Predicting Emerging Themes in Rapidly Expanding COVID-19 Literature with Dynamic Word Embedding Networks and Machine Learning

Ridam Pal[1], Harshita Chopra[4], Raghav Awasthi[1], Harsh Bandhey[1], Aditya Nagori[1,3], Amogh Gulati[1] , Ponnurangam Kumaraguru[1], Tavpritesh Sethi[1,2*]

1. Indraprastha Institute of Information Technology Delhi, India
2. All India Institute of Medical Sciences, New Delhi, India
3. CSIR-Institute of Genomics and Integrative Biology, Delhi, India
4. Maharaja Surajmal Institute of Technology, GGSIPU, New Delhi, India

*tavpriteshsethi@iiitd.ac.in

Abstract

Background. COVID-19 knowledge has been changing rapidly with the fast pace of information that accompanied the pandemic. Since peer-reviewed research is a trusted source of evidence, capturing and predicting the emerging themes in COVID-19 literature are crucial for guiding research and policy. Machine learning, natural language processing and dynamical networks have the potential to enable rapid distillation and prediction of actionable insights for ending the pandemic.

Objective. We hypothesized that emerging COVID-19 research trends can be captured and predicted from networks constructed upon language features. Further, we aimed to detect communities in these networks and used centrality measures to track and predict emerging network modules as dominant themes in a given time period. The goal of our study was to make our findings publicly available as an explainable AI dashboard for researchers and policymakers.

Methods. Abstracts from more than 95,000 peer-reviewed articles from the WHO curated COVID-19 database were used to construct word embedding models. Named entity recognition was used to refine the terms. Cosine similarity between the terms was then used to construct dynamical networks in order to understand the temporal trend of emerging associations over months and visualized as alluvial diagrams. Finally, temporal link prediction between diseases for the subsequent month based on their trends of occurrence in the previous six months was carried out to predict the emergence and disappearance of associations in the rapidly changing pandemic scenario.

Results. Community detection upon dynamical networks clearly demonstrated the emergence of thromboembolic complications as a cluster and dominant theme between March and August, 2020. Forecasting of top-K influential entities further allowed prediction of future trends, such as the emergence of psychiatry theme as a central node by February 2021. XGBoost modeling in our proposed temporal link prediction framework achieved an AUC-ROC score of 0.855 for predicting new dis(associations) one month in advance. Visualization of the underlying word-embedding models allowed interactive querying to choose novel keywords and extractive models summarized the research relevant to the keyword, allowing faster knowledge distillation.

Conclusion: We provide an explainable AI approach for querying, tracking and predicting novel insights in COVID-19 peer reviewed literature. The *EvidenceFlow* web-application is publicly available and emerging trends are updated on a monthly basis. Such approaches will

be crucial to understand and pre-empt actionable research such as vaccine strategies in the ongoing pandemic.

Introduction

COVID-19 pandemic continues to be an enigma with its diverse clinical presentation, controversial evidence for treatment, fast-tracked vaccine development and unclear systemic implications. More than 200 countries have been affected by the pandemic with around 75 million confirmed cases and more than 1.6 million deaths recorded till 22nd December 2020 [1]. The literature around COVID-19 is growing at a similar pace with more than 95,000 research articles peer-reviewed articles made publicly available by the WHO [2]. Knowledge synthesis from peer-reviewed literature will become increasingly difficult for researchers, clinicians, and policymakers alike. Hence understanding COVID-19 in the context of evolving themes is important. Ebadi et al [3] carried out topic modeling and sentiment analysis comparing pre-print with peer reviewed literature over a short time span from January to May 2020. However, we are reporting for the first time, the use of unsupervised word embeddings, networks analysis, link prediction and machine learning to predict emerging themes in COVID-19 literature and making these publicly available as a web-application. The current model is trained upon the 95,000 peer-reviewed articles obtained from the WHO Database and will be updated with new publications and pre-prints as these become available on a monthly basis.

The abstract of articles holds a substantial amount of information about the literature. Named entities play a crucial role in deducing valuable information from large amounts of text and influencing the trends of literature. Models pre-trained on biomedical, scientific and clinical benchmark datasets, can be used for extraction of a variety of clinical entities such as diseases, chemicals, adverse drug reactions from continuous text. By creating dynamic networks of the extracted entities and weighing the links by cosine similarity, we study the shift in flow of importance of each node. We implement a framework for predicting the top-K influential nodes which tend to represent the theme of a given month's literature based on forecasted centrality measures by an autoregression model.

