Machine Learning Analysis of the Side-Effects of Chloroquine-Hydroxychloroquine and Lopinavir–Ritonavir in the Treatment of COVID-19

Abstract

Chloroquine-Hydroxychloroquine and Lopinavir-Ritonavir are two of the many combination of drugs being used to treat COVID-19. These drugs are causing serious heart rhythm problems, blood and lymph system disorders, kidney injuries, and liver problems and failure. Here the Latent Dirichlet Allocation (LDA) model has been used to study the trends in the side-effects of these drugs while being used to treat COVID-19. LDA model is an example of a topic model and belongs to the machine learning toolbox. With the literature on COVID-19 increasing rapidly, it is tough for the medical community to keep up to it. The approach used here is scalable and hence can analyse huge volumes of data. News articles published all over the world, that spoke about these two drugs, were collected to make the dataset. The positives and the negatives of these drugs were analysed. It is found that the number of topics for the negatives of these drugs are increasing and is also becoming more diverse from each other with increase in time, implying that the problems of these drugs are increasing with time and hence more people are in danger of being affected by these drugs. The results here draw an impression of the scenario the news reports are making over time and calls for strict guidelines on the usage of these drugs. Hence this approach tackles the urgent problem facing the medical community of having to analyse huge volumes of data to understand the public response on the treatment for COVID-19.

Keywords

Latent Dirichlet Allocation (LDA) model; side-effects; Chloroquine; Hydroxychloroquine; Lopinavir-Ritonavir; COVID-19.

Introduction

A huge number of news articles related to COVID-19 have been published, most of them are trying to bring to notice the alarming problems the world is currently facing. A lot of news articles are being published regarding the side-effects of the drugs being used to treat COVID-19, many reports suggest that [1] most of the drugs being used have no proof of being able to treat COVID-19. A few of the drugs being used have serious side-effects on the patients, like heart [11], kidney, liver and nervous problems. There has been a widespread adoption of the drugs Chloroquine-Hydroxychloroquine and Lopinavir-Ritonavir to treat COVID-19 in patients who have tested positive. There were no exact medicines known to cure the disease and the number of cases was rapidly increasing. These two drugs showed some good signs on the patients and though there was no exact proof for these drugs to be capable of treating the virus these two drugs had a widespread adoption all over the world to treat the patients.

Here are the prior studies in the context of this research. According to the reports, liver and kidneys can be damaged in patients with COVID-19, which may make reaching the therapeutic dose of the medicines difficult and increase the risk of adverse drug reactions in patients [2]. Chloroquine's lethality and behavioral side effects are significant and may be underestimated. Moderately low

overdosage of chloroquine can result in rapid death. Moreover, therapeutic doses are known to cause psychosis, delirium, personality changes, and depression [3]. Chloroquine can interfere with intracellular functions, inhibits enzyme activity, has anti-inflammatory activity, and effects immune functions [4]. A number of biomedical text mining systems have been developed to extract biologically relevant information directly from the literature, complementing bioinformatics methods in the analysis of experimentally generated data [5], a similar usage of LDA model is seen here. Use of LDA model to prove that the anti-vax communities in social media platforms are attracting more people and are spreading misinformation related to COVID-19 [6], it uses machine learning to solve a problem in a way similar to the one used in this paper. Lopinavir–ritonavir analysis is also presented in this paper and in hospitalized adult patients with severe COVID-19, no benefit was observed with lopinavir–ritonavir treatment beyond standard care [12].

The hypothesis made before the research was, ‘Chloroquine-Hydroxychloroquine and Lopinavir-Ritonavir have severe side-effects and continued usage might be more harmful’. The LDA model was trained on the dataset extracted from all the news reports on these two drugs, the results prove the hypothesis made. Though the results show that there is a significant improvement in the positive effects of these drugs with time, the model has established that the negative impacts are becoming more serious with prolonged usage of these drugs. The results here draw an impression of the scenario the news reports are making over time and this makes it easier, because keeping upto all the news articles is infeasible, for the medical community to make important guidelines on the usage of these drugs and to suggest safer drugs with lesser side-effects. The solution is systematically built up to the proof of the hypothesis, the time intervals are defined, the data scrapping techniques are told, then the methodology in the preprocessing of the data and the application of the LDA model is presented, later the transparency in the selection of optimal number of topics is shown just before the intertopic distance map proves the hypothesis with some explanation.

