

Analysing Covid- 19 Content From Social Media Using Latent Dirichlet Allocation

Submitted in partial fulfillment of the requirements for the award of
Bachelor of Engineering degree in Computer Science and Engineering

by

CHILEKAMPALLI PAVANIGANGA (37110144)

CHIGICHERLA BHAVYASREE (37110141)



**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
SCHOOL OF COMPUTING**

SATHYABAMA

**INSTITUTE OF SCIENCE AND
TECHNOLOGY (DEEMED TO BE
UNIVERSITY)**

Accredited with Grade “A” by NAAC

JEPPIAAR NAGAR, RAJIV GANDHI SALAI, CHENNAI - 600 119

MARCH 2021



SATHYABAMA

INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)

Accredited with Grade “A” by NAAC | 12B Status by UGC | Approved by AICTE
JEPPIAAR NAGAR, RAJIV GANDHI SALAI, CHENNAI -600 119

www.sathyabama.ac.in



DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

BONAFIDE CERTIFICATE

This is to certify that this Project Report is the bonafide work of **CHILEKAMPALLI PAVANIGANGA (Reg. No.37110144)** who carried out the project entitled “**Analysing Covid-19 Content From Social Media Using Latent Dirichlet Allocation.**” Under our supervision from **November 2020 to March 2021.**

Internal Guide

Dr. A. YOVAN FELIX M.E., Ph.D

Head of the Department

Submitted for Viva voce Examination held on _____

Internal Examiner

External Examiner

DECLARATION

I CHILEKAMPALLI PAVANIGANGA (Reg.No.37110144) hereby declare that the Project Report entitled **“Analysing Covid-19 Content From Social Media Using Latent Dirichlet Allocation.”** done by me under the guidance of **Mr.A.YOVAN FELIX, M.E., (Ph.D.)** is submitted in partial fulfillment of the requirements for the award of Bachelor of Engineering degree in Computer Science and Engineering.

DATE:

PLACE: CHENNAI

SIGNATURE OF THE CANDIDATE

ACKNOWLEDGEMENT

I am pleased to acknowledge my sincere thanks to **Board of management of SATHYABAMA INSTITUTE OF SCIENCE AND TECHNOLOGY** for their kind encouragement in doing this project and for completing it successfully . I am grateful to them.

I convey my thanks to **Dr. T.SASIKALA M.E., Ph.D., Professor and Dean, School of Computing** and **Dr.S.Vigneshwari, M.E.,Ph.D., and Dr.L.Lakshman, M.E., Ph.D., Head of the Department of Computer Science And Engineering** for providing us the necessary support and details at the right time during the progressive reviews.

I would like to express our sincere and deep sense of gratitude to our Project Guide **Dr.A.YOVAN FELIX, M.E.,Ph.D., Associative professor,**Department of Computer Science And Engineering for his valuable guidance, suggestions and constant encouragement paved way for the successful completion of my project work.

I wish to express our thanks to all Teaching and Non-teaching staff members of the Department of **COMPUTER SCIENCE AND ENGINEERING** who were helpful in many ways for the completion of the project.

ABSTRACT

A large amount of potentially dangerous COVID-19 misinformation is appearing online. Here we use machine learning to quantify COVID-19 content among online opponents of establishment health guidance, in particular vaccinations (“anti-vax”). We find that the anti-vax community is developing a less focused debate around COVID-19 than its counterpart, the pro-vaccination (“pro-vax”) community. However, the anti-vax community exhibits a broader range of “flavors” of COVID-19 topics, and hence can appeal to a broader cross-section of individuals seeking COVID-19 guidance online, e.g. individuals wary of a mandatory fast-tracked COVID-19 vaccine or those seeking alternative remedies. Hence the anti-vax community looks better positioned to attract fresh support going forward than the pro-vax community. This is concerning since a widespread lack of adoption of a COVID-19 vaccine will mean the world falls short of providing herd immunity, leaving countries open to future COVID-19 resurgences. We provide a mechanistic model that interprets these results and could help in assessing the likely efficacy of intervention strategies. Our approach is scalable and hence tackles the urgent problem facing social media platforms of having to analyze huge volumes of online health misinformation and disinformation.

TABLE OF CONTENTS

CHAPTER NO.	TITLE	PAGE NO.
	ABSTRACT	i
	TABLE OF CONTENT	ii
	LIST OF FIGURES	iv
	LIST OF ABBREVIATIONS	v
1	INTRODUCTION	1
	1.1 Introduction	1
	1.2 Motivation	2
2	LITERATURE SURVEY	4
3	AIM AND SCOPE	6
	3.1 Aim of Project	6
	3.1.1 Objectives	6
	3.2 Scope	6
	3.2.1 Advantages	6
4	SYSTEM DESIGN & METHODOLOGY	7
	4.1 Existing System	7
	4.2 Proposed System	7
	4.3 System Functionalities	7
	4.4 Requirement Specification	9
	4.5 Modules	10

CHAPTER NO.	TITLE	PAGE NO.
5	SOFTWARE DEVELOPMENT METHODOLOGY	12
	5.1 Description of Diagram	12
	5.2 Use Case Diagram	12
	5.3 Architectural Design	13
6	RESULTS AND DISCUSSION	14
	6.1 Screenshots	14
7	CONCLUSION	18
	REFERENCES	19

LIST OF FIGURES

FIGURE NO.	TITLE OF THE FIGURE	PAGE NO.
5.1	Flow chart	12
5.3	Work Flow Implementation	13
6.1	Dataset loading	14
6.2	Processing Data	14
6.3	Extracting topic	15
6.4	Viewing Vaccine Topics	15
6.5	Pro Vax & Anti Vax topic	16
6.6	Graph	16
6.7	pyLDAvis Topic Visualization	17

LIST OF ABBREVIATIONS

ACRONYM

EXPANSION

LDA

Latent dirichlet allocation

CV

Chorence metrix

FB

Facebook

CHAPTER 1

INTRODUCTION

N

1.1 INTRODUCTION

Scientific experts agree that defeating COVID-19 will depend on developing a vaccine. However, this assumes that a sufficiently large proportion of people would receive a vaccine so that herd immunity is achieved. Because vaccines tend to be less effective in older people, this will require younger generations to have very high COVID-19 vaccination rates in order to guarantee herd immunity. Yet there is already significant opposition to existing vaccinations, e.g. against measles, with some parents already refusing to vaccinate their children. Such vaccine opposition increased the number of cases in the 2019 measles outbreak in the U.S. and beyond. Any future COVID-19 vaccine will likely face similar opposition increased the number of cases in the 2019 measles outbreak in the U.S. and beyond. Any future COVID-19 vaccine will likely face similar opposition. Mandatory COVID-19 vaccinations for schoolchildren could trigger a global public health conflict. A better understanding of such opposition ahead of a COVID-19 vaccine is therefore critical for scientists, public health practitioners, and governments.

