corpus was extracted from Reddit's API using PRAW—The Python Reddit API Wrapper. The corpus contained 2996 comments, retrieved from the following subreddits:r/askreddit, r/antivax, r/antivaccine, and r/AntiVaxxers; determinants of Vaccine hesitancy were generated from the corpus using Standard LDA and Mallet LDA models. **Findings:** By applying Latent Dirichlet Allocation, we were able to identify the underlying factors that contribute either directly or indirectly toward vaccine-hesitant behavior. Some of the interesting factors of contribution include, but are not limited to rapture, depopulation agenda, the immigrant crisis at the Southern US border, etc. Given that the dataset we used contained a majority of input from people living in the United States, the results are rational; however, the same factors may or may not be contributors worldwide. The topics generated by the standard LDA were less precise and comprehensible than the topics generated by the mallet LDA model. Although a number of contributions have been made in this specific area i.e., understanding the vaccine-hesitant behavior, none report how political and religious factors contribute to the outcome. At a surface level, even though it is well-known that religious and political factors contribute to vaccine hesitancy, our unique Reddit corpus and the methodology as mentioned earlier let us identify fascinating and novel factors that have not been reported elsewhere. **Novelty:** Research to identify the factors that contribute toward vaccine hesitancy is fairly common, especially while the coronavirus pandemic was at its peak. Existing research works predominantly use surveys and other traditional methodologies to identify the factors that contribute toward the said phenomena. The application of Natural language Processing, viz., Latent Dirichlet Allocation could bring out the best latent

variables which cannot be identified using the aforementioned methodologies. This research is fairly novel by the methodology adopted and by the results obtained.

**Keywords:** Vaccine hesitancy; Coronavirus; Latent Dirichlet Allocation; Bayesian Statistics; Reddit

## 1 Introduction

Natural Language Processing (NLP) empowers intelligent machines by providing a better understanding of the human language for linguistic-based human-computer communication. Recent developments in computational technology and the advent of massive amounts of linguistic data have increased the need and demand for automating semantic analysis using data-driven approaches. Therefore, the utilization of data-driven strategies is pervasive now because of the significant advancements demonstrated by implementing deep learning methods in areas such as Automatic Speech Recognition, Computer Vision, and especially, NLP. This research uses Topic Modeling—an unsupervised machine learning technique that is capable of scanning a set of documents, detecting word and phrase patterns within them, and automatically clustering word groups and similar expressions that best characterize a set of documents—to identify the latent factors that contribute toward vaccine hesitancy and resistance to vaccination.

The hesitancy toward getting COVID-19 vaccines is common worldwide [(1)]. Research works have found multiple variables related to COVID-19 vaccine hesitancy in different domains. The identified factors include socioeconomic and demographic characteristics such as age, sex, residence, occupation, income, and marital status [(2)] constructs of the health belief model [(3)], constructs pertaining to the theory of planned behavior and the 5c psychological antecedents [(4)], vaccines-associated knowledge [(5)], attitude toward COVID-19 vaccination campaigns [(6)], conspiracy beliefs [(7)], trust and confidence [(8)], COVID-19 prevention measures [(9)], and the presumed safety and side effects of the vaccines. Despite the hesitancy toward vaccines, the demand for vaccines keeps increasing over time, and disparities in getting access to vaccines within and across the countries are evident [(10)]. Although the primary drivers of vaccine hesitancy are often context-specific, there are some theories which propose that confidence and trust in the COVID-19 vaccine play an important role in improving acceptance of the vaccines [(11)].

Although studies have found the factors that contribute toward vaccine hesitancy at a surface level, they do not pay attention to the minutiae of the said factors. The reluctance to get vaccinated may pose critical challenges not only for COVID-19 but for also the pandemics of the future. Therefore, to address this gap, we conducted a deep learning-based study to determine the latent variables that contribute to vaccine hesitancy and refusal. Given that the big data retrieved from the internet has enormous potential to enlighten the researchers with information that is not readily available or visible to the naked eye, we used the text analytics approach to rifle through the linguistic corpus associated with vaccine hesitancy and refusal discussions.

## 2 Methodology

We identified the subreddits where vaccine-hesitant discussions were common; the text corpus containing retrieved 2996 comments (n = 2996) was retrieved from the Application Programming Interface of Reddit by using the PRAW—Python Reddit API wrapper. tokenized, lemmatized, and processed for ambiguities. Using a standard LDA model and a mallet LDA model, distinct topics—which in turn indicate the reasons for vaccine hesitancy and refusal—were extracted from the corpus.

