Predicting Emerging Themes in Rapidly Expanding COVID-19 Literature with Unsupervised Word Embeddings and Machine Learning

Ridam Pal, Harshita Chopra, Raghav Awasthi, Harsh Bandhey, Aditya Nagori, Tavpritesh Sethi

Submitted to: Journal of Medical Internet Research
on: October 06, 2021

Table of Contents

Predicting Emerging Themes in Rapidly Expanding COVID-19 Literature with Unsupervised Word Embeddings and Machine Learning

Ridam Pal[1]; Harshita Chopra[2]; Raghav Awasthi[1]; Harsh Bandhey[1, 3]; Aditya Nagori[1, 4]; Tavpritesh Sethi[1, 5] PhD, MBBS

[1]Indraprastha Institute of Information Technology Delhi New Delhi IN
[2]Maharaja Surajmal Institute of Technology, GGSIPU New Delhi IN
[3]Duke University Durham GB
[4]CSIR-Institute of Genomics and Integrative Biology, Delhi New Delhi IN
[5]All India Institute of Medical Sciences, New Delhi New Delhi IN

Corresponding Author:
Tavpritesh Sethi PhD, MBBS
Indraprastha Institute of Information Technology Delhi
Delhi Okhla Industrial Estate, Phase III
New Delhi - 110020
New Delhi
IN

Abstract

Background: Evidence from peer-reviewed literature is the cornerstone for designing responses to global threats such as COVID-19. The collection of knowledge in publications needs to be distilled into evidence by leveraging natural language models and machine learning.

Objective: We aim to show that new knowledge can be captured and tracked using the temporal change in the underlying unsupervised word embeddings of literature. Further imminent themes can be predicted using machine learning upon the evolving associations between words.

Methods: Frequently occurring medical entities were extracted from the abstracts of more than 150,000 COVID-19 articles published on the WHO database, collected on a monthly interval starting from February 2020. Word embeddings trained on each month's literature were used to construct networks of entities with cosine similarities as edge weights. Topological features of the subsequent month's network were forecasted based on prior patterns and new links were predicted using supervised machine learning. Community detection and alluvial diagrams were used to track biomedical themes that evolved over the months.

Results: We found that thromboembolic complications were detected as an emerging theme as early as August 2020. A shift towards symptoms of Long COVID complications was observed during March 2021 and neurological complications gained significance in June 2021. A prospective validation of the link prediction models achieved an AUROC score of 0.87. Predictive modelling revealed predisposing conditions, symptoms, cross-infection and neurological complications as a dominant research theme in COVID-19 publications based on patterns observed in previous months.

Conclusions: Machine learning-based prediction of emerging links can contribute towards steering research by capturing themes represented by groups of medical entities, based on patterns of semantic relationships over time.

(JMIR Preprints 06/10/2021:34067)
DOI: https://doi.org/10.2196/preprints.34067

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✔ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**

Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v

Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Predicting Emerging Themes in Rapidly Expanding COVID-19 Literature with Unsupervised Word Embeddings and Machine Learning

Ridam Pal[1], Harshita Chopra[2], Raghav Awasthi[1], Harsh Bandhey[1,3], Aditya Nagori[1,4], Tavpritesh Sethi[1,5]

[1] Indraprastha Institute of Information Technology Delhi, India

[2] Maharaja Surajmal Institute of Technology, GGSIPU, New Delhi, India

[3] Duke University, Durham NC, USA

[4] CSIR-Institute of Genomics and Integrative Biology, Delhi, India

[5] All India Institute of Medical Sciences, New Delhi, India

Abstract

Background:
Evidence from peer-reviewed literature is the cornerstone for designing responses to global threats such as COVID-19. In the massive and rapidly growing corpuses such as the COVID-19 publications, assimilating and synthesizing information is challenging. Leveraging a robust computational pipeline that evaluates multiple aspects such as network topological features, communities and their temporal trends can make this process more efficient.
Objective:
We aim to show that new knowledge can be captured and tracked using the temporal change in the underlying unsupervised word embeddings of literature. Further imminent themes can be predicted using machine learning upon the evolving associations between words.
Methods:
Frequently occurring medical entities were extracted from the abstracts of more than 150,000 COVID-19 articles published on the WHO database, collected on a monthly interval starting from February 2020. Word embeddings trained on each month's literature were used to construct networks of entities with cosine similarities as edge weights. Topological features of the subsequent month's network were forecasted based on prior patterns and new links were predicted using supervised machine learning. Community detection and alluvial diagrams were used to track biomedical themes that evolved over the months.
Results:
We found that thromboembolic complications were detected as an emerging theme as early as August 2020. A shift towards symptoms of Long COVID complications was observed during March 2021 and neurological complications gained significance in June 2021. A prospective

validation of the link prediction models achieved an AUROC score of 0.87. Predictive modeling revealed predisposing conditions, symptoms, cross-infection and neurological complications as a dominant research theme in COVID-19 publications based on patterns observed in previous months.

Conclusion:

Machine learning-based prediction of emerging links can contribute towards steering research by capturing themes represented by groups of medical entities, based on patterns of semantic relationships over time.

Keywords:

COVID-19; Named Entity Recognition; Unsupervised Word Embeddings; Machine Learning; Natural Language Preprocessing

Introduction

Global health threats such as the COVID-19 pandemic have proved to be an enigma with its diverse clinical presentation, controversial evidence for treatment, fast-tracked vaccine development, and unclear systemic implications. Most countries have been affected by COVID-19, with around 187 million confirmed cases over a short span, with more than 4 million deaths recorded until 13th July 2021[1]. The literature around COVID-19 is growing exponentially, with more than 150 thousand COVID-19 articles vetted by the World Health Organization[2]. Understanding evolving themes in a context such as in COVID-19 is essential as knowledge synthesis from peer-reviewed literature becomes increasingly difficult for researchers, clinicians, and policymakers alike. Methods such as topic modeling and sentiment analysis have been previously carried out comparing pre-print with peer-reviewed literature only over a short period. Ebadi et al[3] studied the temporal patterns of sentiments and similarity between publications from different sources over time, using document embeddings. High-level research topics like oncology, personal protective equipment (PPE), analytics, rehabilitation-panic, high-risk groups and genomics were uncovered using structural topic modeling. Although such analyses reflect an abstract overview of the broad areas of research, they lack to capture the evolving context between distinct domain-specific entities. The objective of our study is to analyze and track word-level semantic similarity among biomedical entities to uncover emerging themes.

Abstracts of articles hold a substantial amount of information about the literature. Named entities within the abstracts play a crucial role in deducing valuable information from large amounts of text and influencing literature trends[4]. Models pre-trained on biomedical, scientific, and clinical benchmark datasets have been used to extract various clinical entities such as diseases, symptoms, chemicals, and adverse drug reactions from the continuous text. The relative context of these entities changes over time, leading to a shift in similarity with other words[5]. Unsupervised word embeddings have previously been used to capture complex science concepts using the semantic relationship signified by cosine similarity[6].

Predicting links between “medical terms” is of high significance to understand the underlying themes within the literature and the phenomenon. Link Prediction is 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[7] primarily in online social media networks where nodes are represented by users and edges by the relationship between them. Supervised learning methods based on

topological proximity measures have been vastly used to capture the shifting of links across time within networks[8,9]. Our paper aims to fill these gaps through our proposed framework, EvidenceFlow[10], an interactive web application for tracking literature trends using alluvial diagrams, projection of influential entities, and network analysis across different months. We propose for the first time the use of diachronic word embeddings, link prediction in dynamic networks of entities, and machine learning to predict emerging themes literature and make these publicly available as a web application. This paper also studies the evolution of literature based on changing cosine similarity between extracted entities in weighted temporal networks and predicts future emerging trends using link prediction.

We have primarily focused on the fast emerging COVID-19 literature to train and validate our architecture for this study. We forecasted semantic and topological proximity features of named entity pairs generated from their temporal trends in prior months. Further, we have used these forecasted features to predict links between clinical entities extracted from textual data over the forecasted time interval using machine learning algorithms. Furthermore, these links were used to create a network weighted by forecasted cosine similarity for detecting communities of entities that tend to reflect on themes of the articles published in that month. To assess the efficacy of our predictive modeling, we validated the proximity features of entity pairs forecasted from ARIMA using Mean Squared Error. We have also evaluated the machine learning algorithm's performance for predicting the links over a time span of three months.

The schematic representation of workflow has been demonstrated (Figure 1). The interactive analysis and results of emerging themes is available publicly at our web application called EvidenceFlow. The details about its working can also be found in the Supplementary Material. This study proposes a framework for capturing and tracking imminent themes formed by medical entities in the temporal space based on networks constructed using word embeddings trained upon the evolving COVID-19 literature.

