**Topic Modeling and Sentiment Analysis of COVID-19 Vaccine Myth**

**Jihoon Jeong** and **Nianyou Zhang**

School of Information Study

Syracuse University

Syracuse, NY 13244 USA

{jijoeng, nzhang01}@syr.edu

**Abstract**

The purpose of this project is to identify people’s tendencies and thoughts about the COVID-19 vaccine. We obtained the Reddit users’ data from Kaggle(www.kaggle.com) and analyzed them. By the application of K-means and LDA topic modeling, we found the most frequently discussed topics about the COVID-19 vaccine. We also noticed the sentiment of the people by the application of the various algorithms.

**1 Introduction**

Ever since the availability of COVID-19 vaccines, there have also been many myths and misinformation spread around them. According to Dr. Christopher Murray, director of the University of Washington's Institute for Health Metrics and Evaluation, “Facebook runs a survey every day ... and that's shown that vaccine confidence in the US has been slowly but steadily going down since February.” He added, “We were at 75% of adults saying they wanted the vaccine. Now we're down to, in those surveys, about 67%.” There’s also a Facebook group called [COVID-19 vaccine Side Effects](https://www.facebook.com/groups/allvaccinesarefake) encouraging members to share their stories about the vaccine.

According to the survey from the Johns Hopkins Medicine office, the most common myths about the COVID-19 vaccine are the same as below:

* The COVID-19 vaccine can affect women’s fertility.
* If I’ve already had COVID-19, I don’t need a vaccine.
* Getting the COVID-19 vaccine gives you COVID-19.
* Researchers rushed the development of the COVID-19 vaccine, so its effectiveness and safety cannot be trusted.
* The side effects of the COVID-19 vaccine are dangerous.
* Getting the COVID-19 vaccine means I can stop wearing my mask and taking coronavirus precautions.

Most experts assert those beliefs are not true and need to clear up confusion with reliable facts. In this project, we focused on identifying people’s concerns about the COVID-19 vaccine and reading their tendency about the vaccination by analyzing Reddit users’ postings and comments about the vaccine.

**2 Method**

**2.1 Clustering / Topic Modeling**

In order to conduct the evaluation of the topics in the Reddit posts/comments about the COVID-19 vaccine, we used K-mean clustering and Latent Dirichlet Allocation (LDA)(Blei et al., 2003) for topic discovery. K-means is a clustering algorithm based on the nearest mean and LDA is an algorithm that can help to analyze the latent topic representation of a given corpus or dataset. It posits that each document is a mixture of a small number of topics and that each word's creation is attributable to one of the document's topics.

**2.2 Sentiment Analysis**

For the sentiment analysis, we applied three methods, Multinomial Naïve Bayes(MNB), Support Vector Machines(SVM), and Textblob, and compared the performance with each other. Since our dataset does not contain a lot of observations, we trained our model with a similar dataset from Kaggle.

**2.3 Data Set**

The dataset we used for this project was downloaded from Kaggle, but it was originally from Reddit, one of the biggest platforms where people share their ideas about a certain topic. In detail, data was extracted by PRAW(The Python Reddit API Wrapper) from VaccineMyths, a subreddit where people discuss various Vaccine Myths.

The data contains 1,489 unique data, including a small percent of harsh language, the posts were not filtered. Data contains both posts and comments and both posts and comments contain the following field.

**Table 1** Dataset fields

| title | title of the posting/comment |
| --- | --- |
| score | impact/number of comments |
| id | unique id |
| url | url of post thread |
| commns\_  num | number of comments to this post |
| created | date of creation |
| body | text of the post or comment |
| timestamp | timestamp |

