International Journal of Engineering Applied Sciences and Technology, 2020
Vol. 5, Issue 5, ISSN No. 2455-2143, Pages 302-307
Published Online September 2020 in IJEAST (http://www.ijeast.com)

EVALUATING COVID-19 HEALTH INFORMATION USING MACHINE LEARNING

Buddavarapu Teja Swaroop,
Computer Science student at SNIST

Abstract - A high volume, variety and variability amount of COVID-19 information (data) is available online which might be inaccurate and incorrect. In this paper we quantify, analyse the available COVID-19 content about health guidance using ML and topic modelling, mainly concerned on vaccinations ("anti-vaccination"). It is found that the anti-vaccination community is less focused part around COVID-19 as compared to counterpart, the pro-vaccination (pro-vaccination) community. However, the anti-vaccination community has a broader range of "flavours" of COVID-19 topics, so it has a broader cross-section of individual opinion and their version in COVID-19 guidance available online, e.g. some of them wary of important fast-tracked COVID-19 vaccine and some developing and studying alternative remedies. Hence the anti-vaccination community looks better than the pro-vaccination community as has more support and awareness in the society. This is increasing since a widespread lack of awareness and adoption of a vaccine which makes the world to face in fall rapidly in providing herd immunity, which results in leaving countries open to future resurgences. We developed a ML model that analyse these data and interprets results and helps in assessing the efficiently provide strategies. In this paper the approach done is a scalable and so, tackles the problem facing due to social media platforms which is spreading variety, various and huge amount of misinformation and disinformation which may panic public and this paper is not resembling that all information available are not of these type but most of them are.

Keywords- Machine learning, topic modelling, social computing, medical information systems, community.

I. INTRODUCTION

Scientific and research experts believe that COVID-19 can be stopped will be depending on a vaccine which is being tried by many experts. However, the vaccine developed will be sufficient to people all over the world by which herd immunity [1] is achieved. As vaccines are generally less effective in older people, by which younger generations requires large COVID-19 vaccination rates in order to provide herd immunity. But there is significant negative response to already available vaccinations, e.g. In present scenario the most dangerous disease measles, where most of the parents are refusing to vaccinate their children. Such vaccine opposition resulted in increase of number of cases in the measles during 2019 outbreak in the U.S. and beyond [2]. In the same way the upcoming COVID-19 vaccine may also face similar opposition [3] [4]. Mainly for school children which might trigger a global public health conflict. Such opposition in COVID-19 vaccine might therefore impact critical situations for scientists and also for the government.

Social media platforms which are available online like Facebook (FB), have become so much popular in fora for vaccine (anti-vaccination) to analyse and share health information through online to spread awareness among people which is not checked and cross verified. Such information cannot be checked so it may arise and endanger society health and in particular individual safety and security. In the same way, people who support vaccine (pro-vaccination) can also spread in such communities to discuss, analyse and prescribe for professional public health guidance. Not only today there was already an intense online debate and conflict about anti-vaccination communities and pro-vaccination communities exist. Among the anti-vaccination communities, the records typically drawn and given rise to misinformation on establishing health guidance and lost trust in the government, medical industry, and also in new technologies [5]. The strength added to this after the 2020 birth of COVID situation which has led to a superfluity of information available in social media surrounding COVID-19 that has direct impact on society which may threatens lives [6]. For example, drinking hot water and by adding some spices, adding additives to fish tank, bleach, or also adding cow urine and dung. Besides, some rumours are also circulating such as people with dark skin are healthy and can also be immune to COVID-19. These rumours might lead to relaxed and less social distancing among some communities and hence results in 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 [8][9]. In addition, the world has witnessed an alarming rise in COVID-19 weapon against the Asian community [10] [11] [12]. It makes clear that such information is not an edge phenomenon, and can lead to widely treated as true among the population [13]. Indeed, a recent study by scientists found that nearly 35% of people in America believe that COVID-19 virus took birth in a laboratory, despite statements from infectious disease experts to the contrary.