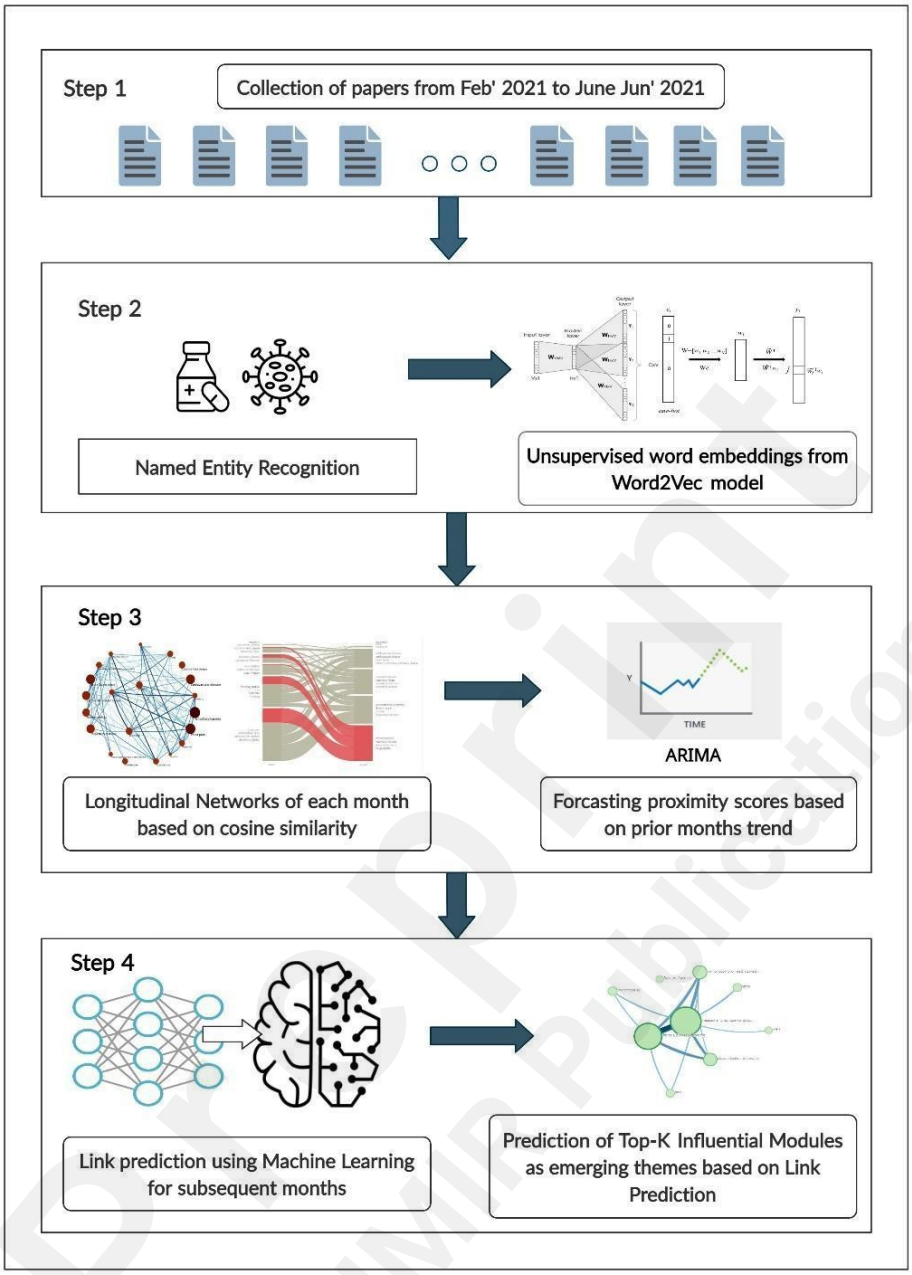


Figure 1: Graphical representation of proposed framework explaining the complete workflow. The pipeline takes abstracts as input from which entities are extracted using Named Entity Recognition. Embeddings are generated, which are used as features for longitudinal networks. These networks are used for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

Methods

Dataset and Text Pre-processing

The dataset was created from abstracts of ~150,000 COVID-19 articles published in the publicly available *WHO Database* [2]from February 2020 to June 2021 (Figure 2A). For every research article, the database contains the corresponding title, authors, source of publication, journal, database, language, type of publication, entry date, country and full text URL. We queried the database on all Full Text articles in the English language, keeping the rest of the fields unfiltered. The frequency of articles concerning specific categories and keywords has

been depicted in Supplementary Table 1. Formatting of text and removing white spaces, punctuation, digits, and stop words were carried out on lowercase converted text using the NLTK package[11]. We list all the softwares and packages used in further analysis along with the corresponding versions and sources in Supplementary Table 6.

Named Entity Recognition

Named Entity Recognition (NER) was used to extract two types of entities: diseases and chemicals, from the original abstracts of vetted research articles using a model pre-trained on BC5CDR corpus by SciSpacy, an open-source project for Biomedical Natural Language Processing[12]. The model identifies the entities with an F1 score of 84.49% [13]. The words extracted under the category of diseases also contained symptoms, adverse effects, conditions, disorders, and syndromes. All of these are collectively referred to as diseases in the other sections. Entities were further used to create networks to study the trends through alluvial diagrams and predict links between nodes across past and upcoming months.

Unsupervised Word Embeddings

Word embeddings were trained upon the abstracts obtained from the WHO Database, updated with new publications and pre-prints as these become available every month. A low-dimensional representation (d=100) for the words present in the corpus of abstracts was learned using the Word2Vec model with the skip-gram algorithm and fixed window size of five, implemented in Gensim[14–16]. Cosine distance between the word vectors of the extracted entities was calculated to analyze the dis(similarity) between entity pairs. Visualization of the word vectors was carried out using TensorFlow Embedding Projector[17] to allow interactive exploration of relationships between diseases and chemicals. To create each month's network of entities, separate Word2Vec models were trained to capture shifts in word similarities in the literature published over time.

Longitudinal Entity Networks and Communities

High cosine similarity represents strong relationships between words. We used diachronic word embeddings to capture the evolving contextual similarities between various diseases and studied its evolution over time. Weighted networks were constructed using similarity between word vectors of extracted entities as edge weights. From each month's corpus of abstracts, top-N(=100) most frequently occurring diseases were extracted, and the pairs having greater than the 90th percentile of cosine similarity based on the corresponding month's word embeddings were used to create a union set of entities across months, preserved as nodes in the temporal networks. Therefore, every month's network had a fixed set of nodes with varying links, labeled as 0 or 1 based on the threshold of cosine similarity, and varying weights, calculated based on the evolving semantic closeness. The mentioned threshold has been chosen empirically based on experimentation; a high threshold has been selected to depict contextual similarity between two words present in the same latent space. For training and evaluation, the fixed set of entity pairs was created from the diseases identified in the abstracts of the papers published from February 2020 to February 2021, using the mentioned procedure. For the subsequent months, the word embedding models were trained on the respective corpora of abstracts, and the links between the fixed set of node pairs were assigned if they appeared in the vocabulary and weighted by the cosine similarity between their word vectors. Community detection was performed over the monthly networks using the Infomap algorithm[18]. Semantic change in the word embeddings led to the formation of communities, which shifted as emerging themes over months. The importance of each node (entities) was tracked using an alluvial visualization

based on PageRank values which changed across different months[19]. Detailed steps with parameters are available in the Supplementary Information.

Time Series Forecasting of Proximity Scores

In order to predict the existence of links between nodes in the networks of subsequent months, we computed five neighborhood proximity scores for the network of each month. Jaccard Similarity, Common Neighbors, Preferential Attachment[20], and Adamic Adar Similarity[21] were used as topology-based features, and Cosine Similarity between the entities represented by the nodes was used as a semantic feature. These proximity scores based upon network topology were calculated using the NetworkX package[22]. The range of Adamic Adar Similarity, Common Neighbors, and Preferential Attachment values lies between $(0.00 - \infty)$, while that of Jaccard Similarity and Cosine Similarity lies in the range of (0.00 – 1.00). To scale the values, we normalized the former three scores in each network to bring them in the range of (0.00 – 1.00).

Every proximity score was modeled as a time series for each node pair, and the value was predicted for the subsequent month using the Auto-Regressive Integrated Moving Average or the ARIMA model [23]. Stationarity of the time series was assessed using Augmented Dickey-Fuller test. First-order autoregressive model (p=1, d=0, q=0) was used for stationary series and non-stationary time series were passed through the random walk order of the model (p=0, d=1, q=0). For validating, proximity scores for the network at timestamp $\tau+1$ were predicted based on their respective past values in the networks till timestamp τ. The model's performance was assessed by comparing the predictions with the original proximity scores in the $\tau+1$ time using the Mean Squared Error. MSE is one of the robust indicators to measure the closeness of forecast outputs to actual values in the time-series setting. To assess its sensitivity to outliers, we analyzed the distribution of errors (Supplementary Figure 5). It was seen that the median of errors was close to zero with minimal influence from outliers. Detailed steps with parameters are available in the Supplementary Information.

Link Prediction between Entities

The proximity scores predicted using the ARIMA model were further used to identify the occurrence of a link between entities in network $G_{\tau+1}$ based on the proximity scores and links in all previous networks [G_1, G_2, G_3, …, G_τ], using supervised machine learning. We experimented with the proposed link prediction approach using Logistic Regression[24], Random Forests[25], SVM[26], AdaBoost[27], XGBoost[28]. For training the models, four proximity scores, Jaccard Coefficient, Preferential Attachment, Adamic Adar Index, and Common Neighbors, were used as features of node pairs at each timestamp till τ. For validation, the forecasted proximity scores of the network at timestamp $\tau+1$ were used to predict links between nodes. Due to the high imbalance between the labels, Area Under Receiver Operator Characteristic Curve (AUROC) score was evaluated to select the optimal threshold for binary classification. While training, validating and testing the model, we did not use Cosine Similarity as a feature as it was the identifier variable for the link. Validation of the model was performed on the predicted proximity scores of April 2021 to June 2021. For Logistic Regression, evaluation of the key assumptions was done using Variance Inflation Factor (VIF) for measuring the degree of multicollinearity, Cook's Distance for detecting the presence of strongly influential outliers, and the scatter plot of log-odds for checking the linearity of independent variables. These tests were not satisfied for the data of most months, hence Logistic Regression was not our preferred model and we did not consider it further in results. Welch's t-test was performed for comparing performance of the machine learning models,

followed by Bonferroni correction [29]. The full details of the algorithm and features are available in the Supplementary Information. We list the parameters set for all the models in Supplementary Table 7.

Community Detection on Predicted Networks

The links between node pairs predicted by the best performing model were used to create networks weighted by cosine similarity scores predicted by the ARIMA model. Infomap algorithm was applied on the predicted and original test network to cluster the nodes into ten modules. The modules were compared using Intersection Over Union (IOU) using the following formula:

$$IOU = \frac{A \cap B}{A \cup B}$$

where A represents a set of nodes in the predicted i^{th} module, $i \in \{1, 2, ..., 10\}$, and B represents a set of nodes in the original j^{th} module, $j \in \{1, 2, ..., 10\}$.

