

Unsupervised Learning Model to Uncover Hidden Knowledge from COVID-19 Vaccines Literature

Tasnim Gharaibeh(✉) and Elise de Doncker

Western Michigan University, Kalamazoo, MI 49008, USA
{tasnim.gharaibeh,elise.dedoncker}@wmich.edu

Abstract. Severe acute respiratory syndrome coronavirus 2 (or SARS-CoV-2) has spread globally, causing a pandemic with, so far, more than 152 million infections and more than three million deaths (as of May 2021). In order to address the COVID-19 pandemic by limiting transmission, an intense global effort is in the development of a safe and effective vaccine, which generally requires several years of pre-clinical and clinical stages of evaluation as well as strict regulatory approvals. However, because of the unprecedented impact of COVID-19 worldwide, the development and testing of a new vaccine are being accelerated. There are currently some authorized, not yet approved, vaccines to fight COVID-19, besides other ones in clinical evaluation or in a pre-clinical stage, and many more being researched. In this work, we used natural language processing and a machine learning model to predict good candidate vaccines. We built an unsupervised deep learning model (CVW2V) to produce word-embeddings using Word2vec from a corpus of published articles, selectively focusing on COVID-19 candidate vaccines that appeared in the literature, to identify promising target vaccines according to their similarity with approved and authorized vaccines.

Keywords: Unsupervised learning · Word embeddings · Word2vec · COVID-19 vaccines

1 Introduction

Machine learning approaches are usually divided into three categories, supervised, unsupervised, and reinforcement learning. Supervised learning deals with labeled data, data that contains both the inputs and the desired outputs, to train the model. It uses linear regression, neural networks, nearest neighbor, or naive Bayes algorithms to predict or classify data. Reinforcement learning allows machines or software agents to automatically decide the required behavior within a specific situation to maximize its performance. Using Q-Learning, the temporal difference (TD), or deep adversarial networks, reinforcement learning algorithms can be applied in robotic hands, computer played board games, and self-driving cars. Unsupervised learning allows the model to work on its own to find structure and discover patterns and information that was previously undetected; mine for rules; and summarize and group the data points, without the need to supervise or lead the model and without any previous insertion knowledge. It mainly deals with the unlabeled data as input [1]. Clustering and association are two types

© Springer Nature Switzerland AG 2021
O. Gervasi et al. (Eds.): ICCSA 2021, LNCS 12950, pp. 544–559, 2021.
https://doi.org/10.1007/978-3-030-86960-1_38

of unsupervised learning. The association rules establish relations among data objects inside large databases, whereas clustering splits the dataset into groups based on their similarities.

Word embedding is one of the most important and versatile examples of unsupervised learning models, leading to a type of language modeling or word representation that allows words with similar meanings to be understood by machine learning algorithms [2]. As a representation of words obtained from an unlabeled large corpus, by embedding both semantic and syntactic meanings, word embedding can be represented as:

$$\mathrm{V} \rightarrow \mathrm{R}^{\mathrm{D}} : w \rightarrow \overrightarrow{w}$$

mapping a word w from a vocabulary V to a real-valued vector $\overrightarrow{w}$ in an embedding space with dimensionality D [33]. To compute the similarities between vectors in the embedding space, the cosine similarity function can be used, which is defined as

$$similarity(w1, w2) = \frac{\overrightarrow{w}1 \cdot \overrightarrow{w}2}{\left\|\overrightarrow{w}1\right\| \left\|\overrightarrow{w}2\right\|}$$

According to this function, the nearest neighbors of a word w are given as a list of words $\mathrm{v} \in \mathrm{V}\backslash\{w\}$, sorted in descending order [33].

This is an important method commonly used in tasks of modern natural language processing (NLP), such as semantic analysis [2], retrieval of information [3], dependency parsing [4–6], query answering [7, 8], and computer translation [7, 9, 10]. In these techniques, individual words are represented as real-valued vectors, often of tens or hundreds of dimensions, in a predefined vector space. Words that are closer in the vector space are expected to be similar in meaning. Each word is mapped to one vector. The vector values are learned in an approach that resembles a neural network, and hence the technique is often lumped into the field of deep learning. Word embedding methods vary and can be grouped into different categories. For example, they can be classified as either paradigmatic models or syntagmatic models, based on the word distribution information [11, 12]. The most important aspect of the syntagmatic model is that words co-occur in the text region, but for the paradigmatic model, it is the similar context that matters. Another grouping depending on how word embedding is generated divides models into two classes, matrix factorization and sliding-window sampling. The first method is based on the word co-occurrence matrix, where word embeddings are obtained from matrix decomposition. In sliding-window sampling, data sampled from sliding windows is used to predict the context word [13]. Word embedding has many important applications. It can be applied in idiomaticity analysis, sentiment analysis, syntax analysis, part of speech (POS) tagging, named entity recognition, as well as textual entailment [14]. Various word embedding models are available, such as Word2vec, Glove, and fastest.

