An Exploratory Study on Temporally Evolving Discussion around COVID-19 using Diachronic Word Embeddings

Avinash Tulasi
IIIT Delhi
NII Japan *
avinasht@iiitd.ac.in

Asanobu Kitamoto
NII Japan
kitamoto@nii.ac.jp

Arun Buduru
IIIT Delhi
arunb@iiitd.ac.in

Ponnurangam Kumaraguru
IIIT Hyderabad
pk.guru@iiit.ac.in

Abstract

COVID-19 has seen the world go into a lockdown and unconventional social situations throughout. During this time, the world saw a surge in information sharing around the pandemic, and the topics shared in the time were diverse. People's sentiments have changed during this period. Given the widespread usage of Online Social Networks (OSN) and support groups, the user sentiment is well reflected in online discussions. In this work, we aim to show the topics under discussion, evolution of discussions, change in user sentiment during the pandemic. We also demonstrate the possibility of exploratory analysis to find pressing topics, change in perception towards the topics, and ways to use the knowledge extracted from online discussions. For our work, we employ Diachronic Word embeddings which capture the change in word usage over time. With the help of the analysis from temporal word usages, we show the change in people's opinion on COVID from being a "conspiracy", to the post-COVID topics surrounding "vaccination".

1 Introduction

Analysing and estimating the impact of changes in social lives of people from time to time is key in digital humanities. With the advent of Online Social Networks (OSNs) it has become easy to access what people have been sharing their lives online, good or bad and seeking support in the online communities. During the pandemic, people seek help and they have also shared information with one another. To understand how people's lives were affected, and to contribute towards digital humanities, we look at conversations on COVID-19 during the pandemic. We listed the wide variety of topics discussed, and the support from each other when they were stuck with the illness.

COVID-19 pandemic has been a huge disaster for humanity. It has changed the way we live forever during the ensued lockdown from the pandemic. We saw a lot of community support and huge participation in supporting each other during the testing times. Similarly, COVID-19 is the first known global pandemic in the Internet age, when the information disposal is rapid and the ability to communicate anywhere on the earth takes seconds. It is of immense significance to see how people communicated around coordinating and how people supported each other during the times, as we can assess the ability of the world wide web in effect.

The information revolution in the form of worldwide web has also come with its own challenges like fake news, conspiracy theories, hate speech, etc. It is of particular interest to understand how the conspiracies and the narrative around COVID-19 has changed ever since the pandemic broke. For example, at the beginning of the pandemic, there has been a lot of speculation that COVID-19 is a hoax, and the virus is just political propaganda. Such narratives can be found around the world, including the United States, India, China, and Japan. However, it did not take long for people to realize that COVID-19 is, in fact, a pandemic. So, during the same period, people started discussing online about the pandemic and the narrative around conspiracy theories. In this work, we aim to understand how the topics involved in COVID-19 have changed since the inception of the pandemic till date.

For this work, we use a data set taken from Reddit, which is a popular social network built around smaller communities called subreddits. Particularly, we choose the subreddit r/COVID19positive [1]. Our data set contains all posts and comments made since the inception of the subreddit, and we use the data to understand the narratives around COVID-19 on how people's perception has

* The work was done as a part of an online internship program with NII Japan.

[1] https://www.reddit.com/r/COVID19positive/

Proceedings of the Workshop on Natural Language Processing for Digital Humanities (NLP4DH), pages 183–190
co-located with ICON-2021, December 16-19, 2021, Silchar, India.
https://doi.org/10.26615/978-952-94-5833-2_021

changed. Specifically, we aim to assess the change in word usages on the basis of proximity among different words in the initial stages of the pandemic till date. The proximity of words and word usage is of importance because, togetherness of words gives us an idea on how people are thinking and what people are talking about. So, to estimate the word usage and proximity, we use robust Machine Learning methods such as Diachronic word embeddings (Kutuzov et al., 2018).

The Diachronic word embeddings (Montariol, 2021) take different embedding models trained on text corpora which are linked together in time. The text corpora are expected to evolve over time, and the temporal usage of words is captured by taking smaller snapshots of the corpus. To employ the technique, we train individual models and then align the models in such a way that the same words stay in relatively similar positions across time; thereby making the change in word embeddings relative. By closely looking at snapshots and the drift of words among snapshots, we can estimate the change in word usage. Multiple works have studied and seen a change in word usage throughout human history with the lens of corpora.

