CALIFORNIA STATE UNIVERSITY
FULLERTON™

Department of Computer Science

This project has been satisfactorily demonstrated and is of suitable form.

This project report is acceptable in partial completion of the requirements for the Master of Science degree in Computer Science.

COVID-19 Vaccine Opinion Analysis using Machine Learning	
Project Title (type)	
Urvashi Singh	
Student Name (type)	
Dr. Bin Cong	
Advisor's Name (type)	
Advisor's signature	Date
Reviewer's name	
Reviewer's signature	Date

Report Prepared By: Urvashi Singh
Under Guidance Of: Dr. Bin Cong
Course: CPSC 597 - 02
Date: May 21, 2021

COVID-19 Vaccine Opinion Analysis using Machine Learning - Project Report

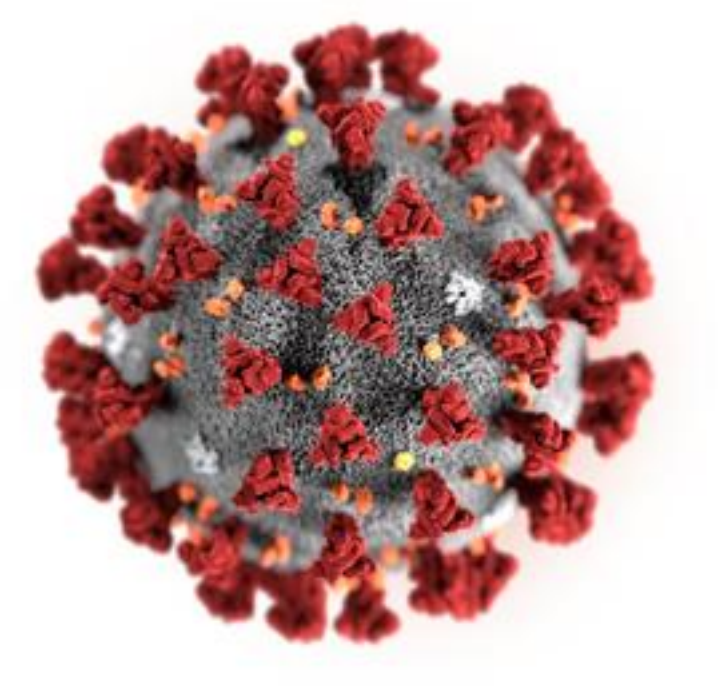

Table of Content

ABSTRACT

Since COVID-19 hit the world, we all have come to a pit stop. The losses are huge and in almost every field, amidst this crisis some people have been using social networking platform to spread misinformation. Dangerous misinformation about COVID-19 and now about COVID-19 vaccine is appearing online. In this project I have tried to quantify and categorize COVID-19 contents as ‘anti-vaccine’ or ‘pro-vaccine’. For this project I have manually gathered publicly available content of these online communities involved in COVID-19 discussion. To identify whether the post is related to ‘anit-vaccine’ or ‘pro-vaccine’, I have used various technologies and algorithms used in the main paper [1] such as NLTK (Natural language processing and Tool Kit) for python programs to work with human language data [5], Gensim’s Word2Vec algorithm that uses large amounts of unannotated plain text, word2vec learns relationships between words automatically [4]. The output are vectors, one vector per word, with remarkable linear relationships, unsupervised machine learning technique called Latent Dirichlet Allocation (LDA) to analyze the emergence and evolution of topics around COVID-19 [2]. The LDA method is a statistical model for discovering the abstract topics aka topic modeling [2], topic Coherence metrics to obtain pairwise user preferences of topical coherences and to determine how closely each of the metrics align with human preferences [3]. Based on the Coherence metrics and pyLDAvis graph, I will conclude if the ML technique LDA was able to correlate data and extract most discussed topics from each Facebook Post.

1. INTRODUCTION

It is a well-proven fact across all the scientific community today that defeating COVID-19 will require the development of a vaccine. Many people across the globe (ideally every human being on earth) need to be vaccinated to attain group immunity. It has also been discovered that the COVID vaccines are less effective in aged people, which is why it will need the younger generation to take vaccination at a much higher rate so that group immunity can be achieved [10]. Even if we do not consider COVID, there is still considerable disagreement on present vaccinations, e.g., against polio, with many parents refusing to vaccinate their children. Similar oppression is in the case of measles, for which the oppression raised the number of cases in 2019, which caused the outbreak in the U.S. and neighboring countries [17]. This trend shows that any upcoming COVID vaccination drive is most tending to face the same opposition. It will raise conflict of interest in the public if COVID vaccination is made mandatory for school children. Hence, there is a need to understand such disapproval from the public beforehand for government, implementing vaccination on ground and scientists, developing the vaccine.

The anti-vaccine communities are becoming well-liked on various social media forums such as Twitter, Instagram, Facebook, etc. but the most popular among these for them to flourish is Facebook (FB), where they are sharing health misinformation. Such kind of misrepresentation of information is causing people's safety and common public health issue [10], [18]. In a similar manner, those who are in support of vaccination (pro-vaccine) are also collaborating in these online portals to spread awareness on public health advice. There were already severe differences between a pro-vaccine and anti-vaccine community fighting online in support of their narrative. The surge of COVID-19 at the beginning of 2020 added fuel to the fire. It has steered common people to the excess of misinformation in social media platforms frightening life [6].

The Section 1 of the report begins with the introduction of the problem statement and describes the current situation of data handling over the social media portals with respect to pro-vaccine and anti-vaccine community. It also mentions how the misinformation about COVID vaccine or using substitute un-proven remedies can be fatal for people. In the two sections I am discussing about the Motivation and

Objective behind implementing this project. I have broadly described two needs of implementing this project

(1) Analyzing the difference between anti-vaccine propaganda vs generic conversation on COVID and

(2) Automating the process of analyzing the anti vs pro vaccine sentiments.

Section 2 describes the literature survey from different online sources as well as research papers from IEEE.

Section 3 describes the detailed approach taken in this project and talks about Data and machine learning analysis. This section describes the methodology of accessing the public Facebook pages content for anti-vaccine and pro-vaccine communities and finding their connection to other related pages. This section describes existing approach as well as proposed approach using the machine learning technique called “Latent Dirichlet Allocation (LDA)” [15] to investigate danger and progression of topics around COVID.

Section 4 talks about the requirement of the project both functional and non-functional requirements, which marks a base line to build this system.

Section 5 covers the design and implementation part where it covers system design, use case diagram, class diagram and sequence diagram. At the end it will cover the implementation flowchart and finally implementation steps and artifacts.

Section 6 will cover ML project testing, overall testing of the project and End result or confusion of the System.

Section 7 will cover summary of the whole project and future work that can be done to improve this system.

Section 8 has all the references or materials used in building this project report.

1.1.Motivation

The motivation for implementing this project comes from the fact that there is an overabundance of erroneous information available on social media portals, causing a ruckus among the common people and causing a threat to the life of people. Some of the instances are life-threatening cures that have been advised on the online portals, such as drinking fish and additives, bleach, etc., along with well-planned threats against public health officials one of the directors of the U.S. National Institute of Allergic and Infectious Diseases [7].

The volume in which false misinformation has been added daily and the speed at which it is spreading across the globe on social media platforms is alarming. This is the reason why social media companies are struggling to comprehend this false misinformation and are not able to filter or analyze them efficiently [19], [20]. On top of it, the condition is getting more critical because people are spending more time than average due to the imposition of lockdown across the globe and social distancing. This will enhance the probability for people of getting exposed to false information flooding on social media, and as a result, they will risk their life, as well as their contacts live.

The work done in this project is motivated by the following two needs:

(1) a medium to analyze the connection between online anti-vaccine propaganda misinformation and the generic online conversation encompassing COVID-19 discussions; and

(2) there is a need to automate the approach to filter and process the online data because every day increasing volume of online data is cumbersome, inaccurate and inefficient is done manually.

(3) I have proposed an automated machine learning approach to achieve the two tasks mentioned above. It has been addressed how online conversation has been changed between pro-vaccine and anti-vaccine over the two months duration in early 2020 and the difference we observed in the anti-vaccine and pro-vaccine narrative.

1.2.Objective

We know with the above introduction that anti-vaccine society or group supporting it unveils a wide range of variety of COVID discussion, and therefore they can convince or charm a wider section of people looking for guidance on COVID online. For example, there are a set of people who are looking for a faster COVID-19 vaccination drive; on the other hand, there are people pursuing substitute medications online. In this case, the anti-vaccine community is in a superior state to appeal new support moving forward compared to the pro-vaccine group. It is concerning because fresh support to the anti-vaccine community will hinder the propagation of COVID vaccine drive and obstruct in achieving herd immunity and can leave continents susceptible to future COVID resurrections.