Online social media platforms, and in particular the built-in communities that platforms like Facebook (FB) feature, have become popular fora for vaccine opponents (anti-vax) to congregate and share health (mis)information. Such misinformation can endanger public health and individual safety. Likewise, vaccine supporters (pro-vax) also congregate in such online communities to discuss and advocate for professional public health guidance. Well before COVID-19, there was already an intense online conflict featuring anti-vax communities and pro-vax communities. Within anti-vax communities, the narratives typically draw on and generate misinformation about establishment medical guidance and distrust of the government, pharmaceutical industry, and new technologies such as 5G communications. Adding fuel to this fire, the January 2020 birth of the COVID-19

“infodemic” has led to a plethora of misinformation in social media surrounding COVID-19 that directly threatens lives. For example, harmful “cures” are being proposed such as drinking fish tank additives, bleach, or cow urine, along with

coordinated threats against public health officials like Dr. Anthony Fauci, director of the U.S. National Institute of Allergic and Infectious Diseases. Moreover, false rumors have been circulating that individuals with dark skin are immune to COVID-19. These may have contributed to more relaxed social distancing among some minorities and hence their over-representation as victims. In Chicago and Louisiana as of early April 2020, 70% of the fatalities were African Americans even though this demographic only makes up 30% of the population. In addition, the world has witnessed an alarming rise in COVID-19 weaponization against the Asian community. It is also clear that such misinformation is not a fringe phenomenon, and can instead be very widely held as true within the general population. Indeed, a recent Pew study found that 30% of Americans believe the COVID-19 virus was likely created in a laboratory, despite statements from infectious disease experts to the contrary.

Unfortunately, the sheer volume of new online content and the speed with which it spreads, means that social media companies are struggling to contain such health misinformation. Making matters worse, people around the world are spending more time on social media due to social distancing imposed during the COVID-19 pandemic. This increases the likelihood that they become exposed to such misinformation, and as a result they may put themselves and their contacts at risk with dangerous COVID-19 remedies, cures and falsehoods.

1.2 MOTIVATION

The present study is motivated by both these needs: the need for a deeper understanding of this intersection between online vaccination opposition and the online conversation surrounding COVID-19; and the need for an automated approach since the sheer volume of new online material every day makes manual analysis a non-viable option going forward. We pursue an automated, machine learning approach that avoids the scalability limitations of manual content analysis. While the present paper is just the first step in a challenging longer-term goal, the automated approach that we present allows the following questions to be addressed: How did COVID-19 change the online conversation within anti-

vaccination and pro-vaccination communities over the two month period in early 2020 when the disease became a global threat.

Unlike many existing works, this study does not use Twitter data since it is known that Twitter is more of a broadcast medium for individual shout-outs, whereas discussions and narratives tend to be nurtured in in-built online community spaces that are a specific feature of platforms like Facebook (e.g., fan page). Twitter does not have such in-built community spaces. In the present methodology, generalized from and data is collected from these online communities, specifically Facebook Pages that support either anti-vaccination or pro-vaccination views. This information is publicly available and does not require any individual details, thereby avoiding any privacy concerns – just as understanding the content of a conversation among a crowd of people in an open, real-world public space does not require knowledge of any personal details about the individuals within that crowd. Details of our approach are given in Sec. II and the Appendix. A third difference between this study and previous ones is that the machine learning findings here are interpreted in terms of a mechanistic model (Sec. IV) that captures the general trend for the coherence in the online conversations over time. Though much work still needs to be done, this study therefore provides a first step toward a fully automated but interpretable understanding of the growing public health debate concerning vaccines and COVID-19.

Online social media platforms, and in particular the built in communities that platforms like Facebook (FB) feature, have become popular fora for vaccine opponents (anti-vax) to congregate and share health (mis)information. Such misinformation can endanger public health and individual safety. Likewise, vaccine supporters (pro-vax) also congregate in such online communities to discuss and advocate for professional public health guidance. Well before COVID-19, there was already an intense online conflict featuring antivax communities and pro-vax communities. Within anti-vax communities, the narratives typically draw on and generate misinformation about establishment medical guidance and distrust of the government, pharmaceutical industry, and new technologies such as 5G communications.

CHAPTER 2

LITERATURE SURVEY

Kata, “A postmodern Pandora’s box: Antivaccination misinformation on the Internet,” *Vaccine*, vol. 28, no. 7, pp. 1709–1716, Feb. 2010, doi: 10.1016/j.vaccine.2009.12.022.

The Internet plays a large role in disseminating anti- vaccination information. This paper builds upon previous research by analyzing the arguments proffered on antivaccination websites, determining the extent of misinformation present, and examining discourses used to support vaccine objections. Arguments around the themes of safety and effectiveness, alternative medicine, civil liberties, conspiracy theories, and morality were found on the majority of websites analyzed; misinformation was also prevalent. The most commonly proposed method of combating this misinformation is through better education, although this has proven ineffective. Education does not consider the discourses supporting vaccine rejection, such as those involving alternative explanatory models of health, interpretations of parental responsibility, and distrust of expertise. Anti- vaccination protestors make postmodern arguments that reject biomedical and scientific “facts” in favour of their own interpretations. Pro-vaccination advocates who focus on correcting misinformation reduce the controversy to merely an “educational” problem; rather, these postmodern discourses must be acknowledged in order to begin a dialogue. With morbidity and mortality from vaccine-preventable diseases having reached record lows, vaccines are one of the most successful tools for biomedical science and public health. Yet paradoxically, the effectiveness of vaccination has led to the reemergence of anti-vaccination sentiments. Vaccines may be seen as unnecessary or dangerous because incidence rates of VPDs in developed countries have plummeted. Vaccine “reactions” – negative health events following vaccination, attributed to the vaccine – then appear to be more common than the diseases themselves. In this way, vaccines can be considered victims of their own success. The media plays a large role in disseminating and sensationalizing vaccine objections. Such objections are part of what has been

called the “anti-vaccination movement”, which has had a demonstrable impact on vaccination policies, and individual and community health. A common sequence to

vaccination scares involves scientific debate about potential vaccine risks, which communication technology transmits via a rhetoric of doubt; parents incorporate this with personal experiences and spread their views to their social groups.