The Word2vec method is based on a shallow neural network with two layers that takes a large corpus of text as its input and produces a vector space, typically of several hundred dimensions, with remarkable linear relationships called analogies. These analogies allow for math operations such as vec(“king”) - vec(“man”) + vec(“woman”) ≈ vec(“queen”) [20]. As an unsupervised learning technique, and a machine learning algorithm used to draw inferences from datasets consisting of input data without labeled responses, it

comprises two techniques: CBOW (continuous bag of words) and Skip-gram. CBOW predicts the probability of a word given a context, while Skip-gram predicts the context given a word.

These models in machine learning have great benefits and a significant impact in many areas of science and technology, including medical research, because the models created by machine learning can learn and generalize the patterns and similarities within the available huge data and can draw conclusions from previously unseen data [15]. On the other hand, the increase in knowledge and understanding of diseases, vaccines, and drugs has been associated with a growth in information. Such big data is ready to deliver the benefits of the application of machine learning (ML) in diagnosis, prognosis, drug development, and vaccine discovery.

Vaccine discovery methods have been costly, and it may take many years to develop an appropriate vaccine against a specified pathogen. In the COVID-19 case, vaccines are essential in reducing morbidity and mortality especially since the virus establishes itself in the population. Thus, the development of a new vaccine should be accelerated to save lives. When most people in a community are vaccinated against a disease, the ability of the pathogen to spread is limited. This also provides indirect protection to people who cannot be vaccinated. Machine learning can accelerate the discovery of effective vaccines, or suggest components of a vaccine after understanding the viral protein structure, and assist medical researchers scouring tens of thousands of related scientific papers [16, 17].

In this research we employ word embedding, as an unsupervised machine language and NLP model, to predict potential vaccines according to their similarity with the authorized vaccines, without any prior explicit insertion knowledge about vaccines, treatment, or biomedical information about COVID-19.

2 Background and Previous Work

2.1 Vaccine Development

A pathogen is a bacterium, virus, parasite, or fungus that can cause disease within the body. Each pathogen has its unique subparts; one important factor is the antigen; which causes the formation of antibodies. The antibodies produced in response to the pathogen's antigen are an important part of the immune system. Usually, vaccines contain weakened or inactive parts of a particular antigen that train the immune system to recognize and fight pathogens, which greatly reduces the risk of infection. The immune system can safely recognize the pathogen when it is encountered naturally; by delivering an immunogen, a specific type of antigen elicits an immune response [15].

To fight virus-infected cells, the host immune system produces antibodies on the surface of B-cells or attacks the virus directly using T-cells. To assist T- and B-cells in attacking and binding with invaders, the human leukocyte antigen (HLA) gene encodes MCH-I and MCH-II proteins, which present epitopes as antigenic determinants [18].

To ensure the safety and efficiency of the new vaccine, many actions and steps are taken, such as pre-clinical studies, phase I clinical trial, phase II clinical trial, phase III clinical trial, phase IV post-marketing surveillance, and human challenge studies. In the pre-clinical studies, the vaccine is tested first in animal studies for efficacy and safety.

Then in phase I clinical trial, small groups of healthy adult volunteers receive the vaccine to test for safety. After that in phase II clinical trial, people who have characteristics for whom the new vaccine is intended are given the vaccine. The next step of testing the vaccine is by giving it to thousands of people and that is in phase III clinical trial. After the vaccine is approved and licensed, many studies are conducted to study future effects of the vaccine in the population in the long term [19].

Since the outbreak of this novel coronavirus, different machine learning approaches have been used to predict potential effective vaccines.

2.2 Machine Learning and the Development of COVID-19 Vaccines

To accelerate the vaccine design process by predicting the design of a multi-epitope vaccine for COVID-19, Zikun et al. implemented a supervised deep learning model (DeepVacPred) in silico. From the available SARS-CoV-2 spike protein sequence, their model was able to predict 26 candidate vaccine subunits, by combining the in silico immunoinformatic and deep neural network strategies. In order to construct a multi-epitope vaccine for the SARS-CoV-2 virus, they identified the top 11 from the 26. According to the best 11 candidates, they suggested a 694aa (consisting of 694 amino acid residues) multi-epitope vaccine that contains 16 B-cell epitopes, 82 CTL epitopes, and 89 HTL epitopes, as a promising vaccine to fight the SARS-CoV-2 viral infection. They also concluded that the designed vaccine can be used successfully with the recent RNA mutations of the virus by investigating the SARS-CoV-2 mutations for RNA [24].