The key focus in this project is the endogenic elaboration of the COVID conversation that occurred at the start of the global pandemic and prior to the first official death reported because of COVID on February 29, 2020 [25]. In this project, I am trying to develop a machine learning algorithm that identifies all the possible topics from a large collection of online posts from different communities surrounding vaccine and COVID-19 discussion. The algorithm will be able to handle a large amount of data effectively, process the data, and shows the results quickly with the use of statistical grouping techniques. This will eliminate the requirement of having slow passed, inaccurate, costly manual human effort.

2. LITERATURE SURVEY

There are lot of existing works that are based on the study of Twitter data as mentioned by D. A. Broniatowsk and Y. Lama in [21], [22]. It is well known fact that Twitter is a broadcast medium for individual choices and response on a topic, but widely the narrative building and discussions are developed in online community pages and therefore in this project I am not using Twitter. I am using Facebook community pages for my study in this project [23]. In the existing approach which has been widespread by N. F. Johnson in [24] and [25], the data is being accumulated from online communities, particularly Facebook pages that provisions either pro-vaccine or anti-vaccine sentiments and views.

3. METHODOLOGY

In this project, we focus on ‘Facebook Pages,’ which are alternatively called ‘clusters.’ The ‘Facebook Pages’ also called ‘fan pages’ or ‘public pages’ are those accounts that symbolize organizations, public figures, communities, and causes. As per Facebook’s policy, any content uploaded on these pages is public and can be viewed by all who visit the page. In this project, we are only analyzing posts on public pages, and we do not touch on personal accounts. The methodology presented in this paper trails [24] and [25] by investigating public posts, opinions, and contexts of Facebook pages for both anti-vaccine and pro-vaccine communities.

Snowball methodology is used to acquire all the publicly available content of the online groups and communities. To start with, a ‘Facebook Page’ is identified where the content points to the discussion related to either pro-vaccine vs. anti-vaccine debate, vaccination, COVID-19, or public policies about vaccination. In the next step, other Facebook pages or Public pages in connection to the parent page are indexed. With the help of ‘computer-assisted filters’ and ‘human coding,’ novel clusters are assessed at each stage. In order to group the clusters,

(1) Anti-vaccine or pro-vaccine or related topics and

(2) includes COVID content or not, and then we check the Page ‘about’ section. The first group, ‘anti-vaccine or pro-vaccine is sub-divided into

   (1) Page’s ‘about’ portion or ‘title’ portrays it as pro-vaccine or anti-vaccine,

   (2) at the minimum two posts were involved in pro-vaccine vs. anti-vaccine debate out of 25 recent posts.

In the next step, each cluster is categorized by two researchers at the minimum impartially, and agreement was reached in each case. In case of the two researchers disagree, a third reviewer would have reviewed the post, and then all three reviewers discussed these cases and reached an agreement. This process allows differentiating between thoughtful content vs. mere spoof. In this project, the study was kept concentrated on English content only; nevertheless, it can also be straightforwardly generalized using a similar method for other languages as well.

The material and data obtained from each of these clusters are grouped together distinctly for the pro-vaccine community and anti-vaccine community, and content of both the data sets were investigated using machine learning. In this project I am using unsupervised machine learning model which is known as ‘Latent Dirichlet Allocation (LDA)’ [15] In order to evaluate the emergency and advent of online discussions ongoing around COVID. The Latent Dirichlet Allocation model is very well described in Wikipedia as [16]. It is a generic statistical model approach which enables observation sets to be elucidated by unobserved group which in turn justifies why some parts of data are alike. Let us take an example to understand this, in case if a document has collection of words and these words are nothing but observations, it postulates that each document is blend of small number of topics and presence of each word can be attributed to one of the document’s topic. This model (LDA) fits in the machine learning toolbox but in broader perspective it relates to the artificial intelligence toolbox.

The topic’s quality of being logical and consistent is measure by coherence score in this project. The coherence scores enable us to have a quantitative way of determining the affiliation of the words in an identified topic which is produced from a distinct algorithm which runs over a trained LDA model. The arithmetic mean of per topic coherence will provide the final coherence score of a single model. In this project I am using A coherence measure which is labeled as [15] to achieve highest correlation within all the available topics. will be discussed in detail in the next part of this section.

As a summary, the machine learning approach used in the project recognizes topics and their narrative from online public Facebook pages with high coherence. This means that the word combination recognized are strongly linked as per the coherence scoring methodology discussed above.

3.1.Existing Algorithm

There are well-known machine-learning algorithms currently present that are applied on topic models such as Latent Semantic Indexing (LSI) [12] and probabilistic Latent Semantic Indexing (pLSI) [11] to disclose secreted or latent thematic structures or we can say topics from vast pools of documents. Without any previous cataloging or annotation, the topic automatically emerges from the statistical properties of the documents or content on Facebook pages. On the other hand, the thematic structures (posts) can be used to inevitably group or condense the documents or online public pages up to a level which is impossible to perform manually. The above-mentioned topic modeling approaches (LSI, pLSI) have demonstrated to be extremely beneficial in illuminating the core idea within a set of documents which in our case is Facebook pages. But they have their limitations as well, next we will briefly go over the existing LSI and pLSI models and their limitations which motivates to use the proposed model called LDA.

3.1.1.Latent Semantic Indexing (LSI) / Latent Semantic Analysis (LSA)

C. “LSA is a technique which examines the correlation between a set of documents and expression they enclose and as an outcome they produce a set of perceptions related to the documents and expressions it contains.

D. One of the most popular application of this model is information retrieval and therefore it is also called LSI (Latent Semantic Indexing). The first paper on LSI was published in 1988 by Dumasi, Furnas, Landauer and Deerwester .

E. LSA/LSI implementation is achieved in four basic steps as mentioned below:

E.A.Term by document matrix (more generically term by context) tend to be thinly dispersed.

E.B.Transform matrix entries to weights, characteristically weight directly by estimated importance in passage and weight inversely by the degree to which knowing word appeared give the knowledge of the passage it showed in.

E.C.Rank reduced SVD (Singular Value Decomposition) performed on matrix which is nothing but the distinctive mathematical breakdown of a matrix into three matrices, two with orthonormal columns and one with singular values on diagonal:

E.C.A.n-highest singular values are set to 10.

E.C.B.creates k-dimension estimate of the original matrix, this is semantic space

E.D.Calculate resemblances among entities in semantic space. “ [12]

3.1.2.Probabilistic Latent Semantic Analysis (pLSA)

“pLSA is nothing but the probabilistic re-casting of LSA (Latent Semantic Analysis or Indexing) algorithm.

It is another kind of improved latent variable model with observed count data and nominal latent variables. It was developed because LSA failed to handle the ‘polysemy’ or noise in the LSA. The pLSA or we can say mainly ‘probabilistic’ system consents for the assessment of schemes under conditions with ambiguity. A homogeneous mechanism for integration can be provided by probabilistic LSA system over heterogeneous information.

pLSA aspect model analysis and formulation is given in steps below:

A. With each observation it associates an unobserved latent class variable z € {}.

B. It describes a mutual probability model over documents and words.

C. Supposes w is independent of d conditioned on z

D. Cardinality of z should be much less than d and w.

E. Generative model formulation of pLSA:

   E.A.Selects document d with probability P(d)

   E.B.Select latent class z with probability P(c|d)

   E.C.Generate a word w with probability P(w|c)

F. This gives expression of joint probability model” [11]

$$P(d,c) = P(d)P(w|d);\ P(w|d) = \sum_{z=Z}^{n} P(c|z)P(z|d)$$

3.2.Proposed Algorithm

Latent Dirichlet Allocation methodology is a multiplicative probabilistic topic scheme which targets to unearth dormant or unseen thematic structures from quantity D. The hidden structures are articulated as topics and topic shares in every document, is represented by unseen variables that LDA postulates onto the quantity. Based on probabilistic selection rules, the multiplicative nature of LDA describes an illusory arbitrary process. Conversely, hidden structures or patterns within the documents (or online public pages) can only be observed by looks at the words, which is, the topics and topic fraction per document, by application of statistical deduction procedures. By employing this we get the subsequent distribution which by observing the document can capture the unknown structures. We can outline the generative process in following steps [7]:

Each topic is a multinomial distribution over the vocabulary V and comes from a Dirichlet distribution. ~ Dir (η). Additionally, every document is represented as a distribution over K topics and come from a Dirichlet distribution~ Dir (α). The Dirichlet parameter α denotes the smoothing of topics within documents, and η denotes the smoothing of words within topics. The joint distribution of all the hidden variables (topics), (per-document topic proportions), (word topic assignments) and observed variables (word in documents).