B. Martin, "Texas anti-vaxxers fear mandatory COVID-19 vaccines more than the virus itself," Texas Monthly, 2020. [Online].

The World Health Organization lists vaccine hesitancy as one of 10 threats to global health. The antivaccine movement uses Facebook to promote messages on the alleged dangers and consequences of vaccinating, leading to a reluctance to immunize against preventable communicable diseases. We would like to know more about the messages these websites are sharing via social media that can influence readers and consumers. What messages is the public receiving on Facebook about immunization? What content is being promoted. We proposed using a social media audit tool and 3 categorical lists to capture information on websites and posts, respectively. The keywords "vaccine," "vaccine 10 truth," and "anti- vax" were entered in the Facebook search bar. A Facebook page was examined if it had between 2500 and 150,000 likes. Data about beliefs, calls to action, and testimonials were recorded from posts and listed under the categories Myths, Truths, and Consequences. Website data were entered in a social media audit template. Users' posts reflected fear and vaccine hesitancy resulting from the alleged dangers of immunization featured on the website links. Vaccines were blamed for afflictions such as autism, cancer, and infertility. Mothers shared testimonies on alleged consequences their children suffered due to immunization, which have influenced other parents to not vaccinate their children. Users denied the current measles outbreaks in the United States to be true, retaliating against the government in protests for fabricating news. Some Facebook messages encourage prevailing myths about the safety and consequences of vaccines and likely contribute to parents' vaccine hesitancy. Deeply concerning is the mistrust social media has the potential to cast upon the relationship between health care providers and the public. A grasp of common misconceptions can help support health care provider practice. Many diseases have been almost. Immunization against disease prevents 2-3 million deaths per year internationally and could prevent with global vaccination.

CHAPTER 3

AIM AND SCOPE

3.1 AIM OF PROJECT

The primary aim of this proposed system is to quantify PRO and ANTI information by utilizing social media public posts where users will post their view on COVID-19 and vaccines and to identify whether post is related to ANTI or PRO author is using various techniques and algorithm.

3.1.1 OBJECTIVES

The anti-vax community exhibits a broader range of “flavors” of COVID-19 topics, and hence can appeal to a broader cross-section of individuals seeking COVID-19 guidance online, e.g. individuals wary of a mandatory fast-tracked COVID-19 vaccine or those seeking alternative remedies. Hence the anti-vax community looks better positioned to attract fresh support going forward than the pro-vax community. This is concerning since a widespread lack of adoption of a COVID-19 vaccine will mean the world falls short of providing herd immunity, leaving countries open to future COVID-19 resurgences.

3.2 SCOPE

This model will present a methodology to identify COVID-19 spreaders using the analysis of the relationship between socio-cultural and economic characteristics with the number of infections and deaths caused by the COVID-19 virus in different countries.

3.2.1 ADVANTAGES

- Accuracy
- Graph visualization using pyLDAvis

CHAPTER 4

SYSTEM DESIGN & METHODOLOGY

4.1 EXISTING SYSTEM

The existing methods using Most models assume a standard SEIR structure. Fraser and colleagues to estimate size but make different changes on the nature of the different compartments and their respective residence times. There are of course many limitations of this study. There are other social media platforms, apart from Facebook, that should be explored but Facebook is the largest. Similar behaviors should arise in any platform where communities can form.

4.2 PROPOSED SYSTEM

The proposed system i.e. we use machine learning to quantify COVID-19 content among online opponents of establishment health guidance, in particular vaccinations ("anti-vax"). We find that the anti-vax community is developing a less focused debate around COVID-19 than its counterpart, the pro-vaccination ("pro-vax") community. However, the anti-vax community exhibits a broader range of "flavors" of COVID-19 topics, and hence can appeal to a broader cross-section of individuals seeking COVID-19 guidance online, e.g. individuals wary of a mandatory fasttracked COVID-19 vaccine or those seeking alternative remedies. Hence the anti-vax community looks better positioned to attract fresh support going forward than the pro-vax community. This is concerning since a widespread lack of adoption of a COVID-19 vaccine will mean the world falls short of providing herd immunity, leaving countries open to future COVID-19 resurgences. We provide a mechanistic model that interprets these results and could help in assessing the likely efficiency of intervention strategies.

4.3 SYSTEM FUNCTIONALITIES

The terms 'Facebook Page' and 'cluster' are used interchangeably here since each Facebook Page is a cluster of people. Facebook Pages, also known as fan pages or public pages, are accounts that represent organizations, causes,

com- munities, or public figures. According to Facebook's policies, "Content posted to a Page is public and can be viewed by everyone who can see the Page". A Facebook

Page is different from a Facebook personal account. Personal accounts represent private individuals, and their posts and interactions are considered more private and targeted to their immediate contacts. This paper does not analyze data from personal accounts. Our methodology follows and by analyzing the public content of Facebook Pages for both anti- vaccination (“anti-vax”) and pro-vaccination (“pro- vax”) communities. The publicly available content of these online communities is obtained using a snowball approach, starting with a seed of manually identified pages discussing either vaccines, public policies about vaccination, or the pro-vs- anti vaccination debate. Then their connections to other fan pages are indexed. At each step, new clusters are evaluated through a combination of human coding and computer-assisted filters.