Methodology

News articles report the different kind of negative impacts these medicines have been having in different places [7]. With the increasing number of articles and them being so varied in nature it becomes increasingly difficult for the medical community, the government and the public to keep up to the problems. These problems are reported in different regions and news is the only way through which the public can get to know about them. The higher rated news articles have been picked and the articles were uniformly distributed over the time interval. A natural language processing model, Latent Dirichlet Allocation (LDA) model has been applied to classify the trends of these news articles and to give a conclusion that a reader would otherwise come to by reading all the published news articles [8] [10]. An LDA model is a generative statistical model that allows sets of observations to be explained by unobserved groups that explain why some parts of the data are similar. For example, if observations are words 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 makes predictions by estimating the probability that a new set of inputs belongs to each class. The class that gets the highest probability is the output class and a prediction is made.

The LDA model uses Bayes Theorem to estimate the probabilities. Briefly Bayes’ Theorem can be used to estimate the probability of the output class (k) given the input (x) using the probability of each class and the probability of the data belonging to each class:

$P(Y=x|X=x) = (P_k * f_k(x)) / sum(P_k * f_k(x))$

Where P_k refers to the base probability of each class (k) observed in your training data (e.g. 0.5 for a 50-50 split in a two class problem).
The f(x) above is the estimated probability of x belonging to the class. A Gaussian distribution function is used for f(x). Plugging the Gaussian into the above equation and simplifying we end up with the equation below. This is called a discriminate function and the class is calculated as having the largest value:

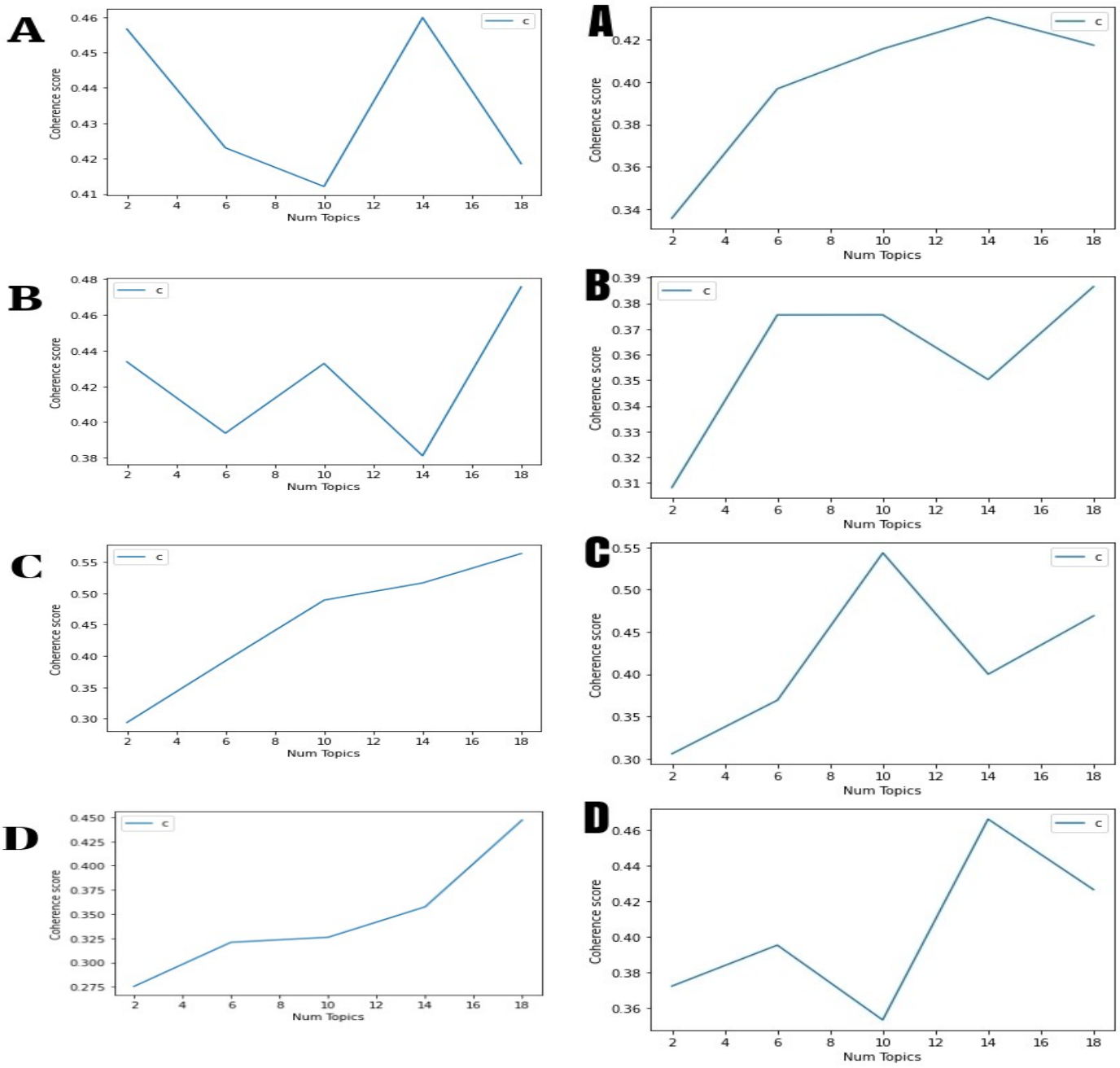
$D_k(x) = x * (\mu_k/\Sigma^2) – (\mu_k\text{^}2/(2*\Sigma^2)) + \ln(P_k)$