In addition to predicting broader themes, a study of resurfacing and diminishing links at individual entity level can also reveal the evolution of research. Link Prediction has been defined as the task of predicting the existence of links between two nodes in a complex network based on a set of topological features. The problem of link prediction in real-world temporal networks has been explored a lot in recent years, primarily in online social media networks where nodes are represented by users and edges by the relationship between them. Bu et al [5] proposed a novel semi-supervised learning framework, which integrates survival analysis and game theory for predicting future links. Peddada et al [6] explored the problem of link prediction using supervised learning methods based on proximity scores to capture the temporal shift. In this paper, we propose a framework to predict reconnected and missing links between clinical entities such as diseases extracted from textual data over T time intervals, using a set of proximity scores derived from associated dynamic networks and word embedding similarity. The co-occurrence of words in a span of text plays a vital role in capturing a high-level semantic relationship. Hence, we label the links based on co-occurrence analysis between entity pairs. Given the vast research found on online social networks, our framework differs from the standard link prediction models as it studies the concept by

applying named entity recognition in the scientific literature. The prediction of links between diseases mentioned in abstracts reflects on accurate and validated insights, hence demonstrating the effectiveness of our proposed approach.

Methods

Data-sets. The dataset was created using more than 95,000 peer-reviewed research articles related to coronavirus present in the *WHO Database* [2] from February 2020 to September 2020 [supp. figure 1(a)].

Text Pre-processing and Exploratory Data Analysis. Formatting of text and removal of white-spaces, punctuation, digits and stop words was carried out on lowercase converted text using NLTK package [18]. Word frequency distributions were visualized as chatter plots using ggplot2 package [7].

Named Entity Recognition. Named Entity Recognition was used to extract two types of entities: diseases and chemicals, from the original abstracts of vetted research articles using a pre-trained model (*en_ner_bc5cdr_md*) from SciSpacy, an open-source project developed for Biomedical Natural Language Processing [13]. Entities were further used for creating networks to study the trends through alluvial diagrams and for predicting temporal links between diseases across past and upcoming months.

Unsupervised Word Embeddings. A low-dimensional representation for the disease and chemical entities was learned using the word2vec model with skip-gram algorithm, one-hot encoding and fixed window size of five, implemented in Gensim [11, 12, 19]. Each word vector obtained from its embedding represents an entity and the distance between word vectors was used to calculate dis(similarity) between entities. Visualization of the word vectors was carried out using Tensorflow Embedding Projector [20] to allow interactive exploration of relationships between disease and chemical entities. Separate word2vec models were trained for each month from February to September, 2020 in order to allow capturing of dynamic changes in word similarities in COVID-19 literature.

Longitudinal Word Vector Networks and Communities. Weighted networks were constructed using similarity scores between word vectors as edge weights. A union of all nodes with top ten percentile similarity scores across February to September, 2020 were preserved as nodes in the networks. Community detection was done over the monthly networks using the Infomap algorithm [17]. Dynamic change in the communities as emerging themes over months was tracked using an alluvial visualization [14]. Detailed steps with parameters are available in the supplementary material.

Time Series Forecasting of Top-K Influential Entities. In order to predict the top-K influential nodes in temporal networks of subsequent months, we evaluated three centrality measures PageRank, Eigenvector Centrality and Degree Centrality of the nodes in the past networks [22].

These centrality values were used to forecast future centralities using the Vector Autoregression (VAR) model [21]. Briefly, the VAR model was fit on a time series of each node's centralities calculated from the networks of February to September, 2020 and predicted the node's centralities for October, 2020. The top-K influential nodes were obtained by sorting the sum of the three forecasted centrality measures in descending order. The performance of forecasts is assessed in comparison with the sum of true centralities in retrospective test data using the ranking metric precision@k.

Temporal Link Prediction between Entities. We predicted the existence of a link between entities at timestamp $\tau+1$ based on computed feature vectors obtained from previous timestamps in the time interval τ. Briefly, 6-month partitions of data starting from February 2020 were used for training models with testing over the subsequent month. Ground truth for presence and weight of link was defined from co-occurrence and cosine similarity respectively. Nine proximity scores based upon network topology were computed using the NetworkX package[15] and a first order difference of these series were taken in order to capture temporal trends. These were further normalized and used as features for predicting the existence of a link at $\tau+1$ using Random Forests [23], SVM [24], AdaBoost [25], XGBoost [26], and LGBM [27] models and the best model for link prediction based on AUC ROC scores was used to predict the subsequent links. The full detail of the algorithm and features are available in the supplementary material.