COVID-19 being two years old, we aim to do a word evolution estimation within these two years for our work. We split our corpus into month-long sub-corpora and train the embedding models on these individual corpora. Later, we align the corpora to maintain the word similarity across time. Once we align the word similarity across time, we estimate the divergence of words or drift from each other. The divergence and drift give us information on how the words have changed. Some interesting observations from our work include the usage of 'conspiracy'; at the beginning of the pandemic, the word and COVID-19 are closely related, and as time progresses, the word conspiracy disappears from the neighborhood of COVID. Similarly, COVID-19 is known for its symptoms such as cough and cold. During the initial phase, these words are seen away from 'COVID' or 'virus'. As time progresses and humanity goes into waves of huge infections, the words come together.

Humanity was lucky enough to produce vaccines in a short time, a narrative around vaccines and their evolution is also interesting to look at. Although a lot of vaccine-related discussions were taking place in the initial days of COVID, the usage of vaccines along with infection has increased recently (as of November - 2021). With our work, we also aim at developing a system that can be used by people with little to no technical knowledge of how word embeddings operate and obtain the information we extract.

In summary, our work focuses on :

1. Did narratives around COVID-19 and discussions around the pandemic have evolved over time? How?

2. The narrative around how getting 'vaccinated' and getting 'positive' are being used together? What does it imply for the community?

3. A system that can assist domain experts such as medical professionals, and journalists in accessing our findings in a visual form.

2 Related Work

In this section, we discuss the works that have studied Social Media websites during the COVID-19 crisis. In an early work, (Liu et al., 2020) the authors have collected Chinese articles about COVID-19 and reported the user sentiment around the pandemic. They have used News articles as the dataset. Similar works are seen in Brazil (de Melo and Figueiredo, 2021), that used Twitter and four European nations (Ghasiya and Okamura, 2021) based on news articles. We see a common pattern in the works which is to collect a publicly available dataset, use a language model and report the findings along side the real-world happenings.

Another widely observed theme in literature is to extract medical information from publicly available articles. The works (Murray et al., 2020; Sarker and Ge, 2021; Wu et al., 2021; Kumar et al., 2021) attempt at extracting medical information such as symptoms, post-COVID medical status, topics around vaccination using OSN data. The OSN of choice in the majority of these works is Reddit. We place our work at an intersection of the above-mentioned literature. We note the absence of literature on the evolution of themes, scientifically extracting and evaluating the change in discussions. Our work bridges the gap in the literature.

3 Methodology

For our work, to study the evolution in perception towards the pandemic related to COVID-19, we choose the subreddit r/COVID19positive as our

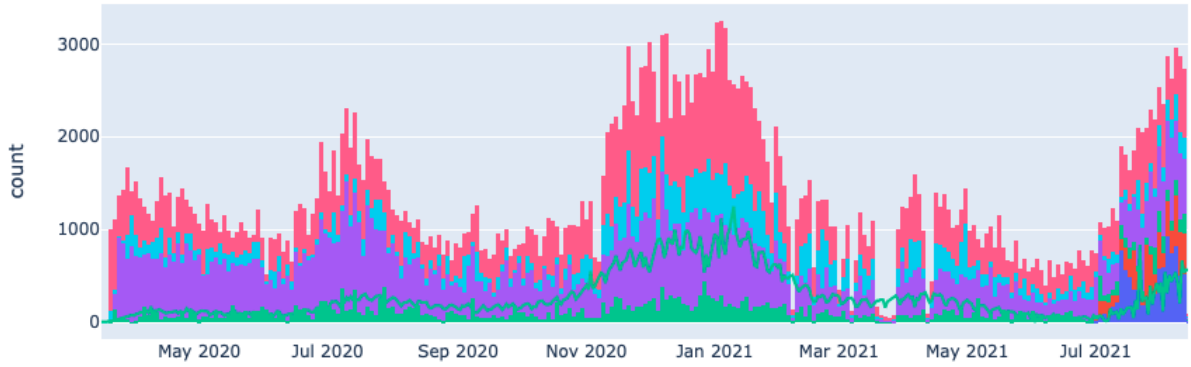

Figure 1: Number of posts per day in our dataset. The colors in the figure represent different flairs on r/COVID19positive subreddit. We can see the emergence of colors *red, green* which represent flairs related to *vaccines, tested positive* in early 2021. The figure also depicts the daily cases in the USA as a line. Our dataset closely follows the number of cases registered in the USA.