Woefully, the sheer volume of new content available which is sometimes misinformation and the speed with which it spreads, shows that online social media is struggling a lot to contain such misinformation. Making things hard and bad, many people all over the world are spending more time on social media, as the social distancing imposed during the COVID-19 pandemic situation [14] [15]. This added and imposed the possibility that such misinformation is spread more among people so that they may put themselves and their family at risk with dangerous COVID-19 remedies, cures and falsehoods. In this paper the below points are discussed- The understanding between the intersections among vaccination opposition and conversation due to COVID-19 and the other need is to adopt an automated approach for the analysis as manual analysis is more complex and a non-viable option to choose.

In this paper the data used is not only from one social media "Twitter", as it only consists of partial data and model accuracy is less. So the data is taken from most famous and widely used "Facebook" which consists of more information than that available in Twitter, as more number of users and more number of posts. In this the individual personal information is not taken as there will be privacy issues associated with the customers. By using this data gave better details to model which gave good results and increased accuracy.

II. DATA AND MACHINE LEARNING ANALYSIS

The terms 'Facebook Pages' and 'group of clusters' are used conversely in this paper (model) as each Facebook Page is a cluster of people individual data. These pages in Facebook are commonly known as fan pages or public pages and these consists of accounts which represents communities, organizations, causes or public figures. As per Facebook policies, "Data present in a Page is public mode and it is visible to anyone who can see the Page". A Facebook Page is different from a persona1 account as personal accounts consists of private individual data and their posts and interactions are privately secured and targeted to their immediate contacts. So in this paper the private data is not analysed i.e., data from personal accounts. In this paper it follows by analysing the public content of Facebook Pages for both anti-vaccination ("anti-vaccination") and pro-vaccination ("pro-vaccination") communities.

The approach used is snowball by which the contents in public mode in communities (online), these discuss either vaccines, public policies about vaccination, or the pro-vs-anti vaccination debate. Then these connections are indexed with other pages. At each step, new clusters are evaluated through a combination of human coding and computer-assisted filters. To classify these clusters into (1) anti-vaccination or pro-vaccination and (2) including COVID-19 content or not, these data is reviewed from the posts and the Page's "about" section. Pro-vaccination and anti-vaccination classifications requires either (a) at least 2 of the most recent 25 posts dealt with the pro-vaccination or anti-vaccination debate, or (b) the page's title or "about" section described it as pro-vaccination or anti-vaccination. In this the researchers from different areas classified each cluster independently. They debated on their suggested classification, so another researcher came into the process and reviewed the posts and then all three reviewers discussed these cases. Agreement was reached in each case. This also helped in differentiating between the serious matters and merely satirical. The special process of removing unnecessary things tendency within Facebook Pages helped to reduce data from chat bots and fake accounts. The study in this paper is not confirmed to one region but throughout the world and every matter necessary which made the accuracy of model to increase and this data is best suited for this project idea.

The cluster contents from Facebook data is then bundled together separately so that we get the anti-vaccination community and the pro-vaccination community, and from these two the results of content were analysed using machine learning. In this way by using unsupervised machine learning technique (Latent Dirichlet Allocation, LDA) the data is analysed and different areas of topics were discovered about Covid-19. In this the LDA method is used to model data documents which is distributed into topics and these topics are futher distributed as bag of words. While training, these distributions are arranged to fit the dataset. In LDA, if data is a collection of words then it is collected into a document, then it predicates into separate document in the form of a small number of topics and each word's presence in the document is attributed as one of the document's topics. LDA is a topic modelling tool used and present in machine learning toolbox and in wider sense present in artificial intelligence toolbox."