### 2.1 Data classification

The text data from Reddit API were retrieved into four documents, namely, documents 1, 2, 3, and 4, thus making the input for the LDA model. The unstructured data with headlines or titles of the posts, comments, and other metadata namely, timestamp and the username. However, excluding the comments, the rest were dropped while processing the corpus.

### 2.2 Data processing

The corpus was normalized, that is the strings were split into tokens; letters were converted from uppercase to lowercase; punctuation, accent marks, and other diacritics were stripped off, followed by the removal of stopwords. In addition to the standard stopwords of the Natural Language Processing Toolkit, we stripped the words "vaccine," "coronavirus," "covid," "covid19," "pandemic," "pfizer", "johnson," "astrazeneca." Our initial observation of the corpus using a word cloud showed that the aforementioned words constituted a major part of the corpus and would be tantamount to "collection words," although we did not use any collection words or query search to collect comments from Reddit's API. We rather used hyperlinks. In addition, we neither stemmed nor lemmatized the corpus as our initial observations indicated that lemmatization of our corpus altered the context of some of the words that we assumed were important for model building. To avoid missing out on information, we used an unlemmatized corpus for analysis.

### 2.3 Hyperparameter optimization

We used Gensim, which uses a fixed symmetric prior per topic [1/number of topics prior]. We did a series of sensitivity tests to determine the Dirichlet Alpha and eta hyperparameters, using both default values of the Gensim library and custom values for both the standard Latent Dirichlet Allocation model and Machine Learning for Language Toolkit model, using different coherence metrics as discussed in the following section.

### 2.4 Coherence Measures

For our evaluation, we consider (i) The UCI measure and (ii) The UMass measure, both of which have been shown to match well with human judgements of topic quality. These measures compute the coherence of a topic as the sum of pairwise distributional similarity scores over the set of topic words, V. This has been generalized as

coherence (V) = $\sum_{vi,vj}$ score $(v_i v_j, \varepsilon)$

where V is a set of words describing the topic and $\in$ indicates a smoothing factor which guarantees that score returns real numbers. The UCI metric defines a word pair's score to be the pointwise mutual information (PMI) between two words, i.e.,

score ($v_i$, $v_j$, $\varepsilon$) = $log \frac{P(vi, vj) + \varepsilon}{p(vi)p(vj)}$

The probabilities of words are computed by counting the co-occurrence frequencies of words in a sliding window over an external corpus, such as Wikipedia. To some extent, this metric can be thought of as an external comparison to known semantic evaluations. On the other hand, the UMass metric defines the score to be based on document co-occurrence:

score ($v_i$, $v_j$, $\varepsilon$) = $log \frac{D(vi, vj) + \varepsilon}{D(vj)}$

where D(x,y) counts the number of documents containing words x and y and D(X) counts the number of documents containing x. More importantly, the UMass metric computes these counts over the original corpus used to train the topic models, rather than an external corpus. This metric is more intrinsic in nature and it attempts to confirm that the models learned data known to be in the corpus.

## 3 Results and Discussion

### 3.1 Descriptive statistics of the processed corpus

The properties of the retrieved corpus before processing were as follows: the documents $d_1$, $d_2$, $d_3$, $d_4$ of the corpus contained, 277704 [$n_1$ = 277704], 283251 [$n_2$ = 283251], 113016 [$n_3$ = 113016], and 127846 [$n_4$ = 127846] words, respectively. When transformed into structured data, $d_1$, $d_2$, $d_3$, and $d_4$ contained 1064, 1077, 554, and 659 rows, respectively, with each row containing a distinct user-generated text or comment. 93 rows in $d_1$, 41 in $d_2$, 56 in $d_3$, and 89 in $d_4$ were found to have missing values and were dropped from the corpus. 27 Non-English entries from $d_1$, 13 from $d_2$, 17 from $d_3$, and 11 from $d_4$ were removed as well. 7 entries from $d_1$ and 4 from $d_4$ were removed for use of explicit verbiage. The number of rows in the documents $d_1$, $d_2$, $d_3$, and $d_4$ after initial processing were as follows: $d_1$, 937; $d_2$, 1023; $d_3$, 481; and $d_4$, 555, with a mean of 208.80 [$\mu 1$ = 208.80], 234.07 [$\mu 2$ = 234.07], 175.44 [$\mu 3$ = 175.44], and 162.96 [$\mu 4$ = 162.96] words per each structured row of the documents. The

descriptive statistics of the processed corpus are given in Figure 1.