Results

46885 distinct diseases and 53375 unique chemicals were identified and top entities are shown in Figure 2(C, D). Anxiety, depression and hypertension were found to be present in top-20 most discussed medical conditions in the research articles. Oxygen and Hydroxychloroquine were followed by nucleic acid and Angiotensin, a peptide hormone that causes vasoconstriction, in the most discussed chemicals. The latent space of word embeddings around the keyword ‘post-covid syndrome’ visualized using TSNE plot (Figure 2B) depicted ‘chronic fatigue’, ‘debilitating’, ‘neurodegenerative disorders’ and ‘vascular complications’ among the closest medical entities in terms of cosine distance. Similar visualization for the term 'mental disorders' can be found in Supplementary Figure 2 and top-10 most similar entities with selected keywords: ‘vaccine’, ‘comorbidity’, ‘adverse effects’, ‘social’ and ‘psychological’ can be found in Supplementary Table 8.

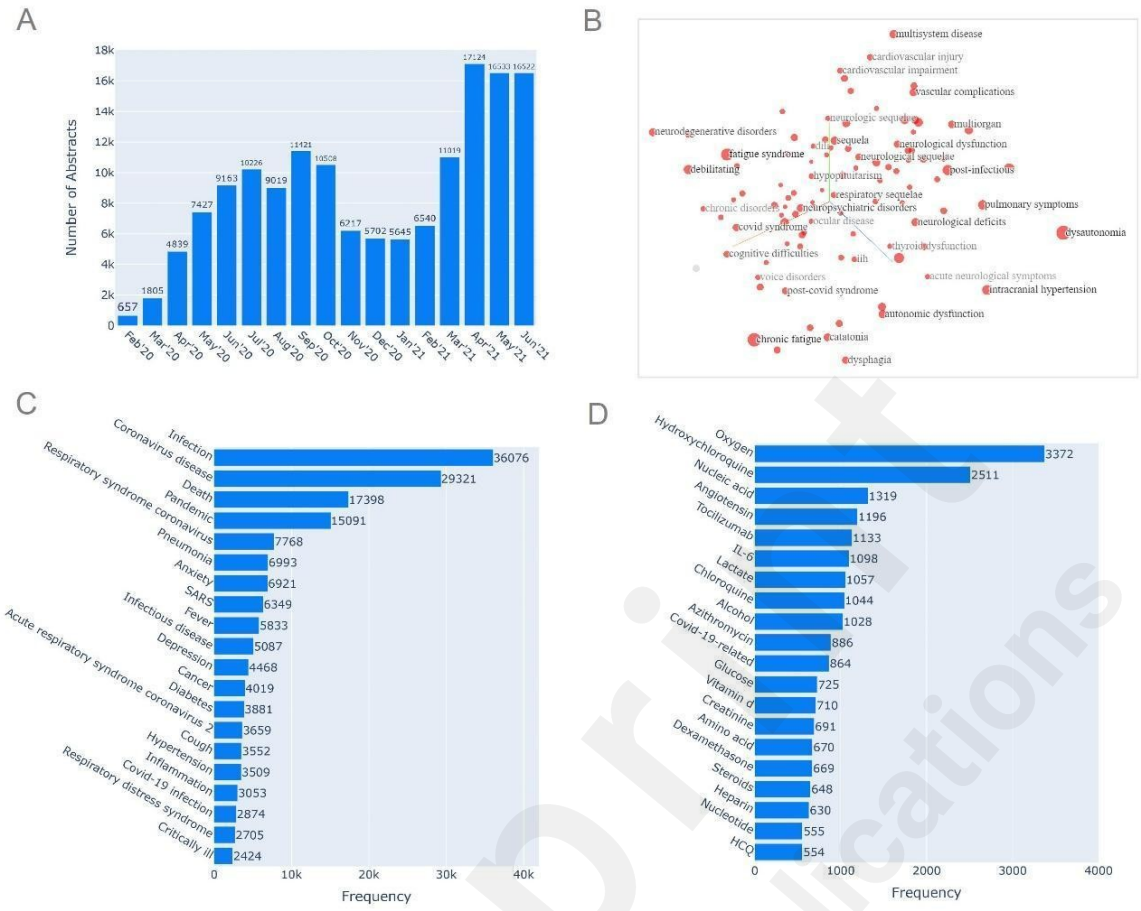


Figure 2: (A) Showing the number of articles occurring each month. The curve depicts that there has been a rampant increase in the number of articles across each month since February 2020. (B) Latent space of word embeddings of diseases visualized around the keyword 'post-covid syndrome', displaying 100 isolated points nearest to it. (C) Bar plot (left) showing the frequency of top diseases in the corpus of abstracts extracted using NER. (D) Bar plot (right) showing the frequency of top chemicals in the corpus of abstracts extracted using NER.

We conducted detailed inference of the alluvial diagram across different months to graphically explore the temporal trends in the literature based on dynamic and homogeneous networks of prevalent medical entities and their associated cosine similarities. Figure 3A represents the flow of themes found in the literature published in 2020. For March 2020, the dominant themes noted were chest pain, acute kidney injury, and lymphocytopenia. While there were lesser traces of thromboembolic complications in literature of early months, it emerged as the most significant theme in August 2020 (Figure 3A). Myocardial injury and cardiovascular diseases surfaced as a crucial cluster of entities in December 2020. Mental health factors such as depression, loneliness, anxiety and burnout gained significance in the literature of the last quarter of 2020. Figure 3B represents the flow of themes found in the literature published in 2021. While thromboembolism, hypoxemia and myocardial infarctions remained major concerns till January 2021, a significant transition towards Long COVID symptoms was found to be a major theme in March 2021. In June 2021, central modules including post effects and neurological complications, stroke, headache and anosmia were found to gain importance, along with newer themes around immunocompromised and chronic diseases. Cross infection related entities gained focus due to the second wave of COVID-19 cases in multiple countries around the world. The importance of mental health effects transitioned from lesser importance in the first quarter to more emerging and prominent links in the second quarter as highlighted

in the alluvial diagram.

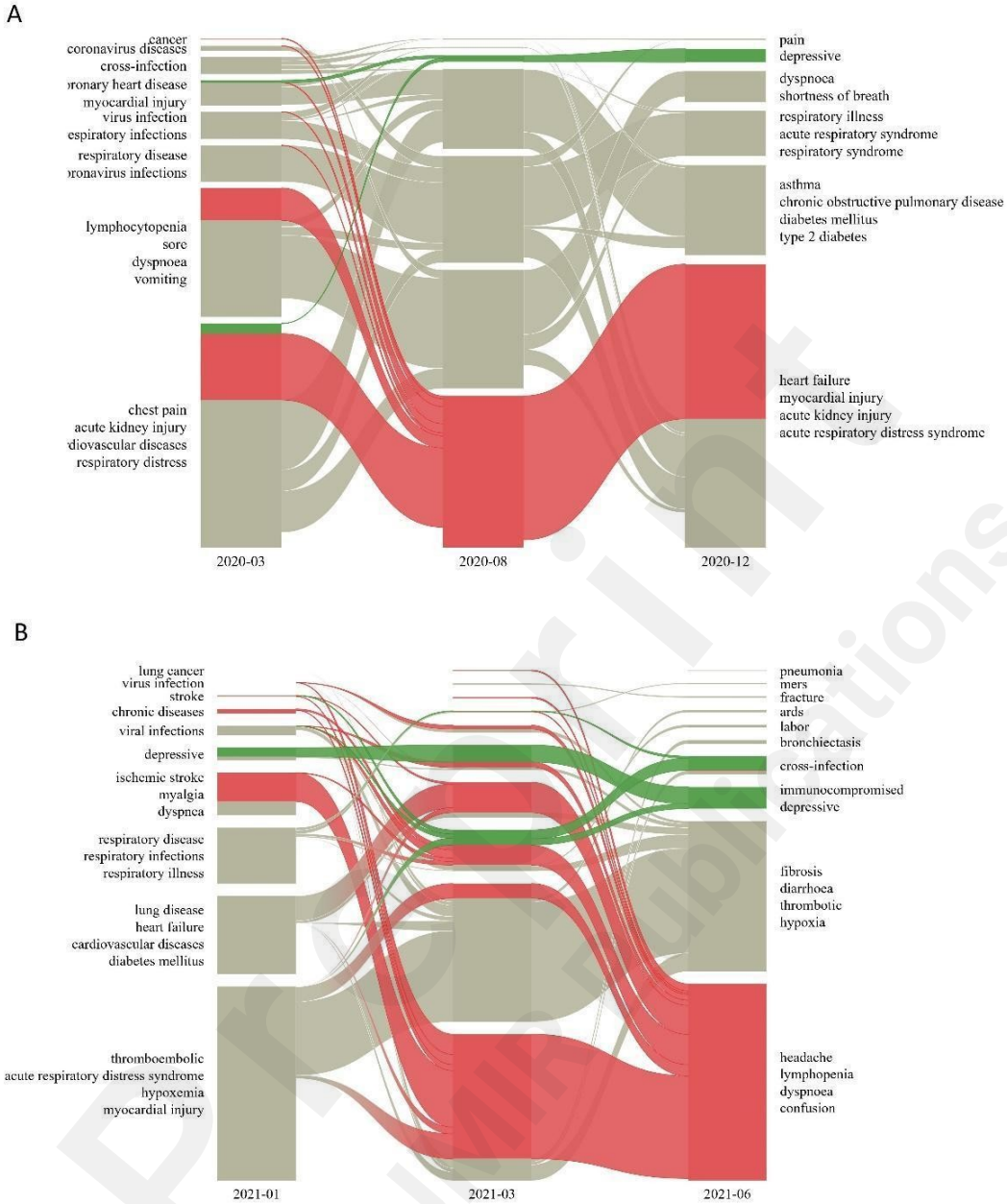


Figure 3: (A) Alluvial diagram for tracking the trends in 2020, from the networks of March, August and December. (B) Alluvial diagram for monitoring the trends in 2021, from the networks of January, March and June. The alluvial diagram eases tracing the temporal dynamics of literature across different time intervals.