To classify a cluster as being anti-vax or pro-vax and including COVID-19 content or not, we reviewed its posts and the Page’s “about” section. Pro-vax and anti-vax classifications required that either (a) at least 2 of the most recent 25 posts dealt with the pro-vax or anti-vax debate, or (b) the page’s title or “about” section described it as pro-vax or anti-vax. At least two researchers classified each cluster independently. If they disagreed on their suggested classification, a third researcher reviewed the posts and then all three reviewers discussed these cases. Agreement was reached in each case. This also enabled us to distinguish between content that is intended to be serious versus merely satirical. The self-weeding tendency within Facebook Pages tends to reduce material from bots and fake profiles. We kept the present study focused on English, though this can be easily generalized using our same procedure. Beyond that, our study was global and not limited to a particular region.

The content of these clusters was then bundled together separately for the anti- vax community and the pro-vax community, and the two resulting sets of content were analyzed using machine learning. Specifically, we used an unsupervised machine learning technique called Latent Dirichlet Allocation (LDA) to analyze the emergence and evolution of topics around COVID-19. The LDA method models documents as distributions of topics and topics as distributions of words. During its training process, these distributions are adjusted to fit the dataset. The LDA method is described correctly in Wikipedia as “[quote] .. a

generative statistical model that allows sets of observations to be explained by unobserved groups that explain

collected into documents, it posits that each document is a mixture of a small number of topics and that each word's presence is attributable to one of the document's topics. LDA is an example of a topic model and belongs to the machine learning toolbox and in wider sense to the artificial intelligence toolbox."

The coherence score provides a quantitative method for measuring the alignment of the words within an identified topic. It is generated from a separate algorithm which is run over a trained LDA model. The overall coherence score of a single model is the arithmetic mean of its per-topic coherences. There are many different coherence metrics to evaluate per-topic coherence. We use CV which is based on a sliding window, one-set segmentation of the top words and an indirect confirmation measure that uses normalized point-wise mutual information and the cosine similarity. It comprises collections of probability measures on how often top words in topics co-occur with each other in examples of the topics. We refer to for a full explanation and discussion of CV.

Machine learning automation can, in principle, help address the significant issues facing social media platforms by mechanically picking out material that requires attention from the huge haystack of online content. While this could help to better curtail online misinformation, one might rightly ask about its accuracy and reliability as compared to human analysts. This has been recently addressed in. We use the same coherence metric (CV) as these authors. They addressed the problem that topic models had previously given no guarantee on the interpretability of their output. Specifically, they produced several benchmark datasets with human judgements of the interpretability of topics and they found results that outperformed existing measures with respect to correlation to human ratings. They achieved this by evaluating 237,912 coherence measures on 6 different benchmarks for topic coherence, making this the biggest study of topic coherences at that time. Separately, we have done our own comparison for the general area of online hate and have found comparable consistency.

4.4 REQUIREMENT SPECIFICATION

This proposed software runs effectively on a computing system that has the minimum requirements. So the main need is to introduce appropriate equipment

for the model. The requirements are split into two categories, namely:

Software Requirements

The basic software requirements to run the program are:

The functional requirements or the overall description documents include the product perspective and features, operating system and operating environment, graphics requirements, design constraints and user documentation. The appropriation of requirements and implementation constraints gives the general overview of the project in regards to what the areas of strength and deficit are and how to tackle them.

- Python idel 3.7 version (or)
- Anaconda 3.7 (or)
- Jupiter (or)
- Google colab

Hardware Requirements

The basic hardware required to run the program are very dependent on the particular 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.

- Operating system : windows, linux
- Processor : minimum intel i3
- Ram : minimum 4 gb
- Hard disk : minimum 250gb

4.5 MODULES

This process leads to a set containing many hundreds of pages for both the anti-vax and pro-vax communities. Before training the LDA models, several steps are employed to clean the content of these pages in a similar way to other LDA analyses in the literature:

Step 1: Mentions of URL shorteners are removed, such as “bit.ly” since these are fragments output by Facebook’s Crowd Tangle API.

Step 2: Many of the posts link to external websites. The fact that these specific

websites were mentioned could itself be an interesting component of the COVID19.

Hence instead of removing them completely, the pieces “.gov”, “.com”, and “.org” were replaced with “ gov”, “ com”, and “ org”, respectively. This operation effectively concatenates domains into a form that will not be filtered out by the later preprocessing steps.

Step 3: The posts are then run through Gensim’s `simple_preprocess` function, which tokenizes the post on spaces and removes tokens that are only 1 or 2 characters long.

This step also removes numeric and punctuation characters.

Step 4: Tokens that are in Gensim’s list of stopwords, are removed. For example, “the” is not a good indication of a topic.

Step 5: Tokens are lemmatized using the WordNetLemmatizer from the Natural Language Toolkit NLTK, which converts all words to singular form and/or present tense.

Step 6: Tokens are stemmed using the SnowballStemmer from NLTK, which removes affixes on words.

Step 7: Any remaining fragments of URLs (other than domain) that are left over after stemming, such as “http” and “www”, are removed.

Steps 5 and 6 help ensure that words are compared fairly during the training process, and that if a particular word is a strong indicator of a topic, its signal is not lost just because it is used in many different forms. These steps rely on words existing in NLTK’s pretrained vocabulary. Any word not in this vocabulary is left unchanged. After this preprocessing, we then train the LDA models on the cleaned data. Specifically, 10 separate LDA models were trained with the “number of topics” parameter ranging from 3- 20, for a total of 180 models in each of the two time intervals T_1 and T_2 . The C_v coherence algorithm was then run over each of these models and the coherence scores were then averaged for each number of topics. To produce the results, multiple trials were run for each number of topics to ensure that the coherence for a particular number of topics is representative of what LDA models tend to find and is not the result of unaccounted noise swaying the model to overfit in other way.