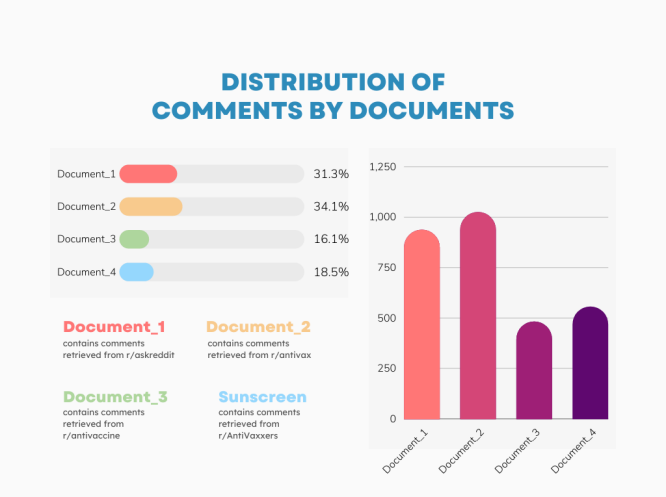


**Fig 1. Distribution of comments by documents**

## 3.2 Unigrams, Bigrams, and Trigrams

The Natural Language Toolkit identified 87891 [14.410%] distinct words from the tokenized corpus [d1, d2, d3, d4]. Frequent unigrams include, but are not limited to, [risk, 439], [clot, 437], [effect, 431], [side, 425], [blood, 423], [infertility, 419], [adverse, 406], [affect, 394], [mercury, 367], [thimerosal, 363], [experimental, 360], [cdc, 359], [sterilization, 345], [depopulation, 342], [surveillance, 320], [microchip, 311], [quantum, 280], [mark, 273], [beast, 273], [revelation, 272], [tribulation, 262], and [forehead, 242]. Similarly, 1118 distinct bigrams were identified by the Language Processing Toolkit. An analysis of the extracted bigrams showed a tight interconnection between the bigram components: most of the bigrams were stable phrases. A representative sample of the identified bigrams from the corpus is given in Table 1.

**Table 1. Representative sample of Bigrams**

| Bigrams and Frequencies | | |
|---|---|---|
| blood, clot, 229 | side, effect, 205 | adverse, risk, 203 |
| impair, fertility, 197 | contain, thimerosal, 193 | birth, defect, 189 |
| big, pharma, 189 | cover, up, 186 | drug, administration, 185 |
| fda, approval, 177 | gene, therapy, 174 | cdc, guidelines, 173 |
| quantum, dot, 170 | mark, beast, 163 | book, revelation, 155 |
| mass, surveillance, 153 | massachusetts, institute, 152 | police, state, 151 |
| mercury, based, 145 | based, preservative, 130 | mmr, autism, 117 |
| quell, population, 114 | depopulation, agenda, 102 | bill, gates, 102 |

In addition to the bigrams listed above, some of the other common bigrams observed in the corpus were [guinea, pig], [lab, rat], [warp, speed], [donald, trump], [anthony, fauci], [crony, capitalist], [invisible, ink], [genetic, experiment], [collateral, damage], [provax, cult], [edward, snowden], [fetal, tissue], [genetic, material], [trial, tribulation], [fast, track], [eugenics, board], and [coerced, sterilization]. Similar to that of the bigrams, the trigrams identified in the corpus showed a tight interconnection between the trigram components and most were stable phrases as well as shown in Table 2.

Other less frequent but informative trigrams observed in the corpus include, but are not limited to, [lack, long, term], [carolina, eugenics, board], [southern, texas, border], [mercury, cause, infertility], [no, miracle, drug], [big, pharma, lobbyist], [store, patient, history], [contain, toxic, ingredient], [lawsuit, against, fda],[implantable, tracking, chip], [immigration, detention, center], [totalitarian, police, state], [rigged, drug, committee], [long, term, research], [united, states, america], [fast, track, approval], [fluorescent, copper, based], [thimerosal, cause, clot], [coerced, hysterectomy, immigrant], [alexandria, ocasio, cortez], [human, rights, abuse], and [potential, side, effect].