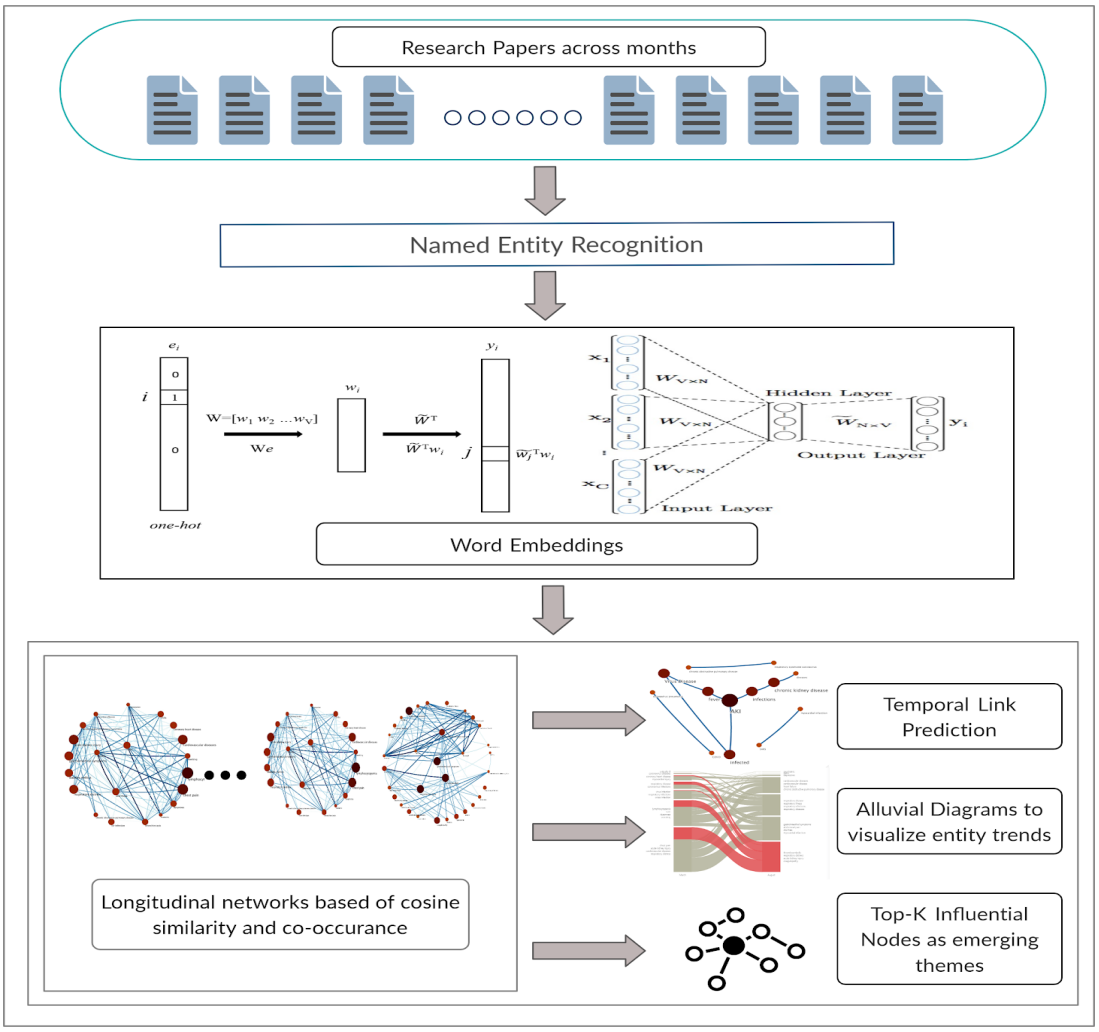

Figure 1: Graphical representation of proposed framework explaining the complete workflow. The pipeline takes abstracts as input from which entities are extracted using NER. Embeddings are generated which are used as features for longitudinal networks. These networks are used for visualizing the trends using alluvial diagrams, temporal link prediction and predicting top-k influential nodes for theme prediction.

Implementation and Availability. *EvidenceFlow*, our web-application with results of online tracking and prediction of emerging themes is available publicly at https://evidenceflow.tavlab.iiitd.edu.in/ .

Results

A total of 21,715 distinct diseases and 19,226 distinct chemicals were identified. Figure 2 shows the top frequent disease and chemical entities identified in the corpus.