data source. Reddit is a community-centric social network where users form communities specific to topics and intents called subreddits. The r/COVID19positive subreddit is "A place for people who came back positive for COVID19 can share your stories, experiences, answer questions and vent!" – according to the subreddit description. As the subreddit contains discussions related to support, experience, and QnA by people tested positive, our data will be rich in capturing diverse contexts while being very specific to our topic of interest. With more than 124K users, the subreddit contains discussions ranging from users seeking support from the loss of a family person, to users discussing the geopolitical issues surrounding the pandemic. Next, we discuss the dataset preparation.

3.1 Dataset Preparation

Firstly, for our data collection, we used the Pushshift-API [2] to collect the ids of all post ever made on the subreddit. We have the earliest post dating back to [April - 2020], and the latest post on [October - 2021]. Once we have the post ids, we use the official Reddit API - PRAW [3] for collecting posts and comments. Our framework takes a post and recursively collects all the comments on a given post along with metadata such as upvotes, awards, flairs, etc.

Having collected all the posts, we first take a look at the topics being discussed in our dataset. On Reddit, moderators of a subreddit can define the topics, and these are similar to hashtags on networks like Facebook and Twitter. The topics are called *flairs*; just like hashtags flairs are displayed in colored boxes at the end of a post making the post *tagged under a topic*. So, we take the help of flairs, which are accepted by community and created by moderators; we identify the topics being discussed on the subreddit r/COVID19positive. By making flairs our choice for analysis at the initial stage, we are also avoiding assumptions on topics under discussion.

A list of all flairs and their meaning on the subreddit r/COVID19positive is as listed below:

1. **Tested positive - {Me, Long Hauler, Family}:** The four flairs related to testing positive. Each flair shows a special case where either the original poster is tested positive *me*, or a family member of the poster is tested positive *family*, or someone related (or themselves) to the poster is battling for a longer time than usual *long hauler*.

2. **Question - {medical, to those who tested positive}** The two flairs represent queries to the communities by users. The first flair *medical* represents questions related to scientific research, particularly vaccination related to COVID 19. While the flair *to those who tested positive* contains posts about someone's plight while they were battling with the infection. The questions result in discussions related to symptoms and conspiracy theories, as we will see further in the paper.

3. **Vaccine - {tested positive, discussion}** The two flairs are related to vaccines and vaccination. These flairs appeared in the later part of

[2] https://github.com/pushshift/api
[3] https://praw.readthedocs.io/en/stable/

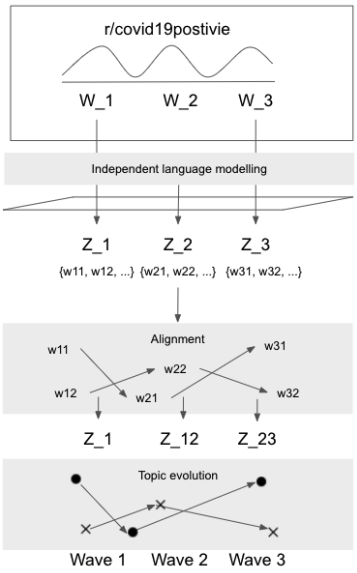

Figure 2: Framework used for our work. We have four distinct steps, where we split the dataset, then we obtain independent language models on the split corpora. Later we align the word embeddings to estimate and present the drift of each word as the final outcome.

the pandemic as shown in Figure 1. The flairs represent a changed infection pattern where people are seeing infections among the vaccinated population. Also, the vaccine-related discussion contains posts around and about different vaccines, vaccination strategies, and conspiracy theories related to vaccines.

The pandemic has a different infection rate in different nations owing to the local laws, control among the population, and other governing policies. We find that majority of our dataset contains posts made by people living in the United States of America, based on the occurrence of geography names. So, we plotted the infection rates in the USA on top of our daily posts, and the figure 1 shows the correlation between infections ranging in the USA and the posts being made on the subreddit r/COVID19positive.

