Towards Containing COVID-19 Pandemic by Mining Knowledge from Scientific Literature and Social Media

Conference Paper · September 2021

CITATIONS
0

READS
304

6 authors, including:

Syeda Jannatus Saba
Shahjalal University of Science and Technology
8 PUBLICATIONS **23** CITATIONS
SEE PROFILE

Biddut Sarker Bijoy
Stony Brook University
5 PUBLICATIONS **9** CITATIONS
SEE PROFILE

Souvika Sarkar
Auburn University
13 PUBLICATIONS **18** CITATIONS
SEE PROFILE

Md Saiful Islam
Shahjalal University of Science and Technology
114 PUBLICATIONS **1,302** CITATIONS
SEE PROFILE

Towards Containing COVID-19 Pandemic by Mining Knowledge from Scientific Literature and Social Media

Syeda Jannatus Saba*
Shahjalal Univ. of Science & Tech.
Sylhet, Bangladesh
syeda06@student.sust.edu

Biddut Sarker Bijoy*
Shahjalal Univ. of Science & Tech.
Sylhet, Bangladesh
biddut12@student.sust.edu

Souvika Sarkar
Auburn University
Alabama, US
szs0239@auburn.edu

Md Saiful Islam
Shahjalal Univ. of Science & Tech.
Sylhet, Bangladesh
saiful-cse@sust.edu

Sheikh Rabiul Islam
University of Hartford
Connecticut, US
shislam@hartford.edu

Md. Ruhul Amin
Fordham University
New York, US
mamin17@fordham.edu

***Abstract*—The ongoing pandemic, COVID-19, has been sweeping the world, affecting millions of people from 221 countries and territories, claiming 3.01 million deaths around the world in an unprecedented way. There is a pressing demand to elicit the unknown facts to contain and eradicate the virus globally. In this work, we apply a semi-supervised text-mining technique on the social media data and research paper archive (i.e., CORD-19 dataset), to uncover new information about the COVID-19 virus transmission. The discovered novel information, demonstrated by an enhanced *Chain of Infection*—an epidemiological term describing the infection process, provides valuable insights about the spread of the virus, which in turn can aid in developing improved policies towards containing the virus.**

***Index Terms*—Chain of Infection, Big Data Analytics, Pandemic, COVID-19**

I. Introduction

One can consider the COVID-19 pandemic as the most crucial global health calamity of the century and the greatest challenge that the humankind faced since the 2nd World War. The pandemic has engulfed all the nations of the world, altered the pace, fabric , and nature of our lives, and cost hundreds of thousands of human lives. The ravages of this global pandemic has resulted into international efforts to understand, track, and mitigate the disease.

To tackle this epidemic, it is necessary to keep the individuals on the front lines of the fight —healthcare practitioners, policymakers, medical researchers, etc., up-to-date with the essential knowledge about the disease and infection process. Therefore, COVID-19, SARS-CoV-2, other Coronaviruses, and related topics have yielded a burgeoning corpus of scientific publications. A handful of recent research work have showcased different search engines for retrieving valuable information from scientific literature related to the disease. Various methodologies such as keyword search, natural language queries, semantic relevancy, and knowledge graphs have been employed in the *Sketch Engine COVID-19* [3], *COVIDSEE* [4], and *Amazon's CORD19 Search* engines [5]. Besides, the social network dataset could be another source of novel information. However, due to the very large volume of datasets and the variety of search results, specific knowledge discovery is arduous, although very crucial and viable. In this article, we present a systematic approach to identify an improved *Chain of Infection*, specifically focusing on the COVID-19 virus.

From the epidemiologic point of view, infectious diseases result from the interaction of the agent, host, and environment. The transmission of the infectious diseases occurs when the infectious agent (e.g., COVID-19 virus) leaves its *reservoir* or *host* (e.g., human) through a *portal of exit* (e.g., nose) using a *means of transmission* (e.g., droplet), and infects a *new/susceptible host* (e.g., human) by entering through an appropriate *portal of entry* (e.g., mouth). These sequences of disease transmission is called the *Chain of Infection* by the epidemiologists. Figure 1 represents these disease transmission stages for different infectious agents depicting individual components of the chain and their interconnections, and Table I represents a partial list of known examples for each component of the chain [1, 2].

We apply a semi-supervised *Human-in-the-loop* text-mining approach, on a collection of scientific literature (CORD-19) and social media (Twitter) data. In other words, our main contribution in this work is the application of a semi-supervised information retrieval technique on two different data sources, a recently collected Twitter dataset and a recently released research paper archive (CORD-19), to uncover new information about the COVID-19 virus regarding its infection and transmission process. An enhanced Chain of Infection is vital for informed decision making by healthcare practitioners, policymakers, and medical researchers, to contain a virus like COVID-19.