Figure 3 represents the LDA model in shield notation [21], where light shaded nodes denote the hidden random variables, the dark nodes represent the noted random variables and edges shows provisional dependencies between the two. The rectangles are called plates representing replication. Assume that over the word distribution there are K topics (K-plate) which are contingent on the Dirichlet parameter η, i.e., ∏ K, k=1 p (| η). Next there is a per document topic proportion subjected on Dirichlet parameter α, for all D documents (D-plates), i.e., ∏ D, d=1 p (| α). Lastly, we figure out per word assignment of topic for all N words (N-plate) of document d ∈ D. The word topic assignment depends on formerly drawn pre document topic and the drawn word depending on per word topic allocation and entire topics

With the expression ∏ N, n=1 {p (|) p (|, we can recover the probability of (row) from (column) within the K X V topic matrix. It is assumed that we have conditioned only on observed variables, words present in the documents in order to infer the hidden structure.

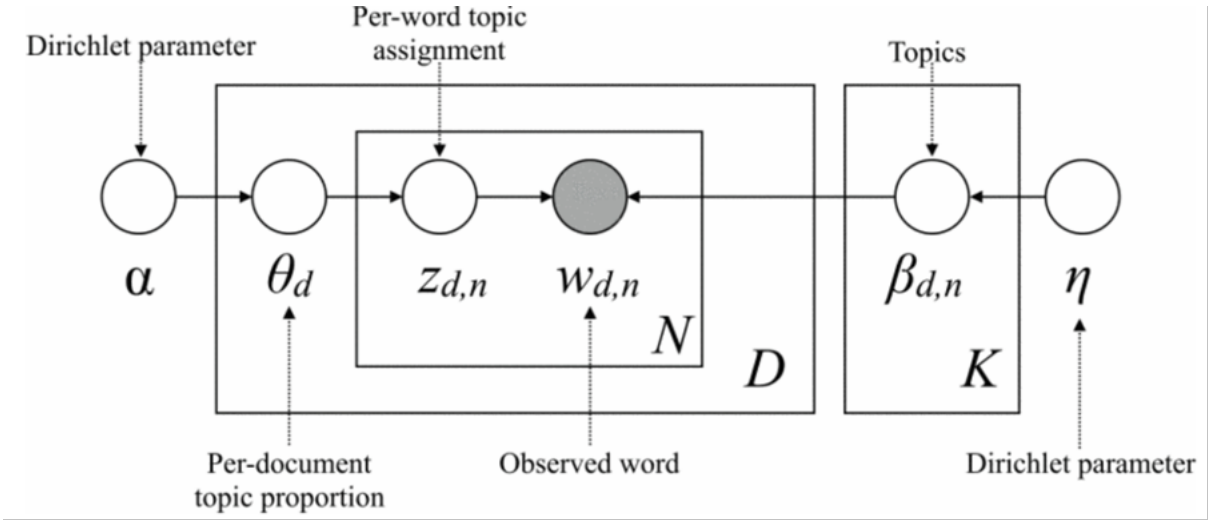

Figure 1: LDA algorithm shown in graphical model [7]

4. REQUIREMENTS

4.1.Functional Requirements

- User should be able to upload Facebook posts in CSV file format.
- User should be able to Process Dataset and prepare it for LDA algorithm.
- User should be able to view Extracted topics based on LDA Algorithm.
- User should be able to view Coherence score graph between anti and pro vaccine topics.
- User should be able to view Coherence score topics via pyLDAvis graph.

4.2.Non Functional Requirements

- The system should be consistent with data provided in Facebook Post and what is extracted as a result.
- The system should be available as long as User wants to view graphs.
- The system should respond in real-time so that Users can view result as soon as its processed via LDA algorithm.

5. DESIGN AND IMPLEMENTATION

Main motivation in this project is the endogenic growth of COVID discussion at the start of pandemic. For this project Facebook public posts are collected for the period from 15th January 2020 – 15th February 2020.

5.1.Software Requirements

“The appropriation of requirements and implementation constraints gives the general overview of the project regarding what the areas of strength and deficit are and how to tackle them.”[14]

A. Python IDEL 3.7 version (or)

B. Anaconda 3.7 (or)

C. Jupiter (or)

D. Google COLAB

5.2.Hardware Requirements

“Minimum hardware requirements are very dependent on the software being developed by a given Enthought Python / Canopy / VS Code user. Applications that need to store large arrays/objects in memory will require more RAM, whereas applications that need to perform numerous calculations or tasks more quickly will require a faster processor.” [13]

A. Operating system : Windows/ Linux/ Mac

B. Processor : Intel i3 (min)

C. RAM : 4 GB (min)

D. Hard disk : 250 GB (min)

5.3. Facebook Posts

This is the sample of Facebook posts that is acceptable by the system.

PS: The file should be in .csv format.

```
Posts
"Hundreds gathered in Boston to protest mandatory flu vaccine shots. The order is for all students under the age
"A Hong Kong study is the first in the world to confirm reinfection of Covid-19, suggesting that it's "unlikely
"In a shift that perplexed some doctors, the US Centers for Disease Control and Prevention has changed its Covid-
"Proper hand washing WILL help prevent Coronavirus."
"you can help slow the spread of COVID-19. Wear a mask and practice physical distancing (staying 6 feet away fro
"Last year my work said mandatory flu vaccine or wear a mask, most got it because they didn't want to wear a mas
"I'm afraid to trust anyone anymore. I feel that the CDC will use the " flu" shot to practice on the COVID shot.'
"flu vaccine increases chances of getting COVID. Already been seen in the military protest."
"Even if you get the flu & get tested the test come out positive as COVID-19 they need to fix this problem."
"I will not be taking a shot for the flu or for the coronavirus that is how it is spreading in the first place"
"I stopped and protest getting the flu vaccine around 15 years ago. Never had the flu since. It actually made me
"I can not take the Flu shot and protest ......what can I do to protect myself?? Been asking everyone but no ans
"Yep, coupled with the flu season; difficult to diagnose ( guessing game?) fever, cough, body ailments, etc.,,,,
"If people actually wear a mask stay six feet apart and wash their hands, I'll bet there will be less flu this y
"Show us the test results for what is in the Flu Shot and it may be considered. Without that we don't even know
"You're pushing the flu vaccine like"
"It helps or something. We can't create a flu vaccine that helps prevent so many deaths but yet we're going to r
"To whom gets the flu shot and death. My family will continue to not get the flu shot nor wear masks."
"some of the spread of flu and covid is due to laziness People who are too lazy to cough and sneeze into their a
"Well, that's what viruses usually do - spread through the population. I knew it, the fear mongering would sprea
"So this is a false statement why hasn't fb taken it down!!! EVEN DR.OZ SAID GETTING THE FLU VACCINE MAKES MORE
"I haven't had the flu shot in 20-30 years and protest 1'll be darned if 1 take your "Covid-19" shot either!!! O
"Avoid mold and vaccines and you should be fine. Lack of poison is essential to a healthy immune system. Breathi
```

Figure 2: Sample CSV file format of Facebook Posts

The upload window should appear as soon as you click on the 'Upload Facebook Post Dataset'

Should display posts.csv loaded, if the post was uploaded otherwise display error.