Moving further in our study, we will now focus on the change in perception, the context of discussions, and key focus points with time during the pandemic. So, as our dataset is closely related to the USA infection statistics, we split the entire timeline into three parts, with the part 1 being from [Mar - Oct 2020] - where we see a dip in the infection numbers; part 2 from [Nov 2020 - Apr 2021], as we see another short dip in the infection numbers, and the remaining discussions are our part 3 which contains data from [May 2021 - till date]. Another rationale behind making the split is to capture the new emergence of infections among the vaccinated population, we make the third split when the use of flairs *vaccinated - tested positive* increase along with the infections numbers in the USA.

Once we make the split, we now have three corpora of our dataset. Additionally, we consider the posts and comments as independent documents because we are studying the overall evolution of topics during the pandemic. We call the three corpora C_1, C_2, C_3. The corpora contain special symbols, emoticons, etc. As a standard NLP pre-processing practise, we do lemmatization, cleaning of text with regular expressions, etc. Our resulting dataset contains 95,799 documents in C_1; 46,529 documents in C_2 and 55,696 documents in C_3. After extracting the common words being used in all of the three corpora, we work with 9,805 words. Having prepared the data and given an overview of the entire dataset and the potential topics in our dataset; we now proceed to discuss the framework using which we will answer our research questions.

3.2 Algorithm

With the dataset prepared, we now proceed to introduce Diachronic word embeddings (Montariol, 2021) and further processing of the social media data.

3.2.1 Introduction to Diachronic Word Embeddings

Traditional embedding models, when trained independently, do not capture information or knowledge present in another word embedding model. If a Corpus is evolving over time, then taking a word embedding or training a model on snapshots is not enough to capture the temporal evolution. By training multiple word models, we can capture the smaller snapshots and differences in nuances of the words in short intervals. However, these models are not comparable with subsequent models in a different time slice. To overcome the difficulty and ensure that words capture the meaning over time among multiple models, researchers have proposed Diachronic word Embeddings.

One way to overcome the challenge (which we adapt in our work) (Szymanski, 2017; Hengchen and Tahmasebi, 2021) is to align words among different models which were independently trained; keeping similar words in similar positions. The constraint being, words in subsequent models should

be aligned to maintain the neighborhood. By aligning models, we can ensure that words do not drift far away and that the difference between words is comparable across independent word models. For the alignment, we use pearson's distance as proposed in the original work.

3.2.2 Interpretation

The resulting word embeddings capture the temporal evolution of word usage. However, the evolution can be represented in the form of distance movement and neighbourhood change. As time progresses, the words keep moving in the latent space; with the movement the distance between words among subsequent timestamps may be large, or it may be small. If the distance is large, it means the words have changed their meaning significantly; if it is small, the words have stayed mostly the same. Additionally, another way to capture the temporal evolution of words is to use the neighborhood of words. Neighborhood reflects the context in which a word is used. Hence, by looking at the neighborhood, we understand what other words are more likely to co-occur with the given word. Using a similar approach, we study the contextual difference resulting from the contextual evolution of a given word over time. In this work, we employ these two methods and we present the results using the distance as well as the neighbourhood.

3.2.3 Representation Extraction

As discussed in an earlier Section 3.2.1 Diachronic word embeddings capture the changes in word usage over time. To answer our questions, we need to estimate how some words might drift away from other words as time progresses, indicating an evolution or change in user sentiment, topics of choice with time. With our framework, we apply the word embedding techniques, and then we align words w.r.t existing Diachronic embedding methods. We then discuss the implications of word drift and enhance the results with the use of *flairs* from earlier. Then we present our findings in detail. Now, we proceed to present the framework in detail.

Firstly, we train three independent word embedding models. For this part, we choose word2vec. Each corpus C_i gives an embedding model represented by Z_i, and the word representation in each embedding model is given by $w_{i,j}$ where i is the model and j is the word. Secondly, we proceed to align the word embeddings. For this part, we choose the method used by (Huang and Paul, 2019). The alignment is done pair-wise. We take the C_1, C_2 pairs, and we find the common vocabulary C_{12} among the two corpora. Then, we sort the words in the common corpus by the frequency of occurrences and align the words in the decreasing order of frequency in the target corpus. We then perform an SVD operation between the embedding matrices Z_1, Z_2 which only contain the common vocabulary. As a result of the operation we obtain an aligned embedding matrix Z_{12}. Similarly, we align other pairs of embedding matrices to obtain Z_{23}. Once the aligning operations are complete, we have three sets of embeddings Z_1, Z_{12}, Z_{23} that represent each wave, respectively. The entire framework is shown in Figure 2.