The coherence score which is used as a quantitative method for measuring the alignment of the words within the document as an identified topic. It is produced from an algorithm which is executed on a trained model of LDA. The overall coherence score of a single model is calculated as the arithmetic mean of its per-topic coherences. There are many techniques available to evaluate per-topic coherence. In this model we used *C*V which is based on a sliding window protocol, one-set segmentation of the top words and an indirect confirmation measure that uses normalized point-wise mutual information and the cosine similarity. It incorporates collection of probability measures in which top layer words in topics can co-occur with each other. ML automation helps in addressing the remarkable issues surrounding online social media platforms by insensible topics out from the huge contents taking birth in available data online. This might also help in reducing the insensible online misinformation, which can also be helpful in increasing the accuracy and reliability when compared with manually identifying. In this paper the same coherence metric (*C*V) is used as many intellectuals. They labelled the problem that topic modelling previously has not given interpretability results in many of their research output. Significantly, these people also produced some benchmarks to the datasets with human analysis and ML analysis on these like topics and they found that the results produced outperformed existing results with respect to correlation to human analysis ratings. They achieved this by evaluating 237,912 coherence measures on 6 different benchmarks for topic coherence, which made it the biggest study of topic coherences as compared with present times. In this paper our own comparisons are also done for this general area of online hate and have found comparable consistency in the results.

Hence in this paper, the machine learning approach used identifies content topics in social media online with good high coherence i.e., the word clusters are identified which are strongly related with each other by using an approach coherence scoring as mentioned previously. This is how data is collected and analysed where human analysed words seemed to have less accuracy and reliability and the model generated by LDA technique have showed good results.

III. RESULTS

The crucial area we need to focus mainly is that endogenetic development of COVID-19 information during the pandemic situation at early period and in prior during the start of deaths trolled in major countries. So we mainly gathered social media online data from most widely used Facebook of public posts during February and April period. To update the changes over period to period, the whole period is divided into small time intervals. When we have more time intervals will indirectly results in smaller amounts of data within each period and consequently more variations in the model, so the model is divided into two time intervals, *P1* and *P2*. The first time-interval is from 2/17/2020-3/17/2020 (*P1*) which contains of a total of 1123 pro-vaccination posts and comments, and 4670 total anti-vaccination posts and comments. The second time-interval is taken from 3/27/2020-4/27/2020 (*P2*) which consists of a total of 976 pro-vaccination posts and replies, and 4200 total anti-vaccination posts and replies. So these two time slots or simply windows are equal and consists of similar amounts of data.

The results obtained are relatively robust when compared with other time intervals. Fascinatingly, P1 time interval roughly corresponds to the time where COVID-19 situation is taken seriously and largely considered as a problem in Asia, while *P2* time interval is corresponds to the time during which Covid-19 situation has become as a serious issue in Europe and some other countries. For more quality and proof that the data taken about the COVID-19 conversation during these intervals, we also compared this data within the articles that counts from worldwide Anglophone newspapers and worldwide Google trends.

The LDA model was trained over the posts taken in the following distinct groups: anti-vaccination posts in *P1*, anti-vaccination posts in *P2*, pro-vaccination posts in *P1*, and pro-vaccination posts in *P2*. For each of these sets, 12 separate LDA models were trained with parameter of number of topics ranging from 3-20, for a total of 180 models in each of the four groups. Later the *C*V coherence algorithm was executed with each of these models and respective coherence scores were taken and then mean of these scores were taken for each number of topics. These averaged scores are plotted in Figure. 1 which shows the result of the same procedure applied to all posts in our dataset, and to all anti-vaccination posts, and to all pro-vaccination posts.

The coherence score *C*V for the entire period of study (i.e. *P1*+*P2*) in Fig. 1, shows it is consistently larger across the number of topics for pro-vaccination than for anti-vaccination, suggesting that the pro-vaccination community and

International Journal of Engineering Applied Sciences and Technology, 2020
Vol. 5, Issue 5, ISSN No. 2455-2143, Pages 302-307
Published Online September 2020 in IJEAST (http://www.ijeast.com)

overall it is more focused discussion around COVID-19 than the anti-vaccination. This consistency with the pro-vaccination community shows a more rigid outcome around public health mainly about how it is focused on advising and giving suggestions to people in order to follow professional medical guidance.