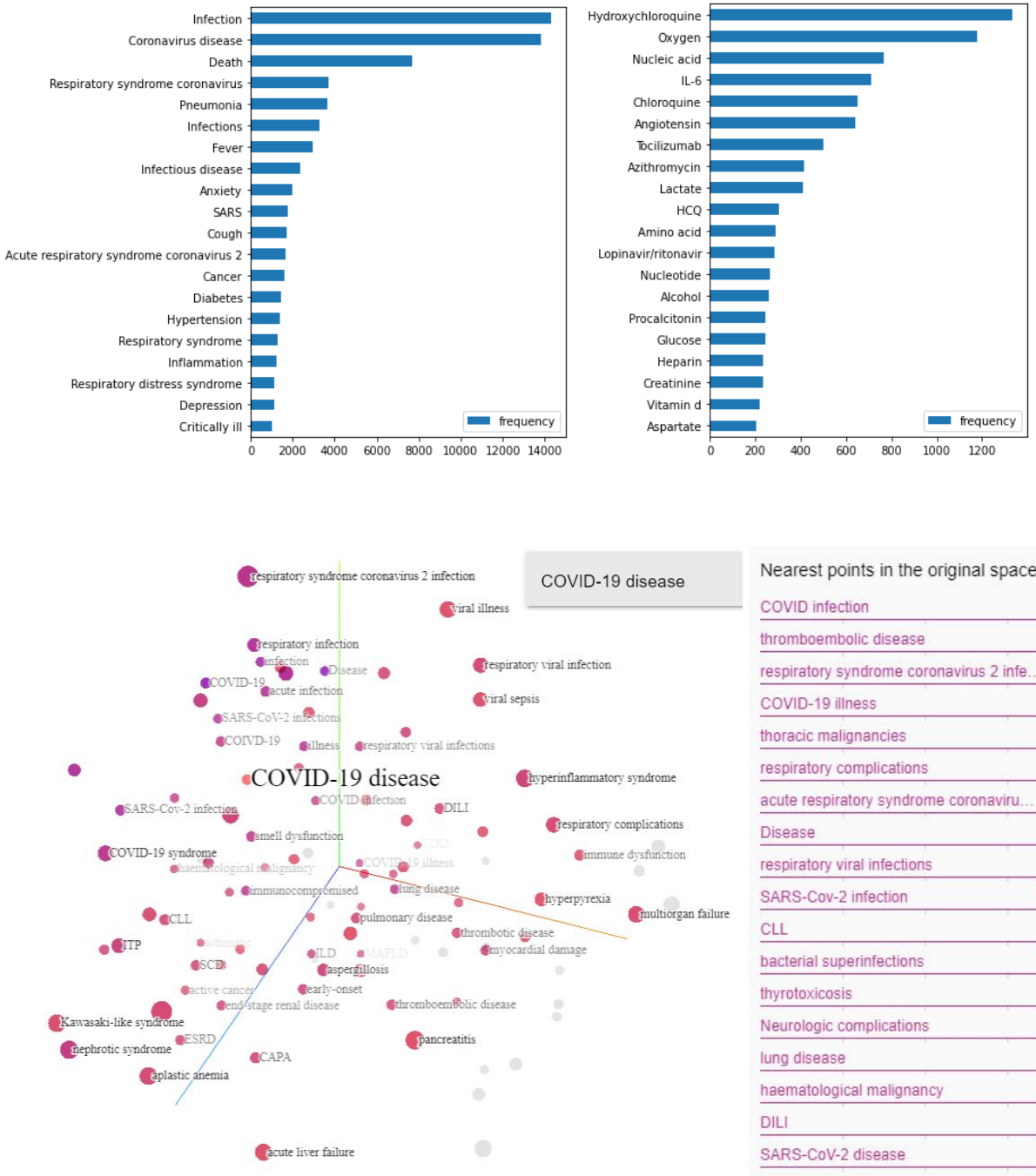

Figure 2: (a) Bar plot (left) showing frequency of top diseases in the corpus of abstracts extracted using Named Entity Recognition. (b) Bar plot (right) showing frequency of top chemicals in the corpus of abstracts extracted using Named Entity Recognition. c) Latent space of word embeddings visualized around the keyword 'COVID-19 disease', displaying 100 isolated points nearest to it. d) Entities nearest to 'COVID-19 disease' in terms of cosine distance in the original space.