CHAPTER 5

SOFTWARE DEVELOPMENT METHODOLOGY

5.1 DESCRIPTION OF DIAGRAM

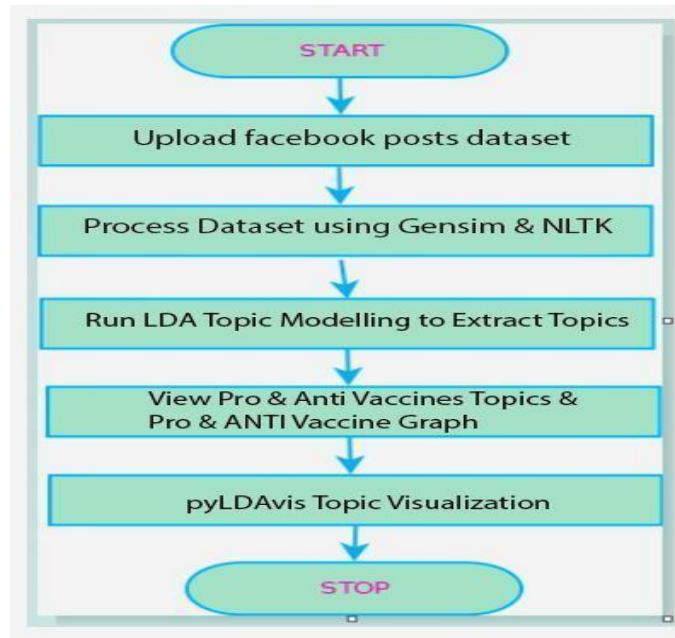
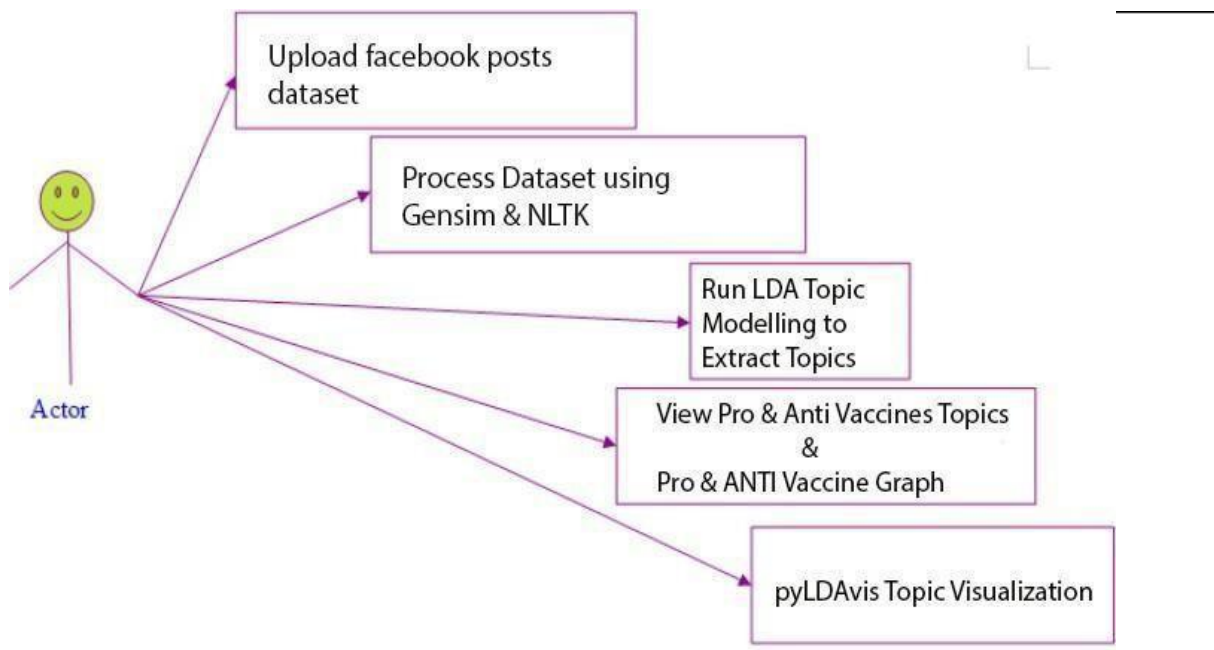