**3 Analysis Procedures**

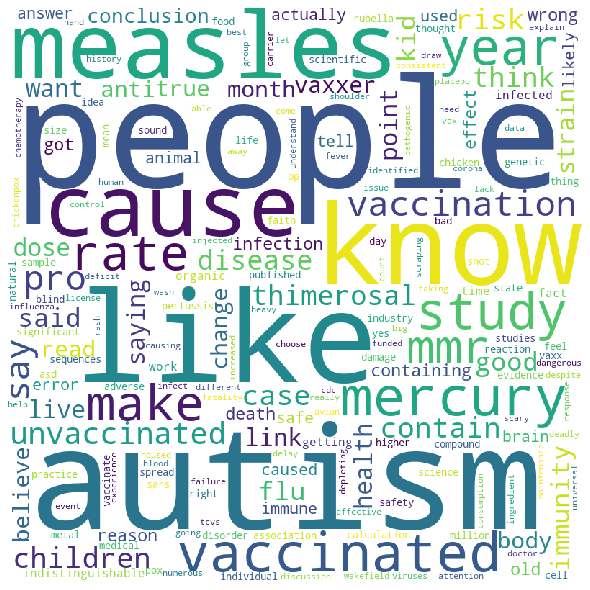
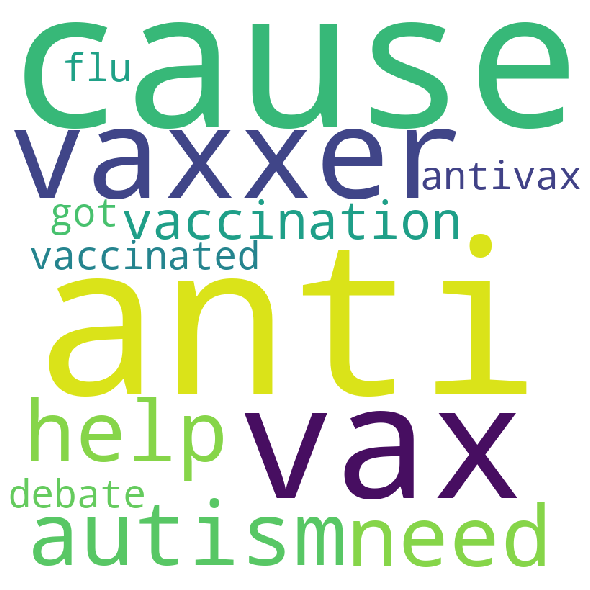
**3.1 Data Preprocessing**

We only focused on the ‘title’ and ‘body’ columns of the given dataset since these are directly related to the people’s opinion about the COVID-19 vaccine. Though this dataset had 1,489 unique observations, more than 300 hundred observations were NAs in the ‘body’ column, which needed to be dropped. It also included some unnecessary phrases and hyperlinks that needed to be removed before the main process.

**3.2 Exploratory Data Analysis**

In order to scan the whole dataset contents briefly, we created word clouds for the “title” and “body” columns. Since all data were about the COVID-19 vaccine, we excluded some expected frequent words, like a vaccine, vaccines, virus, and covid by adding them to the stopwords list.

**Figure 1** Word Clouds Results  
 (left: title / right: body)

****

One interesting point from the word clouds above was that both title and body columns had a common but unique word, autism. We speculated many people worried about the risk of having autism as a side effect of the COVID-19 vaccine.

**3.3 Clustering and Topic Labeling**

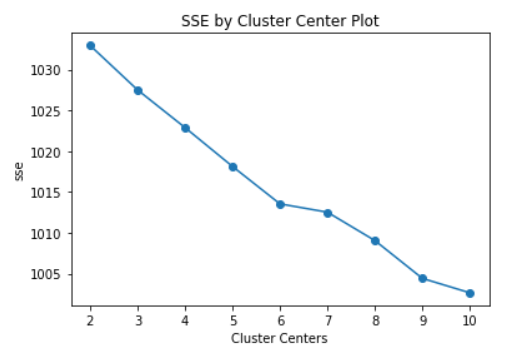
We used TF-IDF vectorizer for this analysis because this could penalize and weight the word counts by a measure of how often they appear. In other words, we regarded some words which appeared frequently only in a few posts/comments could not represent the general users’ opinions about the COVID-19 vaccine.