**Table 2. Representative sample of Trigrams**

| Trigram | Frequency | Trigram | Frequency |
|---|---|---|---|
| cause, blood, clot | 99 | risk, side, effect | 98 |
| adverse, risk, reaction | 94 | long, term, effect | 89 |
| high, risk, group | 89 | mmr, cause, autism | 81 |
| human, guinea, pig | 77 | food, drug, administration | 70 |
| crony, capitalist, greed | 69 | quell, population, growth | 63 |
| nsa, surveillance, program | 56 | collect, personal, information | 53 |
| bible, book, revelation | 48 | Invisible, ink, tattoo | 47 |
| quantum, dot, dye | 43 | north, carolina, eugenics | 43 |

## 3.3 Results of Hyperparameter Optimization

We tested the Standard Latent Dirichlet Allocation model and Amherst's Machine Learning for Language Toolkit model [Mallet] for different values of alpha [symmetric, auto, 0.5] while keeping our eta as 0.01 [$\eta$ = 0.01] for all the implementations. The symmetric alpha for standard LDA is measured by dividing 1.0 by the total number of topics the model takes as the input, while the symmetric alpha for MALLET LDA is measured by dividing 5.0 by the total number of input topics. The results are given in Table 3.

The symmetric alpha for standard LDA is 0.125 for all the topics as the value is obtained by dividing 1.0 by the total number of topics [k = 8], that is [1.0/8 = 0.125], and the symmetric alpha is equally 0.625 for all the topics of mLDA as the value is obtained by dividing 5.0 by the total number of topics [5.0/8 = 0.625]. Further, as could be seen in Table 3, Gensim generated different "auto" alpha values for each topic of the standard LDA model with a mean of 0.2733 and a standard deviation of 0.0901. Likewise, the mean alpha of the mallet LDA is 0.26225 and a standard deviation of 0.142. We tested our LDA models for different hyperparameter values; however, we chose "auto" alpha over symmetric alpha because the latter may reduce the number of very small, poorly estimated topics, but may disperse common words over several topics. In addition, rather than deciding on fixed hyperparameters for the entire collection (with each topic having a similar probability in the model, and each word has a similar probability in each topic), it makes much more sense to allow for some differentiation between overall topic probabilities in a model: after all, it makes perfect sense that some topics are more general, and therefore widespread, whereas others are more specific and therefore less common. This intuition is implemented in the hyperparameter optimization function of Mallet.

**Table 3. Results of Hyperparameter Optimization**

| Model | Alpha[$\alpha$] | $k_1$ | $k_2$ | $k_3$ | $k_4$ | $k_5$ | $k_6$ | $k_7$ | $k_8$ |
|---|---|---|---|---|---|---|---|---|---|
| | symmetric | 0.125 | 0.125 | 0.125 | 0.125 | 0.125 | 0.125 | 0.125 | 0.125 |
| sLDA[c_v] | auto | 0.220 | 0.143 | 0.273 | 0.423 | 0.324 | 0.217 | 0.230 | 0.357 |
| | $\alpha$ = 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 |
| mLDA[c_v] | symmetric | 0.625 | 0.625 | 0.625 | 0.625 | 0.625 | 0.625 | 0.625 | 0.625 |
| | auto | 0.161 | 0.245 | 0.147 | 0.439 | 0.331 | 0.158 | 0.492 | 0.125 |
| | $\alpha$ = 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 |

## Results of Model Evaluation

Table 3 shows the coherence by a number of topics for standard LDA and machine learning for language toolkit models evaluated using c_v and UMass metrics. We tested the models for different values of k between 1 and 25, whereas the hyperparameters alpha and eta., were set as default. We observed that graphs of both standard and mallet LDA models evaluated using c_v metric were quite similar, and the graphs of standard LDA and mallet LDA models evaluated using UMass metrics were similar to each other as shown in Figure 2.

In c_v metric, the maximum value indicates the optimal topic coherence, whereas in the case of UMass metric, the value close to zero indicates the highest coherence. The highest coherence value estimated by the standard LDA model using c_v metric was 0.717 for the number of topics, k = 7. Likewise, the highest coherence value evaluated by the Machine learning for language toolkit model using c_v metric was 0.720 for the number of topics, k = 8. On the flipside, the closest value to 0 in the list of coherence values generated by sLDA model using UMass metric was 0.242 and the corresponding number of optimal topics

suggested by the model was 10 [k = 10]. The value closest to zero in the list of coherence values generated by the mLDA model was 0.018, for the number of topics, k = 8. Figure 2 shows how coherence values vary for different values of k [between 1 and 25]. We chose k= 8 as the optimal input for our LDA topic models based on our previous observations from hyperparameter optimization and coherence evaluation. Using the above criteria, we built a standard LDA model and a machine learning for language toolkit model [both using c v as the coherence metric], to predict the k number of topics and their corresponding word probabilities from our tokenized corpus. The results are discussed in the following section.