Fig 5.1: FLOW CHART

5.2 USE CASE DIAGRAM



5.3 ARCHITECTURAL DESIGN

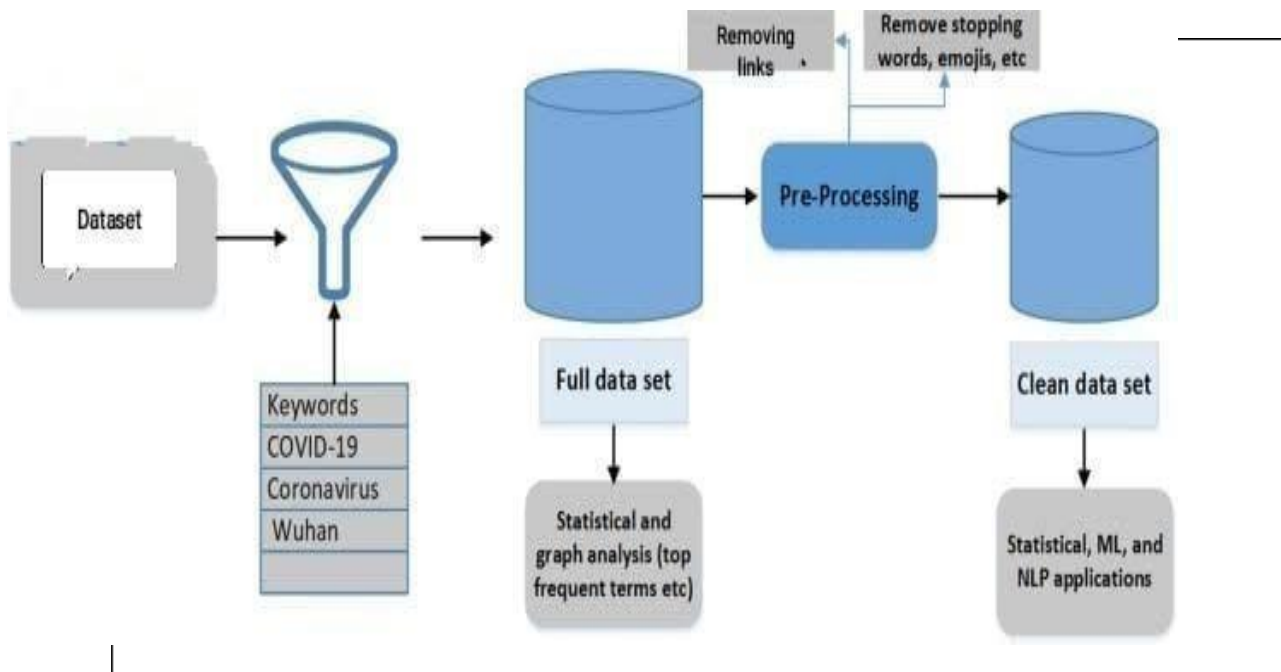


Fig 5.3. Work Flow Implementation

The architecture clearly explains about the functioning of the project. Initially a dataset is taken and trained. Later all the keywords, stop word such as “bit.ly” since these are fragments output by social media. The fact that these specific websites were mentioned could itself be an interesting component of the COVID-19 conversation. Hence instead of removing them completely, the pieces “.gov”, “.com”, and “.org” were replaced with “ gov”, “ com”, and “ org”, respectively. This operation effectively concatenates domains into a form that will not be filtered out by the later pre-processing steps. The posts are then run through Gensim’s simple preprocess function, which tokenizes the post on spaces and removes tokens that are only 1 or 2 characters long. This step also removes numeric and punctuation characters. Tokens that are in Gensim’s list of stopwords, are removed. For example, “the” is not a good indication of a topic. Tokens are lemmatized using the Word Net Lemmatizer from the Natural Language Toolkit NLTK, which converts all words to singular form and/or present tense. Any remaining fragments of URLs (other than domain) that are left over after stemming, such as “http” and “www”, are removed. Hence the output is predicted based on the post data from social media.

CHAPTER 6

RESULTS AND DISCUSSION

6.1 SCREENSHOTS

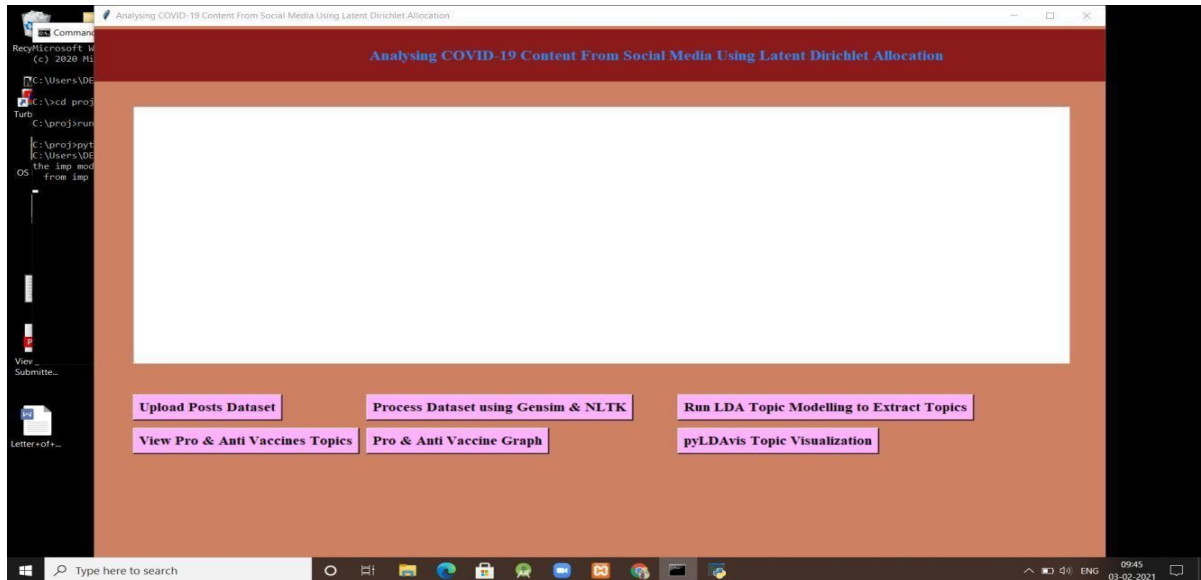


Fig 6.1. Dataset loading

Fig 6.1. is the home page of the project. Click on 'Upload Facebook Posts Dataset' button to load dataset and we upload 'Posts.csv' file and then click on 'Open' button to load dataset.



Fig 6.2. Processing Data

Fig 6.2. Now clicking on ‘Process Dataset using Gensim & NLTK’ button is used to read dataset and to clean dataset.



Fig 6.3. Extracting topics

Fig 6.3. After cleaning will get above text from all posts and now clicking on ‘Run LDA Topic Modelling to Extract Topics’ button to extract topics from all posts.

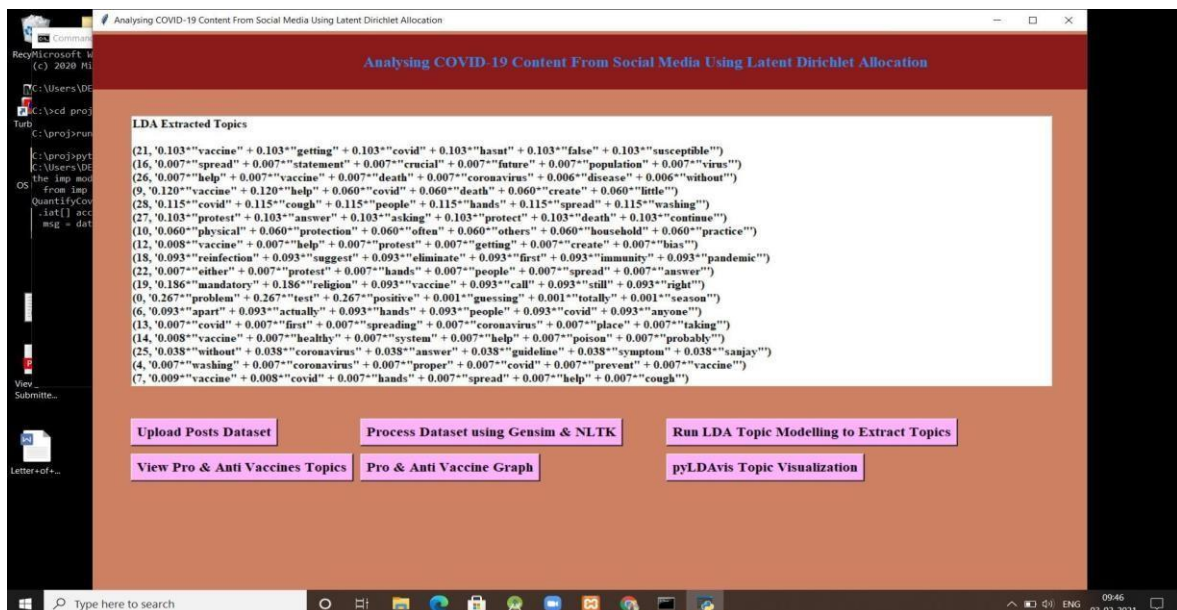


Fig 6.4. Viewing Vaccine Topics

Fig 6.4. TOPIC and its probability score for each word from POSTS and now clicking on 'View Pro & Anti Vaccines Topics' button to view all topics with coherence score.