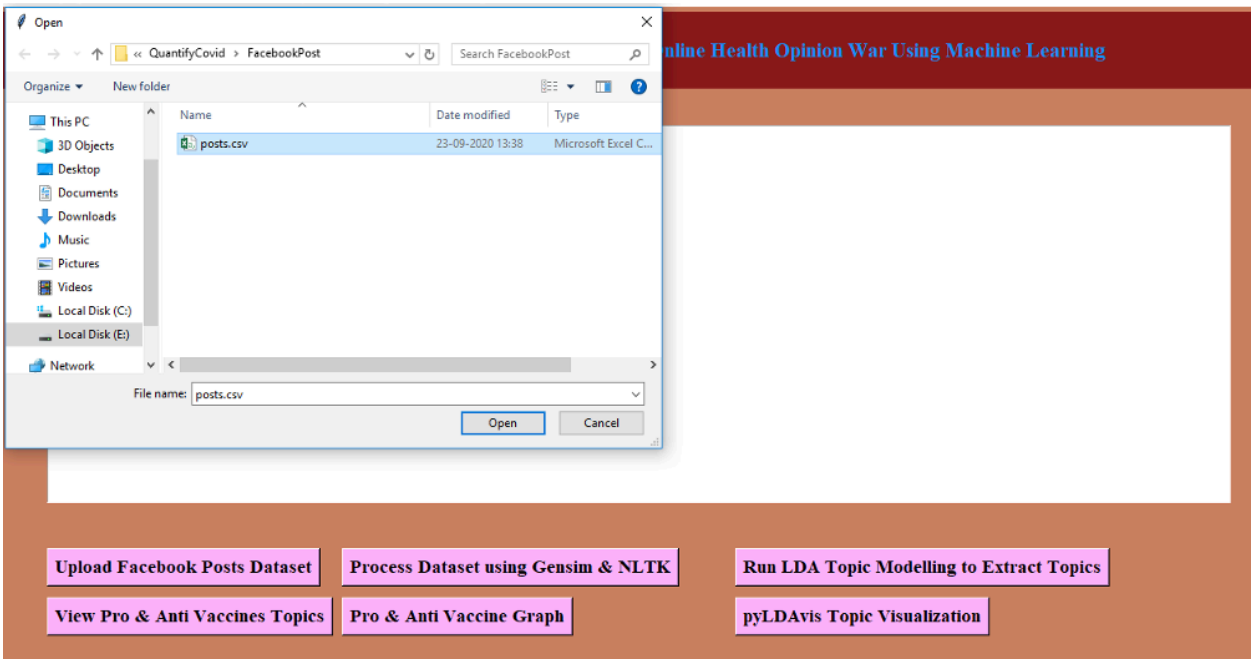

Figure 3: Upload Window

5.4. Data Pre-processing

Once User clicks on ‘Process Dataset using Gensim & NLTK’ button, the system will pre-process the uploaded data. The pre-processing includes removing ‘stop words’, changing plurals to singular form, extracting adjectives, removing spaces, etc. After this pre-processing the data will be displayed on screen for user to view.

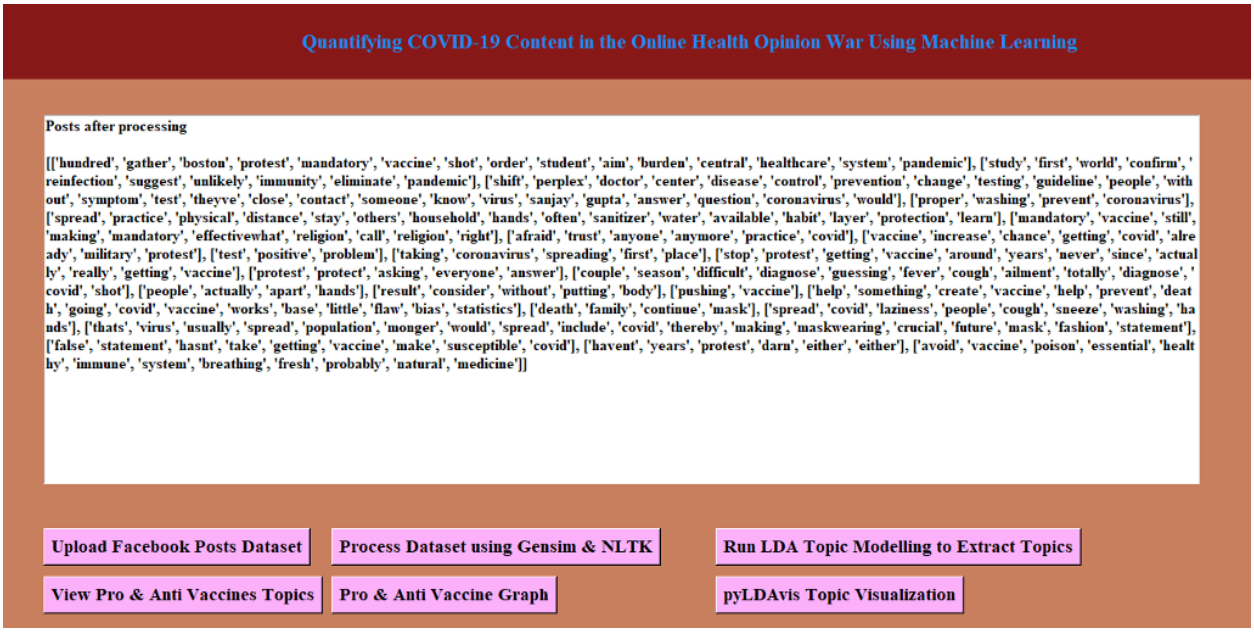

Figure 4: Pre-processed data displayed

5.5.Topic Coherence Measurement

Once the LDA's distribution is assessed, K topics are characterized by multinomial distributions over V. Different probability is allocated to each word that every topic distribution contains. Those words that occur more repeatedly have high probability of words within topics. The top 10 high probability words are used to label the topics. Nevertheless, number of topics output by LDA is defined by K. High K represents uninterpretable topics whereas low K depicts in fewer or broad topics. Hence selecting the correct value of K is very important in modeling the topic algorithm, comprising LDA.

It is required to assess the property of generated topics to quantity the projected probability of held-out data, but this inversely impacts human interpretability. Due to this reason researchers have recommended topic coherence methodology, that is a qualitative approach to unearth the coherence of the topic [23] where the fundamental ideation is based in the "distributional hypothesis of linguistics" [15] that demonstrates that the similar meaning words incline to appear in alike contexts. For a topic to be coherent all or most of the words (say top N words) are interrelated. The main computational challenge is to acquire a process which associates favorably with human topic ranking data. Generally human topic ranking data is gold standard and if a method correlates satisfactory with it, it is considered as suitable gauge for topic interpretability.

In this project a new coherence measure is found .This achieves highest level of coherence when compared with human topic ranking data. This is the reason why is adopted as the coherence measure in this project and implemented in the source code discussed in the results and discussion section of the report. coherence is divided in four segments:

1) Word sets are created by segmenting the data,

2) Calculating probability of occurrence of word or word sets,

3) Computing how firmly a word pair supports another word pair and

4) Calculating final coherence measure by accumulating every confirm measure.

5.5.1.“Anti” and “Pro” Vaccine Topics

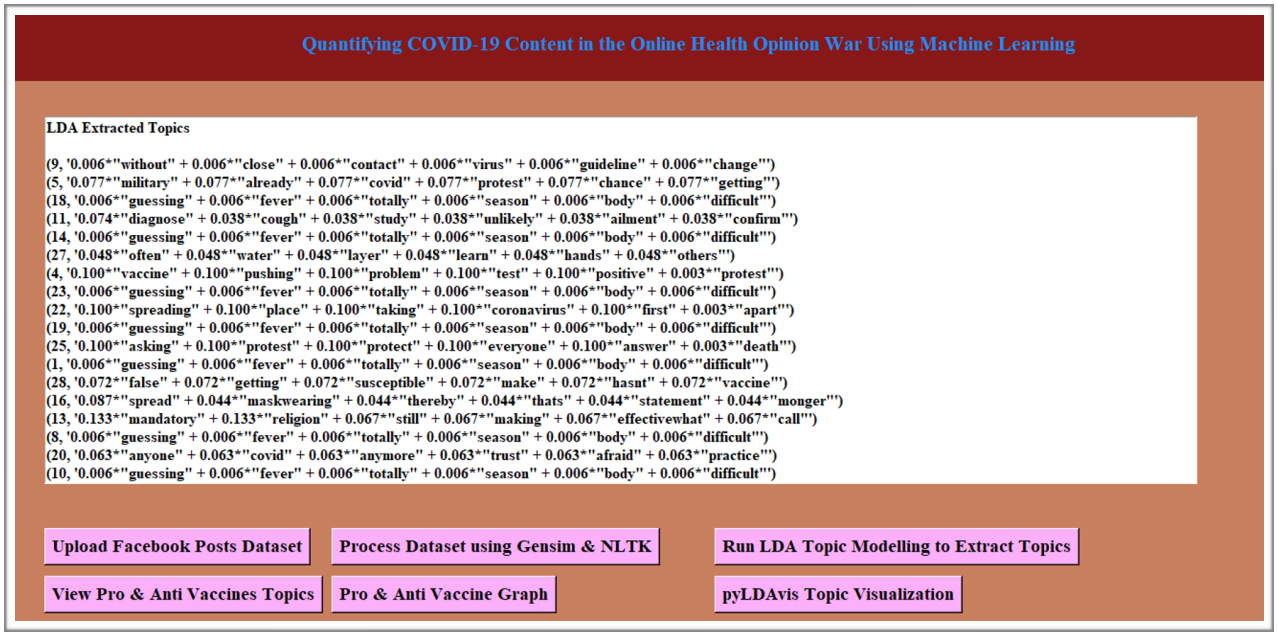

Figure 5 : “Anti” and “Pro” Topics with their Probability Score

As we hit, “Run LDA Topic Modeling to Extract Topics” tab we can fetch all the “anti” and “pro” topics with their Probability Score, and it can be seen plotted below based on their coherence score.