We further advanced the analysis of trends to predicting links between entity pairs for the upcoming months. Our proposed framework for Temporal Link Prediction effectively forecasted five proximity scores, including semantic and topological measures, between node pairs by modeling its time series using the ARIMA model. Mean Squared Error in the predictions of each proximity score for April 2021, May 2021, and June 2021 was shown in Figure 4A (Supplementary Table 2). The associations between diseases for the successive month were predicted as links using supervised learning based on dynamic networks belonging to the previous months. Our results show that among the four classifiers (Supplementary Table 3), the AdaBoost model with 50 estimators and learning rate of 0.1 classified links with a mean AUROC score of 0.871 (all P<.001, statistically significant at a Bonferroni-corrected significance level of .0167) in the test data of June 2021 (Figure 4B, 4C).

Comparison among other classifiers is shown in Supplementary Table 9. The predicted links weighted by forecasted cosine similarity showed a high intersection with the original modules, hence validating the proposed architecture (Table 1). Supplementary Table 4 shows the clusters detected in the original network vs the predicted network. The ARIMA model was used for forecasting proximity scores for subsequent months based on the trends in node pair proximity measures retrieved from the prior months (Feb 2020 - June 2020). Our findings suggest that themes of predisposing conditions and risk factors, studies on cross-infection and neuro-psychiatric manifestation will assume a higher centrality in the upcoming quarter of 2021 (Supplementary Table 5).

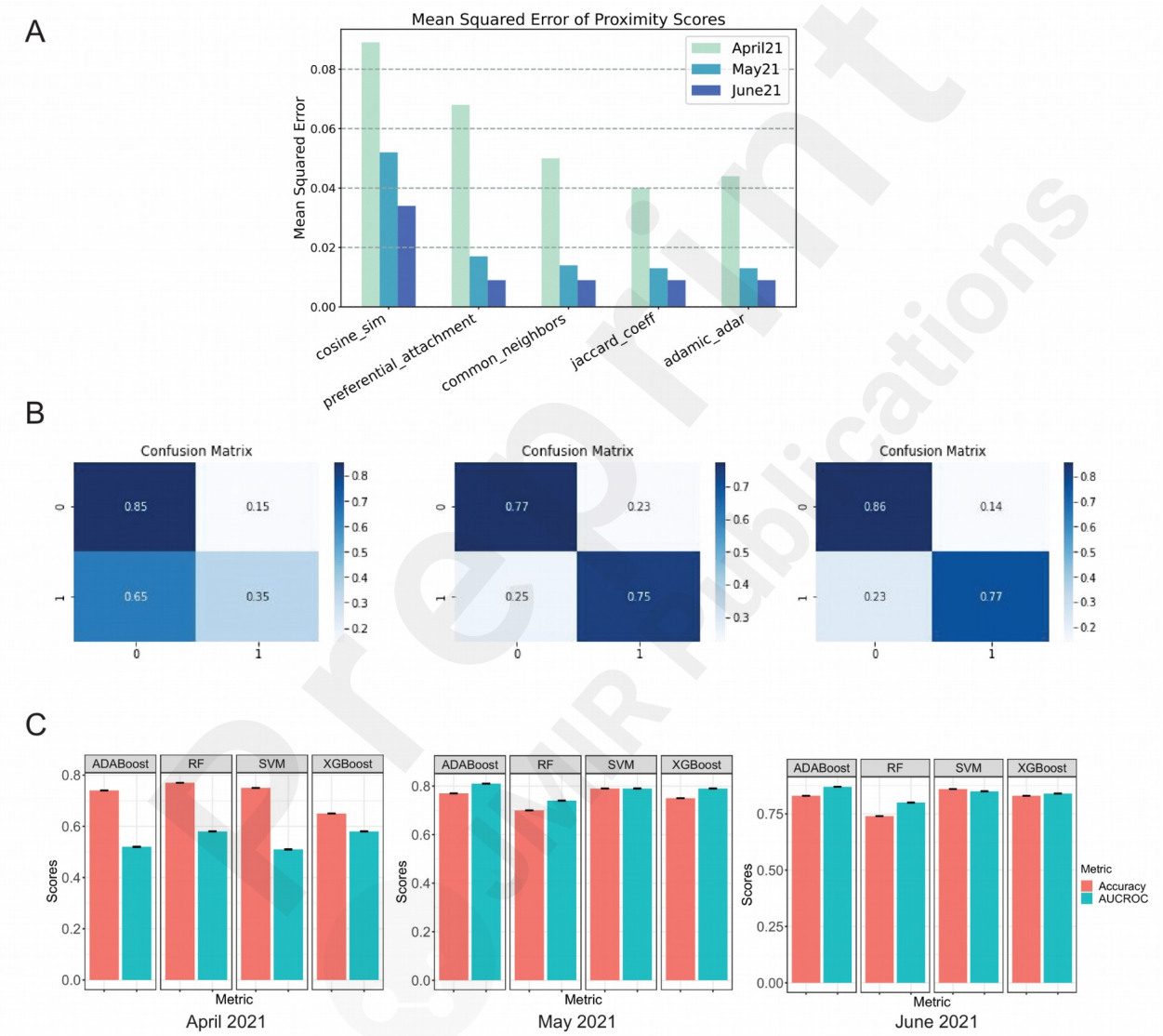


Figure 4: (A) Evaluation of Mean Squared Error (MSE) between original and predicted proximity scores for the network of April 2021, May 2021, June 2021. (B) Confusion Matrix with normalized values of results from AdaBoost classifier across the months of April 2021l, May 2021, and June 2021. AdaBoost has been the best-performing model across all three months. (C) Results of link prediction between disease entities from March 2021 to June 2021, with a margin of error for 95% Confidence Intervals. The mean value of metrics has been recorded by testing the models on a resampled test set.

The intersection of nodes between predicted and original modules was analyzed to prospectively validate the effectiveness of the proposed prediction framework. Table 1 depicts the top nodes in the different modules along with their respective IOU scores for January and

June 2021. The collection of intersecting nodes have been interpreted to represent broad themes. Organ damage like acute kidney injury and pulmonary embolism associated with COVID-19 was the most central theme in literature from January 2021, followed by cardiovascular diseases, respiratory infections and psychological effects. Interestingly, major themes in June 2021 shifted towards conditions related to Long COVID and neurological symptoms. Headache, encephalitis and confusion were predicted to be the central nodes, and showed a high IOU score when compared with the original network. Supplementary Figure 1B shows the percentage of articles published in June 2021 mentioning entities from each module for the predicted network. Supplementary Figure 1A demonstrates the same for true networks. A subset of nodes belonging to different modules from both predicted and true networks have been demonstrated in Supplementary Table 4.

Module ID	**January 2021**		**June 2021**	
	Top Nodes	***IOU***	***Top Nodes***	***IOU***
1	acute kidney injury, ARDS, coagulopathy, myocardial injury, pulmonary embolism	0.45	headache, lymphopenia, dyspnea, confusion, encephalitis, nausea	0.71
2	cardiovascular disease, diabetes mellitus, COPD, hypertension	0.66	fibrosis, coagulopathy, thrombotic, hypoxia, inflammation, delirium	0.70
3	respiratory infection, MERS, respiratory diseases	0.55	comorbidity, asthma, COPD, hypertension, dementia, diabetes	0.64
4	depression, insomnia, anxiety, loneliness	0.71	traumatic, anxiety, depression, loneliness, burnout, insomnia	0.81
5	myalgia, lymphopenia, headache, anosmia, dyspnoea	0.43	immunocompromised, chronic diseases like tuberculosis	0.33

Table 1: Clusters or Modules of diseases from the predicted network of January 2021 and June 2021. The given Intersection over Union (IOU) was computed between clusters of predicted and original networks of the respective months. A subset of top intersecting nodes in each cluster is mentioned, which collectively signify themes.

Analysis of networks constructed upon chemical entities revealed the evolution of various drugs studied in the COVID-19 literature. During February 2020, the major module contained entities such as paracetamol, tofacitinib, thalidomide, vitamins, zinc and other linked chemicals. Another relevant module included central entities such as doxycycline, ruxolitinib, heparin and ivermectin, which were discussed in the scientific research on treatment and prevention of COVID-19. In contrast, our recently updated models showed the emergence of evidence for various immunosuppressive drugs such as Tacrolimus and anti-inflammatory drugs such as Glucocorticoids and Colchicine during November 2021 (Supplementary Figure 3). These relatively less important entities in earlier months started to become more prominent as the literature expanded. Evidence around ‘statins’ also gained centrality over recent months. Our findings show that the proposed framework captures the dynamic changes in the importance of entities based on their evolving relationship with neighboring entities.