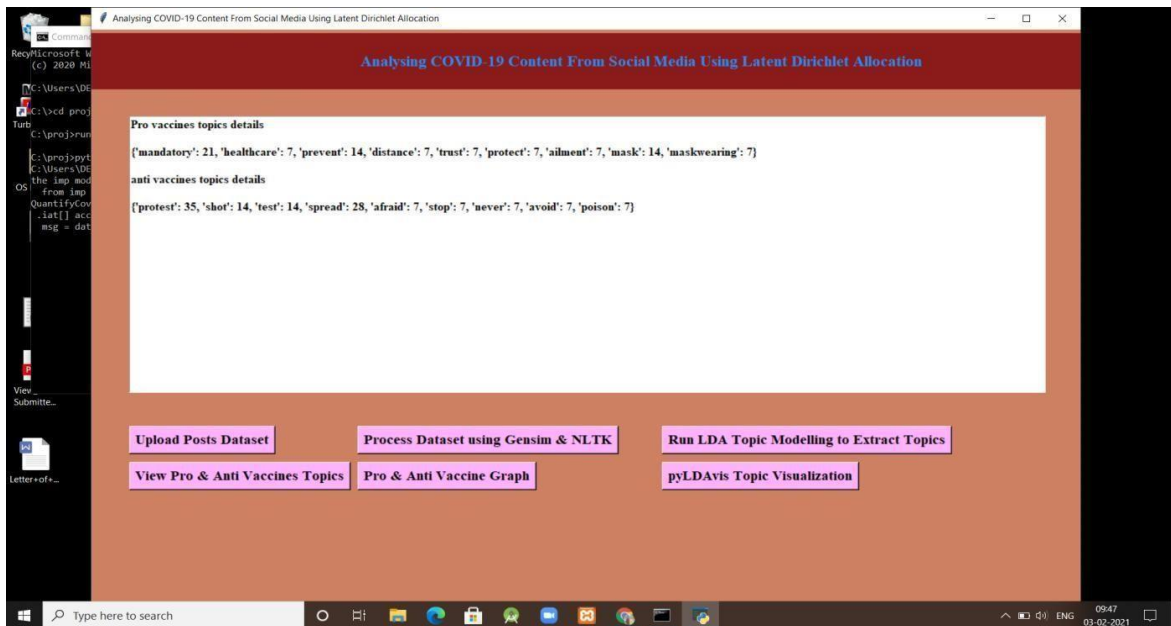


Fig 6.5. Pro Vax & Anti Vax topics

Fig 6.5. All topics of PRO and ANTI with count value and now clicking on 'Pro & ANTI Vaccine Graph' button to get below graph and to analyse number of peoples are in favour of PRO or ANTI.

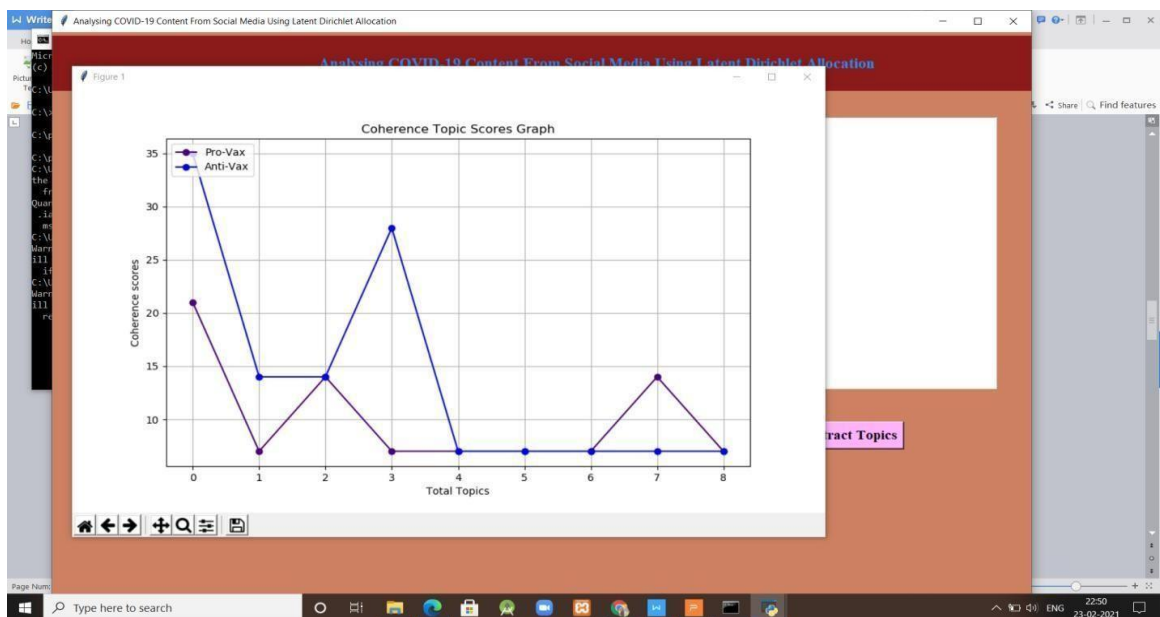


Fig 6.6. Graph

Fig 6.6. In above graph x-axis represents number of topics and y-axis represents coherence score and in above graph blue line refers to ANTI and indigo colour line refers to PRO vaccine and from above graph we can conclude more peoples are

discussing ANTI topics about vaccine. Now click on 'pyLDavis Topic Visualization' to get visualization of all topics in browser.