5.6.System Design
5.6.1.Use Case Diagram

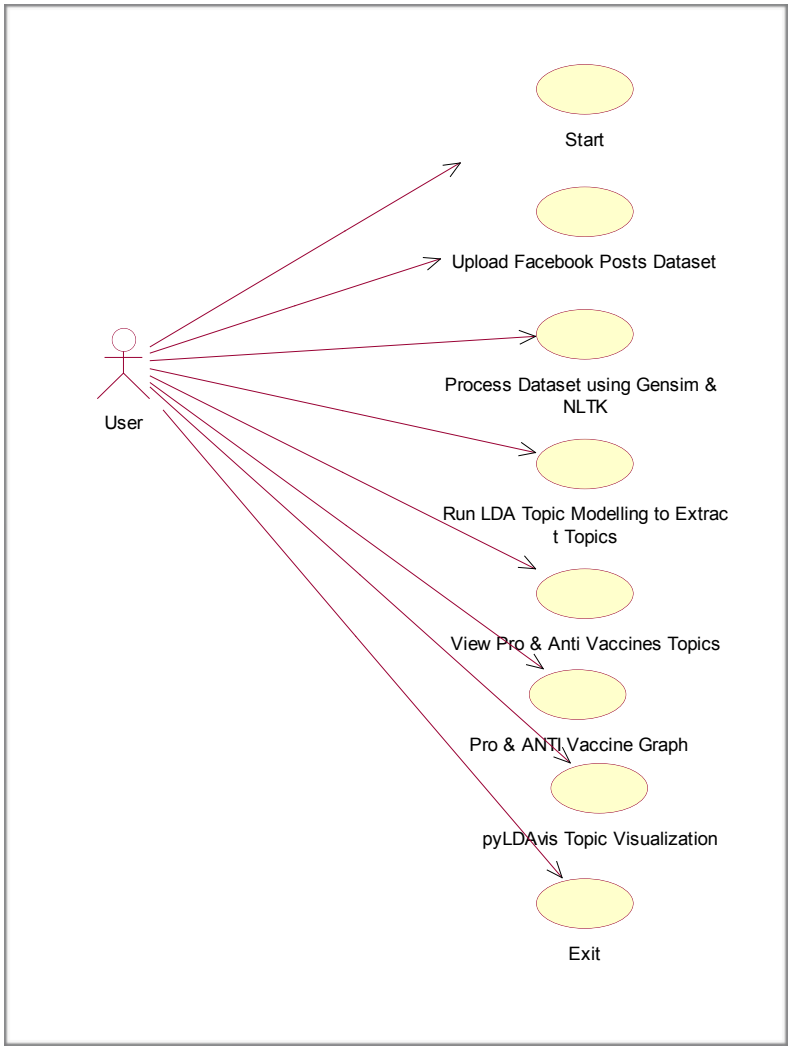

Figure 6: Project Use Case Diagram

5.6.2.Class Diagram

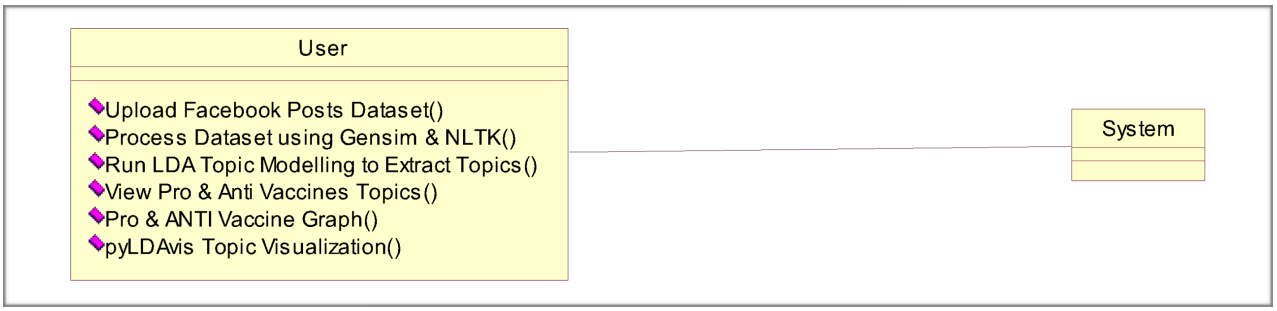

Figure 7: Project Class Diagram

5.6.3.Sequence Diagram

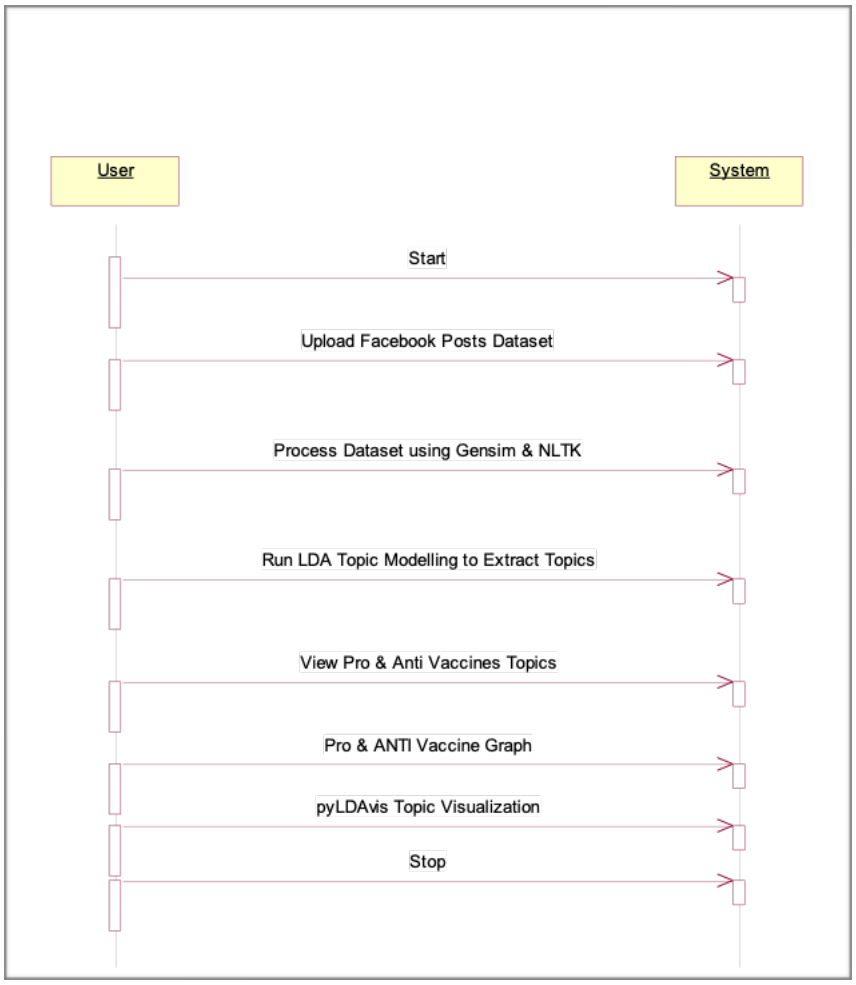

Figure 8: Project Sequence Diagram

5.7.Project Flowchart

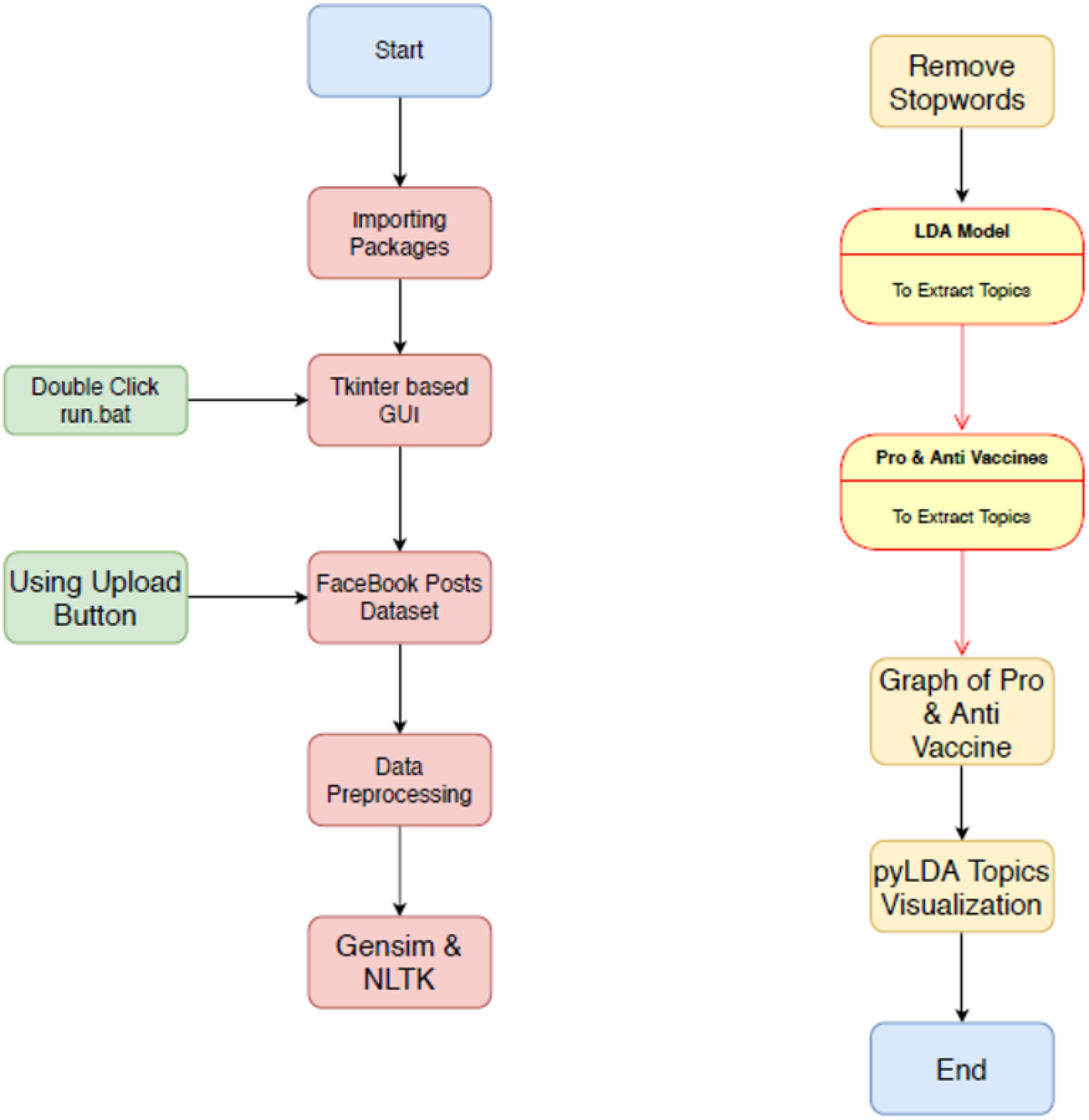

Figure 9: Project Flow Chart

5.8.Project Prototype Design

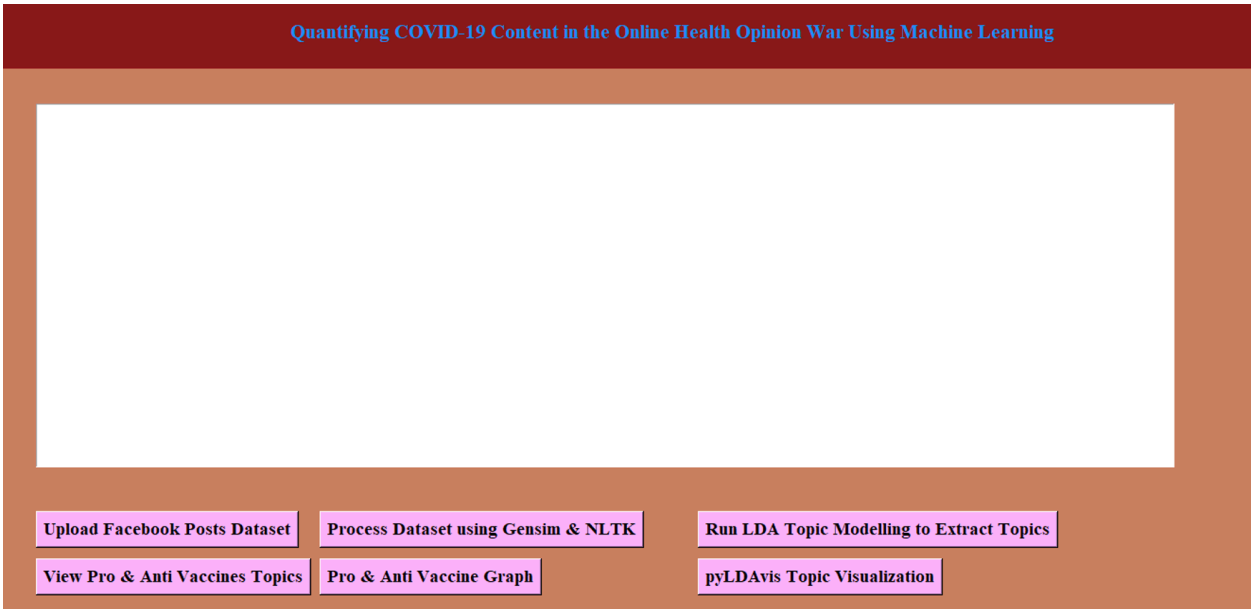

Figure 10: Project Prototype Design

The Project Prototype is a an interface built using Python tkinter package to demonstrate the underlying implementation of each Project Activity.

6. TESTING AND RESULT

In this section, we will go through the test strategies used to identify issues or bugs in the system, I have also discussed about the end result with snapshots of each artifact.

6.1.ML Model Testing and Result

To be cost effective, the testing should be focused on areas where it will be most effective. Figure 5 shows the software testing process we should follow for our project.