$D_k(x)$ is the discriminate function for class k given input x, the μ_k, Σ^2 and P_k are all estimated from the data. This is just an insight as to how the LDA model works.

This study includes news articles for Chloroquine-Hydroxychloroquine from 16/ 03/ 2020 to 01/ 07/ 2020. To understand the change in the trend over time the analysis was made for two different time intervals, the first time interval T1 was from 16/ 03/ 2020 to 15/ 05/ 2020 and the second time interval T2 was from 16/ 05/ 2020 to 01/ 07/ 2020. This study is slightly inclined more on the latest news, when this work was published, so that the results would be more appropriate to the current situation and hence the time interval T2 is slightly smaller so that both the time intervals T1 and T2 have the same number of articles. The news articles were divided into four categories, negative articles in T1, negative articles in T2, positive articles in T1 and positive articles in T2. For the news articles on Lopinavir-Ritonavir the study was made from 16/ 03/ 2020 to 24/ 06/ 2020 and appropriate divisions in time intervals were made similarly.

To scrape the content from the websites in which the required news articles were published a python program was developed. Given the link of a website this program will scrape all the text under the paragraph tags in that website and then converts all the text to a csv file. Before the data was trained using the LDA model the data was preprocessed. LDA model works best when only the nouns and the verbs in the text are fed into it. Efforts were put into remove the non-useful words, called stop words. Though not one hundred percent a lot of it were removed. Lemmatization is a process of converting the words into their root forms, for example, ‘locomotive’ is made ‘locomot’. This is made so that the LDA model doesn’t recognize a word in different forms as different words. Removing punctuations, NaNs, emails, new line characters and single quotes were also the part of data preprocessing. Bigrams and trigrams were identified and added as new words, for example, when the model encountered ‘machine learning’ the model shouldn’t treat them as two different words, we used prebuilt bigrams and trigrams to search for their occurence in our dataset and then after such words were identified they were joined using an underscore, like ‘machine_learning’, and were added to the dataset as new words. After all this all the remaining words in the dataset were split based on blank spaces to get a bag of words. Then all the words were sorted alphabetically, all the words were mapped to the number of times they had appeared and all the words were given a unique id. This was a dictionary that was created, every word would have corresponding tuple in which the first element would be the id of the word and the second element would be its frequency. This dictionary was fed into the LDA model while training it and LDA model proves the hypothesis made. The news articles were divided into four categories, negative articles in T1,negative articles in T2, positive articles in T1 and positive articles in T2. LDA model was trained separately for all the categories to analyse the trends.