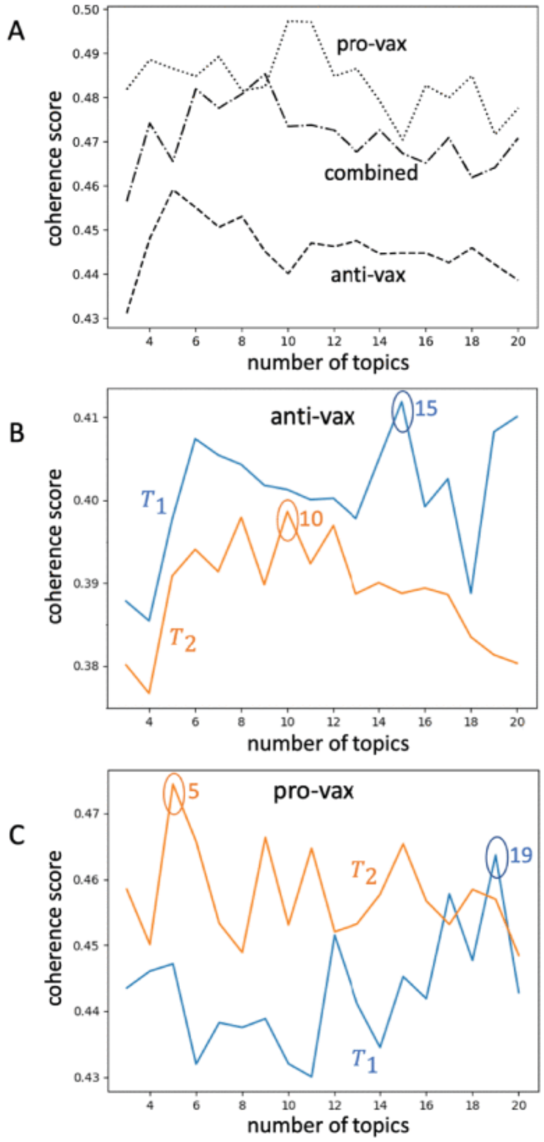

Fig. 1.Experimental Results

The concerning thing about pro-vaccination community is that the overall coherence obtained is higher and it is not well settled to engrossed with the wide variety of more unfocused areas of COVID-19 contents says that these are spreading more in online. This shows that a significant potential disadvantage for the pro-vaccination community which results in low attraction and attention among people all over the world who are now emerging into online area in search for a particular refinement 'palates' of COVID-19 which describes that revives to the users. These users could consequently be pulled toward the anti-vaccination cause.

Figure 1B and 1C represents the graph between the coherence scores on the topics count compared with respect to changes over time which gives some curves over the time periods *P1* and *P2*. The curve moves up from *P1* to *P2* for the pro-vaccination community and the optimal number of topics shows a dramatic decrease in the graph. This is compatible with the conception that the pro-vaccination community is working for a common exposition and describes with comparably less Covid-19 'palates' of discussion and exposition than the anti-vaccination community. This gives strength as it suggests that the pro-vaccination community is less enhancing over some period of time as many types of people of all kinds all over the world, who are in search of their perspective COVID-19 'palates'. By contrast, the curves for the anti-vaccination community from *P1* to *P2* has shown a smaller reduction in the number of topics which are optimal (17 to 11) and the curves moved down, in the direction reverse to the pro-vaccination. Hence the anti-vaccination curve has shown a slow increase in focus by which there is a coherence reduction overall, i.e. these 11 number of topics for *P2* are effectively more focused as compared to the original 17 for *P1*, and so is the overall anti-vaccination community has became and shown more accommodating to the manifold population of new additions coming into the online health space over period of time.

The changes which has shown in the pro-vaccination community during the time period *P1* to *P2* is such that the optimal number of topics decreases (i.e., the number of circles decreases from 18 to 6 which is observed in Fig. 1C) and the number of topics generated are mainly located in the same portion of the space (i.e., toward the right-hand side in Fig. 2D). As analysing Fig. 1B, the change in the anti-vaccination community from time period *P1* to *P2* is such that the optimal number of topics starts off slightly smaller than the pro-vaccination and also shown a decreases over time (i.e., the number of circles decreases) which implies there are more topics (i.e., more circles) than for the pro-vaccination in time period *P2*. The number of topics are also more scattered in space in Fig. 2B when it is compared with Fig. 2D. The results when observed shows that these are linear as compared with earlier expounding that the pro-vaccination community is more focused as compared with anti-vaccination community in terms of COVID-19 topics, and that the pro-vaccination community is developing towards COVID-19 exposition and chronicle with a less diverse other than that of anti-vaccination community.