We start with a background of related work. Then we

*Both authors contributed equally to this research.

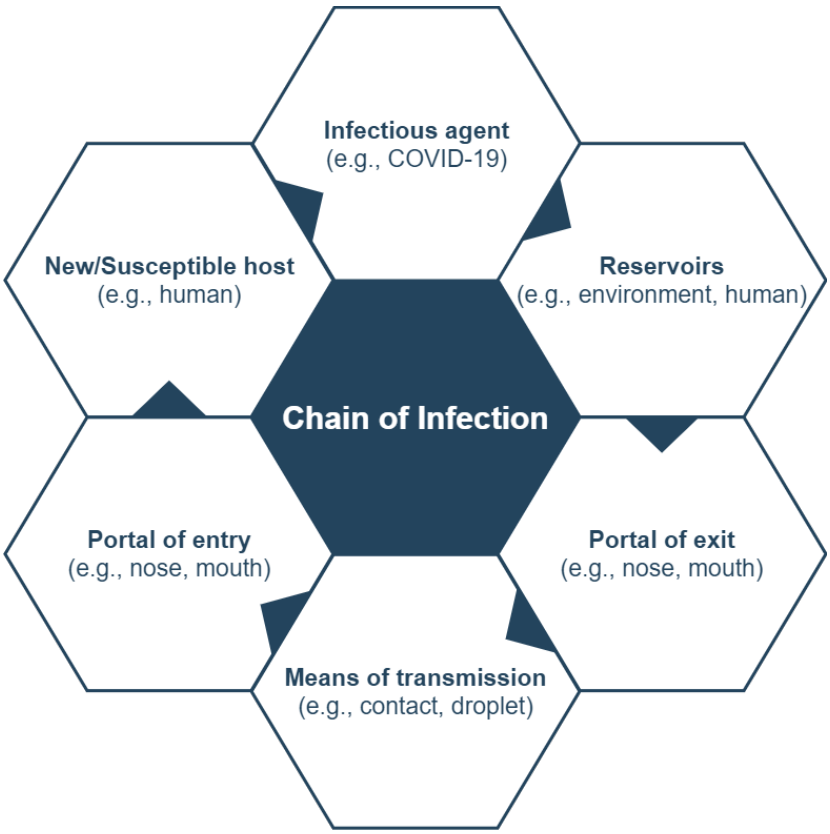

Fig. 1. Chain of infection [1, 2]

describe our experiments, followed by the results section which contains a discussion on results from the experiments. We conclude with some limitations and future work.

II. Background

With the isolation of people from physical public spaces in response to the COVID-19 pandemic, online platforms have become even more prominent tools to analyze public opinion and concerns using social network discussion. To connect the machine learning community with biomedical domain experts and policymakers, research collaborations are taking place around the world to identify effective treatments and management policies for Covid-19. Various corpus/datasets [6, 7, 8] have been published including some questions answering dataset: [9]. The CORD-19 coronavirus-related corpus [6], primarily from PubMed, published in 2020.

Search engines such as [10, 11, 12, 13] were launched, to provide users with high-quality information from the scientific literature, to inform evidence-based decision making and to support insight generation. Question Answering sytems [14, 15, 16, 17] were developed to shrink the gap of knowledge, and spread information widely.

Individuals, organizations, and governments are using social media for awareness and information to communicate with each other on a number of issues relating to the pandemic. Hence a school of researchers performed a sentiment analysis of Twitter posts [18, 19] to identify public emotions. Another line of work analyzed conspiracy theories propagated over Twitter [20, 21].

Since the spread of the Coronavirus disease, uninhibited misinformation is also spreading over traditional and social media at a rapid pace. A few attempts have been made to identify fake/incorrect information [22, 23, 24].Analysis on various topics such as government support, reports on newly infected cases and deaths, and panic buying have been performed by a few group of researchers [25, 26].

Textrank: We used *TextRank* algorithm [27], with optimizations features in the similarity function [28], to filter out vocabularies for Twitter data. The *TextRank* algorithm is a graph-based ranking algorithm for text processing. Here, the primary data model used for it is a graph. The original document is split into words, and these words represent nodes in the graph. The edges inside the graph are assigned weights based on the similarity scores between the words. An edge is set up between two different words if they occur together in a text. The similarity score is higher if these two words co-occur more frequently in the input text. The Pagerank algorithm [29] was applied to the graph in order to derive each word's importance. Words with higher importance are considered as top words or keywords in the document.