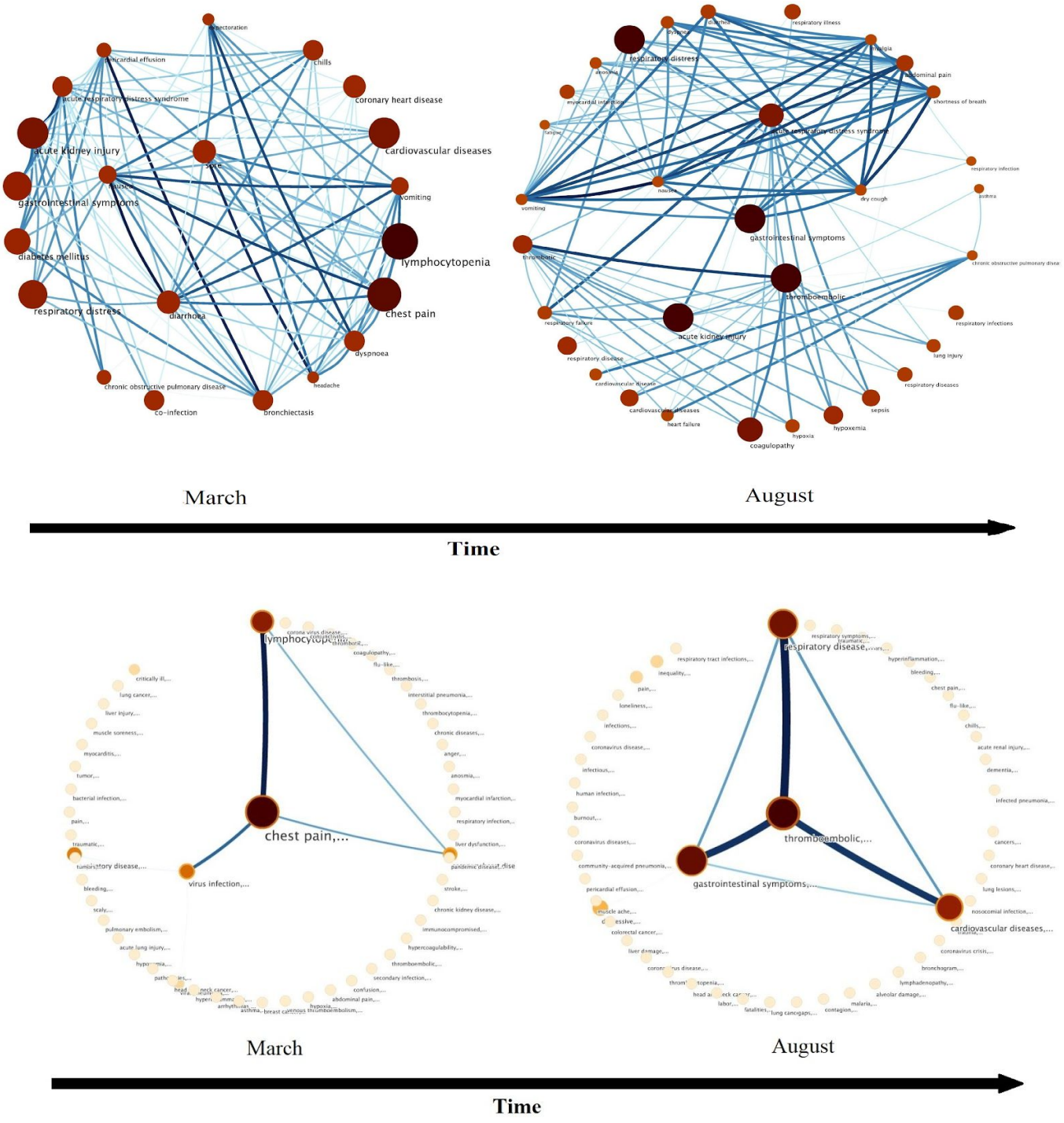

Figure 3: a) Longitudinal source network from March 2020 to August 2020. b) Longitudinal Community network from March 2020 to August 2020. Emergence of Thromboembolic complications as a major theme by August, 2020 have been illustrated by source and community network.