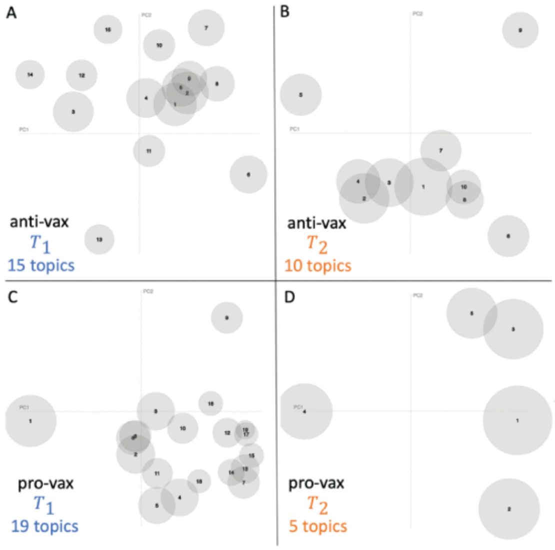

Fig. 2. ML analysis

IV. TOWARD A ARBITRARY MODEL INTERPRETATION

For this an arbitrary model corresponding to future support with factual findings and which is minuscule using machine learning for the results. Concretely, in this paper computer simulated method is generated by which the online contents from the overall community, each of which is designated by a vector, for each component *xi* resembles the robustness for a topic which resembles the online contents. The components obtained from the communities of online data can be a bag of words or may sometimes be a group of short phrases. The arbitrary model chosen for this topic is very simple, which indeed reflect the factual observations and belles-lettres surrounding the topic of online discussions, as listed above are studied in detail by Kata (a detailed pattern). Later we choose the components which are selected randomly to develop the content topics.

The Components are later clustered together as if the *x* values are sufficiently similar (Fig. 3A) or different (Fig. 3B). The output of the model is shown in Fig. 3 which is a one-dimensional view of result. When we compare the two-dimensional curves for the same data also gave merely similar results, though it is visually more complex as it has a third dimension of component time. The results obtained also gave similar curves as in Fig. 2 when viewed visually. When we analyse the curves in Fig. 3, has a confluent which is quicker when compared with the Figs. 1 and 2 for the pro-vaccination community. By observing, the heterophony is slower to gel, it is more consistent with the anti-vaccination community in Figs. 1 and 2. The dotted red coloured line in Figs. 3A and 3B represents the area of miniature which is more consistent when compared with the Figures. 2D and 2B, these represent's the pro-vaccination and anti-vaccination communities respectively.

The time gap in the gel period is analysed in Fig. 3B for heterophony (anti-vaccination) when analysed with homophile (pro-vaccination) in Fig. 3A, the analysed result obtained using mathematical statistical analysis. We can analyse that the gel formation time is inversely proportional to the probability average between the two randomly taken components from one cluster, in which the homophile is larger when compared with heterophony and so the gel formation period is later for heterophony than homophile exactly as observed in Fig. 3. This can also be visualized by a mathematical way where the gel formation sizes i.e., the circles present in Fig.2, which is larger for homophile as compared with heterophony which can be observed in Fig.3,

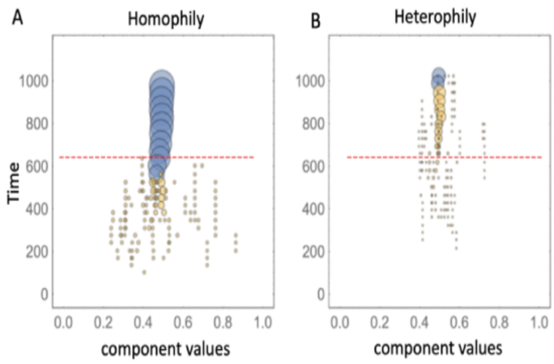

Fig.3 Comparing model

Inspite this on the bright side for the pro-vaccination community, the counterfeit of this arbitrary model explain that the homophile (pro-vaccination) is less stable to absorb an in pouring in new users with a range of *x* values, when seen with the heterophony (anti-vaccination). The anti-vaccination community is more appealing to new data of people and hence the anti-vaccination is good and more comfortable to users and can have a long span run as compared with the pro-vaccination case.