Semi-supervised Topic Modeling: Topic modeling is a

TABLE I
PARTIAL LIST OF EXAMPLES UNDER EACH ELEMENT OF THE *Chain of infection* [1, 2]

Element	Examples
Infectious agent	microorganism
Reservoirs	human, plant, animal, environment, food, water, feces,
Portal of entry/ exit	mouth, nose, anus, urinary diversion, blood, tissue, mucosa,
Means of transmission	biting, touching, kissing, sexual intercourse, sneezing, coughing, spitting, singing, talking, eye, conjunctiva, toys, soiled clothes, eating utensils, handkerchiefs, surgical instruments or dressings, stethoscopes, droplets, dust,
New/ Susceptible host	human, young or older population with weak immune system

statistical modeling technique that discovers the abstract "topics" that appear in a collection of documents; it clusters a collection of words in a way such that each soft-cluster represents a topic in a document. Latent Dirichlet Allocation (LDA) is a probabilistic, generative topic modeling technique that transforms bag-of-words counts into a topic space of lower dimensionality. It builds a topic-per-document model and a word-per-topic model, modeled as Dirichlet distributions. However, these models are unsupervised in general, making the content of topics challenging to control. Intrigued by this possibility, a new class of topic modeling, the semi-supervised topic model was introduced that allows researchers to draw on their domain knowledge to guide the model in the right direction. Here, researchers can provide the model with an initial set of "anchor/ seed words". These groups of seed words represent potential and expected topics the model should attempt to find. The semi-supervised topic model [30], that we use in our experiment, implements latent Dirichlet allocation (LDA) using collapsed Gibbs sampling [31]. For each Document **(D)** and for each word **(W)** in that document, the probability of that word belonging to each topic **(Z)** is calculated using the following Formula 1.

$$\mathbf{P(Z|W,D)} = \frac{\mathbf{zn_w} * \mathbf{zc_d}}{\mathbf{tc_z}} \tag{1}$$

Here, zn_w is the count of W in topic Z , zc_d is the count of words in D that belongs to Z, tc_z is the total count of words in Z.

Despite many works on the topic modeling, the background study reveals that our knowledge about COVID-19 *Chain of Infection* has a lot of rooms for improvement. So in this paper, we attempt to enhance the *Chain of Infection* from a recently crawled social media dataset and research article archive related to COVID-19.

III. METHODS AND EXPERIMENTS

Figure 2 depicts a high-level flow of activities under our methods and associated experiments. The process starts with data cleaning and preprocessing, and then gradually enhance the initial seed words, towards an enhanced Chain of Infection, with a Guided LDA coupled with active human supervision.

Generally, *Chain of Infection* describes the process of infection, which implies how a pathogen moves into the aforementioned chain (see Figure 1) to spread disease within a community. In accordance with Figure 1, we used six topics: *infectious agent, reservoirs, portal of exit, means of transmission, portal of entry, and new / susceptible host* in our proposed model, for extending our knowledge about this chain. Given the COVID-19 pandemic, there is a pressing demand to improve over the current knowledge of virus infection. Indeed, for a better perception of the COVID-19 pandemic and to break the propagation cycle (i.e., chain), we need a comprehensive and up-to-date *Chain of Infection*.

A. Datasets and Pre-processing

We use two standard and publicly available datasets, one from the scientific literature, and another from the social media. We applied the topic modeling on both datasets, and compared the results with each other.

1) Scientific Literature Dataset:: The COVID-19 Open Research Dataset (CORD-19) [32] consists of scientific papers, news articles on COVID-19, and related historical research regarding coronavirus and its variants, e.g., sers, mers, etc. It was prepared by the White House and a coalition of leading research groups in response to the COVID-19 pandemic.

This dataset comprises of nearly 128,025 scholarly articles, including over 70,372 papers that follow standard research paper format (i.e., Paper ID, Abstract, Full Text, etc.). Initially, we collect 128025 JSON files located in the CORD-19 Dataset folder. Then we filter out the JSON files that do not have a proper format for *Paper id, Abstract* and *Body*. Later, we eliminate the duplicate papers from the corpus. After filtering, we find 70,372 research papers with full text. Next, we separate all papers based on language. And we find these distributions based on languages: 'en': 67944, 'de': 939, 'es': 553, 'fr': 445, 'nl': 325, 'it': 53, etc. Finally, the total number of papers that we use in our experiments is 67,944, and they are written in English language. Then we preprocess the corpus using spaCy [33]. We use the en_core_sci_lg mode, a full spaCy pipeline for biomedical data with a larger vocabulary and 600k word vectors, for this task. Further, we turn each word into lower case letters and filtered them by punctuation.