Discussion

Principal Findings

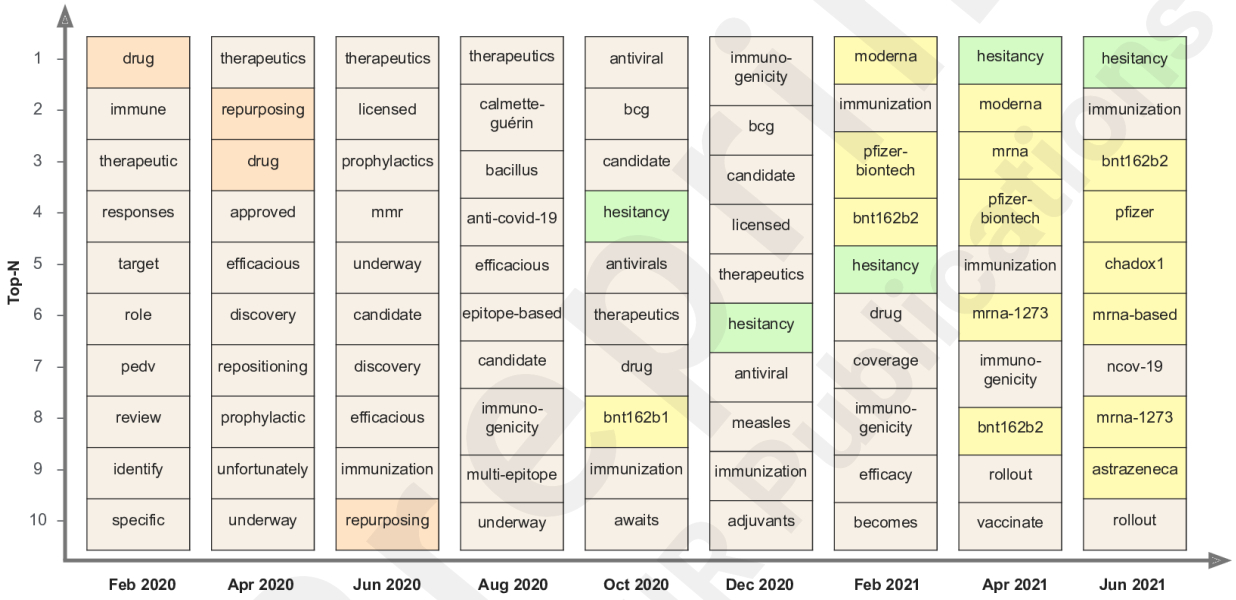
In this paper, we demonstrate a computational approach, EvidenceFlow, in which a user will interact with the rapidly expanding COVID-19 literature to derive and predict emerging themes. The proposed framework tracks patterns of changing semantic and topological proximity between entity pairs across months. Further, it predicts links and network communities that may emerge in the future months. Hence users can follow the papers that contribute to emerging communities of themes, e.g. literature around thromboembolic complications which was captured as early as August 2020 and mental health factors during the end of 2020. Interacting with the clusters on the interactive interface of the EvidenceFlow model revealed that symptoms of Long-COVID, such as fatigue, headache, myalgia, cough, and anosmia, were found to be forming a central cluster during March 2021. This early signal for accumulating evidence was later validated in large prospective and retrospective cohorts of COVID-19 patients[30–32]. Another way in which users can interact with EvidenceFlow is to gain an understanding of the evolution of themes that goes beyond the current approaches such as topic modeling and sentiment tracking[3]. An example is the early finding of imminent themes around neurological complications, such as confusion, psychiatric illness, stroke, and mental health factors such as anxiety, depression, PTSD, burnout and insomnia, in June 2021. Our violin plot analysis (Supplementary Figure 5) showed that despite the mean error being centered on zero, there were some outlier node pairs whose predicted associations deviated from the ground truth. Future scope of this work will involve an analysis of such associations and insights gained by an interactive analysis of such pairs on the *EvidenceFlow* application.

Prediction of the themes represented by rising centrality of entities can assist in formation of promising research hypotheses. The dynamics of literature reveal the emergence of central themes as a combination of pre-existing themes in recent times[6]. For example, the alluvial diagram (Figure 3A) demonstrated how entities from multiple modules in March 2020 merged into a major cluster of thromboembolic complications. Similarly, the flow of importance of psychological disorders over the months indicate their contemporary relevance in the COVID-19 literature and their links with other entities in the cluster. Our framework can potentially help researchers in monitoring the existing themes and directing their studies based on the trends and predictions.

We conducted an analysis on the trends of PageRank centrality of selected chemical and disease entities. Statins, a class of lipid-lowering medications, were found to be gaining centrality in late 2021 as compared to earlier values (Supplementary Figure 4A). Numerous studies discussed statins for having anti-inflammatory and immunomodulatory effects that may reduce the severity of COVID-19[33,34]. Glucocorticoids, a class of steroid hormones that reduce inflammation and suppress the immune system, also emerged as an entity with rising entity (Supplementary Figure 4B). Depression and other mental health disorders started becoming a prominent topic of research during the middle of 2020 and gained higher importance in subsequent months (Supplementary Figure 4C). COVID-19 has also been largely discussed in the context of a thromboembolism and our model captured its emerging evidence as a theme till late 2020. However, the trends showed that its centrality in the literature relatively decreased in 2021 (Supplementary Figure 4D). Discovering such trends from a large corpus is indeed possible using manual curation and analysis by experts. However, our EvidenceFlow pipeline provides an efficient lens to discover, track and predict emerging trends. This framework will enable faster synthesis of evidence, which then can be validated by experts.

To explore the potential of unsupervised word embeddings and changing cosine similarity among words, we analyzed the trends of terms having maximum similarity with selected

keywords. For example, we analyzed the temporal shift in the context of "vaccine" over the months by finding the top-10 terms most similar to *vaccine* in the latent space of word embeddings trained on the abstracts from each month (Figure 5). From February to August 2020, the research on COVID-19 vaccines was underway and the studies revolved around "therapeutics", "prophylactics", "drug-repurposing" and associations with MMR (measles-mumps-rubella) vaccine and BCG (Bacillus Calmette–Guérin) vaccine. As the clinical trials of certain vaccine candidates became prominent after August 2020, a theme of vaccine *hesitancy* emerged in October 2020 and gained higher similarity in subsequent months. Additionally, as the literature evolved in 2021, a wide range of COVID-19 vaccines such as BNT162b1, Pfizer-BioNTech, AstraZeneca, ChAdOx1, mRNA-1273 or Moderna were found to be majorly discussed in the context of research on vaccines. Terms such as *immunogenicity* and *efficacy* further suggested high association with vaccine trials and rollouts. Recently updated models showed the emergence of 'booster' doses from August 2021 onwards. Such retrospective evaluation of development of evidence from literature over time can assist the research community in



deriving detailed insights leveraging the applications of word embeddings.

Figure 5: Temporal evolution of the context of the term "vaccine" across alternate months. Top-10 most similar words based on cosine similarity using monthly Word2Vec embeddings are plotted. Origin and evolution of drug repurposing in early months, hesitancy and vaccine candidates in later months are highlighted.

Limitations

Our study has some limitations. Firstly, although the WHO database is built using a detailed search strategy for COVID-19 literature, it does not explicitly report the exact purpose or accuracy of the search and decision process. The documentation [35] mentions screening done by expert reviewers and an attempt to remove duplicates, but further details are lacking. For example, the process doesn't clarify if redundancy across various publishers was taken care of. Further, the frequent use of 'OR' combination of keywords may have led to inclusion of less relevant articles, while other forms of literature, such as patent applications, which can add value to the study were not included in this database. Nonetheless, we chose the WHO COVID-19 database as it provides a large collection of articles that are updated regularly from searches of multiple bibliographic databases [2]. This, combined with curated expert-referred scientific articles which wouldn't be readily accessible on a custom search, was useful for building the

EvidenceFlow pipeline. Future work with this framework will include potential extension to databases curated both through generic queries and expert vetting, thus facilitating targeted evidence synthesis from a variety of databases.
Further, we are currently using abstracts of research articles to extract named entities and may be missing on the details contained in the full-text of the article while training word embeddings. Therefore, future work may build upon the framework to include the full text of articles and full text, wherever available. The NER model used in our study has been reported to achieve an F1 score of 84.49% on a benchmark dataset[13]. Despite the limitations of F1 score such as equal weightage given to precision and recall[36,37], F1 remains one of the most widely reported performance indicators. We chose this metric in the absence of other metrics reported for this NER model. For forecasting, we utilized a relatively basic model such as an AR approach as our goal was to capture robust patterns, however, further research is possible for the use of more complex time-series approaches with higher order difference and lags. Moreover, as the number of timestamps and data points increase, advanced architectures such as RNN and LSTM[38,39] can be used for handling complex trends in the time series efficiently. Further experiments with larger networks can reveal themes that were not found with top-100 entities. Importantly, our model is supporting early detection of emerging trends, but it cannot capture themes on which no evidence has been accumulating.

Conclusion

Consortia across the globe were formed for the advancement of research related to COVID-19. The global attention has led to widespread increase in the scientific literature to study and prevent the disease from spreading, hence understanding the disease from multiple perspectives. We introduced a framework built upon COVID-19 specific literature vetted by the WHO and deployed as a dashboard called EvidenceFlow[10]. The dashboard allows the user to unravel the literature with an interactive map of embeddings based on the visualization provided by Tensorboard. It aims to track literature trends using alluvial diagrams, multi-level community detection, and projection of influential entities through network analysis across different months. This study presented how machine learning-based prediction of emerging links can contribute towards analyzing research by capturing themes represented by groups of medical entities, based on patterns of semantic relationships over time.

Acknowledgments

We acknowledge the support from the Center of Excellence in Healthcare and the Center of Excellence in Artificial Intelligence at IIIT-Delhi.

Conflict of Interest

None declared.

Author Contributions

R.P. and H.C. designed and implemented the computational framework, interpreted the results, and wrote the paper. H.B. contributed to writing and created the associated dashboard. R.A. and A.N. interpreted the results and provided feedback on statistical methods. T.S. designed the study, analyzed the results, and contributed to writing. All authors read and approved the final paper.