V. LIMITATIONS OF THE STUDY

There might be many limitations occurring in future and the data from other social media platforms available of which Facebook is widely used and has more data from any other online media, so the data from Facebook is choosen which covers most of the topics around Covid-19. It will also be interesting, to compare the results with other data where studies on Twitter, where data is available more in the form of short phrases of individual statements. There is also a chance for influence of external factors and agents. However, the social media online communities behave like if they are bots or show troll behaviour. Further more data and analysis is to be done for the details of the content. This might need surpass not only just text and even LDA also, since online data may also consists of memes and images which also contains data and this trend is followed by many communities. The model generated outputs needs to be validated in detail for the time-evolution of topics. Finally more research is also to be done for formulating the result outputs across various platforms in detail, applicative out-run for policy makers. These limitations will be addressed in future work.

VI. CONCLUSION

These topics suggest that the online anti-vaccination community is developing miscellaneously and accommodating discussion around COVID-19 when compared with the pro-vaccination community. As a result, the pro-vaccination community is at edge by means of itself appealing in ecology of heterogeneous with potential to new data by the people in the COVID-19 online community of discussion, and may result in concerns, questions online and may possibly come to conclusions for predetermined conception, misinformation and even may result in fabrication also.

The analysis by machine learning techniques provides a first step for the due course either by restoring, or by means of augmenting the non-flexible attempts of human mediator process of pinpointing online misinformation. In addition, the arbitrary model as in Fig.3 might be helpful for the situation for testing the speed at which coherence score is developing and what might be the effect that the coherence might cause in certain number of topics. This is obtained by using the empirical analysis and the results are shown in Fig.2, and it is repeated for some time intervals so as to identify the development of topics and new topics which are becoming popular as a remedy with bleach as best example. This might also result in Facebook in the form of advertisements posts that contains these new topics and phrases.

Overall, the method resembles from a machine-learning algorithm LDA which identifies feasible number of topics within the data collected from the posts of online social media communities around and about vaccine and COVID-19 related subject. In order to handle large volumes of online data, which results in emerging statistical tools like clustering and grouping techniques, which is more reliable and robust when compared with human analysis.

VII. REFERENCES

[1] New Kata, "A Pandora box: Anti-vaccination information in the Online Internet", Vaccine, vol. 28, pp. Feb. 2010.

[2] Global measles cases surge amid stagnating vaccinations, New York, NY, USA: NBC News, Apr. 2019

[3] B. Martin, Texas Anti-Vax Fear Mandatory COVID-19 Vaccines More than the Virus Itself, Austin, TX, USA: Texas Monthly, 2020

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

[5]R Scientists Brand 5G Claims ’Complete Rubbish, London, U.K: BBC News, Apr. 2020.

[6] Covid-19 Disease Advice for the Public: Myth Busters, Geneva, Switzerland: WHO, Apr. 2020.

[7] After Threats Anthony Fauci to Receive Enhanced Personal Security, New York, The New York Times, Apr. 2020

[8] H. Yan and M. Holcombe, Covid-19 Pandemic is hitting more in African-American Communities very Hard, New York, NY, Apr. 2020

[9] A. Maqbool, Why Virus Hit African Americans so Hard, London, U.K: BBC News, Apr. 2020

[10] J. Guy, East Asian Student Assaulted in ’Racist’ Coronavirus Attack in London, Apr. 2020

[11] Korean Interpreter, M. Rajagopalan, ’Chinese Tried to Punch Her Off Her Bike’, New York, USA: BuzzFeed News, Apr. 2020