**Table 2** Vectorizer Settings

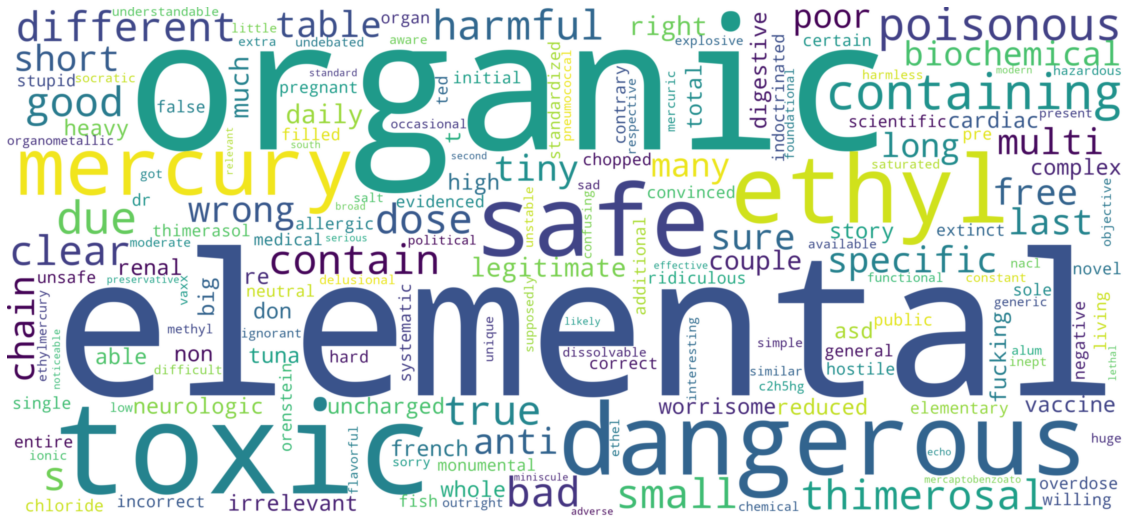
| type | TF-IDF vectorizer |
| --- | --- |
| min\_df | 3 |
| stop\_words | A customized list |
| ngram\_range | (1, 2) |

In order to determine the number of clusters, we applied the elbow method and found 6 as the optimal one. Similar to the EDA process, we excluded meaningless frequent words by adding them to the stopword list.