Figure 13: Type of manual Software Testing

In our project we are taking the advantage and simplicity off Black box testing and it is applied to Machine Learning models. It is used to test the intrinsic features of the algorithm used to create the model. Main contest, on the other hand is to find test vision that can confirm the outcome of the test against anticipated values known before. Figure 6 below shows the absence of test vision while testing Machine Learning models because contrary to the conservative software development and testing, there is no expected values known beforehand because Machine Learning model yields a kind of prediction values. Due to this it is hard to equate the prediction with known expected values.

In this case in order to test our machine learning model, a concept of pseudo-oracles is introduced which showcases the conditions where outputs from given set of inputs

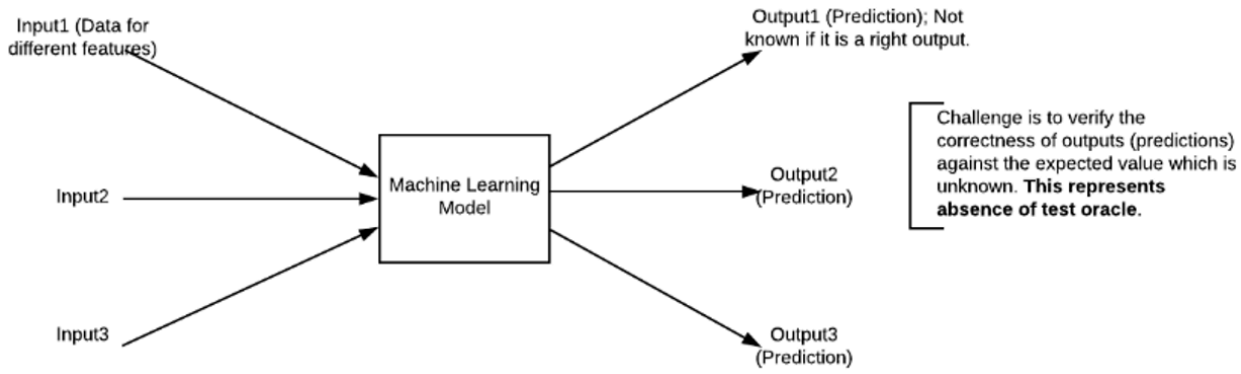

Figure 14 The absence of test oracles in ML

are compared and correctness is determined.