2) Twitter Dataset:: The Twitter dataset [34] comprises of more than 515 million tweets metadata. This dataset includes CSV files that contain IDs and sentiment scores of the tweets related to the COVID-19 pandemic. The tweet IDs have been collected by monitoring the real-time Twitter feed for coronavirus-related tweets using 90+ different keywords and

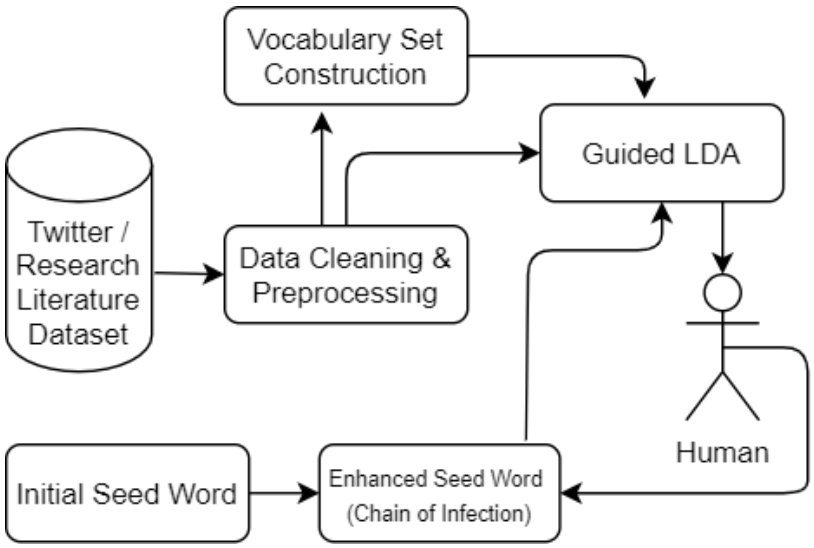

Fig. 2. High level flowchart for mining novel factors related to *Chain of Infection* for COVID-19 virus.

hashtags. This dataset only contains tweets IDs, because of twitter's content redistribution policy[1]. We have collected the data from March 19, 2020, to September 15, 2020. We have fetched these tweets using Twarc [35] and Hydrator [35]. Using these tools, we have collected more than 462 million tweets which is a significant amount (3TB in size) for quality data mining tasks. Due to some ongoing adjudications, we use Twitter Developer API[2] along with Twarc.

After downloading, each tweet object was in the JSON format embedded in a JSONL file. The information inside these tweet objects is arranged in a multi-level hierarchy, making it tedious for researchers to access necessary information. We flatten the data skeleton to a single level hierarchy, filter out unnecessary data from tweet objects, and store them in CSV files. We took a random sample, totalling 7% of the tweets, from the entire dataset, since the collected dataset was too big to process. We focus only on the tweet attribute *text*. The raw tweets' texts are highly unstructured and contain uninformative and misleading words. The downloaded data had to be preprocessed in order to make it suitable for reliable analysis. To process Twitter data, we apply standard data cleaning and preprocessing techniques. We change all text to lowercase. We convert the informal words and phrases into standard forms with the help of abbreviations and slangs collected from the Internet, e.g., *ty* replaced by *thank you*, *lol* replaced by *laughing out loud*, *btw* replaced by *by the way*, etc. We remove all the hashtags' signs and convert the text in each hashtag into a single word. We use the tweets-preprocessor [36] module to clean the tweet texts by removing all types of smiley, emoji, mentions, twitter reserved words (*rt, fav*), and URLs. Stopwords provide little or no information to the text analysis, introduce noise to our analysis, and slow down the process. To filter out the stopwords, we use the stop word list provided by NLTK [37]. We apply spaCy [33] lemmatizer on words, in the processed text, to avoid the inclusion of different variations of a single word as distinct words in the analysis— for example, *cough, coughed, and coughing* are different forms of the single word *cough*. Table II represents a brief description the COVID-19 Tweets Dataset used in our analysis.