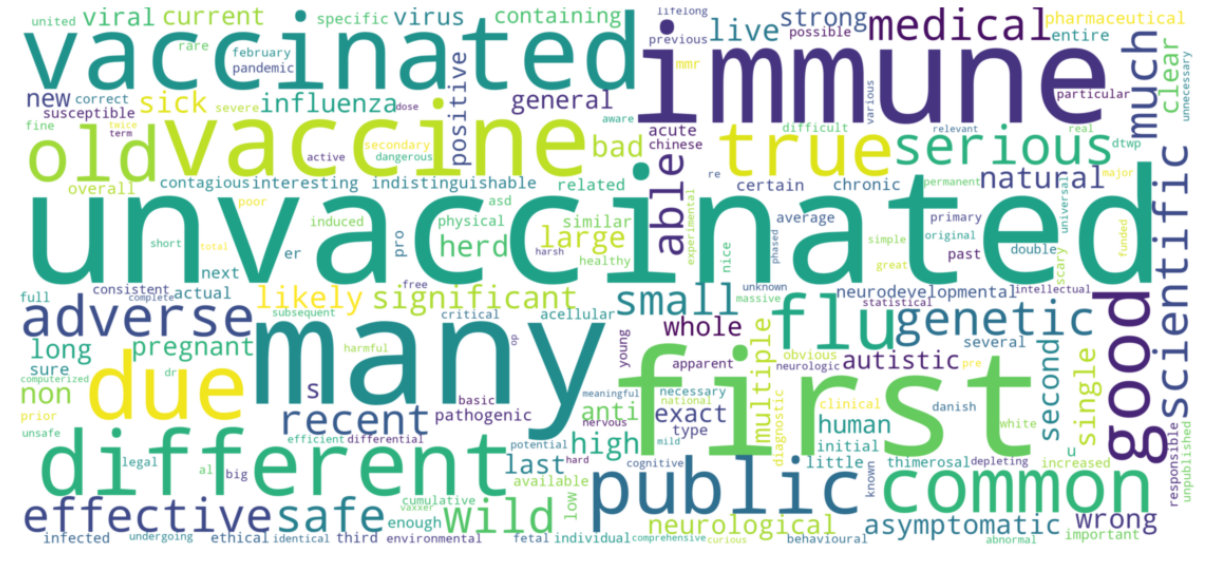
**Figure 2** Elbow Method result



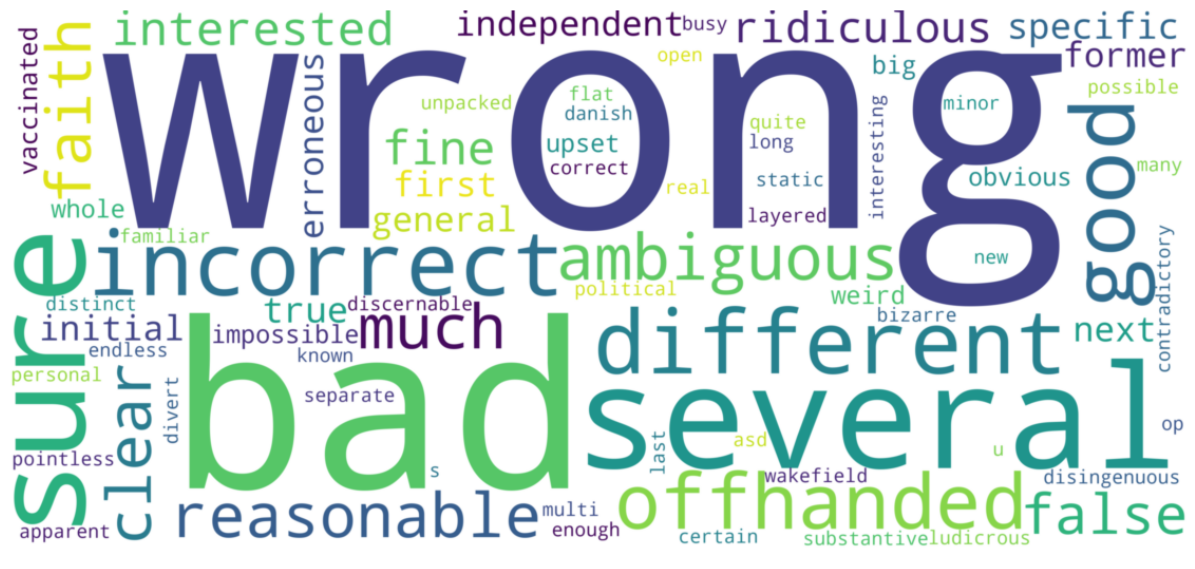
**Figure 3** Topic 0: Word Clouds Results



**Figure 4** Topic 1: Word Clouds Results



**Figure 5** Topic 2: Word Clouds Results



**Figure 6** Topic 3: Word Clouds Results



**Figure 7** Topic 4: Word Clouds Results



**Figure 8** Topic 5: Word Clouds Results



From Figure 3 to Figure 8 are the results of K-means clustering. Most frequent words found in each cluster were relevant to the COVID-19 vaccine and made sense, but we could not find common topics clearly in each cluster.

**Table 3** Topic modeling results

| **Topic** | **Key Words** |
| --- | --- |
| 0 | haha dude joke coincidence excellent stroke flat appreciate gets yes yep immunology misinformation careful breathing really de earth debate wait |
| 1 | welcome dumb explain rules aluminum big deal real science feeling possible deal quack misinformation big bodies downvote sure chracter allergic google kind |
| 2 | like one autism said measles vaccinated read wrong mercury yes cause point still immunity children tell going really anti actually |
| 3 | idiot x200b odd northwest someone pacific northwest pacific trolling f\*\*k bit hope vaxfact site assumed hello posts troll fraudulent bias |
| 4 | lol oh bruh hell her sodium logical shut antivax water says sorry article dihydrogen monoxide joke coat f\*\*king heavy |
| 5 | much mercury something change mean anti vaxxers anti vaxxers compounds elements comments gonna reddit appreciated sub one indeed play stroke compliment |

Table 3 is the result of LDA topic modeling. We also couldn’t find topics apparently from the results, thus it was hard to label them. These ambiguities were caused by the numerous pointless words like, haha, dude, etc. Since this data was extracted from the posts and comments from the Reddit users, it was understandable to contain a lot of informal words like slang and swear words. However, we needed to remove them to find the commonalities(topics) in each group.

**3.4 Sentiment Analysis**

Our Reddit Vaccine data was unsupervised ones, so we developed two different strategies to approach the sentiment analysis task:

1. We found a very similar data set from Kaggle.com which has sentiment labels. We then trained models with popular algorithms using the labeled data. As the last step, we made sentiment predictions on our unlabeled Reddit Vaccine data using models developed with a similar data set.