Results



(I) **(II)**

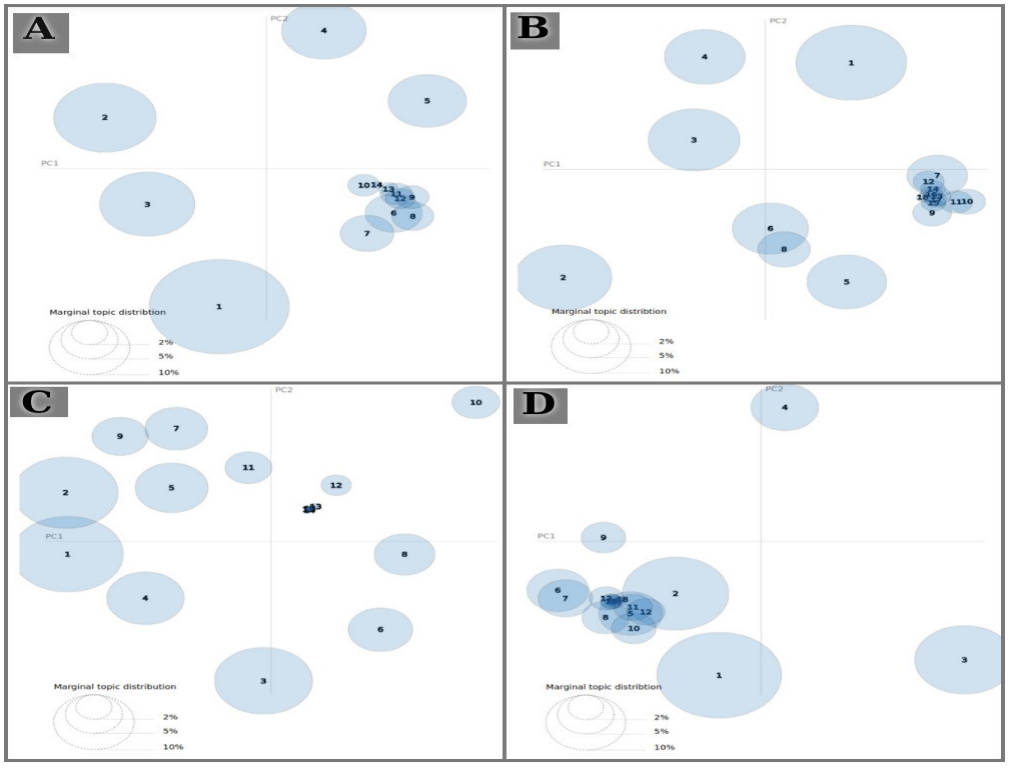
Fig. 1

(I) is for Chloroquine-Hydroxychloroquine and (II) is for Lopinavir-Ritonavir.
In both the cases the convention is as follows:
(A) is the coherence values for negative reports in T1
(B) is the coherence values for negative reports in T2
(C) is the coherence values for positive reports in T1
(D) is the coherence values for positive reports in T2