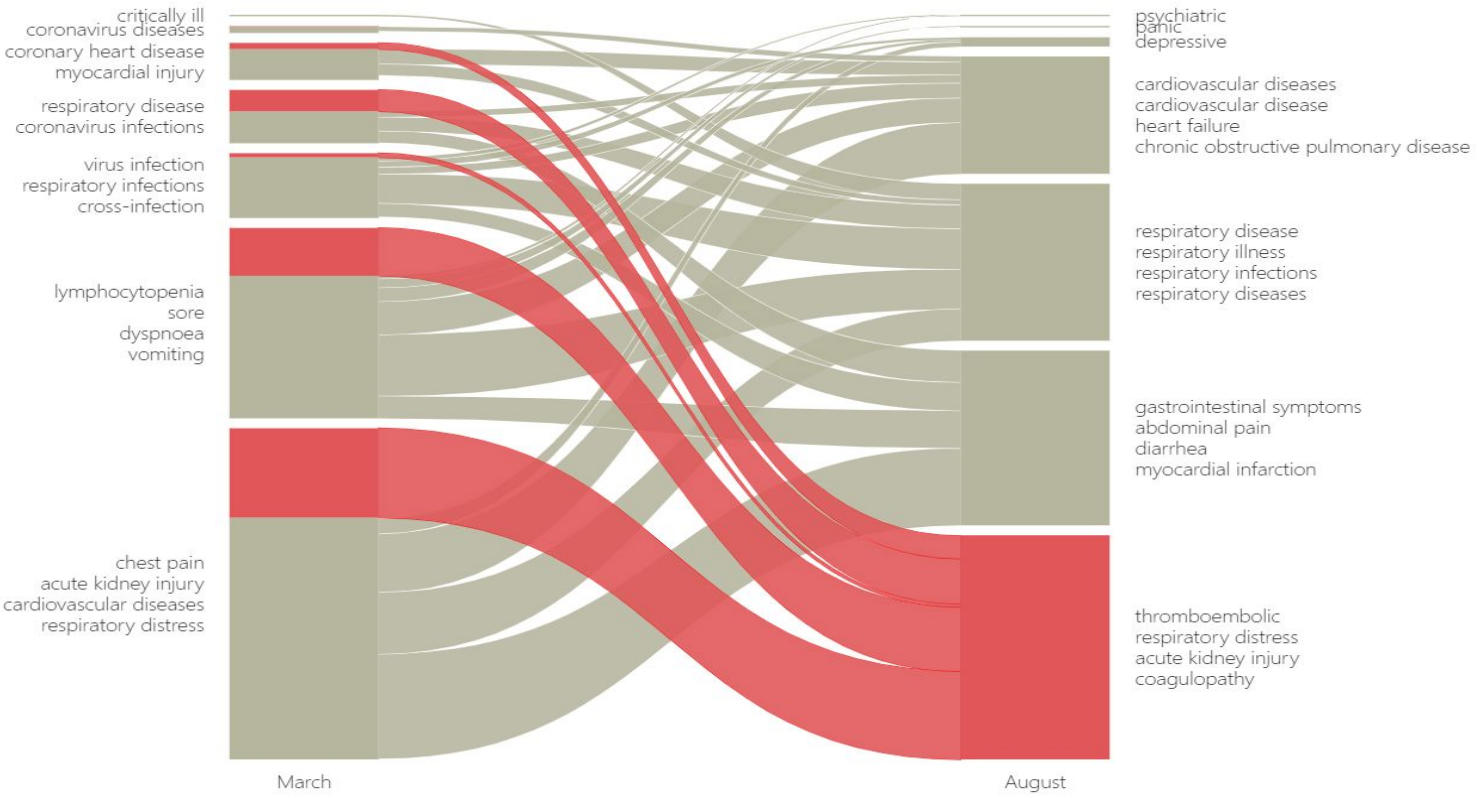

Figure 4: Alluvial diagram for tracking the trends from March 2020 to August 2020. The alluvial diagram eases tracing the trends of temporal dynamics of literature across different months. The diagram clearly illustrates the emergence of thromboembolic complications as a major theme by August. The Vector Autoregression model trained upon the network centralities predicted that the "psychiatric" streamline, seen here as a relatively unimportant module in August, would assume a higher centrality in February, 2021.

The association among the most prevalent diseases is represented graphically using an alluvial diagram. A detailed inference of the alluvial diagram across the month of March and August depicted the emergence of thromboembolic complications as the most important module. In March, the dominant modules were lymphocytopenia, chest pain and acute kidney injury, depicting lesser traces of thromboembolic complications in initial months. The network of August evidently captures the rising influence of nodes linked to thromboembolism, gastrointestinal symptoms, respiratory and cardiovascular diseases. Our study depicts how word embeddings generated from Word2Vec model trained on each month's literature can support the creation of entity networks. The importance of a node is influenced by its topological position and centrality measures along with Pagerank. Our time series analysis presents how the dynamic networks of entities can further be leveraged to efficiently forecast the most influential nodes which can represent a broad theme of a given month's research.

Node-1	**Node-2**	**Paper (titles)**
COVID	AKI	1. COVID-19 and the kidney 2. Targeting acute kidney injury in COVID-19 3. Acute Kidney Injury in COVID-19 Patients: An Inner City Hospital Experience and Policy Implications

		4. Incorporation of urinary neutrophil gelatinase-Associated lipocalin and computed tomography quantification to predict acute kidney injury and in-hospital death in covid-19 patients 5. Successfully treating three patients with acute kidney injury secondary to COVID-19 by peritoneal dialysis: Case report and literature review. 6. Acute Kidney Injury Associated with COVID-19: A Retrospective Cohort Study 7. Prevalence of Acute Kidney Injury in Severe and Critical COVID-19 Patients in Wuhan, China
COVID	immunocompromised	1. Telemedicine as the New Outpatient Clinic Gone Digital: Position Paper From the Pandemic Health System REsilience PROGRAM (REPROGRAM) International Consortium (Part 2) 2. Coronavirus Disease 2019 Viremia, Serologies, and Clinical Course in a Case Series of Transplant Recipients 3. Diet, Gut Microbiota and COVID-19 4. Contribution of Nanotechnology in the Fight Against COVID-19 5. Biomarkers of Cytokine Release Syndrome Predict Disease Severity and Mortality From COVID-19 in Kidney Transplant Recipients. 6. COVID-19 Severity and Outcomes in Patients With Cancer: A Matched Cohort Study. 7. A Case of Guillain-Barré Syndrome Associated With COVID-19.
respiratory syndrome coronavirus	hypoxia	1. Clotting disorder in severe acute respiratory syndrome coronavirus 2
inflammation	thromboembolic	1. Active smoking and severity of COVID-19 infection in cancer patients
pneumonia	chronic obstructive pulmonary disease	1. Clinical Factors Associated with Progression and Prolonged Viral Shedding in COVID-19 Patients: A Multicenter Study
diabetes	depression	1. Sarcopenia during COVID-19 lockdown restrictions: long-term health effects of short-term muscle loss.
cough	sepsis	1. Clinical characteristics and 28-day mortality among patients with solid cancers and COVID-19 in a tertiary hospital 2. Real-world outcomes in thoracic cancer patients (pts) with severe acute respiratory syndrome coronavirus 2 (COVID-19): Single UK institution experience

Table 1: This table depicts the new links between entity pairs (Node1 and Node2) which were not present in the previous three months, as predicted by the XGBoost model for the month of October using the testing set of the previous six months (March to September). This is a subset of the correctly predicted links, that were found to be not present for the previous two months. 'AKI' has shown emerging links with 'SARS' and 'COVID'. 'Chronic obstructive pulmonary disease' has been predicted to show links with 'pneumonia' and 'death'. The links represent the frequent co-occurrence of entities at the given timestamp based on networks of previous months. The Papers mentioned are from October and they represent the validation of our model's temporal link prediction as their abstracts talk about the given two entities, hence verifying the concept of co-occurrence.

Discussion

An open-source dashboard called *EvidenceFlow*, has been built, which can act as a template for collection research articles for a specific disease or in adverse scenarios, to propagate proper information related to research at faster access. The dashboard also allows the user to unravel the literature with a dynamic map of embeddings based on the visualization provided by Tensorboard. The dashboard aims to track literature trends using alluvial diagrams, projecting influential entities, and network analysis across different months.