References

1. Coronavirus Disease (COVID-19) Situation Reports [Internet]. [cited 2022 Jan 19]. Available from: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports
2. Global research on coronavirus disease (COVID-19) [Internet]. [cited 2022 Jan 19]. Available from: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/global-research-on-novel-coronavirus-2019-ncov
3. Ebadi A, Xi P, Tremblay S, Spencer B, Pall R, Wong A. Understanding the temporal evolution of COVID-19 research through machine learning and natural language processing. Scientometrics 2021 Jan 1;126(1):725–739. [doi: 10.1007/s11192-020-03744-7]
4. Cho H, Lee H. Biomedical named entity recognition using deep neural networks with contextual information. BMC Bioinformatics 2019 Dec 27;20(1):735. [doi: 10.1186/s12859-019-3321-4]
5. Kutuzov A, Øvrelid L, Szymanski T, Velldal E. Diachronic word embeddings and semantic shifts: a survey. Proc 27th Int Conf Comput Linguist [Internet] Santa Fe, New Mexico, USA: Association for Computational Linguistics; 2018 [cited 2022 Jan 19]. p. 1384–1397. Available from: https://aclanthology.org/C18-1117
6. Tshitoyan V, Dagdelen J, Weston L, Dunn A, Rong Z, Kononova O, Persson KA, Ceder G, Jain A. Unsupervised word embeddings capture latent knowledge from materials science literature. Nature 2019 Jul;571(7763):95–98. [doi: 10.1038/s41586-019-1335-8]
7. Bu Z, Wang Y, Li H-J, Jiang J, Wu Z, Cao J. Link prediction in temporal networks: Integrating survival analysis and game theory. Inf Sci 2019 Sep 1;498:41–61. [doi: 10.1016/j.ins.2019.05.050]
8. Özcan A, Öğüdücü ŞG. Supervised temporal link prediction using time series of similarity measures. 2017 Ninth Int Conf Ubiquitous Future Netw ICUFN 2017. p. 519–521. [doi: 10.1109/ICUFN.2017.7993838]
9. Güneş İ, Gündüz-Öğüdücü Ş, Çataltepe Z. Link prediction using time series of neighborhood-based node similarity scores. Data Min Knowl Discov 2016 Jan 1;30(1):147–180. [doi: 10.1007/s10618-015-0407-0]
10. EvidenceFlow [Internet]. EvidenceFlow. [cited 2022 Jan 19]. Available from: https://evidenceflow.tavlab.iiitd.edu.in/index
11. Bird S, Loper E. NLTK: The Natural Language Toolkit. Proc ACL Interact Poster Demonstr Sess [Internet] Barcelona, Spain: Association for Computational Linguistics; 2004 [cited 2022 Jan 19]. p. 214–217. Available from: https://aclanthology.org/P04-3031
12. Neumann M, King D, Beltagy I, Ammar W. ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing. Proc 18th BioNLP Workshop Shar Task 2019;319–327. [doi: 10.18653/v1/W19-5034]
13. scispacy [Internet]. scispacy. [cited 2022 Jan 19]. Available from: https://allenai.github.io/scispacy/
14. Ma L, Zhang Y. Using Word2Vec to process big text data. 2015 IEEE Int Conf Big Data Big Data 2015. p. 2895–2897. [doi: 10.1109/BigData.2015.7364114]
15. Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J. Distributed Representations of Words and Phrases and their Compositionality. Adv Neural Inf Process Syst [Internet] Curran Associates, Inc.; 2013 [cited 2022 Jan 19]. Available from: https://papers.nips.cc/paper/2013/hash/9aa42b31882ec039965f3c4923ce901b-Abstract.html
16. Rehurek R, Sojka P. Gensim–python framework for vector space modelling. NLP Cent Fac Inform Masaryk Univ Brno Czech Repub 2011;3(2).
17. Smilkov D, Thorat N, Nicholson C, Reif E, Viégas FB, Wattenberg M. Embedding projector:

Interactive visualization and interpretation of embeddings. ArXiv Prepr ArXiv161105469 2016;
18. Bohlin L, Edler D, Lancichinetti A, Rosvall M. Community detection and visualization of networks with the map equation framework. Meas Sch Impact Springer; 2014. p. 3–34.
19. Rosvall M, Bergstrom CT. Mapping Change in Large Networks. PLOS ONE Public Library of Science; 2010 Jan 27;5(1):e8694. [doi: 10.1371/journal.pone.0008694]
20. Barabasi A-L, Albert R. Emergence of scaling in random networks. Science 1999 Oct 15;286(5439):509–512. [doi: 10.1126/science.286.5439.509]
21. Adamic LA, Adar E. Friends and neighbors on the Web. Soc Netw 2003 Jul 1;25(3):211–230. [doi: 10.1016/S0378-8733(03)00009-1]
22. Hagberg A, Swart P, S Chult D. Exploring network structure, dynamics, and function using networkx [Internet]. Los Alamos National Lab. (LANL), Los Alamos, NM (United States); 2008 Jan. Report No.: LA-UR-08-05495; LA-UR-08-5495. Available from: https://www.osti.gov/biblio/960616
23. Zhang GP. Time series forecasting using a hybrid ARIMA and neural network model. Neurocomputing 2003 Jan 1;50:159–175. [doi: 10.1016/S0925-2312(01)00702-0]
24. Wright RE. Logistic regression. American Psychological Association; 1995;
25. Breiman L. Random Forests. Mach Learn 2001 Oct 1;45(1):5–32. [doi: 10.1023/A:1010933404324]
26. Hearst MA, Dumais ST, Osuna E, Platt J, Scholkopf B. Support vector machines. IEEE Intell Syst Their Appl 1998 Jul;13(4):18–28. [doi: 10.1109/5254.708428]
27. Freund Y, Schapire RE. A Short Introduction to Boosting. Proc Sixt Int Jt Conf Artif Intell Morgan Kaufmann; 1999. p. 1401–1406.
28. Chen T, He T, Benesty M, Khotilovich V, Tang Y, Cho H. Xgboost: extreme gradient boosting. R Package Version 04-2 2015;1(4):1–4.
29. Etymologia: Bonferroni correction. Emerg Infect Dis 2015 Feb;21(2):289. PMID:25786274
30. Taquet M, Dercon Q, Luciano S, Geddes JR, Husain M, Harrison PJ. Incidence, co-occurrence, and evolution of long-COVID features: A 6-month retrospective cohort study of 273,618 survivors of COVID-19. PLOS Med Public Library of Science; 2021 Sep 28;18(9):e1003773. [doi: 10.1371/journal.pmed.1003773]
31. López-León S, Wegman-Ostrosky T, Perelman C, Sepulveda R, Rebolledo PA, Cuapio A, Villapol S. More than 50 Long-Term Effects of COVID-19: A Systematic Review and Meta-Analysis [Internet]. Rochester, NY: Social Science Research Network; 2021 Jan. Report No.: ID 3769978. [doi: 10.2139/ssrn.3769978]
32. Blomberg B, Mohn KG-I, Brokstad KA, Zhou F, Linchausen DW, Hansen B-A, Lartey S, Onyango TB, Kuwelker K, Sævik M, Bartsch H, Tøndel C, Kittang BR, Bergen COVID-19 Research Group, Cox RJ, Langeland N. Long COVID in a prospective cohort of home-isolated patients. Nat Med 2021 Sep;27(9):1607–1613. PMID:34163090
33. Daniels LB, Ren J, Kumar K, Bui QM, Zhang J, Zhang X, Sawan MA, Eisen H, Longhurst CA, Messer K. Relation of prior statin and anti-hypertensive use to severity of disease among patients hospitalized with COVID-19: Findings from the American Heart Association's COVID-19 Cardiovascular Disease Registry. PLOS ONE Public Library of Science; 2021 Jul 15;16(7):e0254635. [doi: 10.1371/journal.pone.0254635]
34. Peymani P, Dehesh T, Aligolighasemabadi F, Sadeghdoust M, Kotfis K, Ahmadi M, Mehrbod P, Iranpour P, Dastghaib S, Nasimian A, Ravandi A, Kidane B, Ahmed N, Sharma P, Shojaei S, Bagheri Lankarani K, Madej A, Rezaei N, Madrakian T, Los MJ, Labouta HI, Mokarram P, Ghavami S. Statins in patients with COVID-19: a retrospective cohort study in Iranian COVID-19 patients. Transl Med Commun 2021 Jan 25;6(1):3. [doi: 10.1186/s41231-021-00082-5]

35. who-covid-19_sources_searchstrategy_20211012.pdf [Internet]. [cited 2022 Jan 23]. Available from: https://www.who.int/docs/default-source/coronaviruse/who-covid-19-database/who-covid-19_sources_searchstrategy_20211012.pdf
36. Hand D, Christen P. A note on using the F-measure for evaluating record linkage algorithms. Stat Comput 2018 May 1;28(3):539–547. [doi: 10.1007/s11222-017-9746-6]
37. Powers DMW. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. ArXiv201016061 Cs Stat [Internet] 2020 Oct 10 [cited 2022 Jan 19]; Available from: http://arxiv.org/abs/2010.16061
38. Sherstinsky A. Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) network. Phys Nonlinear Phenom 2020 Mar 1;404:132306. [doi: 10.1016/j.physd.2019.132306]
39. Greff K, Srivastava RK, Koutník J, Steunebrink BR, Schmidhuber J. LSTM: A Search Space Odyssey. IEEE Trans Neural Netw Learn Syst 2017 Oct;28(10):2222–2232. [doi: 10.1109/TNNLS.2016.2582924]

