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

Samuel Duraivel P*, DR Lavanya R, DR Samuel Gideon George P*****

*Department of Media Sciences, CEG - Anna University, Chennai
** Department of Media Sciences, CEG - Anna University, Chennai
***Department of Pharmacy Practice, Krupanidhi College of Pharmacy, Bangalore

Abstract- This research paper explores the underlying factors that contribute toward vaccine hesitancy, resistance, and refusal. Using Latent Dirichlet Allocation (LDA), an unsupervised generative-probabilistic model, we generated latent topics from user generated Reddit corpora on reasons for Vaccine hesitancy. Although we hoped to explore the grounds for vaccine hesitancy across the globe, our findings suggest that the corpus used for analysis had been generated by users living predominantly in the United States.Observation of the topics generated by the LDA model led to the discovery of the following latent factors: (i) fear of risks and side effects, (ii) lack of trust in policymakers, (iii) related to religious belief, (iv) related to mass surveillance theories, (v) perception of vaccination as a precedence to totalitarianism, (vi) racial background pertaining to retrospective events of racial injustice, such as selective sterilization, (vii) depopulation agenda fueled by theories affiliated to Global warming and extinction rebellion, (viii) and perception of vaccination as a campaign to quell immigrant population growth, fueled by reports of coerced sterilization of immigrants in the ICE detention.

Index Terms- Vaccine hesitancy; coronavirus; latent dirichlet allocation; bayesian statistics.

I. INTRODUCTION

Identifying and understanding Coronavirus vaccine hesitancy within distinct populations may be a hard task that requires a fair amount of experience in the field of psychology and human behavior. However, research on this particular area of study may aid future public health messaging. Hesitancy and resistance toward vaccines has been a subject of various studies throughout the past [1], [2], [3] and in the recent times of the coronavirus pandemic [4]. Survey has been the preferred method to observe and discover the factors that contribute to vaccine hesitant behavior among populations living within specified geographic locations. Using surveys, hypotheses are tested through analysis of the participant responses. However, discovering the latent contributors of an event or an outcome is almost minimal and difficult to attain through analysis of survey responses, if at all. In this research, we explore the underlying factors of vaccine hesitancy through application of Latent Dirichlet Allocation to user generated Reddit corpora on vaccine hesitancy and refusal. The Internet being the virtual cosmopolitan society aids and simplifies the information retrieval process from populations of diverse socio-demographic backgrounds; and recent advancements in the field of computational linguistics and Natural Language Processing favor a computerized approach [5] to analyze the massive data that remains available at large. Latent Dirichlet Allocation [6] is a Bayesian hierarchy topic model that generates topic keywords from the text corpus with efficiency and reduced complexity at the same time [7]. Besides, the LDA model characterizes the possibility that a document might have multiple topics, whilst unigram models assume the possibility that a given document has nothing more than a topic. In other words, the Latent Dirichlet Allocation model assumes a collection of K "topics." Each topic defines a multinomial distribution over the vocabulary and is assumed to have been drawn from a Dirichlet, $\eta_k \sim \text{Dirichlet}(\eta)$. Based on the topics, LDA assumes the following generative process for each document d. Foremost, the model draws a distribution over topics $\theta_d \sim \text{Dirichlet}(\alpha)$. Second, for each word i in the document, the model draws a topic index $z_{di} \in \{1,\ldots,K\}$ from the weight of the topics $z_{di} \sim \theta_d$ and draws the observed word ω_{di} from the selected topic, $\omega \sim \beta_{zdi}$. For the purpose of simplicity, symmetric priors are assumed on θ and β, but this assumption is easy to be relaxed [8]. Thus, in simple terms, LDA helps to explain the similarity of data by clustering features of the data into unobserved sets. A combination of these sets then constitute the observable data. The method can be applied to solve various tasks including, but is not limited to, topic identification [9], entity resolution [10], and Web spam classification [11].

II. RELATED WORK

According to a Canadian survey, although only 3 percent of parents refused all vaccines for their children, 19 percent consider themselves to be vaccine hesitant [12]. Vaccine-hesitant parents are a larger and more attentive group compared with vaccine refusers [13],[14]. Sixty-three percent of Canadian parents look for information about immunization on the Internet; of these, close to half perform a Google search [15]. A large number of antivaccine websites exist that propagate a range of anti-vaccine messages [16]. Much of the existing literature on vaccine resistance and hesitancy primarily focus on the explicit reasons why individuals choose not to get a particular vaccine or defy vaccination programmes in general [17], [18], [19], [20]. Survey has remained the preferred methodology to assess the underlying factors that contribute toward vaccine resistance. However,