The potential of the word embeddings as well as NER was leveraged to extract insights regarding the topmost similar diseases or chemicals with selected keywords. Vaccine, which has been a rising topic lately, had the highest cosine similarity with Ad26.COV2.S (also known as Ad26) and mRNA-1273, which are few of the most discussed candidate vaccines for COVID19 in the literature. 'Comorbidity' is found to have a high similarity with hyperlipidemia, diabetes mellitus, heart as well as kidney diseases. A number of long-term effects have been reported post recovery due to a weakened immune system. Exploring 'adverse effects' as a keyword depicted correlations with cardiac adverse events, maladaptive anxiety and humoral immunodeficiency. People with dementia-related neuropsychiatric symptoms have been affected adversely as well. 'Social' factors were found to have the highest similarity with connectedness and an increase in family violence and psychological damage. It also highlights the existing gap between rural and urban communities. The economic recession caused by the pandemic has also led to loneliness and social anxiety which was captured by the language models as well. 'Psychological' health has been negatively impacted due to worry and stress over the coronavirus, which is characterized by aggravation of conditions such as PTSD and an intuitively high cosine similarity with 'eating disorder'. Hence, Natural Language Processing techniques were effectively used to capture latent associations among general keywords and named entities. Detailed descriptions of the analysis can be found in the supplementary table [supp. table 2].

Exploring the evolution of literature based on themes helps reveal insightful trends using Natural Language Processing. In recent years, many methods have been put forward to predict centrality in dynamic networks on the basis of past values [4]. Forecasting top-K influential entities based on centrality measures can also assist in steering research while understanding the temporal dynamics of themes represented by them. The method presented in this paper has provided a novel approach for identifying critical nodes in entity networks weighted by cosine similarity, and can be extended to various other dynamical analyses such as the impact of entities on expanding dynamic knowledge graphs.

The study conducted on the resurfacing links for October 2020 depicts the efficacy of the model by accurately predicting more intense possibilities of coronavirus being linked with critical infection in the body which alludes to acute kidney injury (AKI). Links have also been predicted between chronic obstructive pulmonary disease and pneumonia. The diminishing links for the month of October [supp. table 6] also reveals various inferences about latent space of literature. Inference of the model was done for the month of November [supp. table 7]. Links between anxiety and depression, and many other links with words 'anxiety' and 'trauma' are predicted to diminish. This helps us to infer that in the month of November, mental health has not been discussed often. Awareness regarding mental health has been

raised in multiple ways including research work and the topic seems to dilute with the normalizing situation worldwide, as the topic of vaccination has been rising. The temporal shift of links captured by computing the difference between the normalized proximity scores of node pairs in each monthly interval lets the model track links by taking into account language features as well as topological attributes of the networks.

However, one limitation of the current tool is that the networks have been analysed with limited frequency of entities, primarily diseases. The future work in this direction can expand to include other medical entities such as genes, drugs or adverse drug effects. Advancement in the architecture of temporal link prediction can include larger data and complex models like RNN and LSTM to predict the links. We do not train deep learning models as the training data points are limited and these models would tend to overfit. However, the dashboard has surfaced as a great tool for a high-level study of the COVID-19 literature.

Conclusion

The COVID-19 literature has been expanding at an exponential pace since the beginning of 2020. We examined an approach to take advantage of dynamic and homogeneous networks of medical entities and their associated cosine similarities to explore the trends in literature using alluvial diagrams. Our proposed time series analysis of top-K most influential entities correctly forecasts 8 out of top-10 influential nodes in the month of October [supp. table 3]. We used the model based on past centrality measures to further predict the importance of entities in January and February [supp. table 4]. The inference suggests that entities linked to ‘psychiatric’ themes shall emerge along with major influence of respiratory conditions, thromboembolism and malignancy in the literature for these months.

We further advance the analysis of trends to predicting links between entity pairs for the upcoming months. Our proposed framework for Temporal Link Prediction effectively captures reconnecting and diminishing links between diseases present in the scientific literature for the successive month on the basis of dynamic networks belonging to the previous six months. Our results show that the XGBoost model is able to classify links with an AUC-ROC score of 0.855 in the test set [supp. table 5].

We validated our results by mentioning the papers that contain the excerpts pertaining to the co-occurrence of disease-pairs whose links were correctly predicted for the month of October. The proposed frameworks make use of NLP based networks and surface as an efficient tool for querying, tracking and predicting insights from COVID-19 peer reviewed literature.

Acknowledgement

This work was partially supported by the Wellcome Trust/DBT India Alliance Fellowship IA/CPHE/14/1/501504 awarded to Tavpritesh Sethi. We also acknowledge support from the Center of Excellence in Healthcare and the Center of Excellence in Artificial Intelligence at IIIT-Delhi

Funding
None

Conflict of Interest

None

References

1. World Health Organization. "Coronavirus Disease (COVID-19) Situation Reports". (2020). https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports.

2. "WHO Global research Database on coronavirus disease(COVID-19)." https://www.who.int/emergencies/diseases/novel-coronavirus-2019/global-research-on-novel-coronavirus-2019-ncov.