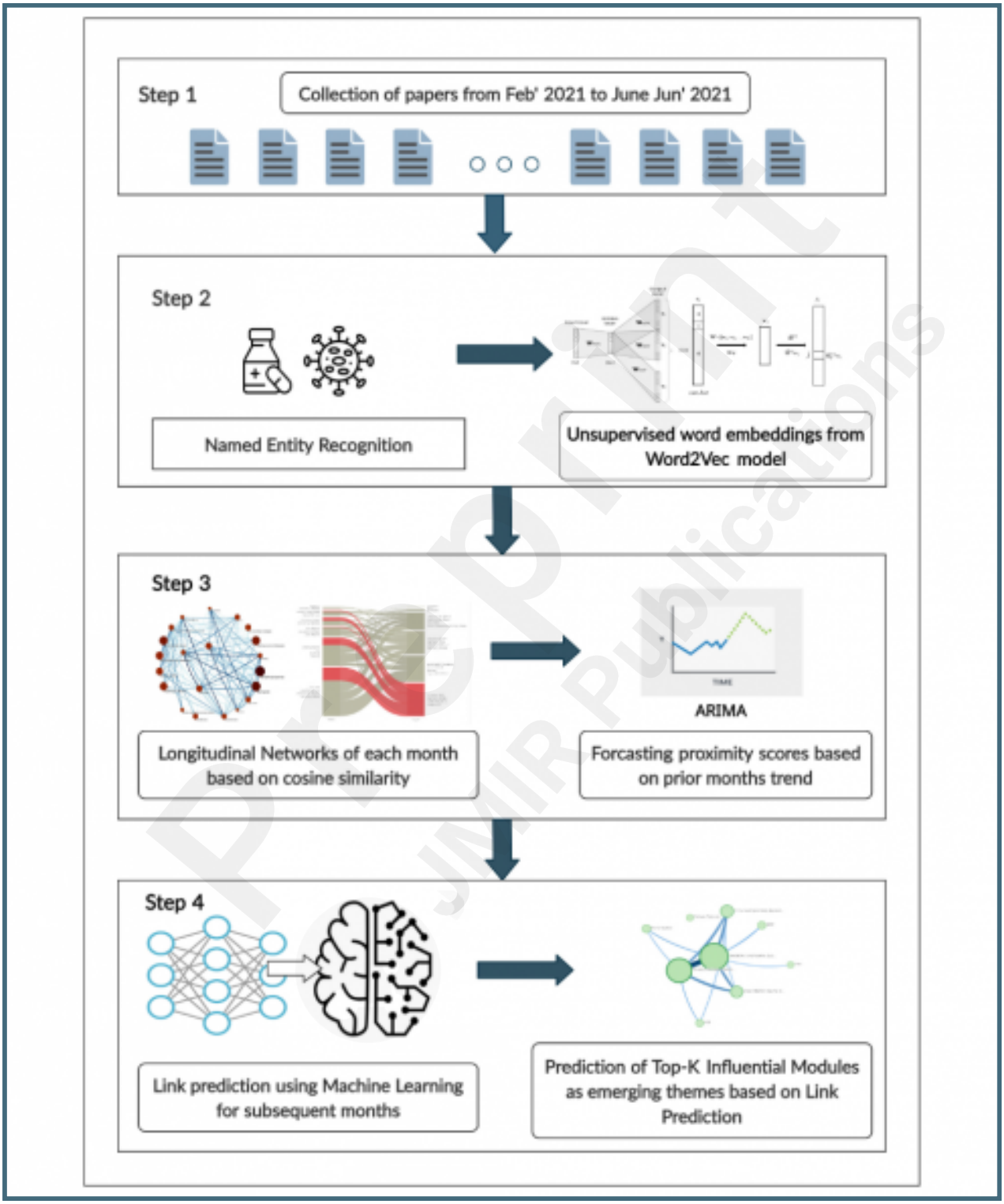
Abbreviations

AUROC: Area Under Receiver Operator Curve
IOU: Intersection Over Union
ARIMA: Autoregressive Integrated Moving Average
NER: Named Entity Recognition
MSE: Mean Squared Error

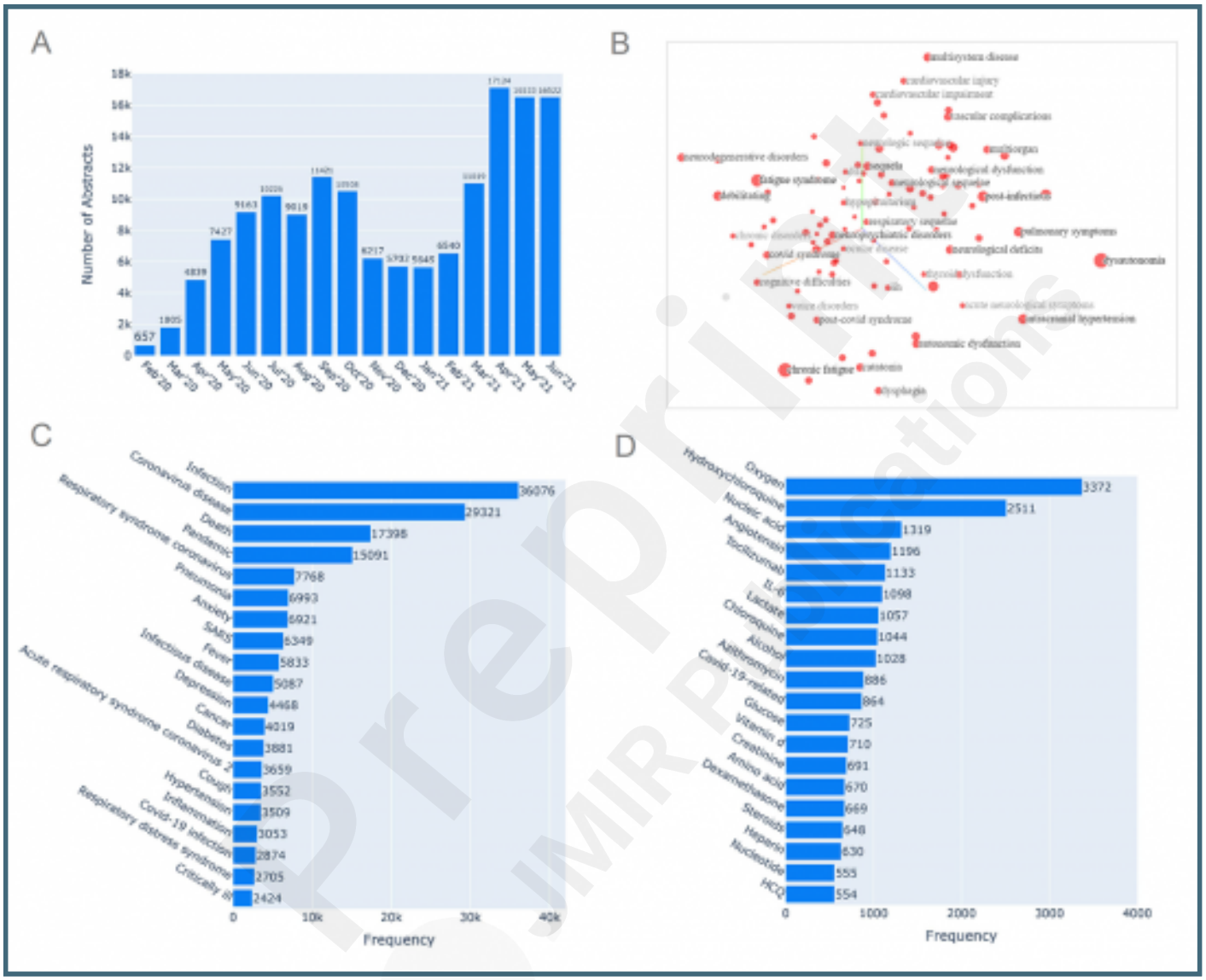
Supplementary Files

Figures

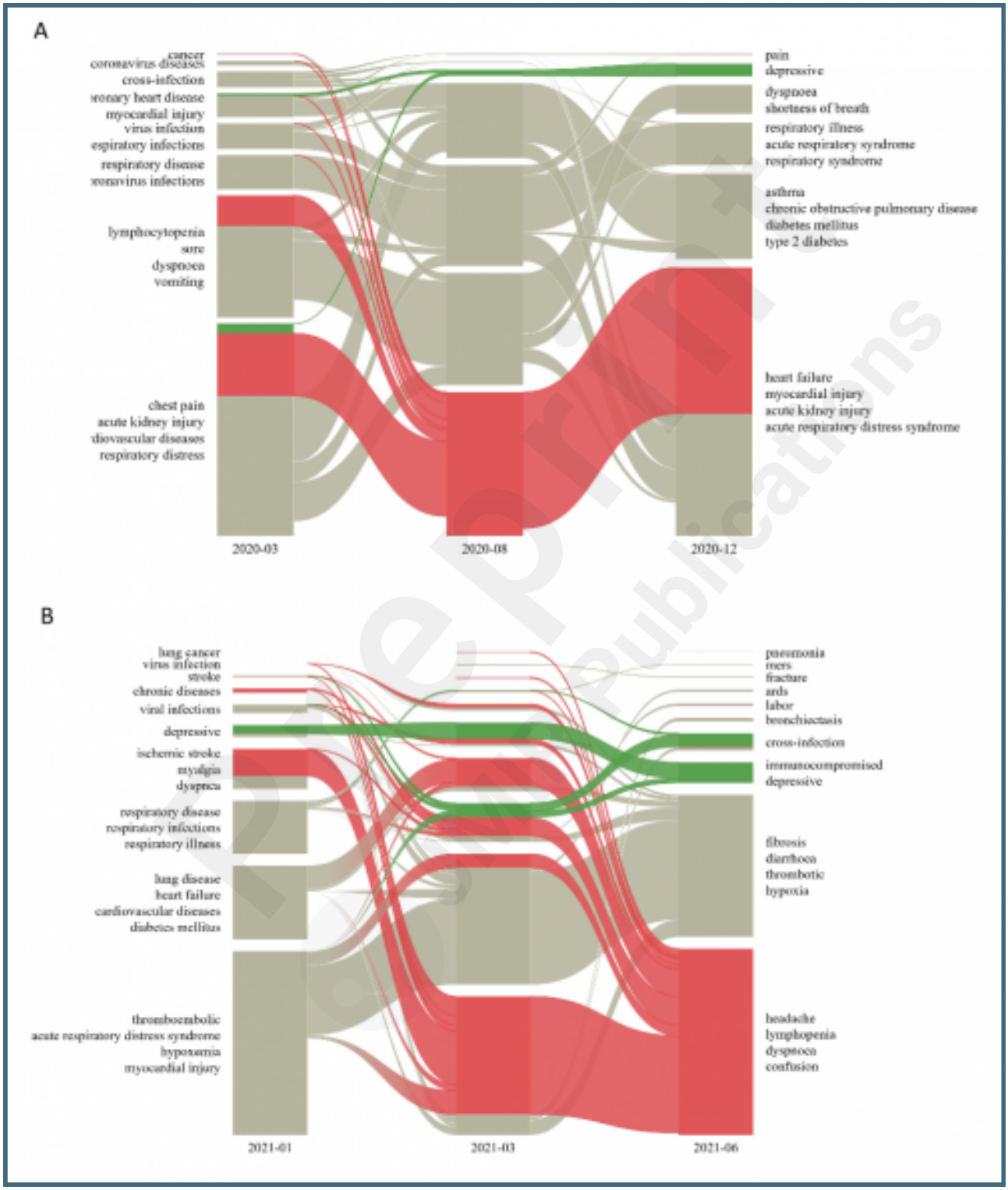
Graphical representation of proposed framework explaining the complete workflow. The pipeline takes abstracts as input from which entities are extracted using NER. Embeddings are generated, which are used as features for longitudinal networks. These networks are used for visualizing the trends using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