exploratory analysis of opinionated text using Natural Language Processing techniques widens the horizon, leading to identification of latent factors that are less noticeable to the naked eye [21]. Analysis of lexical bundles to observe word combinations or co-occurrence of words, also referred to as "collocation" or "collocability" [22] has been used successfully in the past for information retrieval from text corpora [23], [24], [25]. In the context of machine learning and translation, lexical bundles or collocations are referred to as n-grams or Multi-word expressions (MWEs) [26] and are used in the weighting of topic models in mixture language model adaptation [27]. Internet web forums and social media platforms are a major resource of user generated text data, which when properly analyzed would result in discovery of latent, underlying factors that are otherwise obscure to human knowledge. Although the majority of the mainstream social media platforms censor controversial information related to vaccines, Reddit neither censors nor shuns users out of the platform for unpopular opinion related to vaccines. A goldmine of information, both bizarre and useful, can be found on the platform related to vaccines and a lot more other controversial topics. Unlike other social media platforms that rely on individuals connecting and interacting with people they know in the offline world, Reddit lets people connect based on things they care about [28]. This feature lets like minded people to discuss anonymously about things they care about, which they cannot in real life without being "cancelled" or "ostracized" for holding an unpopular opinion. Anti-vaccine discussions are rampant in Reddit with active subreddits dedicated to bringing vaccine hesitant people together from across the globe. Besides, controversial information spreads faster and further than non-controversial information in Reddit [29], thus attracting a wide variety of comments from users from diverse backgrounds.

III. METHODS

Methods of data collection, processing, and analysis of the corpus are discussed in this section. We used standard libraries of Python.

A. Data Retrieval

We collected comments from subreddits (r/askreddit, r/antivax, r/antivaccine, r/AntiVaxxers) that specifically discussed "the reasons not to get the vaccine." The Data were retrieved from Reddit using PRAW (Python Reddit API Wrapper), a Python package that allows access to the Application Programming Interface of Reddit [30]. The text data from Reddit API were retrieved into four documents, namely, documents 1,..,4, 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.

B. 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 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. Also, 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 important for model building. To avoid missing out information, we used an "unlemmatized" corpus for analysis.

C. Data Processing

The parameters of the prior are called hyperparameters. In LDA, the distribution of topics over documents and words have priors that are represented with alpha and beta respectively. The alpha parameter specifies prior beliefs about topic sparsity or uniformity in the documents and the beta hyperparameter controls the distribution of words per topic. Different packages use different notations for these hyperparameters and in Gensim they are denoted by alpha and eta. Besides, gensim 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.

D. Data Processing

Topic Coherence measures score a single topic by measuring the degree of semantic similarity between high scoring words in the topic. These measurements help distinguish between topics that are semantically interpretable and topics that are artifacts of statistical inference. For our evaluation, we consider (i) The UCI measure [31] and (ii) The UMass measure [32], 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

$$coherence(V) = \sum_{v_i, v_j} score(v_i, v_j, \in)$$

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.,

$$score(v_i, v_j, \in) = log \frac{p(v_i, v_j) + \in}{p(v_i)p(v_j)}$$

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:

$$score(v_i, v_j, \in) = log \frac{D(v_i, v_j) + \in}{D(v_j)}$$

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.

IV. RESULTS

The properties of the retrieved corpus before processing were as follows: the documents d1, d2, d3, d4 of the corpus contained, 277704 [n1 = 277704], 283251 [n2 = 283251], 113016 [n3 = 113016], and 127846 [n4 = 127846] words respectively. When transformed into structured data, d1, d2, d3, and d4 contained 1064, 1077, 554, and 659 rows respectively, with each row containing a distinct user-generated text or comment. 93 rows in d1, 41 in d2, 56 in d3, and 89 in d4 were found to have missing values and were dropped from the corpus. 27 Non-English entries from d1, 13 from d2, 17 from d3, and 11 from d4 were removed as well. 7 entries from d1 and 4 from d4 were removed for use of explicit verbiage. The number of rows in the documents d1, d2, d3, and d4 after initial processing were as follows: d1, 937; d2, 1023; d3, 481; and d4, 555, with a mean of 208.80 [μ1 = 208.80], 234.07 [μ2 = 234.07], 175.44 [μ3 = 175.44], and 162.96 [μ4 = 162.96] words per each structured row of the documents. The descriptive statistics of the processed corpus is given in Table 1.

Summary	Number of comments	Number of words	Words per row
Document_1	937	195646	208.80
Document_2	1023	239454	234.07
Document_3	481	84387	175.44
Document_4	555	90443	162.96
Sum	2996	609930	781.27
Mean [μ]	749.0	152482.5	195.31
Standard deviation [σ]	234.45	66919.77	27.95

Table 1: Descriptive statistics of the processed corpus

4.1.1. 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 2.