[1]https://developer.twitter.com/en/developer-terms/agreement-and-policy
[2]https://developer.twitter.com/en/products/twitter-api

B. Vocabulary Set Construction

As with any text-mining task, we started by creating a reasonable vocabulary set for our *Chain of Infection* mining task. Tweets are generally not written in a standard form of language. There are many variations of the same words, for instance, shortened or elongated stop words, e.g., *yahoooo*, *thnx*, etc., that provide little or no importance to the text analysis and make the vocabulary list bulky, slow down the experiment, and introduce noise in the results. To get around this difficulty, we used the Summa-TextRank [27, 28] algorithm to extract top keywords from tweet texts. At first, we made chunks of 300 tweets, and extracted top keywords from that chunk. Tweets are processed in chunks to make the keyword extraction and the associated graph building process more meaningful, as the text of a single tweet is sometimes too short to yield a meaningful result. If a chunk produces more than 300 top keywords, we only consider the first 300 words as keywords and discard the rest. The resultant keyword list is then appended to the final vocabulary.

In the research paper dataset, we use Summa-TextRank to filter out unnecessary words and extract top keywords from each research paper. Here, the ratio of produced top keywords with respect to all the keywords is set to 0.2. To extract keywords from lengthy research papers, we considered the first 10,000 words of these papers due to the huge computational complexity of the generated graph with too many nodes. Only

TABLE II
SUMMARY STATISTICS OF TWITTER DATASET

Attribute	Summary
Starting timestamp	Mar 19,2020 19:52:18
Ending timestamp	Sep 15, 2020 04:24:54
Total Number of tweets	462,702,229
Number of tweets by verified users	9,851,045
Language	English (EN)
Coverage	Global

keywords with length greater than two are added to the final vocabulary list.

C. Topic Modeling Experiments

Initially, we applied the basic Latent Dirichlet Allocation (LDA) [38, 39] method, a popular topic model, on our text data to uncover new factors related to the six stages of the *Chain of Infection* of the COVID-19 virus. Here, we treated each stage of the infection as individual topics, where each topic is characterized by a distribution over the entire vocabulary of words. However, as traditional topic models are fully unsupervised, the extracted topics were not aligned with the infection stages, as expected.

To overcome this challenge, we used a variation of topic modeling—semi-supervised topic modeling that allowed us to insert prior knowledge about the infection stages into the topic model. Here, we employed COVID-19 related seed word sets (see Table I), collected from [1, 40], that guided / constrained the model to group seed words of each stage into the same topic; thus making the topics more specific and semantically related to the seeds which can ultimately help break this chain. This way, semi-supervised topic modeling method gives us the flexibility to control the alignment of topics extracted from the topic model and thus, can discover fine-grained keywords related to the COVID-19 infection more effectively. Noteworthy, we used the same seed word list (see Table I) for both the research paper and twitter dataset.

To implement the semi-supervised learning objective mentioned above, we first constructed the required vocabulary-set as described in *Vocabulary Set Construction* subsection. The next step of our experiment was similar for both the research paper and the Twitter dataset, where we considered n-grams with n = 1, 2 as the extended vocabulary. Essentially, we assumed that bigrams are powerful enough to represent any factor of the infection stages. Next, we use GuidedLDA [30] model to conduct semi-supervised topic modeling. In GuidedLDA model, number of topics n is set to 6, number of sampling iterations n_iter is set to 100, and refresh is set to 10. Then, we tune seed_confidence as 0.85 which denotes the amount of extra boost that should be given to a term in the seed list (between 0 to 1) and fit our model with the document-term matrix produced by the countvectorizer and the seed word list.

After completing one iteration, the model gives a list of words regarding topics in descending order of weight. We revised this list of topic related words by actively involving humans in the loop. To be more specific, we revised the semi-supervised topic models over multiple iterations in an interactive fashion with active engagement from humans. In each iteration, we appended coherent and topic-related words from each topic's resultant word distribution to the corresponding seed word list of that topic. We achieved this through manual inspection of top-m (a user-defined threshold) keywords for each topic and then, active selection by a human volunteer.

As mentioned above, we presented the top m words with the highest probability scores to the humans for evaluation and active feedback. This process helped us leverage the power of the machine and human intelligence together, for creating a continuous feedback loop allowing the algorithm to give every time better results. At the end, we collect words selected through human feedback and add them to the seed word list of the corresponding topic. After each iteration, GuidedLDA learns new things with the help of *human-in-the-loop* (HITL) style active supervision. In this type of simulation, a human is always part of the simulation and consequently influences the outcome in such a way that incrementally improve results in the subsequent iterations. At each iteration, we performed HITL in GuidedLDA with the latest expanded set of seed words. This process is visually depicted in Figure 2, where, the process gradually expands the initial seed words with active human supervision to yield new knowledge about *Chain of Infection*.

After repeating this process for multiple iterations, the list of seed words expanded into a more interpretable and coherent list. Simultaneously, each topic got more aligned with the corresponding stages of the *Chain of Infection*. This method allows us to analyze the resultant top-ranked words for each topic, and to explore unknown information about the virus transmission. The loop continues until humans can no longer find any new meaningful word.