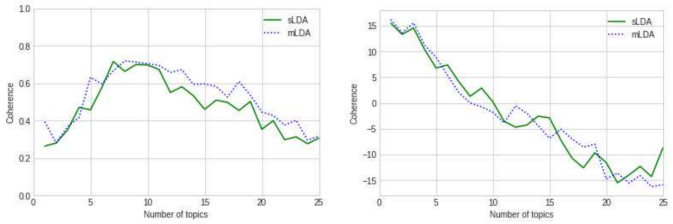


**Fig 2. Results of Hyperparameter Optimization**

## 3.5 Evaluation of generated topics

The properties of our topic model were as follows: a number of topics, k = 8; hyperparameters [alpha and eta] = set as default/auto, and coherence metric set as c_v. Topics generated by the standard LDA model are given in Table 4 and the topics generated by machine learning for the language toolkit model are given in Table 5. Close observation of the results generated by the models indicated that mLDA has outperformed standard LDA, in generating topics from the corpus.

**Table 4.** Topics generated by standard Latent Dirichlet Allocation model

| Topics | Probabilities of Words |
|---|---|
| Topic_1 Fear of risks and side effects | 0.0463*"risk" + 0.0457*"defect" + 0.0451*"clot" + 0.0436*"effect" + 0.0413*"birth" + 0.0367*"mmr" + 0.0367*"blood" + 0.0362*"side" + 0.0342*"contain" + 0.0310*"cause" + 0.0275*"serious" + 0.0212*"infertility" + 0.0154*"autism" + 0.0144*"mercury" + 0.0132"toxic" |
| Topic_2 Undefined | 0.0368*"women" + 0.0362*"miracle" + 0.0357*"guinea" + 0.0322*"border" + 0.0320*"cult" + 0.0320*"luciferase" + 0.0312*"quantum" + 0.0280*"fertility" + 0.0267*"toxic" + 0.0262*fast" + 0.0249*"program" + 0.0206*"pharma" + 0.0206*"lobbyist" + 0.0172*"fda" 0.0155*"risk" |
| Topic_3 Undefined | 0.0315*"administration" + 0.0302*"mercury" + 0.0277*"microchip" + 0.0267*"risk" + 0.0259*"people" + 0.0255*"committee" + 0.0247*"paternalism" + 0.0223*"operation" + 0.0171*"sterilization" + 0.0171*"near" + 0.0159*"forehead" + 0.0138*"totalitarian" + 0.0123*"research" + 0.0117*"christ" + 0.0101*"term" |
| Topic_4 Undefined | 0.0282*"birth" + 0.0261*"contain" + 0.0236*"federal" + 0.0222*"effect" + 0.0213"therapy" + 0.0198"track" + 0.0196"melinda" + 0.0196"based" + 0.0178"global" + 0.0162"america" + 0.0157"cause" + 0.0141"population" + 0.0139"choice" + 0.0109"million" + 0.0100"days" |
| Topic_5 Undefined | 0.0313*"thimerosal" + 0.0279*"preservative" + 0.0247*"monitor" + 0.0246*"revelation" + 0.0245*"copper" + 0.0241*"store" + 0.0238*"approval" + 0.0234*"growth" + 0.0234*"warp" + 0.0220*"infertility" + 0.0215*"free" + 0.0180*"history" + 0.0146*"era" + 0.0111*"fda" + 0.0106*"northrup" |
| Topic_6 Lack of trust in policymakers | 0.0342*"big" + 0.0332*"greed" + 0.0301*"pharma" + 0.0298*"food" + 0.0289*"administration" + 0.0287*"rig" + 0.0264*"lobby" + 0.0250*"approval" + 0.0231*gene" + 0.0178*"drug" + 0.0159*"trial" + 0.0137*"capitalism" + 0.0133*"cdc" + 0.0127*"mmr" |
| Topic_7 Related to Evangelicalism | 0.041*"mark" + 0.036*"book" + 0.033*"beast" + 0.032*"revelation" + 0.030*"bible" + 0.030*"forearm" + 0.029*"tribulation" + 0.023*"end" + 0.022*"rapture" + 0.021*"jesus" + 0.020*"" + 0.020*"forehead" + 0.018*"heaven" + 0.017*"earth" + 0.016*"submission" |
| Topic_8 Undefined | 0.0390*"blood" + 0.0388*"program" + 0.0339*"abuse" + 0.0329*"gates" + 0.0300*"tattoo" + 0.0297*"thimerosal" + 0.0284*"totalitarian" + 0.0268*"ingredient" + 0.0267*"long" + 0.0258*"impair" + 0.0254*"clot" + 0.0237*"dye" + 0.0233*"texas" + 0.0226*"affect" + 0.0220*"computer" |