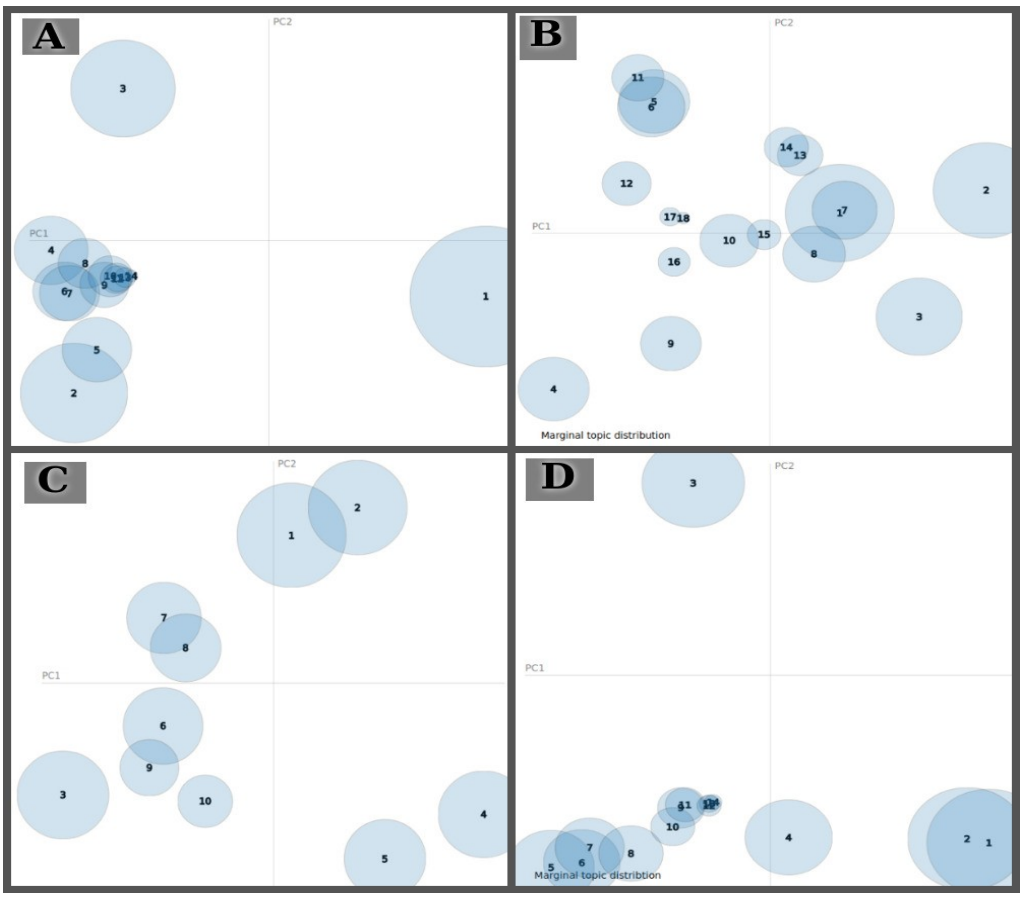
The LDA model was trained several times by changing the number of topics every time as per the suggestions of the coherence value graph, to eventually get a highly coherent intertopic distance map. The results out of a model are said to be coherent if the topics generated by a model support each other, if all the topics are related then the model is 100% coherent but that is not always the case. For instance, the medicines used might have shown different types of side-effects like heart problems in some cases, kidney problems in other and nervous system disorders in some other case. It is important to know all the kinds of problems and hence the goal is not to attain 100% coherence.Here the LDA model is primarily being used to highlight the more frequent problems, hence the model was trained several times by changing the number of topics for the dataset to be classified on and the coherecnce value of the model was measured using the ***C_v*** measure, which is one of the many coherence measures.

The LDA model ranks all the words from all the documents in the dataset. It just doesn't rank them based on the frequency of the appearance of the words, it uses several metrics like the weightage to the word will be reduced if it appears in more number of documents and will be increased everytime the word appears. Then the topics are ranked based on the weightage given to the words that are present in the document. The coherence value of the model is calculated by taking the average of the coherence values of all the topics in the model, this way the model predicts how related all the topics in the model are.

Fig. 1A shows that the highest coherence value for the negative reports in T1 is 0.46 and the optimal number of topics is 14, Fig. 1B shows that the highest coherence value is 0.48 for the negative reports in the interval T2 and the optimal number of topics are 18. Most of the reports spoke about the similar drawbacks but that's not the case with the reports that spoke good about the drugs. The coherence value is increasing as the number of topics are increasing in the Figures 1(I)(C) and 1(I)(D) because there's not much of similarity in the positive reports. A wide range of topics suggests that the topics are not strong enough, a topic is strong only when more reports talk about the same topic. Having a lot of topics would make the graphs look shabby and one wouldn't be able to infer it, hence we have limited the number of topics to an optimal number.



(I)



(II)

Fig. 2

(I) is for Chloroquine-Hydroxychloroquine and (II) is for Lopinavir-Ritonavir. In both the cases the convention for the Intertopic distance maps is as follows:

(a) negatives in T1 and T2 (A and B respectively)

(b) positives in T1 and T2 (C and D respectively)