IV. RESULTS

To increase the accuracy of topic predictions, we introduce continuous learning of topics through a semi-supervised topic model based on the Human-in-the-loop (HITL) approach. We present the explored novel information for each component from the *Chain-of-Infection* as follows.

1) Infectious Agents:: In the first phase of *Chain of Infection*, infectious agent, no new information is discovered from both Twitter and the *research paper analysis*, although we used several terms, e.g., microorganism, covid-19, etc mentioned in [2, 1] as seed-words. But some structural information regarding the coronavirus, such as *capsid, spike protein*, and *inhibitor* are revealed by the *research paper analysis*.

2) Reservoirs:: Our model, on the research paper dataset, successfully captures many of the non-living reservoirs of COVID-19, such as *water, surface, plastic, steel, and device*. While the results produced by the Twitter dataset contain some racial words, e.g., *black people, americans, chinese, group muslim, and muslims graveyard* that point out the racially biased opinion of Twitter users. Twitter users tend to presume

TABLE III
NEWLY DISCOVERED INFORMATION FOR THE *Chain of Infection*. INFORMATION FROM THE TWITTER ANALYSIS RESULTS ARE IN THE *italic* FORMAT, AND INFORMATION FROM THE RESEARCH PAPER ANALYSIS ARE IN THE NON-ITALIC FORMAT.

Element	Examples
Infectious agent	capsid, spike protein, inhibitor
Reservoirs	water, surface, plastic, steel, device, *black people, americans, chinese, group muslim, muslimas graveyard, rally, campaign*
Portal of entry/ exit	infection, inflammatory, inflammation, injury, open wound, needle puncture, *mask, eye nose, eyes , face*
Means of transmission	travel, market, group, exposure, work, ppe, *travel, party, vote, protest, restaurant, beach, bar, school, class, teacher, exam*
New/ Susceptible host	old, male, comorbid, cancer, asthma, illness, diabetes, hypertension, *doctor, police, essential worker, nurse, worker, student, poor*

the people of some certain races or religions as a spreader or source of COVID-19 virus. The result also contains words, such as *rally, and campaign* that imply the involvement of events like *American Presidential Election* and *Black Lives Matter Movement* in spreading COVID-19 virus.

3) Portal of Entry/Exit:: In our model on the Twitter data, *mask* appears as a high-weight word under the category *Portal of Entry/Exit*. In other words, *mask* could pose a high threat and act as a portal of entry/exit when it is not produced, cleaned, and worn properly. Here is a representative tweet:

> *Wearing a mask improperly can actually increase your risk of getting disease. However wearing a proper mask will decrease the spread of COVID and help prevent you from getting COVID.*

The result also contains some other words, such as *eye nose, eyes, and face*, which is known as portal of entry/exit by researchers. In the *research paper analysis* we found words like, *infection, inflammatory, inflammation, injury, open wound, needle puncture*, etc. These words imply *open wound*, and *infection* might act as a *Portal of entry/exit*. However, no source has proven or confirmed the validity of this information yet, and needs further investigation. Furthermore, the *research paper analysis* was unable to capture enough new information about the portal of entry/exit; maybe that is because of the presence of old research articles in the dataset.

4) Means of transmission:: Means of transmission topic in *research paper analysis*, contains words like *travel, market, group, exposure, work* etc., while the words found in the *Twitter analysis* are *travel, party, vote, protest, restaurant, beach, bar* etc. Here is a representative tweet:

> *Health officials say a person who spent several hours at a bar during the Sturgis Motorcycle Rally has tested positive for COVID-19 and may have spread it to others.*

According to both analyses, these words indicate that *traveling, outdoor social activities, and meetings* play a significant role in accelerating the transmission among people. The *Twitter analysis* also brought some words related to education, e.g., *school, class, teacher, and exam*. So we assume twitter users think that reopening educational institutions, and taking the classes and exams might speed up the transmission among students and teachers. The *Research paper analysis* contains words like *ppe* as a mean of transmission which is surprising yet convincing. Though we use the protective instruments to prevent transmission of COVID-19, not following the hygiene guidelines and inappropriate use of these instruments might cause the same things to work the other way around.