The standard LDA model, despite a high coherence [coherence(c_v) = 0.717], did not generate coherent topics, except for three as shown in Table 5. The topics we observed to be coherent were as follows: fear of risks and side effects, lack of trust in policymakers, and related to Evangelicalism. The words in Topic 1 are fit to be collectively classified as "Fear of Risks and Side

Effects." Similarly, the words observed in Topics 4 and 5 are fit to be collectively categorized as "Lack of Trust in Policymakers" and "Related to Evangelicalism," respectively. Close observation of other topics indicates that some of the topics are partially coherent, whereas some are erratic with words mixed up with zero possibility of any coherence at all. On the flipside, the Machine Learning for Language Toolkit model surprisingly did a fair job of generating topics from our topics as shown in Table 5.

**Table 5.** Topics generated by Machine Learning for Language Toolkit Model

| Topics | Probabilities of Words |
|---|---|
| Topic_1 Fear of risks and side effects | [('risk', 0.03651699416016036), ('cause', 0.03416314345622569), ('adverse', 0.03376193394548789), ('toxic', 0.03349653981483784), ('defect', 0.03161460060918715), ('effect', 0.031116584045796432), '(clot', 0.030452128545862475), ('infertility', 0.026381535925209865), ('mercury', 0.024102725540928408), ('thimerosal', 0.02363135165762211), ('side', 0.023254910018593124), ('autism', 0.02031834265271079), ('birth', 0.014355831788711413), ('ingredient', 0.012817913680383061) ('reproductive', 0.01139122654179521)] |
| Topic_2 Lack of trust in policymakers | [('fraud',0.04160710513392156), ('rig', 0.040107046775601604), ('greed', 0.03630838271380726), ('cdc', 0.03543552241527582), ('lobbyist', 0.03358137292511565), ('pharma', 0.030721936883064085), ('administration', 0.027592281554903772), ('fda', 0.023385528944218165), ('drug', 0.018388490208626714), ('trial', 0.015362626970138533), ('approval', 0.013498542393280726), ('dollar', 0.012126684594666027), ('corporate', 0.01212055499909571), ('capitalist', 0.011005494125115884), ('big', 0.010318057191469734)] |
| Topic_3 Related to Evangelicalism | [('bible', 0.03852640089073605), ('book', 0.03802561366730776), ('christ', 0.036664853712672474), ('revelation', 0.03609395879758897), ('forehead', 0.03528351400888228), ('end', 0.034986601595521666), ('luciferase', 0.032973928728931096), ('satanic', 0.031928489929576004), ('mark', 0.02847247757264254), ('time', 0.024648387573605473), ('quantum', 0.022898259127506287), ('beast', 0.01934600128171001), ('tribulation', 0.0191886422323628), ('eschatology', 0.015536698207378994), ('rapture', 0.01206712536914099)] |
| Topic_4 Related to mass surveillance | [(surveillance, 0.04144496396347914), (track, 0.039090446901758766), (monitor, 0.03575479401654916), (collect, 0.03509909203313708), (personal, 0.029749268527953884), (information, 0.028418763034266187), (privacy, 0.02736910148176199), (right, 0.02535369022187812), (microchip, 0.02435556514431767), (nsa, 0.020991112927326326), (record, 0.0202667487078337), (snowden, 0.01869790794572588), (quantum, 0.01643426126592837), (citizen, 0.01572125082992414), (implant, 0.014661837846194781)] |
| Topic_5 Related to repression / authoritarianism | [('government', 0.039549297094926085), ('country', 0.03617621621697432), ('totalitarian', 0.03105098044403766), ('fascist', 0.03070493029799759), ('citizen', 0.03036062769542068), ('civil', 0.028329645909330348), ('liberty', 0.025787531384076363), ('autonomy', 0.023734561097302598), ('society', 0.02215329734816137), ('state', 0.020582283923119844), ('control', 0.0186593024465695), ('choice', 0.01834693980804135), ('personal', 0.016893070145626243), ('free', 0.014566149115439317), ('body', 0.012745287982758199)] |
| Topic_6 Related to population control | [('population', 0.041174128576727434), ('depopulation', 0.040962414094801676), (overpopulation, 0.04082622818035195), ('planet', 0.032320997235782446), ('reduce', 0.03164397857099357), ('quell', 0.030979481926013845), ('genealogy', 0.027243514580361214), ('sterilization', 0.0259916959773087), ('balance', 0.025568607349296262), ('eugenics', 0.022687870609774508), ('global', 0.021501305810269038), ('hysterectomy', 0.0199533606694739), ('agenda', 0.019623140811601488), ('warming', 0.014493937622447509), ('dna', 0.0134608059538107 2] |
| Topic_7 Related to race / racism / racial justice | [('african', 0.040608374821904006), ('american', 0.039760957547884015), ('black', 0.03903646933072922), ('people', 0.03786954068106863), ('women', 0.03525278482418869), ('latina', 0.03411672143718666), ('hispanic', 0.031438716015916905), ('xenophobic', 0.025478906689980867), ('klan', 0.023505808387450325), ('navajo', 0.02159963144636494), ('eugenics', 0.020151612196361857), ('carolina', 0.017912616570644246), ('paternalism', 0.013770685631721936), ('ableism', 0.013751592758233788), (sterilization, 0.010600351651251706)] |
| Topic_8 Related to immigration | [('immigration', 0.03724440723004234), ('immigrant', 0.036703318521539), ('border', 0.03306569400756612), ('ice', 0.0318088731617815), ('detention', 0.02969042320096986), ('asylum', 0.027163550842475105), ('center', 0.025214504060601422), ('processing', 0.02282206302145553), ('women', 0.018615667855827592), ('daca', 0.016693866884981846), ('refugee', 0.015647741688634035), ('southern', 0.015496586549257266), ('coerced', 0.012139762603643352), ('deport', 0.011989055708302492), ('hysterectomy', 0.010353391058588948)] |