2. We applied generic sentiment labeling tools on our Reddit Vaccine data set.

The main evaluation metric was F-measure. We labeled 75 data-point reviews by hand, before comparing the labels with predictions. We also did error analysis to better understand mislabeled data.

**3.4.1 Developing models from similar dataset**

We found sentiment analysis for medical drugs dataset from the same website where we downloaded original data, kaggle.com (<https://www.kaggle.com/arbazkhan971/analyticvidhyadatasetsentiment?select=train_F3WbcTw.csv>). This dataset was the best dataset we found as the similar dataset to our original data, because it was about medical reviews, and had clear labels. This dataset had 5279 data points (review text) and had labels 0(617), 1(837), 2(3825). This dataset was heavily skewed: 72.46% data points are labeled 2.

We extracted review text column and label column, removed NAs and duplicates before applying appropriate vectorizer and building models.

We tested two algorithms for building sentiment prediction models: Multinomial Naïve Bayes(MNB), Support Vector Machines(SVM).

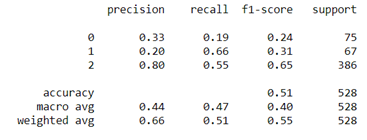
**Table 4** Vectorizer Settings

| type | Count vectorizer |
| --- | --- |
| min\_df | 5 |
| binary | True |
| ngram\_range | (1, 2) |

We chose uni & bigram boolean count vectorizer with minimum document frequency of 5. We decided not to remove stop words because stop words could express sentiment.

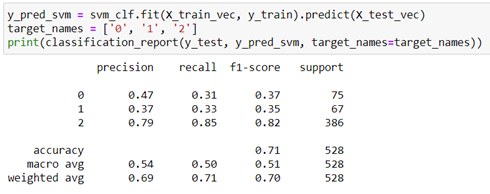
***MNB*** With default parameters, the accuracy was 0.5395 with 5-fold cross validation and the details were the same as below:

**Figure 9** MNB Result



***SVM*** With default parameters, the accuracy was 0.7274 with 5-fold cross validation and the details were the same as below:

**Figure 10** SVM Result



**3.4.2 Apply generic sentiment labeling tool: TextBlob**

TextBlob is one of the most popular sentiment analysis tools. This is the brief explanation for its sentiment property from TextBlob official website:

*The sentiment property returns a namedtuple of the form Sentiment(polarity, subjectivity). The polarity score is a float within the range [-1.0, 1.0]. The subjectivity is a float within the range [0.0, 1.0] where 0.0 is very objective and 1.0 is very subjective.*

In our analysis, we only used the polarity score.

We tested different boundaries for negative, positive and neutral sentiments. The dataset was heavily skewed toward neutral, thus the narrow boundary between -0.1 and 0.1 was a proper range to determine the negative and positive reviews.

**4 Results**

**4.1 Topics labeling result**

The final result of topic labeling with extended stopwords is the same as below. (Table 4) Though it still had some ambiguous groups, it was more obvious compared to previous results. Two groups, topic 0 and topic 4 contain some words about chemical materials like mercury, sodium, aluminum, etc. We concluded these word groups came from the postings about the ingredients of the vaccine. We also found another group, topic 2, containing medical & disease terms, for example, autism, measle, immunity, etc. We concluded this group of words was brought from the Reddit user’s concern about the side effects of the COVID-19 vaccine.

**Table 5** Improved topic labeling results

| **Topic** | **Key Words** |
| --- | --- |
| 0 | **Chemical Components** |
| **mercury sodium elements compounds thimerosal chloride compound** hear brain contains sodium brain damage organic contain indeed elemental consumption food daily toxic |
| 1 | **Unknown** |
| stroke odd assume salt excellent science northwest live risk kids playing pacific controlled approval nurse smith walker children |
| 2 | **Diseases / Medical** |
| like **autism measles** wrong cause read anti really point **vaccinated** saying bad time actually children **immunity** way **immune** someone made |
| 3 | **Agreement** |
| **ok** tell misinformation **welcome** change reason **yes possible** quack unfortunately needle vial spreads lying multi assumed title **loved** allergic ones |
| 4 | **Chemical** |
| day **aluminium ethylmercury** comments questions **mercury dihydrogen monoxide** worry name relevant like appreciated heavy eat bet big deal started immunology |
| 5 | **Unknown** |
| yes joke dumb coincidence food compliment vaccinated youre work rules thought works antivaxxers god damn myth consequences feeling kid bias |