Both the drugs selected, Chloroquine-Hydroxychloroquine and Lopinavir-Ritonavir, happen to show similar behaviours. The explanation will particularly be based on Chloroquine-Hydroxychloroquine and the inferences of the analysis of Lopinavir-Ritonavir would be obvious. The **figure 2(I)(A)** talks about the side-effects of chloroquine and hydroxychloroquine, one of the frequent medicines being used to treat COVID-19. This figure is talking about the side effects in the time phase T1 (i.e., 16/ 03/ 2020 to 15/ 05/ 2020). The circles are dense in one region and are sparsely spread in other areas indicating that most of the news articles are talking about the same side effects of these two drugs. A few studies have claimed that these two drugs have no use in treating COVID-19 and there are many serious side-effects like heart rythm problems, blood and lymph system disorders, kidney injuries and liver problems. Most of the news articles report similar such side-effects of these drugs and hence the LDA model has created a dense region of topics in the graph. The optimal number of topics for the negative topics in this interval happen to be 14 as is supported by **figure 1(I)(A)**. The size of the circle corresponding to a topic depends on the words in the topic and the weightage of the words, the words are weighed with respect to the entire model. Hence if a topic has words which are heavily weighed then the topic will have a bigger circle representing it.

Figure 2(I)(B) is the intertopic distance map for the side-effects of these drugs in the period T2 (i.e., 16/ 05/ 2020 to 01/ 05/ 2020). As **figure 1(I)(B)** says there are 18 topics in this map. In both

the **figures 1(I)(A)** and **1(I)(B)** there is a dense region in the maps showing that these drugs have significant side-effects because most of the reports have reported similar side-effects. But there is an increase in the number of topics in the second interval T2 despite having a dense region in the map, impying that the reports have reported more number of similar problems i.e., the side-effects of these drugs are increasing as it is continued to be used for longer durations and on more patients.

The **figures 2(I)(C)** and **2(I)(D)** have 18 topics representing the news articles that spoke positively about these two drugs in the time intervals T1 (16/ 03/ 2020 to 15/ 05/ 2020) and T2 (16/ 05/ 2020 to 01/ 07/ 2020) respectively. As seen in the map in **figure 2(I)(C)** the positive topics in T1 are widely spread and there is no significant cluster formed implying that all the reports are talking about different positives, though a few topics are heavily weighed they are unlike each other. If many news articles talk about the same topic then the topic is considered to be significant and a dense region would be formed in the intertopic distance map, but there's no such dense region here and hence there is no significant positive effect reported of these two drugs in the time interval T1. In the time period T2 a dense cluster is formed because many of the news articles in T2 have reported similar positive effects of these two drugs. This shows that the positive effects of these drugs have increased from T1 to T2, it can be because of various reasons like the medical community might have learnt the right dosage of these drugs to be given to the patients. It is great that the topics in **figure 2(I)(D)** are not widespread, it means that most of the news articles talk about the same positive effect, but as told earlier getting 100% coherence is not the only objective because there can be many important positives or negatives of the drugs which could be unlike each other. From **figures 2(I)(A)** and **2(I)(B)** it is seen that from T1 to T2 talking about the negatives of these two drugs, the number of topics have increased, the density of the cluster formed is increased and topics are spread wider. The few topics which are far from the dense region are unlike what most articles say but they are heavily weighed, i.e., have big circles, and they mean that more problems are appearing from these drugs. For instance, most of the articles might be talking about heart problems and these farther topics could be talking about kidney and liver problems because of these drugs [16].

To acknowledge the limitations of this work, this research analysis does not include all the news articles ever published from the onset of COVID-19. Even in the timespan specified, all the possible published news articles haven't been used. The topic was searched in Google search and every article in the search result was carefully examined by one of the team members before it was selected for the dataset, this way articles from many pages of the search results were picked to ensure that the dataset was a true reflection of the trend in the news articles being published. The LDA model forming a dense region implies that it could capture the trend in the news articles, this was possible because the dataset was well picked. Anyone who repeated the same methodology can reproduce the same results.

Conclusion

The results support the hypothesis, the negatives in T2 have worsened in all aspects when compared to T1. The optimal number of topics for the negative reports are increased in T2 when compared to T1, the dense region has become denser and the topics have become more widespread indicating that newer side effects are being formed with continued usage of these drugs. The positive reports in T2 are good but there is increasing side-effects and the positive reports have no consistency in T1 and T2 unlike the negative reports. There are other safer drugs with lesser side-effects and same or even better positive impacts, hence it is very important that the medical community impose strict guidelines on the usage of these two drugs.