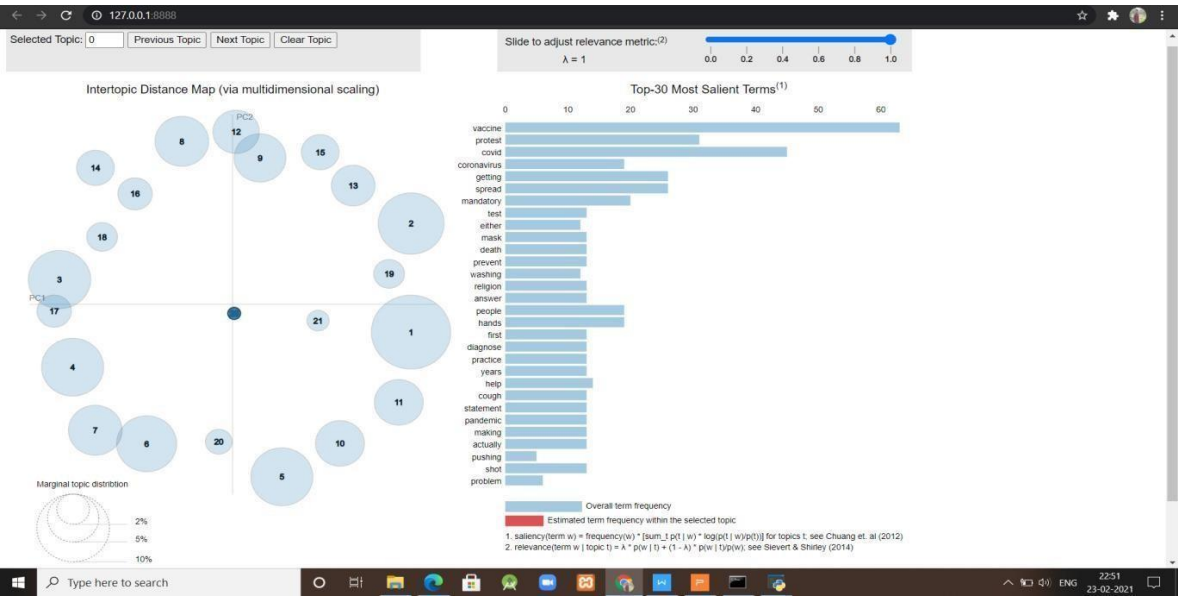


Fig 6.7. pyLDavis Topic Visualization

Fig 6.7. Similarly we can see that the In above graph large circle refers that topic occurs more number of time and if occur less number of times then its circle will be in small size. In right size we can see all topics from that posts and u can click on 'Next Topic' button from top side of window to get next topic visualization. In above graph each circle represents 1topic.

CHAPTER 7

CONCLUSIO

N

7.1 CONCLUSION

The output predicts whether the input given from social media content VAX or antiVax using the LDA algorithm. A methodology to identify COVID-19 spreaders using the analysis of the relationship between socio-cultural and economic characteristics with the number of infections and deaths caused by the COVID-19 virus in different countries using 5-layer multiplex network.

These findings suggest that the online anti-vax community is developing a more diverse and hence more broadly accommodating discussion around COVID- 19 than the pro-vax com- munity. As a result, the pro-vax community runs the risk of making itself less engaging to the heterogeneous ecology of potential new users who join the online COVID-19 discussion, and who may arrive online with a broad set of concerns, questions, and possibly preconceived notions, misinformation and even falsehoods.

The analysis in this paper also provides a first step toward eventually either replacing, or at least supplementing, the non-scalable efforts of human moderators tasked with iden- tifying online misinformation. In addition, the mechanistic model (Fig. 3) could be used for what-if scenario testing of how quickly coherence develops and what the impact would be of breaking up the coherence around certain topics.

Overall, this approach shows that a machine-learning algo- rithm, the LDA algorithm, identifies plausible topics within collections of posts from online communities surrounding the vaccine and COVID-19 debate. In addition to being able to handle large quantities of data, its results emerge quickly using statistical grouping techniques, instead of having to rely on potentially biased, slow and costly human labelling.

REFERENCES

- [1] C. Sievert and K. Shirley, "LDAvis: A method for visualizing and interpreting topics," in Proc. Workshop Interact. Lang. Learn., Visualizat., Interface, 2014, pp. 63– 70, doi:10.3115/v1/W14-3110.
- [2] Facebook Policies. (2020). Pages, Groups and Events Policies. Accessed: Apr. 13, 2020. [Online]. Available: https://www.facebook.com/policies/pages_groups_events
- [3] Harrison Jr, D. and Rubinfeld, D.L., 1978. Hedonic housing prices and the demand for cleanair.
- [4] N. F. Johnson, R. Leahy, N. J. Restrepo, N. Velasquez, M. Zheng, P. Manrique, P. Devkota, and S. Wuchty, "Hidden resilience and adaptive dynamics of the global online hate ecology," Nature, vol. 573, no. 7773, pp. 261– 265, Sep. 2019, doi:10.1038/s41586-019-1494-7.
- [5] N. F. Johnson, M. Zheng, Y. Vorobyeva, A. Gabriel, H. Qi, N. Velasquez, P. Manrique, D. Johnson, E. Restrepo, C. Song, and S. Wuchty, "New online ecology of adversarial aggregates: ISIS and beyond," Sci- ence, vol. 352, no. 6292, pp. 1459–1463, Jun. 2016, doi:10.1126/sci- ence.aaf0675.
- [6] P. D. Manrique, M. Zheng, Z. Cao, E. M. Restrepo, and N. Johnson, "Generalized gelation theory describes onset of online extremist support," Phys.Rev.Lett., vol. 121, no. 4, Jul. 2018, Art.no.048301, doi:10.1103/PhysRevLett. 121.048301.
- [7] Shinde, N. and Gawande, K., 2018, October. Survey on predicting property price. In 2018 International Conference on Automation and Computational Engineering (ICACE) (pp. 1- 7). IEEE.
- [8] S. Syed and M. Spruit, "Full-text or abstract? Examining topic coherence scores using latent Dirichlet allocation," in Proc. IEEE Int. Conf. Data Sci. Adv. Analytics (DSAA), Oct. 2017, pp. 165–174, doi: 10.1109/DSAA.2017.61.