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	mercury, based, 145	mmr, autism, 117
quell, population, 114	depopulation, agenda, 102	bill, gates, 102

Table 2. Representative sample of Bigrams

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 3.

Trigrams and Frequencies			
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	million, dollar, business	68
operation, warp, speed	65	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

Table 3. Representative sample of Trigrams

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].

4.1.2. Results of Hyperparameter Optimization

We tested the Standard Latent Dirichlet Allocation model and Amhert's Machine Learning for Language Toolkit model [Mallet] for different values of alpha [symmetric, auto, 0.5] while keeping our eta as 0.01 [η = 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 [33]. The results are given in Table 4.

Model	Alpha[α]	k_1	k_2	k_3	k_4	k_5	k_6	k_7	k_8
sLDA [c v]	symmetric	0.125	0.125	0.125	0.125	0.125	0.125	0.125	0.125
	auto	0.220	0.143	0.273	0.423	0.324	0.217	0.230	0.357
	α = 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
	α = 0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5

Table 4. Results of Hyperparameter Optimization

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] [26]. Further, as could be seen in Table 5, 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 having 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 while others are more specific and therefore less common [11]. This intuition is implemented in the hyperparameter optimization function of Mallet [34].

4.1.3. Results of Model evaluation

Table 4 shows the coherence by 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, while 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 Figures 1 and 2.

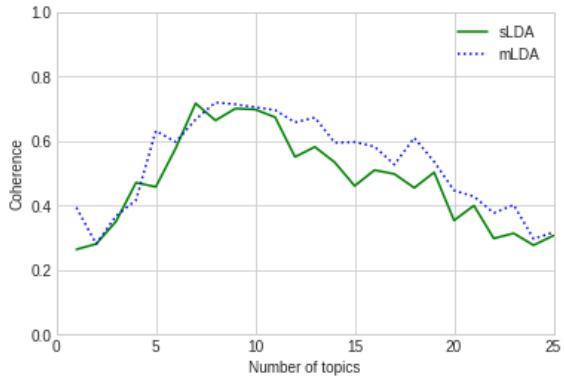

Fig 1. Topic coherence obtained using c v metric

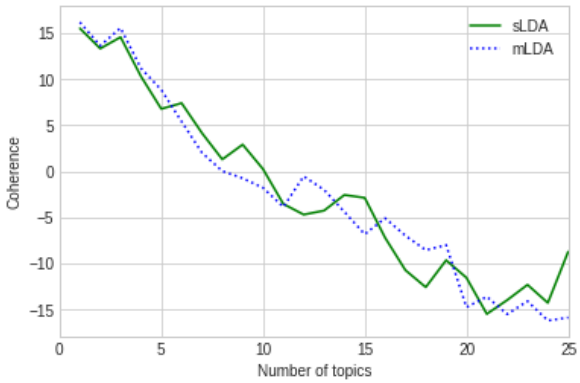

Fig 2. Topic coherence obtained using UMass metric

In c v metric, the maximum value indicates the optimal topic coherence [35], while in the case of UMass metric, the value close to zero indicates the highest coherence [35]. 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].

Number of Topics sLDA	sLDA (c v)	sLDA (UMass)	mLDA (c_v)	mLDA (UMass)
1	0.264	15.512	0.394	16.182
2	0.281	13.319	0.281	13.629
5	0.458	6.774	0.632	8.848
7	**0.717**	4.152	0.667	2.007
8	0.664	1.289	**0.720**	**0.018**
10	0.698	**0.242**	0.705	-1.793
11	0.674	-3.548	0.696	-3.872
15	0.461	-2.881	0.597	-6.837
20	0.354	-9.656	0.447	-14.762
25	0.307	-8.727	0.317	-15.881

Table 5. Results of Model Evaluation

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 1 and 2 show 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.

4.1.4 Evaluation of generated topics

The properties of our topic model were as follows: 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 6 and the topics generated by machine learning for the language toolkit model are given in Table 7. Close observation of the results generated by the models indicated that mLDA has outperformed standard LDA, in generating topics from the corpus.

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 6. The topics we observed to be coherent were, 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 6 and 7 are fit to be collectively categorized as "Lack of Trust in Policymakers" and "Related to Evangelicalism" respectively. Close observation of other topics indicate that some of the topics are partially coherent, while 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 7.