3. Ebadi, Ashkan, Pengcheng Xi, Stéphane Tremblay, Bruce Spencer, Raman Pall, and Alexander Wong. "Understanding the temporal evolution of COVID-19 research through machine learning and natural language processing." Scientometrics (2020): 1-15.

4. Kim, Hyoungshick, John Tang, Ross Anderson, and Cecilia Mascolo. "Centrality prediction in dynamic human contact networks." Computer Networks 56, no. 3 (2012): 983-996.

5. Bu, Zhan, Yuyao Wang, Hui-Jia Li, Jiuchuan Jiang, Zhiang Wu, and Jie Cao. "Link prediction in temporal networks: Integrating survival analysis and game theory." Information Sciences 498 (2019): 41-61.

6. Peddada, Amani V., and Lindsey Kostas. "Users and Pins and Boards, Oh My! Temporal Link Prediction over the Pinterest Network."

7. Wickham, Hadley, and Winston Chang. "ggplot2: An implementation of the Grammar of Graphics." R package version 0.7, URL: http://CRAN. R-project. org/package= ggplot2 3 (2008).

8. Lohmann, Steffen, Jürgen Ziegler, and Lena Tetzlaff. "Comparison of tag cloud layouts: Task-related performance and visual exploration." In *IFIP Conference on Human-Computer Interaction*, pp. 392-404. Springer, Berlin, Heidelberg, 2009.

9. Foltz, Peter W. "Latent semantic analysis for text-based research." *Behavior Research Methods, Instruments, & Computers* 28, no. 2 (1996): 197-202.

10. Kherwa, Pooja, and Poonam Bansal. "Latent Semantic Analysis: An Approach to Understand Semantic of Text." In *2017 International Conference on Current Trends in Computer, Electrical, Electronics and Communication (CTCEEC)*, pp. 870-874. IEEE, 2017.

11. Ma, Long, and Zhang, Yanqing. "Using Word2Vec to process big text data." In *2015 IEEE International Conference on Big Data (Big Data)*, pp. 2895-2897. IEEE, 2015.

12. Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. "Efficient estimation of word representations in vector space." *arXiv preprint arXiv:1301.3781* (2013).

13. Neumann, Mark, et al. "Scispacy: Fast and robust models for biomedical natural language processing." arXiv preprint arXiv:1902.07669 (2019).

14. Rosvall M, Bergstrom CT. “Mapping Change in Large Networks” PLoS ONE, 5:e8694, 2010.

15. Aric A. Hagberg, Daniel A. Schult and Pieter J. Swart. “Exploring network structure, dynamics, and function using NetworkX”, in Proceedings of the 7th Python in Science Conference (SciPy2008)

16. Ruopp, Marcus D., Neil J. Perkins, Brian W. Whitcomb, and Enrique F. Schisterman. "Youden Index and optimal cut‐point estimated from observations affected by a lower limit of detection." Biometrical Journal: Journal of Mathematical Methods in Biosciences 50, no. 3 (2008): 419-430.

17. Bohlin, Ludvig, Daniel Edler, Andrea Lancichinetti, and Martin Rosvall. "Community detection and visualization of networks with the map equation framework." In Measuring scholarly impact, pp. 3-34. Springer, Cham, 2014.

18. Loper, Edward, and Steven Bird. "NLTK: the natural language toolkit." arXiv preprint cs/0205028 (2002).

19. Rehurek, Radim, and Petr Sojka. "Gensim–python framework for vector space modelling." NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic 3, no. 2 (2011).

20. Smilkov, Daniel, Nikhil Thorat, Charles Nicholson, Emily Reif, Fernanda B. Viégas, and Martin Wattenberg. "Embedding projector: Interactive visualization and interpretation of embeddings." arXiv preprint arXiv:1611.05469 (2016).

21. Johansen, Søren. "Modelling of cointegration in the vector autoregressive model." Economic modelling 17, no. 3 (2000): 359-373.

22. Oldham, Stuart, Ben Fulcher, Linden Parkes, Aurina Arnatkevic˘iūtė, Chao Suo, and Alex Fornito. "Consistency and differences between centrality measures across distinct classes of networks." PloS one 14, no. 7 (2019): e0220061.

23. Breiman, Leo. "Random forests." Machine learning 45, no. 1 (2001): 5-32.

24. Hearst, Marti A., Susan T. Dumais, Edgar Osuna, John Platt, and Bernhard Scholkopf. "Support vector machines." IEEE Intelligent Systems and their applications 13, no. 4 (1998): 18-28.

25. Freund, Yoav, Robert Schapire, and Naoki Abe. "A short introduction to boosting." Journal-Japanese Society For Artificial Intelligence 14, no. 771-780 (1999): 1612.

26. Chen, Tianqi, and Carlos Guestrin. "Xgboost: A scalable tree boosting system." In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, pp. 785-794. 2016.

27. Ke, Guolin, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. "Lightgbm: A highly efficient gradient boosting decision tree." In Advances in neural information processing systems, pp. 3146-3154. 2017.