3.3 Representation of temporal word evolution

Now that we have created the three aligned word embedding models, with the help of Diachronic Embedding framework (NTAM), we have all the words in comparable positions. Particularly, with the help of plots and target keywords, we will decipher the user sentiment across the subreddit and the topic evolution from time to time. For making the plots, we take the position of different words in the three models, and then we also obtain their 10 nearest neighbours. As discussed in an earlier Section 3.2.1, neighbourhood and the near/far movement of words are key indicators of capturing temporal changes in topics.

We present the temporal evolution of words in our dataset in the form of two visual representations:

- **Neighborhood evolution** To assess the change in the context of word usage, we plot the words from each model on a single 2D plane. As mentioned in section 3.2.1 the neighborhood represents the context of word usage, and words in the neighborhood. These neighborhood words are more probable to occur co-occur with the base word, making them relevant. Similarly, the closeness of a tight-knit neighborhood shows a distinctive topic, and a neighborhood in which the distances are not tight-knit can be attributed to a neighborhood disintegrating. In our work, we plot the neighbourhood of a given word. For interpretation, if the neighborhood is small, we want the readers to understand that the topic

is concrete and the word usage belongs to a single context. Similarly, if the neighborhood is large, the reader should understand that the word is drifting away from a given neighborhood, and the frequency of word usage has decreased.

- **Word Drift** Another key aspect in studying the difference in the temporal evolution of topics is the word drift. "How much does a word move for away from itself over time?". By comparing two different words and the drift along time, we can show how the usage of the word has changed comparative to another word. In our work, we look at 'cancer' and 'cold' for example. Cancer was not widely used at the beginning of the pandemic however it started being frequent post pandemic. By showing the difference between words, we highlight the word importance in the global context.

Figures 3, 4, 5 show the difference between different words as seen in the neighborhood and the drift.

4 Results and Discussion

Having established the methodology and visualizations that aid us in understanding word evolution, we now discuss the results.

4.1 Narrative evolution during COVID-19

Conspiracy: Sentiments and discussions around COVID-19 have seen a vast change. Firstly there was some speculation that the virus outbreak could be a hoax, and governments are trying to push propaganda with a narrative around the Coronavirus. This has resulted in a lot of conspiracy theories and agenda against the government. However, during the later parts, particularly after the first wave of infections, people have started believing that COVID-19 is indeed a real threat. The later parts show a different sentiment among the users. With the help of words drifting apart from each other and by looking at the membership of words such as "conspiracy", "government" and "hoax", we demonstrate that there is indeed a change in the neighborhood among these words. The change in the neighbourhood also indicates that people have moved away from believing in a hoax. An increase in the popularity of scientific articles that have been shared and discussed on the COVID-19

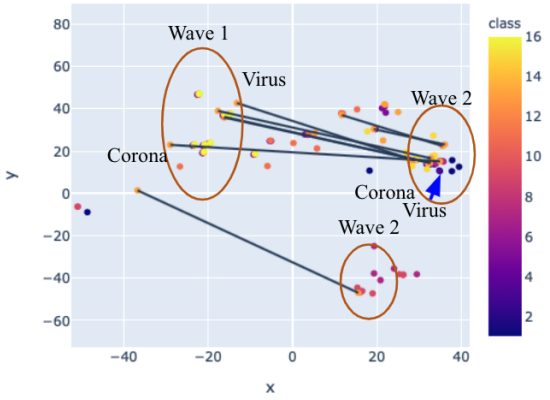

Figure 3: Figure showing the difference between wave 1, and 2. The words in the picture are 'corona', 'pandemic' 'virus', etc.

positive subreddit is also present. We show the nearest neighbours of conspiracy words before the first year and later in the second wave in Figure 3.