5) Susceptible Host:: Words like *old, male, comorbid, cancer, asthma, illness, asthma, diabetes, hypertension,* etc were discovered under the susceptible host topic in the *research paper analysis*. Therefore, *older adults, males, and people with comorbidities* are more likely to be hosts. The result was also able to point out the probable comorbidities of susceptible hosts. Furthermore, the *Twitter analysis* results contain words that indicate the class, race, or profession of a person, e.g., *doctor, police, essential worker, nurse, worker, student, poor* etc. Therefore, the people from the professions who worked in the front-line during this pandemic (doctor, nurse, police, essential workers), and people with financial distress (i.e., poor people) are more likely to be COVID-19 hosts, according to Twitter users' opinion.

A list of newly discovered information is presented in the Table III. It is vital to note that we evaluate the results based on how they are related to COVID-19, which is our primary target.

V. CONCLUSION AND FUTURE WORK

In this paper, we apply a semi-supervised text mining technique on two datasets, a recently crawled Twitter dataset on COVID-19 and a research literature archive related to COVID-19, for further extending the *Chain of Infection* with refined information from the collection of explored novel information. We are able to explore some novel information about the COVID-19 virus which was previously unknown to the research community and practitioner. We expect that the enhanced *Chain of Infection* will be helpful to the healthcare practitioners, policymakers, and medical researchers for more informed decision making. Going forward, we want to validate the enhanced *Chain of Infection* with an epidemiologist and incorporate a domain expert in the human-in-the-loop approach

to validate the findings with the experts from the medical domain.

References

[1] *Introduction to Epidemiology by Centers for Disease Control and Prevention*. 2020 (Accessed November 16, 2020). URL: https://www.cdc.gov/csels/dsepd/ss1978/lesson1/section10.html.

[2] Chris E. Patterson. *Basic principles of infection control*. 2020 (Accessed Nov 16, 2020). URL: https://cdn.journals.lww.com/nursingmadeincrediblyeasy/Fulltext/2015/05000/Basic_principles_of_infection_control.7.aspx.

[3] Andy Way et al. "Rapid development of competitive translation engines for access to multilingual covid-19 information". In: *Informatics*. Vol. 7. 2. Multidisciplinary Digital Publishing Institute. 2020, p. 19.

[4] Karin Verspoor et al. "COVID-SEE: Scientific Evidence Explorer for COVID-19 related research". In: *arXiv preprint arXiv:2008.07880* (2020).

[5] Taha A Kass-Hout and Ben Snively. *AWS launches machine learning enabled search capabilities for COVID-19 dataset*. https://aws.amazon.com/blogs/publicsector/aws - launches - machine - learning - enabled - search - capabilities-covid-19-dataset/. 2020.

[6] Lucy Lu Wang et al. "CORD-19: The Covid-19 Open Research Dataset". In: *ArXiv* (2020).

[7] Emily Chen, Kristina Lerman, and Emilio Ferrara. "Covid-19: The first public coronavirus twitter dataset". In: *arXiv preprint arXiv:2003.07372* (2020).

[8] Emanuele Pepe et al. "COVID-19 outbreak response, a dataset to assess mobility changes in Italy following national lockdown". In: *Scientific data* 7.1 (2020), pp. 1–7.

[9] Raphael Tang et al. "Rapidly Bootstrapping a Question Answering Dataset for COVID-19". In: *arXiv preprint arXiv:2004.11339* (2020).

[10] Edwin Zhang et al. "Rapidly deploying a neural search engine for the covid-19 open research dataset: Preliminary thoughts and lessons learned". In: *arXiv preprint arXiv:2004.05125* (2020).

[11] Andre Esteva et al. "Co-search: Covid-19 information retrieval with semantic search, question answering, and abstractive summarization". In: *arXiv preprint arXiv:2006.09595* (2020).

[12] Debasmita Das et al. "Information retrieval and extraction on covid-19 clinical articles using graph community detection and bio-bert embeddings". In: *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*. 2020.

[13] Kirk Roberts et al. "TREC-COVID: Rationale and Structure of an Information Retrieval Shared Task for COVID-19". In: *Journal of the American Medical Informatics Association* (2020).

[14] Jinhyuk Lee et al. "Answering questions on covid-19 in real-time". In: *arXiv preprint arXiv:2006.15830* (2020).

[15] David Oniani and Yanshan Wang. "A Qualitative Evaluation of Language Models on Automatic Question-Answering for COVID-19". In: *arXiv preprint arXiv:2006.10964* (2020).

[16] Carmen Riggioni et al. "A compendium answering 150 questions on COVID-19 and SARS-CoV-2". In: *Allergy: European Journal of Allergy and Clinical Immunology* (2020).

[17] Dan Su et al. "CAiRE-COVID: A Question Answering and Multi-Document Summarization System for COVID-19 Research". In: *arXiv preprint arXiv:2005.03975* (2020).