We named the topics with appropriate labels as shown in Table 5 . Although few unrelated words were observed in Topics 7 and 8, the majority of the other words indicate that the topics are related to the racial system and immigration, respectively.

Both standard and MALLET LDA models generated topics related to "risks and side effects," "lack of trust in the policymakers," and "Evangelicalism."

However, the results of the standard LDA model indicate that words are mixed up except for three topics, and it gets erratic at the end. However, observation of the bigrams and trigrams indicate that the words coexist in the corpus, like "immigration" and "sterilization," which together make phrases and sentences that talk about the sterilization of immigrants in the ICE detention, etc. Although sterilization and immigration are totally different topics, their frequent coexistence of them in the corpus might have influenced the output generated by the standard LDA model. On the flipside, the topics generated by machine learning for the language toolkit model [Mallet] are less erratic and more precise in terms of outcome, leading to the discovery of eight latent topics from the tokenized corpus. Our results are fairly coherent with the previous studies that identified potential causes of vaccine hesitancy in the historical, political, and sociocultural contexts. Social science research has shown that vaccination decision-making should be understood in a broader sociocultural context [(12)]. In addition, Seth Mnookin—Journalist—explains how vaccination has become a source of fear and a target for misinformation [(13)]. Our research adds more context and empirical proof to the statement. Kata has shown that anti-vaccination websites shared common characteristics and used similar arguments and strategies to disseminate their message [(14)]. In this research, we identified the arguments that were made on Reddit to propel anti-vaccine propaganda that is specifically related to novel coronavirus vaccination campaigns.

## 4 Conclusion

We used Latent Dirichlet Allocation, an unsupervised generative-probabilistic machine learning model to discover the latent factors that contribute to vaccine hesitancy and resistance. Although our research focused on finding factors from populations across the world, our results indicate that the analyzed Reddit corpus has been generated by users predominantly from the United States. We used a standard LDA model and a MALLET-LDA model for a topic generation. The outcome of the standard LDA model was less precise and erratic when compared to the results of the mLDA model. We named the latent factors generated by the mLDA model with appropriate labels as shown in Table 5. We conclude that the primary contributors to vaccine hesitancy are fear of risks and side effects, lack of trust in policymakers, religious belief and background, conspiracy theories namely, mass surveillance, vaccination as precedence to totalitarianism, and depopulation agenda. Besides, an interesting finding was immigration deterrence and racial hate crime contribute toward vaccine hesitancy among the immigrant and minority population in conjunction with retrospective events of racial bias and injustice.