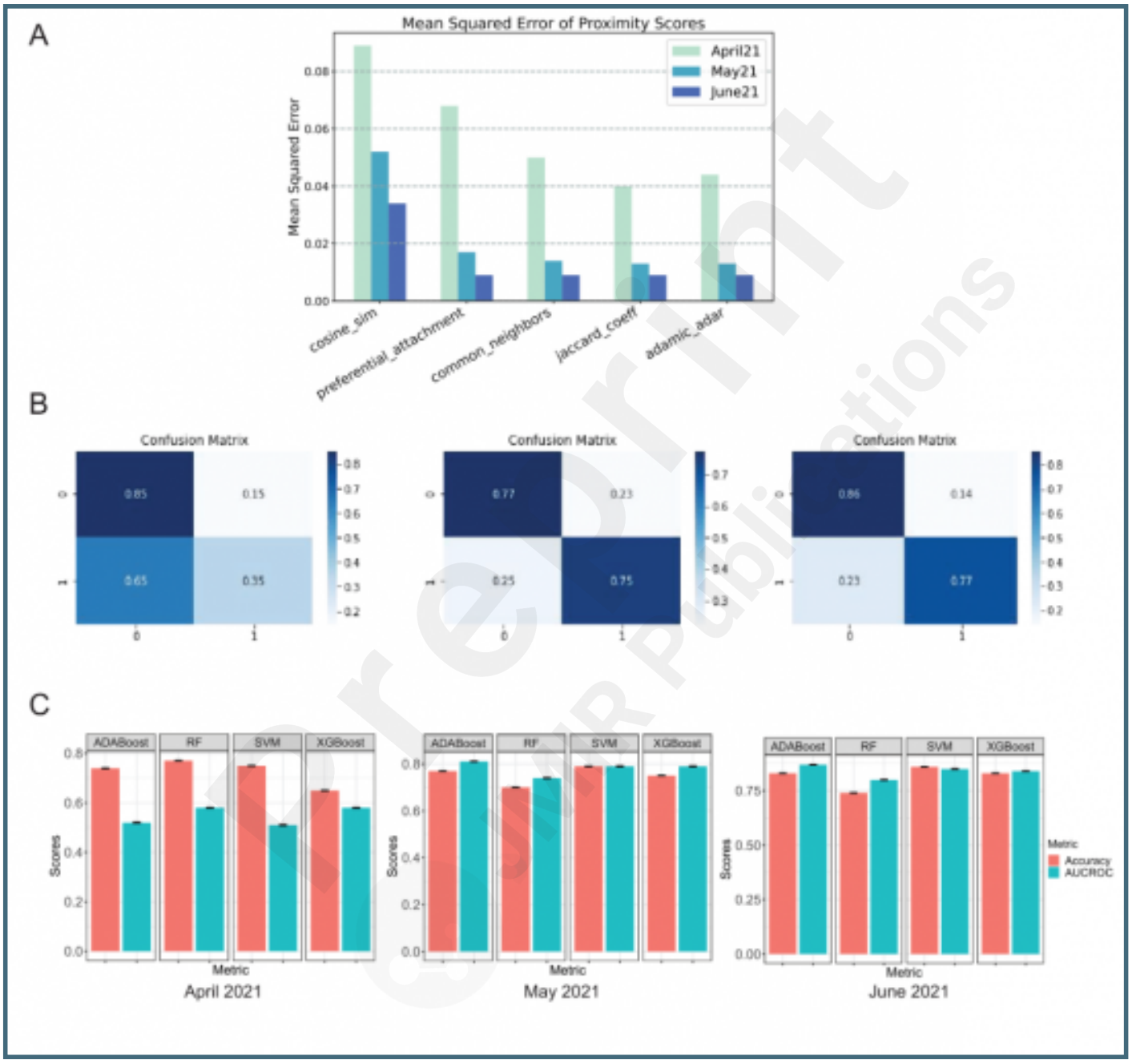
(A) Showing the number of articles occurring each month. The curve depicts that there has been a rampant increase in the number of articles across each month since February 2020. (B) Latent space of word embeddings of diseases visualized around the keyword 'post-covid syndrome', displaying 100 isolated points nearest to it. (C) Bar plot (left) showing the frequency of top diseases in the corpus of abstracts extracted using Named Entity Recognition. (D) Bar plot (right) showing the frequency of top chemicals in the corpus of abstracts extracted using Named Entity Recognition.



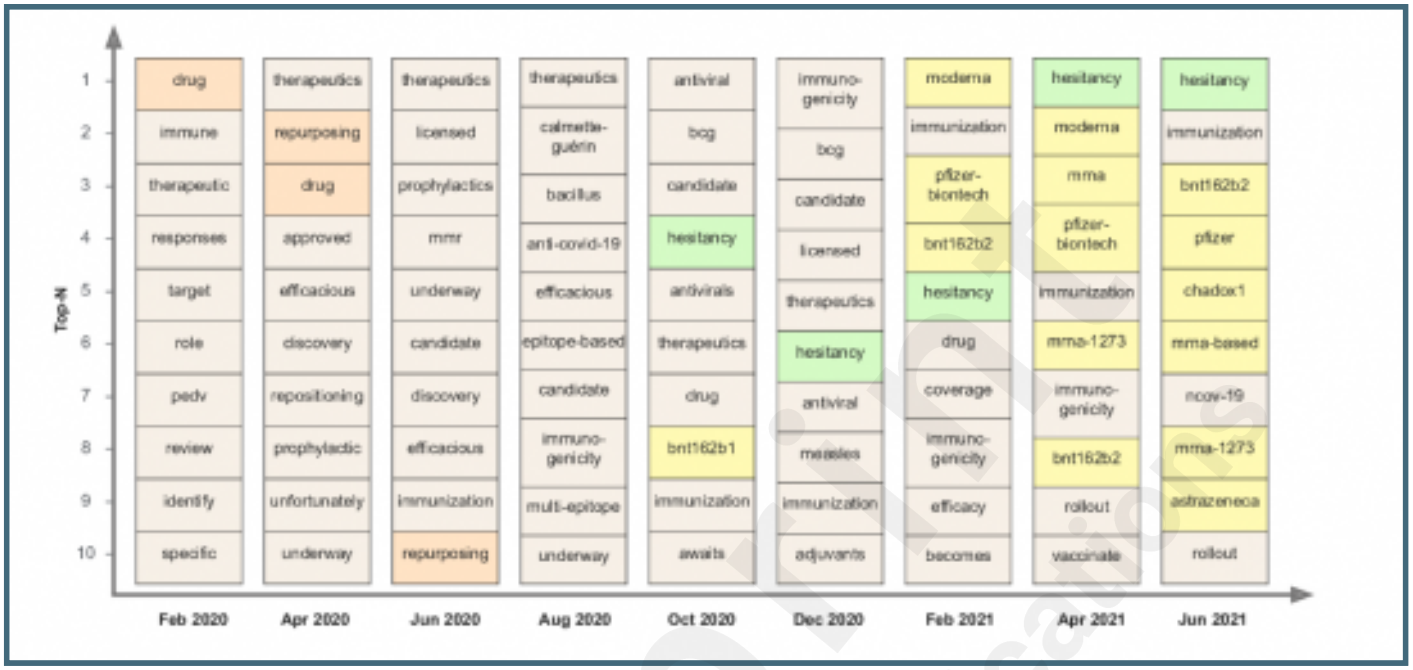
(A) Alluvial diagram for tracking the trends in 2020, from the networks of March, August and December. (B) Alluvial diagram for monitoring the trends in 2021, from the networks of January, March and June. The alluvial diagram eases tracing the temporal dynamics of literature across different time intervals.



(A) Evaluation of Mean Squared Error (MSE) between original and predicted proximity scores for the network of April 2021, May 2021, June 2021. (B) Confusion Matrix with normalized values of results from AdaBoost classifier across the months of April 2021l, May 2021, and June 2021. AdaBoost has been the best-performing model across all three months. (C) Results of link prediction between disease entities from March 2021 to June 2021, with a margin of error for 95% Confidence Intervals. The mean value of metrics has been recorded by testing the models on a resampled test set.



Temporal evolution of the context of the term “vaccine” across alternate months. Top-10 most similar words based on cosine similarity using monthly Word2Vec embeddings are plotted. Origin and evolution of drug repurposing in early months, hesitancy and vaccine candidates in later months are highlighted.



Multimedia Appendixes

Supplementary Material.
URL: http://asset.jmir.pub/assets/57c5573b9d2967e89949879ffcb3d3d1.docx