Let’s revisit the key points from all the sections. Once COVID-19 started no one were trained to handle the new situations that would arise, especially the medical community. When the number of cases were increasing rapidly and there was no proved medicine to treat the disease a few drugs threw some light to the situation. As a few drugs showed some positive signs in treating the ones affected by the virus there was a widespread adoption of these drugs. These drugs started showing severe side-effects on the patients, but they were rare in the beginning and were increasing slowly. News is one of the best mediums to understand these side-effects and everyone couldn’t know about these side-effects because every other person using the drug was not affected. Hence a text mining technique was used in this research to depict the trends in the negatives these drugs were causing. The news articles about the good and the bad about these drugs were web scrapped for two consecutive time intervals. This data was preprocessed and trained on the Latent Dirichlet Allocation model to show the trends. The transparency in the selection of the optimal number of topics was shown, later the trends were explained using the intertopic distance maps for both the time intervals.

To summarise, the analysis was made on four datasets. The analysis on negative news in T1, negative news in T2, positive news in T1 and positive news in T2. It was seen that most of the reports were talking about the same negative topics in T1 and in T2 though most of the reports were talking about the same topic there were other side-effects emerging, implying that the prolonged usage of these drugs could be more dangerous. The positive reports weren’t streamlined in T1, however the positive effects increased in T2. With time there are safer drugs being discovered and as the negative effects are increasing it is important that strict guidelines are imposed on the usage of Chloroquine-Hydroxychloroquine and Lopinavir-Ritonavir. A few governments have already made strict guidelines on the use of these drugs [13] [14] [15].

The benefits of this solution is that the medical community can understand the trend without having to go through all the rapidly increasing literature. But the setback of this solution is that it doesn’t exactly specify the content of the trending topics. It is difficult to specify the content of the top topics because it’s not easy to make the machine construct meaningful phrases and summarize all the top topics. Reading a few articles one can get to know the trending problems because most of the articles talk about the similar content but this solution shows the emergence of newer side-effects and one wouldn’t get that by reading just a few articles. For those who aren’t even aware that side-effects of this magnitude exists this solution highlights the alarming problem. The benefits of the methodology is that the results are reproducible by any similar analysis of the side-effects of these drugs. However, the setback is that it is not feasible to consider all the news articles that’s been published but a fair selection of the dataset is crucial for the model to give authentic results.

Future Work

This research considered only the text content of all the news, developing methods to analyse the images and videos as well would give better insights. This solution proves that the negatives of these drugs are very serious and that they are increasing rapidly. In this solution the machine couldn’t construct meaningful phrases to specify the content of the top topics, this could be a further area of research. News articles published worldwide have been taken into account here, further research could be done analysing the trends in only specific regions. That would provide better insights for the policy makers to take actions specific to their region. Similar analysis can be made on social media platforms like Twitter [9], Facebook, Instagram, etc. Analysing the tweets, posts on facebook and the communities that are formed in these social media platforms.

References

1) Awadhesh Kumar Singh, Akriti Singh, Altamash Shaikh, Ritu Singh and Anoop Misra. Chloroquine and hydroxychloroquine in the treatment of COVID-19 with or without diabetes: A systematic search and a narrative review with a special reference to India and other developing countries, Elsevier Public Health Emergency Collection.

2) Rismanbaf A, Zarei S.Liver and Kidney Injuries in COVID-19 and Their Effects on Drug Therapy; a Letter to Editor. ArchAcad Emerg Med. 2020; 8(1): e17.

3) Good MI, Shader RI. Lethality and behavioral side effects of chloroquine. Journal of Clinical Psychopharmacology. 1982 Feb;2(1):40-47. DOI: 10.1097/00004714-198202000-00005.

4) Gordon DA, Klinkhoff AV. Second line agents.In: Harris ED, Budd RC, Firestein GS, GenoveseMC, Sergent JS, Ruddy S, et al, editors. Kelley'stextbook of rheumatology. 7th ed. Philadelphia:Elsevier; 2005: 877-99.

5) Krallinger M., Leitner F., Valencia A. (2010) Analysis of Biological Processes and Diseases Using Text Mining Approaches. In: Matthiesen R. (eds) Bioinformatics Methods in Clinical Research. Methods in Molecular Biology (Methods and Protocols), vol 593. Humana Press.