Symptoms: As we have seen, discussions around conspiracy have dropped down; also, conversations saw a surge in discussions around symptoms. In this part, we study the COVID-19 symptoms such as *cough, fever*. We show that these words were very close to each other in the second half of our time split(C_2). However, as time progressed, the symptoms are seen very close to the words such as *COVID, coronavirus* and *infection* which shows that users are thoroughly using symptoms along with discussions key to the pandemic. The key change to notice during the third wave of infections (C_3) is the symptoms being nearest-neighbors to chronic diseases such as *cancer* and *heart failure*. This change is particularly important because, even though patients are tested negative; for patients who have recovered from a COVID-19 infection, there is a chance of potential hazard. We show the the neighborhood change around words representing symptoms in Figure 4.

Post-COVID Symptoms: Chronic diseases like *cancer* and *heart failure* have seen a surge in usage as well as they have moved closer to the keywords. The change can be attributed to the fact that people with pre-existing conditions are more vulnerable to COVID-19. A long-standing battle with the disease is correlated with a lot of deaths reported after tested negative; with a chronic existing condition. We have also seen the closeness of word alongside these diseases, which is once again factor that plays an important role in the community health. Old people being a potential target

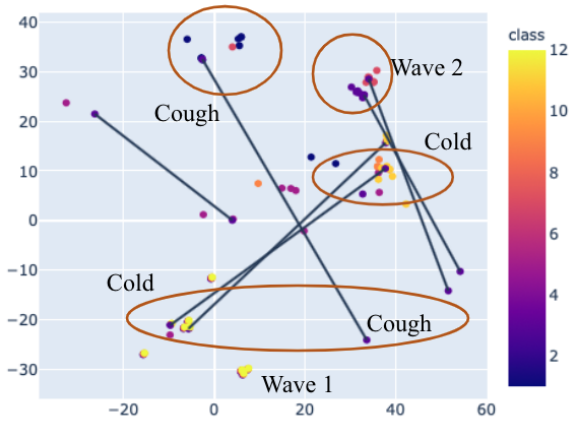

Figure 4: Figure showing the difference in symptoms during the pandemic. The evolution of symptoms here shows a change in user perspective towards the same.

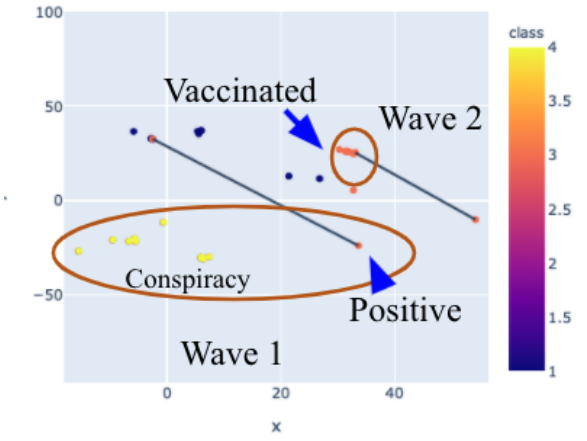

Figure 5: Figure showing the discussions around vaccinated and getting positive.

for COVID-19, could not recover completely from the infection and eventually getting some other disease was seen. we show the neighborhood change of COVID symptoms post-COVID symptoms in Figure 5.

4.2 Vaccinated - tested positive

Having seen the comparison between the first and the second wave and studied how word usage has changed; we have also established - during the same time, there is a surge in usage of words such as cancer and heart failure. However, another important aspect to consider is the possibility of getting tested after being vaccinated. We have captured the phenomenon in our data set. With the help of flairs we now study how the post-pandemic and the post-COVID community is dealing with getting vaccinated.

Getting vaccinated does not make a person immune to the infection. But, the infection will not have a long-standing effect on a person. Vaccination is there to eradicate the effect of COVID. In the United States, a lot of people have been vaccinated, and they are also getting an infection. So, in this context, we look at the *vaccinated* flair, and we find that a lot of people are recovering within their homes; for example, refer to the link [4]. Similarly, post-pandemic syndrome effects are very clear in old people and people with long-standing health conditions. The same is reflected in post with flair "tested positive family". Please note that the flair is representative of a family being infected. In this flair, particularly during the post-pandemic period we find that a lot of people are getting tested positive and being admitted for long-standing diseases such as *cancer* or *heart disease*. Old people in particular are getting infected after getting vaccinated.

With this subsection, we shed light on the future of COVID-19 pandemic after the wave 3. We can expect the vaccinations to stand as a guard against a flood of infections. However, we cannot expect infections to go down. Similarly, people with bad health conditions are more vulnerable even after getting vaccinated.