**4.2 Sentiment Analysis Result**

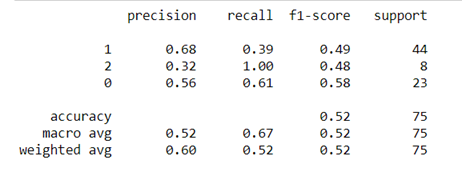
**4.2.1 Sentiment analysis with similar dataset**

From the previous step, SVM algorithm achieved better accuracy and F-measure. We then applied the SVM model to our data, and the result was not good. The model failed in the predictions and classified every review text as sentiment label 2(neutral). This was caused by the heavily skewed training dataset. This result will be further discussed in the challenge section.

**4.2.2 Sentiment analysis with TextBlob**

The F-measure for TextBlob method was 0.52 (using 75 hand-labeled data points).

**Figure 11** TextBlob result



**5 Conclusion and Challenges**

**5.1 Clustering / Topic Modeling**

Though the clustering result was not clear enough to label all groups, we found the people’s common concern about the COVID-19 vaccine. Some people worried about the harmful ingredient of the COVID-19 vaccine and others worried about the side effects (risk of getting diseases) from the vaccinations.

**5.2 Sentiment Analysis**

TextBlob was the preferred method for sentiment analysis for our dataset since it showed the best prediction results.

However, method 1 with a different dataset could show better results if the similar dataset was not skewed one. The ideal dataset would be a vaccine review with a similar skewness from the same source, Reddit.

**5.2.1 Error Analysis**

(1: Positive / 2: Negative / 0: Neutral)

* [Actual: 2 / Predict: 0] *That's not how chemistry works, Dr. Dumbass. If this were the case, table salt would literally kill you upon ingesting.*
* [Actual: 2 / Predict: 1] *You get more of those minerals by breathing for 1 hour than you do from 1 vaccine.*

We found two patterns of error from the error analysis result.

1. Failure to catch sentiment from the fact-based posts
2. Failure to read sentiment from the sole post, not a dialogue

**5.3 Challenges**

**5.3.1 Evaluation Metric for sentiment analysis**

The challenge of sentiment analysis is the evaluation metric. Datasets about real reviews are usually heavily skewed. Majority of the data points are neutral.

Accuracy is not appropriate in this situation because the dataset is heavily skewed. For example, if 80% data points in our dataset is neutral, then a model that predicts all data points to be neutral would have 80% accuracy. However, this model would be useless because it does not predict any sentiment.

Usually, the solution for the above problem is to proceed with binary classification and use area under ROC curve (AUC) as the evaluation metric.

**5.3.2 Reddit users’ arguments**

Reddit posts and comments have a format of dialogue. It means we need to identify the back and forth situation to read the user’s sentiment to the vaccine, instead of one post(comment). Therefore, the sentiment analysis of each title or body has some limitations to catch the sentiment.

Some Reddit users also write the list of the scientific facts(knowledge), instead of the emotional negation sentences when rebutting another user’s posts. Therefore the algorithm classifies these posts/comments as neutral ones and fails to predict right.

**References**

[1] Reddit Vaccine Myth (Dataset)  
*https://www.kaggle.com/gpreda/reddit-vaccine-myths*

[2] COVID-19 Vaccines: Myth Versus Fact  
*https://www.hopkinsmedicine.org/health/conditions-and-diseases/coronavirus/covid-19-vaccines-myth-versus-fact*

[3] Tutorial: Quickstart-Sentiment Analysis *https://textblob.readthedocs.io/en/dev/quickstart.html*

[4] Madhu, Shrija(2018, August) An approach to analyze suicidal tendency in blogs and tweets using Sentiment Analysis, 2018, *International Journal of Scientific Research & Management Studies*, 34-36

[5] They claimed the Covid-19 vaccine made them ill. Then they went viral  
*https://www.wired.co.uk/article/covid-vaccine-misinformation-facebook*