6) Richard F. Sear, Nicolás Velásquez, Rhys Leahy, Nicholas Johnson Restrepo, Sara El Oud, Nicholas Gabriel, Yonatan Lupu, Neil F. Johnson. Quantifying COVID-19 Content in the Online Health Opinion War Using Machine Learning, IEEE Access, IEEE, 2020, Volume 8.

7) Jo EllenStrykerPh.D. Reporting Medical Information: Effects of Press Releases and Newsworthiness on Medical Journal Articles' Visibility in the News Media, Elsevier, Preventive Medicine, Volume 35, Issue 5, November 2002, Pages 519-530.

8) 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.

9) Kouzy R, Abi Jaoude J, Kraitem A, El Alam MB, Karam B, Adib E, Zarka J, Traboulsi C, Akl EW, Baddour K. Coronavirus Goes Viral: Quantifying the COVID-19 Misinformation Epidemic on Twitter. Cureus. 2020 Mar 13;12(3):e7255. doi: 10.7759/cureus.7255. PMID: 32292669; PMCID: PMC7152572.

10) Chenghua Lin and Yulan He. 2009. Joint sentiment/topic model for sentiment analysis. In *Proceedings of the 18th ACM conference on Information and knowledge management* (*CIKM '09*). Association for Computing Machinery, New York, NY, USA, 375–384. DOI:https://doi.org/10.1145/1645953.1646003

11) Elissa Driggin, Mahesh V. Madhavan, Behnood Bikdeli, Taylor Chuich, Justin Laracy, Giuseppe Biondi-Zoccai, Tyler S. Brown, Caroline Der Nigoghossian, David A. Zidar, Jennifer Haythe, Daniel Brodie, Joshua A. Beckman, Ajay J. Kirtane, Gregg W. Stone, Harlan M. Krumholz, Sahil A. Parikh. Cardiovascular Considerations for Patients, Health Care Workers, and Health Systems During the COVID-19 Pandemic. J Am Coll Cardiol. 2020 May, 75 (18) 2352-2371.

12) in Cao, M.D., Yeming Wang, M.D., Danning Wen, M.D., Wen Liu, M.S., Jingli Wang, M.D., Guohui Fan, M.S., Lianguo Ruan, M.D., Bin Song, M.D., Yanping Cai, M.D., Ming Wei, M.D., Xingwang Li, M.D., Jiaan Xia, M.D., et al. "A Trial of Lopinavir–Ritonavir in Adults Hospitalized with Severe Covid-19", The NEW ENGLAND JOURNAL of MEDICINE. March 18, 2020.

13) Calleri, G. Malaria prophylaxis and guidelines. *Infection* **42,** 913–916 (2014). https://doi.org/10.1007/s15010-014-0658-5.

14) Panda, B K; Diwan, Arundhati; Marne, Sourabh R; Singh, Priyanka. Drug Utilization Study of Anti-malarial in a Tertiary Care Teaching Hospital. Journal of Current Pharma Research; Satara Vol. 2, Iss. 3, (Apr-Jun 2012): 527-532.

15) Howard M. Cann, Henry L. Verhulst. FATAL ACUTE CHLOROQUINE POISONING IN CHILDREN. Pediatrics Jan 1961, 27 (1) 95-102.

16) Gu, J., Han, B. and Wang, J., 2020. COVID-19: gastrointestinal manifestations and potential fecal–oral transmission. *Gastroenterology*, *158*(6), pp.1518-1519.

17) Kome Gbinigie, Kerstin Frie. Should chloroquine and hydroxychloroquine be used to treat COVID-19? A rapid review. BJGP Open 2020; 4 (2): bjgpopen20X101069. DOI: 10.3399/bjgpopen20X101069.

Authors

1) Bhavin G Chennur (Corresponding author)
Department of Computer Science, PES University Electronic City Campus, Bengaluru, India.

Academic email address: bhavingchennur@pesu.pes.edu

Postal address: #771, 17th B Cross, 6th phase, J P Nagar, Bengaluru – 560078, Karnataka, India.

2) Nishanth S Shastry
Department of Computer Science, PES University Electronic City Campus, Bengaluru, India.