5 Concluding Remarks

In this work, we have used Diachronic word embeddings to assess the temporal evolution of topics around COVID-19 pandemic. For one and half years, we take the help of r/COVID19positive subreddit. In order to conduct our study, we have captured the users throughout the three waves of the United States. We have closely followed the pandemic in the United States because our conversations have followed the peaks from the United States. In our work, we split the Corpus into three, and compared the three sub-corpora; each subcorpora capturing a different phases of the pandemic. In the first wave, we've had a lot of uncertainty around COVID, and a lot of conspiracy theories were floating around. We find that by the second wave, people got serious, and there was a lot of support going on in the community. People were discussing vaccines in the latter half of the second wave. However, during the third wave, we have seen that people were discussing being vacci-

[4] https://www.reddit.com/r/COVID19positive/comments/qu26z4/tested_positive_twice_in_4_months_fully_vaxxed/

nated and getting a positive result on their COVID test after vaccination. We also demonstrate an important aspect of post-COVID symptoms, which are related to long-standing illness such as *cancer* and *heart failure*. With this work, we demonstrate the utility of Diachronic embeddings in the real world. We also demonstrated the application of Diachronic embeddings on a very short time period unlike any previous literature or research.

In the future we aim to build a dashboard that can be easily used by domain experts, medical professionals and journalists to access the information we extracted in our work.

Acknowledgement

The authors would like to thank National Institute of Informatics (NII), Japan for providing the online internship opportunity. Also, the authors found the thesis (Montariol, 2021) extremely resourceful for their research.

References

Piyush Ghasiya and Koji Okamura. 2021. Investigating covid-19 news across four nations: A topic modeling and sentiment analysis approach. *IEEE Access*, 9:36645–36656.

Simon Hengchen and Nina Tahmasebi. 2021. A collection of swedish diachronic word embedding models trained on historical newspaper data. *Journal of Open Humanities Data*, 7.

Xiaolei Huang and Michael J. Paul. 2019. Neural temporality adaptation for document classification: Diachronic word embeddings and domain adaptation models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4113–4123, Florence, Italy. Association for Computational Linguistics.

Navin Kumar, Isabel Corpus, Meher Hans, Nikhil Harle, Nan Yang, Curtis McDonald, Shinpei Nakamura Sakai, Kamila A Janmohamed, Weiming Tang, Jason L Schwartz, et al. 2021. Covid-19 vaccine perceptions: An observational study on reddit. *medRxiv*.

Andrey Kutuzov, Lilja Øvrelid, Terrence Szymanski, and Erik Velldal. 2018. Diachronic word embeddings and semantic shifts: a survey. *arXiv preprint arXiv:1806.03537*.

Qian Liu, Zequan Zheng, Jiabin Zheng, Qiuyi Chen, Guan Liu, Sihan Chen, Bojia Chu, Hongyu Zhu, Babatunde Akinwunmi, Jian Huang, et al. 2020. Health communication through news media during the early stage of the covid-19 outbreak in china: digital topic modeling approach. *Journal of medical Internet research*, 22(4):e19118.

Tiago de Melo and Carlos MS Figueiredo. 2021. Comparing news articles and tweets about covid-19 in brazil: sentiment analysis and topic modeling approach. *JMIR Public Health and Surveillance*, 7(2):e24585.

Syrielle Montariol. 2021. *Models of diachronic semantic change using word embeddings*. Ph.D. thesis, Université Paris-Saclay.

Curtis Murray, Lewis Mitchell, Jonathan Tuke, and Mark Mackay. 2020. Symptom extraction from the narratives of personal experiences with covid-19 on reddit. *arXiv preprint arXiv:2005.10454*.

Abeed Sarker and Yao Ge. 2021. Long covid symptoms from reddit: Characterizing post-covid syndrome from patient reports. *medRxiv*.

Terrence Szymanski. 2017. Temporal word analogies: Identifying lexical replacement with diachronic word embeddings. In *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 2: short papers)*, pages 448–453.

Wei Wu, Hanjia Lyu, and Jiebo Luo. 2021. Characterizing discourse about covid-19 vaccines: A reddit version of the pandemic story. *arXiv preprint arXiv:2101.06321*.