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"

Table 6. Topics generated by standard LDA 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', 0.012067125369140 99)]
Topic 4 Related to mass surveillance	[(surveillance, 0.04144496396347914), (track, 0.039090446901758766), (monitor, 0.03575479401654916), (collect, 0.03509090203313708), (personal, 0.029749268527953884), (information, 0.028418763034266187), (privacy, 0.02736910148176199), (right, 0.02535369022187812), (microchip, 0.024355565144317 67), (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', 0.01449393762244739), ('dna', 0.01346080595381072)]
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', 0.03143871601591690), ('xenophobic', 0.025478906689980867), ('klan', 0.023505808387450325), ('navajo', 0.021599631446364944), ('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', 0.022822063021455553), ('women', 0.018615667855827592), ('daca', 0.016693866884981846), ('refugee', 0.015647741688634035), ('southern', 0.015496586549257266), ('coerced', 0.012139762603643352), ('deport', 0.011989055708302492), ('hysterectomy', 0.010353391058588948)]

Table 7. Topics generated by mLDA Model

We named the topics with appropriate labels as shown in Table 7. Although few unrelated words were observed in Topic 7 and 8, the majority of the other words indicate that the topics are related to 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 sterilization of immigrants in the ICE detention, etc. Although sterilization and immigration are totally different topics, the 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.

V. CONCLUSION

We used Latent Dirichlet Allocation, an unsupervised generative-probabilistic machine learning model to discover the latent factors that contribute toward 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 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 7. We conclude that the primary contributors of 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 a 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 [36, 37].

Conflict of Interests

The authors declare no conflict of interest.

REFERENCES

[1] [1] Dube, E., Laberge, C., Guay, M., Bramadat, P., Roy, R., Bettinger, J.A. (2013) Vaccine hesitancy. Human Vaccines Immunotherapeutics. 9:8. 1763-1773.

[2] [2] MacDonald, N.E., the SAGE Working Group on Vaccine Hesitancy. (2015). Vaccine. 33. 4161- 4164.

[3] [3] Jacobson, R.M., Sauver, J.L., Rutten, L.F. (2015) Vaccine Hesitancy. Mayo Clinic Proceeding. 90. 1562-1568.

[4] [4] Murphy, J., Vallieres, F., Bentall, R.P., Shevlin, M., McBride, O., Hartman, T.K., McKay, R. Bennett, K., Mason, L., Miller, J.G., Levita, L., Martinez, A.P., Stocks, T.V., Karatzias, T., Hyland, P. (2021) Psychological characteristics associated with COVID-19 vaccine hesitancy and resistance in Ireland and the United Kingdom. 12, Article number: 29 (2021).

[5] [5] President, E.D. (2005) Enhanced Text Retrieval Using Natural Language Processing. Bulletin of the American Society for Information Science and Technology. 24. 14-16.

[6] [6] Fei-Fei, L., & Perona, P. (2005) A Bayesian hierarchical model for learning natural scene categories. Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 05). 2. 524–531.

[7] [7] Blei, D., Ng, A., Jordan, M., (2003) Latent Dirichlet allocation. Journal of Machine Learning Research. 3. 993–1022.

[8] [8] Wallach, H., Mimno, D., & McCallum, A., (2009) Rethinking lda: Why priors matter. Advances in Neural Information Processing Systems. 22. 1973–1981, 2009.

[9] [9] Griffiths, T.L., & Steyvers. (2004) Finding Scientific Topics. Proceedings of the National Academy of Sciences of the United States of America. 101. 5228-5235.

[10] [10] Bhattacharya, I., & Getoor, L. (2006) Proceedings of the Sixth SIAM International Conference on Data Mining. 47-58.

[11] [11] Biro, I., Siklosi, & DSzabo. (2009) Proceedings of the Fifth International Workshop on Adversarial Information Retrieval on the Web. 37-40.

[12] [12] Dube, E., Bettinger, J.A., Fisher, W.A., Naus, M., Mahmud, S.M., Hilderman, T. (2016) Vaccine acceptance, hesitancy and refusal in Canada: Challenges and potential approaches. Canada communicable disease report = Relevé des maladies transmissibles au Canada. 42(12). 246-251.

[13] [13] Leask, J. (2011) Target the fence-sitters. Nature. 473. 443-445.

[14] [14] Busby, C., Jacobs., Muthukumaran, R. (2017) In Need of a Booster: How to Improve Childhood Vaccination Coverage in Canada. Commentary 477. Toronto, ON: CD Howe Institute.

[15] [15] Ekos Research Associates Inc. (2011) Survey of parents on key issues related to immunization. Final report. Ottawa, ON: EKOS Research Associates Inc. Available from: www.ekospolitics.com/articles/0719.pdf.

[16] [16] 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. 30. 3778-3789.