[18] Gopalkrishna Barkur and Giridhar B Kamath Vibha. "Sentiment analysis of nationwide lockdown due to COVID 19 outbreak: Evidence from India". In: *Asian journal of psychiatry* (2020).

[19] FA Binti Hamzah et al. "CoronaTracker: worldwide COVID-19 outbreak data analysis and prediction". In: *Bull World Health Organ* 1 (2020), p. 32.

[20] Wasim Ahmed et al. "COVID-19 and the 5G conspiracy theory: social network analysis of Twitter data". In: *Journal of Medical Internet Research* 22.5 (2020), e19458.

[21] Henna Budhwani and Ruoyan Sun. "Creating COVID-19 Stigma by Referencing the Novel Coronavirus as the "Chinese virus" on Twitter: Quantitative Analysis of Social Media Data". In: *Journal of Medical Internet Research* 22.5 (2020), e19301.

[22] Ramez Kouzy et al. "Coronavirus goes viral: quantifying the COVID-19 misinformation epidemic on Twitter". In: *Cureus* 12.3 (2020).

[23] Karishma Sharma et al. "Covid-19 on social media: Analyzing misinformation in twitter conversations". In: *arXiv preprint arXiv:2003.12309* (2020).

[24] J Scott Brennen et al. "Types, sources, and claims of Covid-19 misinformation". In: *Reuters Institute* 7 (2020), pp. 3–1.

[25] Thirunavukarasu Balasubramaniam, Richi Nayak, and Md Abul Bashar. "Understanding the Spatio-temporal Topic Dynamics of Covid-19 using Nonnegative Tensor Factorization: A Case Study". In: *arXiv preprint arXiv:2009.09253* (2020).

[26] Yuyu Luo et al. "DeepTrack: Monitoring and exploring spatio-temporal data: a case of tracking COVID-19". In: *Proceedings of the VLDB Endowment* 13.12 (2020), pp. 2841–2844.

[27] Rada Mihalcea and Paul Tarau. "Textrank: Bringing order into text". In: *Proceedings of the 2004 conference on empirical methods in natural language processing*. 2004, pp. 404–411.

[28] Federico Barrios et al. "Variations of the Similarity Function of TextRank for Automated Summarization". In: *CoRR* abs/1602.03606 (2016). arXiv: 1602.03606. URL: http://arxiv.org/abs/1602.03606.

[29] Sergey Brin and Lawrence Page. "The Anatomy of a Large-Scale Hypertextual Web Search Engine". In:

COMPUTER NETWORKS AND ISDN SYSTEMS. 1998, pp. 107–117.

[30] Jagadeesh Jagarlamudi, Hal Daumé III, and Raghavendra Udupa. "Incorporating Lexical Priors into Topic Models". In: *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*. Avignon, France: Association for Computational Linguistics, Apr. 2012, pp. 204–213. URL: https://www.aclweb.org/anthology/E12-1021.

[31] Thomas L Griffiths and Mark Steyvers. "Finding scientific topics". In: *Proceedings of the National academy of Sciences* 101.suppl 1 (2004), pp. 5228–5235.

[32] Lucy Lu Wang et al. "CORD-19: The Covid-19 Open Research Dataset". In: *ArXiv* (2020).

[33] Matthew Honnibal and Ines Montani. "spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing". In: *To appear* (2017).

[34] Rabindra Lamsal. *Coronavirus (COVID-19) Tweets Dataset*. 2020. DOI: 10.21227/781w-ef42. URL: https://dx.doi.org/10.21227/781w-ef42.

[35] Alex Galarza. *Documenting the Now*. 2018.

[36] Vasisouv, Alextsil, Idimitriadis. *Tweets Preprocessor [Twitter Preprocessor Module]*. Version Latest. (Accessed on 01/10/2020). URL: https://github.com/vasisouv/tweets-preprocessor.

[37] Edward Loper and Steven Bird. "NLTK: the natural language toolkit". In: *arXiv preprint cs/0205028* (2002).

[38] David M Blei, Andrew Y Ng, and Michael I Jordan. "Latent dirichlet allocation". In: *Journal of machine Learning research* 3.Jan (2003), pp. 993–1022.

[39] Jonathan K Pritchard, Matthew Stephens, and Peter Donnelly. "Inference of population structure using multilocus genotype data". In: *Genetics* 155.2 (2000), pp. 945–959.

[40] Chris E Patterson. "Basic principles of infection control". In: *Nursing made Incredibly Easy* 13.3 (2015), pp. 28–37.