## References


1) Lin C, Tu P, Beitsch LM. Confidence and Receptivity for COVID-19 Vaccines: A Rapid Systematic Review. *Vaccines*. 2020;9(1):16–16. Available from: https://doi.org/10.3390/vaccines9010016.
2) Hossain MB, Alam MZ, Islam MS, Sultan S, Faysal MM, Rima S, et al. COVID-19 vaccine hesitancy among the adult population in Bangladesh: A nationwide cross-sectional survey. *PLOS ONE*;16(12):e0260821–e0260821. Available from: https://doi.org/10.1371/journal.pone.0260821.
3) Lin Y, Hu Z, Zhao Q, Alias H, Danaee M, Wong LP. Understanding COVID-19 vaccine demand and hesitancy: A nationwide online survey in China. *PLOS Neglected Tropical Diseases*;14(12):e0008961–e0008961. Available from: https://doi.org/10.1371/journal.pntd.0008961.
4) Hossain MB, Alam MZ, Islam MS, Sultan S, Faysal MM, Rima S, et al. Health Belief Model, Theory of Planned Behavior, or Psychological Antecedents: What Predicts COVID-19 Vaccine Hesitancy Better Among the Bangladeshi Adults? *Frontiers in Public Health*;9. Available from: https://doi.org/10.3389/fpubh.2021.711066.
5) Ruiz JB, Bell RA. Predictors of intention to vaccinate against COVID-19: Results of a nationwide survey. *Vaccine*. 2021;39(7):1080–1086. Available from: https://doi.org/10.1016/j.vaccine.2021.01.010.
6) Paul E, Steptoe A, Fancourt D. Anti-vaccine attitudes and risk factors for not agreeing to vaccination against COVID-19 amongst 32,361 UK adults: Implications for public health communications. *medRxiv*. Available from: https://doi.org/10.1101/2020.10.21.20216218.
7) Sallam M, Dababseh D, Eid H, Al-Mahzoum K, Al-Haidar A, Taim D, et al. High Rates of COVID-19 Vaccine Hesitancy and Its Association with Conspiracy Beliefs: A Study in Jordan and Kuwait among Other Arab Countries. *Vaccines*. 2021;9(1):42–42. Available from: https://doi.org/10.3390/vaccines9010042.
8) Rozek LS, Jones P, Menon A, Hicken A, Apsley S, King EJ. Understanding Vaccine Hesitancy in the Context of COVID-19: The Role of Trust and Confidence in a Seventeen-Country Survey. *International Journal of Public Health*. 2021;66:48–48. Available from: https://doi.org/10.3389/ijph.2021.636255.
9) Latkin CA, Dayton L, Yi G, Colon B, Kong X. Mask usage, social distancing, racial, and gender correlates of COVID-19 vaccine intentions among adults in the US. *PLOS ONE*. 2021;16(2):e0246970–e0246970. Available from: https://doi.org/10.1371/journal.pone.0246970.
10) Kothari A, Pfuhl G, Schieferdecker D, Harris CT, Tidwell C, Fitzpatrick KM, et al. The Barrier to Vaccination Is Not Vaccine Hesitancy: Patterns of COVID-19 Vaccine Acceptance over the Course of the Pandemic in 23 Countries. *medRxiv*. 2021. Available from: https://doi.org/10.1101/2021.04.23.21253857.
11) King I, Heidler P, Marzo RR. The Long and Winding Road: Uptake, Acceptability, and Potential Influencing Factors of COVID-19 Vaccination in Austria. *Vaccines*. 2021;9(7):790–790. Available from: https://doi.org/10.3390/vaccines9070790.
12) Streefland P, Chowdhury A, Ramos-Jimenez P. Patterns of vaccination acceptance. *Soc Sci Med*. 1999;49:239–246. Available from: http://dx.doi.org/10.1016/S0277-9536(99)00239-7.
13) Mnookin S. The Panic Virus: A True Story of Medicine, Science, and Fear. New York. 2011.
14) Kata A. Anti-vaccine activists, Web 2.0, and the postmodern paradigm – An overview of tactics and tropes used online by the anti-vaccination movement. *Vaccine*. 2012;30(25):3778–3789. Available from: http://dx.doi.org/10.1016/j.vaccine.2011.11.112.