6.2. Overall Project Testing and Result

Test Case ID	Test Case Name	Test Case Description	Test Procedure	Test Value		Test Case Status	Test Priority
				Expected	Actual		
1	Start Application	Application should start on running the executable Covid19.py file	1. Right click on COVID19.py file and Run	Terminal shows Success Message	Terminal shows Success Message after some Error	Successful	High
2	Facebook Post	Upload Facebook Post from your computer	1. Click on 'Facebook Post' Button 2. Select file from the dialog box	User should be able to upload post and message should display 'Success'	User was able to upload data and message displayed was 'Success'	Successful	High
3	Process Dataset	Process uploaded dataset using Gensim and NLTK Python ML Libraries	1. Click on 'Process Dataset' Button	User should be able to view processed data with stopwords and other unrelated words removed from the post	User was able to view processed dataset	Successful	High
4	Run LDA	Run LDA Topic Modelling to Extract Topics	1. Click on 'Run LDA' Button	User should be able to click on Run LDA button and it will process dataset based on LDA algorithm	User was able to trigger LDA algorithm using 'Run LDA' button	Successful	High
5	View Topics	View Anti-Pro Vaccine Topics	1. Click on 'View Topics'	User should be able to view anti and pro topics that has been fed to the system, this will be used to analyse Anti and Pro inclination	User was able to view anit- and pro topics fed to the system	Successful	High
6	View Coherence Graph	View Coherence Score graph	1. Click on 'View anti-pro score graph'	User should be able to view anti and pro coherence score graph in a spearate window	User was able to view anti-pro coherence score graph in a separate window	Successful	High
7	View pyLDAvis graph	View pyLDAvis graph in browser	1. Click on 'View pyLDAvis graph' button	User should be able to see pyLDAvis graph in browser	User was able to view graph and each topics occurence	Successful	High

Table 1: System Test Cases

6.3.Implementation Result and Conclusion

6.3.1.Coherence Score Graph

The coherence Topic Score Graph is showing that the Anti-vax topics are discussed more as compared to Pro-tax topics. Hence based on the Facebook posts provided to this system, most of the People are NOT willing to take COVID-19 vaccine shots.

The Score graph is calculated based on the LDA extracted topics that we fetched and their weightage of occurrence in the Facebook Posts.

That data was compared with the keywords, I provided to the system with segregation that keyword like ‘anti’, ‘fear’, ‘vaccine’ were ‘anti-vax’ topics and keywords like ‘coronavirus’, ‘vaccine’ were ‘pro-vax’ topics.

Based on the above provided data, this coherence score was plotted showing that ‘anti-vax’ topics mentioned more that ‘pro-vax’ topics.

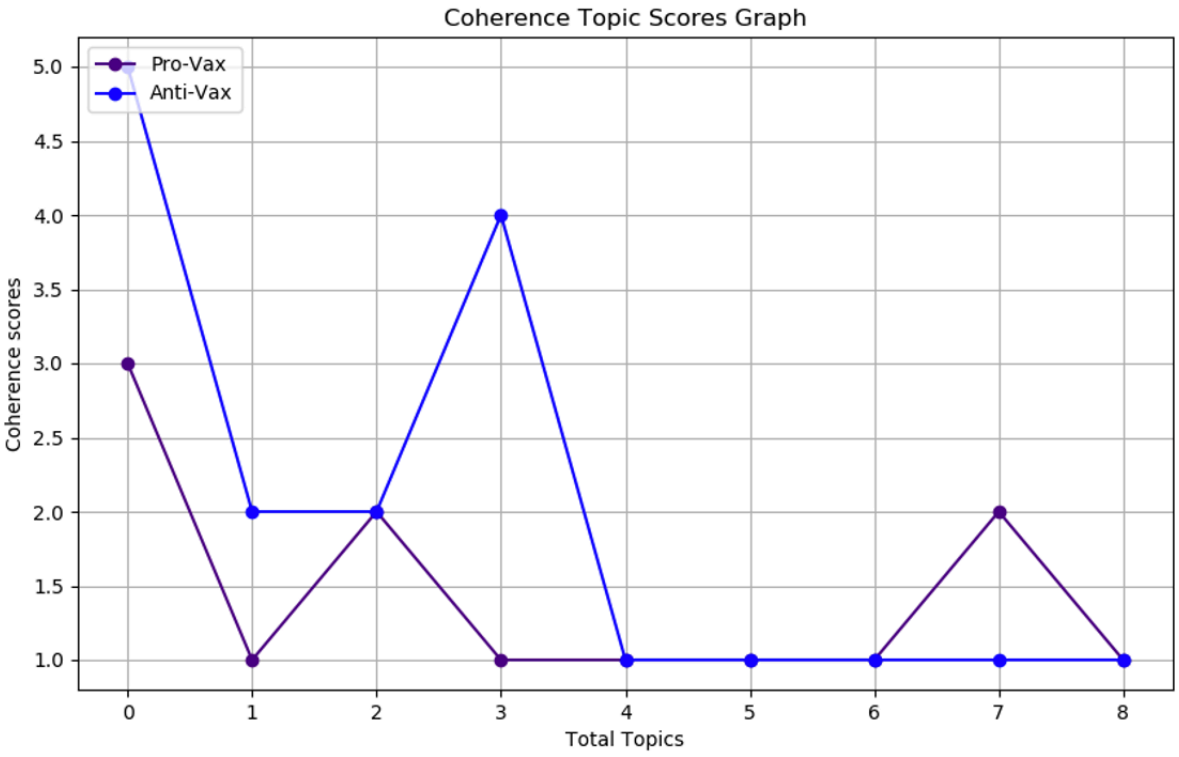

Figure 5: Project Coherence Topic Score Graph

6.3.2.pyLDAvis Graph

To Visualize our coherence score based on the LDA extracted topics, I have used pyLDAvis graph.

In pyLDAvis graph below are points to notice:

- The large circles refers to the topics that have been discussed more number of times in the Facebook Posts.
- The Circles closely concentrated means those topics have cooccurred, I.e they have been mentioned one after another.
- To view topic-vise analysis, we can skip to next topic with the given Next Topic tab above.
- If we hover over the Most Salient Terms, we can see their occurrences in each topic, highlighted with red circles.

Hence based on Visualization graph, ‘vaccine’ and ‘protest’ have been discussed more and cooccurrence with each other. Which means the discussion is inclined more towards, ‘anti-vax’.

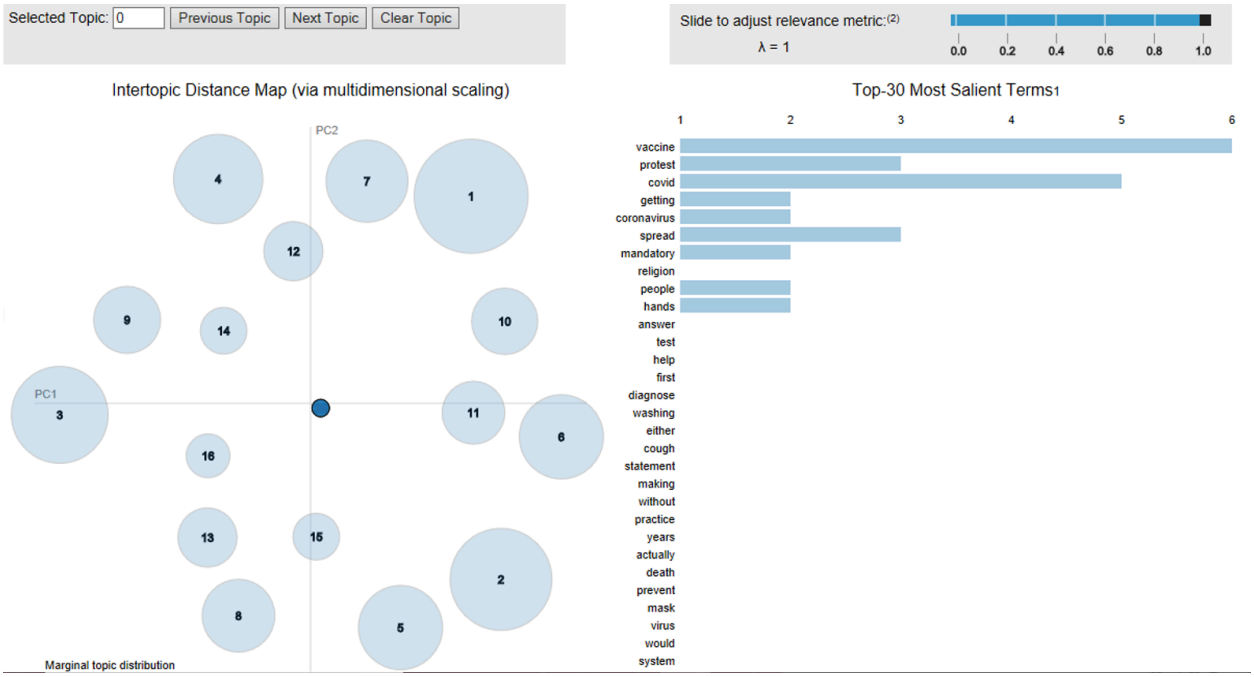

Figure 15 :Project Final pyLDAvis Graph

7. SUMMARY AND FUTURE WORK

The final finding of the project is that with the given set of Facebook Posts, the community is inclined more towards “anti-vaccine” discussion, as the LDA extracted topics weightage of occurrence is more over ‘anti-vax’ topics.