[17] [17] Schmid, P., Rauber, D., Betsch, C., Lidolt, GDenker, M.L. (2017) Barriers of Influenza Vaccination Intention and Behavior – A Systematic Review of Influenza Vaccine Hesitancy, 2005 – 2016. PloS One. 12, e0170550.

[18] [18] Larson, H. J., Jarrett, C., Eckersberger, E., Smith, D. M. Paterson, P. (2014) Understanding vaccine hesitancy around vaccines and vaccination from a global perspective: a systematic review of published literature, 2007–2012. Vaccine 32, 2150–2159.

[19] [19] Siddiqui, M., Salmon, D.A. (2013). Epidemiology of vaccine hesitancy in the United States. Human vaccines Immunotherapeutics. 9. 2643-2648.

[20] [20] Marti, M., Cola, M.D., MacDonald, N.E., Dumolard, L., Duclos, P. (2017) Assessments of global drivers of vaccine hesitancy in 2014. Looking beyond safety concerns. PLOS ONE. 12(3):e0172310.

[21] [21] Manning, C.D., Utze, H.S., Lee, L. (2000) Foundations of Statistical Natural Language Processing. Firth, J.R. (1951) Modes of meaning essays and studies (The English Association).

[22] [22] Biber, D., Conrad, S., Cortes, V (2004) If you look at... Lexical bundles in university teaching and textbooks. Applied linguistics. 25. 371-405.

[23] [23] Eisenmann, B.H., Wagner, D., Cortes, V. (2010) Lexical bundle analysis in mathematics classroom discourse: the significance of stance. Educational studies in Mathematics. 75. 23-42.

[24] [24] Cortes, V. (2013) The purpose of this study is to: Connecting lexical bundles and moves in research article introductions. Journal of English for Academic Purposes. 12. 33-43.

[25] [25] Calzolari, N., Fillmore C.J., Grishman, R., Ide, N., Lenci, A., MacLeod, C., Zampolli, A. (2002) Towards Best Practice for Multiword Expressions in Computational Lexicons. Proceedings of the Third International Conference on Language Resources and Evaluation (LREC 02). L02-1259.

[26] [26] Haidar, M.A., O'Shaughnessy, D. (2011) Unsupervised language model adaptation using n-gram weighting. 24th Canadian Conference on Electrical

and Computer Engineering (CCECE). 000857- 000860, doi: 10.1109/CCECE.2011.6030578.

[27] Squirell, T. (2018) Reddit is the best social media site because it gets community right. Qz.com.

[28] [28] Jasser, J., Garibay, I., Scheinert, S., Mantzaris, A.V. (2020) Controversial information spreads faster and further than non-controversial information in Reddit. Journal of computational social science. https://doi.org/10.1007/s42001-021-00121-z

[29] [29] Boe B. PRAW: The Python Reddit API Wrapper. 2012-, https://github.com/praw-dev/praw/

[30] [30] Newman, D., Lau, J.H., Grieser, K., Baldwin, T. (2010) Automatic Evaluation of Topic Coherence. Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL. 100-108.

[31] [31] Mimno, D., Wallach, H.M., Talley, E., Leenders, M., McCallum, A. (2011) Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing. 262-272.

[32] [32] Mimno, D. (2004) Machine Learning with MALLET. http://mallet.cs.umass.edu/mallet-tutorial.pdf

[33] [33] McCallum, AK. (2002) Mallet: A machine learning for language toolkit. http://mallet. cs. umass. edu

[34] [34] Zhang, H., Zhu, Bingshan., Zheng, Changmeng., Yang, Kai., Wong, R.C.W., Li, Q (2020) Incorporating Concept Information into Term Weighting Schemes for Topic Models. International Conference on Database Systems for Advanced Applications. 227-244.

[35] [35] Rose, J. (2011) A Brutal Chapter In North Carolina's Eugenics Past. npr.org.

[36] [36] Dickerson, C., Wessler, S.F., Jordan.M. (2020) Immigrants Say They Were Pressured Into Unneeded Surgeries. The NewYork Times.

AUTHORS

First Author – Samuel Duraivel P, PhD Scholar, CEG - Anna University, duraivelsamuel@gmail.com

Second Author – DR Lavanya R, Assistant Professor, CEG - Anna University, lavanyaa2@gmail.com

Third Author – DR Samuel Gideon George P, Assistant Professor, Krupanidhi College of Pharmacy, psgsamuel@gmail.com

Correspondence Author – Samuel Duraivel P, PhD Research Scholar, Department of Media Sciences, CEG - Anna University, Chennai. Email - duraivelsamuel@gmail.com