This project was developed based on the base paper [1], to analyze and see how LDA can efficiently segregate “anti” and “pro” topics. As we can see that other than some minor errors, like inclusion of unwanted words, the LDA algorithm was able to analyze each topic and the number of times any term has been discussed.

7.1.How User can leverage this system?

Users can input as many number of Facebook Posts they want, to make this model learn and provide results. It is not necessary to use this system only for COVID-19 vaccine inclination, but it can also be used to understand popular discussion topics.

For that feed in posts from Facebook or Twitter or from any online Social platforms into the the system from ‘Upload Posts’ button. Then use ‘pyLDAvis’ graph to visualize the most discussed topics, instead of viewing Coherence Score graph.

7.2.How readers can improve this system?

The analysis in this project however is hugely based on manual moderators, tasked with segregate certain unnecessary words. The future work can be to remove those manual moderators and make use of algorithms to support LDA work efficiently in identifying plausible topics within collection of posts from any online community, not just Facebook.

Also, as you can notice, this project was fed Facebook Topics through a document that was created manually by going through the Facebook Communities and discussion going on on Covid. The future, work could be to work out an algorithm that can be efficiently used to automate fetching FB posts and then analyze where the major discussion are inclined.

Another improvement is to make a new tab for uploading, segregation topics based on which the LDA model will build its Coherence Score graph. Also, provide a tab to build

graph based on parameters for x-axis and y-axis. This way we can make this system more generalized to understand other topics other than ‘Covid-19 Vaccine’

This system can also be modified to display a more User-friendly conclusion message, like ‘The discussion was inclined towards anti-vaccine’. This way it will be more user-friendly.

8. REFERENCES

[1] R. F. Sear et al., "Quantifying COVID-19 Content in the Online Health Opinion War Using Machine Learning," in IEEE Access, vol. 8, pp. 91886-91893, 2020, doi: 10.1109/ACCESS.2020.2993967.

[2] https://towardsdatascience.com/latent-dirichlet-allocation-lda-9d1cd064ffa2

[3] Fang A., Macdonald C., Ounis I., Habel P. (2016) Topics in Tweets: A User Study of Topic Coherence Metrics for Twitter Data. In: Ferro N. et al. (eds) Advances in Information Retrieval. ECIR 2016. Lecture Notes in Computer Science, vol 9626. Springer, Cham. https://doi.org/10.1007/978-3-319-30671-1_36

[4] https://radimrehurek.com/gensim/auto_examples/tutorials/run_word2vec.html [5] https://www.nltk.org

[6] (2020). Latent Dirichlet Allocation. Accessed: Apr. 13, 2020. [Online]. Available: https://en.wikipedia.org/wiki/ Latent_Dirichlet_allocation

[7] M. Röder, A. Both, and A. Hinneburg, “Exploring the space of topic coherence measures,” in Proc. 8th ACM Int. Conf. Web Search Data Mining (WSDM), 2015, pp. 399–408, doi: 10.1145/2684822.2685324.

[8] C. Sievert and K. Shirley, “LDAvis: A method for visualizing and inter- preting topics,” in Proc. Workshop Interact. Lang. Learn., Visualizat., Interface, 2014, pp. 63–70, doi: 10.3115/v1/W14-3110.

[9] Coronavirus Disease (COVID-19) Advice for the Public: Myth Busters, W. H. Organization, Geneva, Switzerland, 2020. Accessed: Apr. 13, 2020. [Online]. Available: https://www.who.int/emergencies/diseases/novel- coronavirus-2019/advice-for-public/myth-busters

[10] A. Kata, “A postmodern Pandora’s box: Anti-vaccination misinforma- tion on the Internet,” Vaccine, vol. 28, no. 7, pp. 1709–1716, Feb. 2010, doi: 10.1016/j.vaccine.2009.12.022.

[11] pLSA and LDA - people.cs.pitt.edu. https://people.cs.pitt.edu/~milos/courses/cs3750-Spring2020/Slides/class13.pdf

[12] Dumais, S. T., Furnas, G. W., Landauer, T. K. and Deerwester, S. (1988), "Using latent semantic analysis to improve information retrieval." In Proceedings of CHI'88: Conference on Human Factors in Computing, New York: ACM, 281-285.

[13] Enthought Python Minimum Hardware Requirements – Enthought https://support.enthought.com/hc/en-us/articles/204273874-Enthought-Python-Minimum-Hardware-Requirements

[14] The Only Software Requirements Document Template You Need. https://arkenea.com/blog/software-requirements-document-template/

[15] S. Syed and M. Spruit, "Full-text or abstract? Examining topic coherence scores using latent Dirichlet allocation", *Proc. IEEE Int. Conf. Data Sci. Adv. Analytics (DSAA)*, pp. 165-174, Oct. 2017.

[16] *Latent Dirichlet Allocation*, Apr. 2020, [online] Available: https://en.wikipedia.org/wiki/Latent_Dirichlet_allocation

[17] L. Givetash, Global measles cases surge amid stagnating vaccinations, New York, NY, USA:NBC News, Apr. 2019, [online] Available: https://www.nbcnews.com/news/world/global-measles-cases-surge-amid-decline-vaccinations-n1096921

[18] H. J. Larson, "Blocking information on COVID-19 can fuel the spread of misinformation", *Nature*, vol. 580, no. 7803, pp. 306, Apr. 2020.

[19] R. Iyengar, The Coronavirus is Stretching Facebook to its Limits, New York, NY, USA:CNN Business, Apr. 2020, [online] Available: https://www.cnn.com/2020/03/18/tech/zuckerberg-facebook-coronavirus-response/index.html.

[20] S. Frenkel, D. Alba and R. Zhong, Surge of Virus Misinformation Stumps Facebook and Twitter, Apr. 2020, [online] Available: https://www.nytimes.com/2020/03/08/technology/coronavirus-misinformation-social-media.html.

[21] D. A. Broniatowski, A. M. Jamison, S. Qi, L. AlKulaib, T. Chen, A. Benton, et al., "Weaponized health communication: Twitter bots and russian trolls amplify the vaccine debate", Amer. J. Public Health, vol. 108, no. 10, pp. 1378-1384, Oct. 2018.

[22] Y. Lama, T. Chen, M. Dredze, A. Jamison, S. C. Quinn and D. A. Broniatowski, "Discordance between human papillomavirus Twitter images and disparities in human papillomavirus risk and disease in the united states: Mixed-methods analysis", J. Med. Internet Res., vol. 20, no. 9, Sep. 2018.

[23] T. Ammari and S. Schoenebeck, "Thanks for your interest in our facebook group but it's only for dads: Social roles of Stay-at-Home dads", Proc. 19th ACM Conf. Comput.-Supported Cooperat. Work Social Comput., pp. 1361-1373, 2016.

[24] N. F. Johnson, R. Leahy, N. J. Restrepo, N. Velasquez, M. Zheng, P. Manrique, et al., "Hidden resilience and adaptive dynamics of the global online hate ecology", Nature, vol. 573, no. 7773, pp. 261-265, Sep. 2019.

[25] N. F. Johnson, M. Zheng, Y. Vorobyeva, A. Gabriel, H. Qi, N. Velasquez, et al., "New online ecology of adversarial aggregates: ISIS and beyond", Science, vol. 352, no. 6292, pp. 1459-1463, Jun. 2016.