Since the T-cell and B-cell epitopes could be used to produce good vaccines as well as recognize neutralizing antibodies [28], Fast et al. tried to identify 2019-nCoV T-cell and B-cell epitopes based on viral protein antigen presentation and antibody binding properties, using computational tools from structural biology and two supervised neural networks (MARIA, a multimodal recurrent neural network [29] and NetMHCpan4, a trained neural network in [30]). Their results show good antigen presentation scores for 405 potential viral peptides, for both human MHC-I and MHC-II alleles. Their models also identified two potential neutralizing B-cell epitopes close to the 2019-nCoV spike protein receptor-binding domain (440–460 and 494–506). They also concluded that the spike protein for the SARS-CoV-2 virus could be a potential vaccine candidate. This conclusion was the result of their analysis of mutation profiles of 68 viral genomes spread through four continents that showed no mutations are present near the spike protein receptor-binding domain and d 96 coding-change mutations occurred only in regions with good MHC-I presentation scores [28].

By using the same method, the recurrent neural networks (RNN), and a training set consisting of alpha, beta, gamma, and delta coronavirus spike sequences, Crossman simulated the spike protein sequences of coronaviruses for the previously known sequences and tested their characteristics. This simulation could help in the development of vaccine design to predict alternative possible spike sequences that could arise in the future [31].

In another study that introduces a computational framework from several computational approaches such as reverse vaccinology and immunoinformatics tools with deep learning, Abbasi et al. predicted a potentially suitable protein vaccine candidate against SARS-CoV-2. Reverse vaccinology uses the expressed genomic sequences to find new

potential vaccines, whereas immunoinformatics tools such as PSORTb, FungalRV, SignalP, TargetP, IEDB, BLASTp, ProtParam, Vaxijen, etc. were modified to classify proteins into positive and negative datasets. As a result, one protein (Spike S- Surface Glycoprotein with accession no. QHD43416.1) was shortlisted as a potential vaccine candidate out of the 10 proteins of SARS-CoV-2. Out of 41 B-cell epitope sequences that were predicted using BcePred, only 15 peptides were identified to be highly antigenic as specified by the Vaxijen server, where a peptide "DLCFTNVY" is predicted as the highest-ranking peptide. For T-cell epitopes, they consider peptides with a 100% conservancy rate, so 45 epitopes were identified by the ProPred-I tool to be conserved out of the 46 predicted MHC class-I binding epitopes, where KIADYNYKL was found with the highest antigenicity score and 90 out of 94 MHC- II allele binding T-cell epitopes predicted by the ProPred tool with VKNKCVNFN as the highest antigenicity score [32].

3 Methods

3.1 Data Collection and Processing

From Science Direct at Elsevier, we obtained around 54,427 full-text published papers and articles primarily focusing on COVID-19 in the years 2019, 2020, and 2021. In order to perform the text mining process, we used Science Direct application programming interfaces (APIs) (https://dev.elsevier.com/), interactive APIs, and some Python (3.6) code [21]. We searched the Science Direct database using specific search terms as a first step in cleaning the data. These terms combined "COVID-19 OR 2019-nCoV OR COVID19 OR nCoV-19 OR Sars-CoV-2" with vaccines, biomedical, medication, antibodies, immunity, immunology, and drugs. Extensive cleaning processes were needed to reduce the noise and unwanted words in the corpus, speed up model training, and improve model efficiency. We deleted words with a length of more than 20 letters; usually, these words are URLs or descriptions related to the metadata. Then we converted all uppercase characters in the corpus to lowercase, so Vaccine and vaccine will have one vector. Finally, we deleted the stop words and punctuation from the corpus. We started with a data file of size 2.74 GB and obtained a reduced file of size 2 GB.

3.2 Word2vec Training

To train our model (CVW2V), a combination of the Word2vec implementation in genism [22] and the open-source code implemented by Vahe Tshitoyan et al. [23] were used with some modifications. Several parameters and tuners that affect both training speed and efficiency are recognized by Word2vec such as word count, vector size, workers, negative sampling, the model type (CBOW or skip-gram), learning rate, and window size. We have more than 298 million different words in the data, but we consider only 1,679,911 words. These are words that occurred more than five times because words that appear in a billion-word corpus only once or twice or less than 5 times are probably uninteresting typos and garbage. Furthermore, there is not enough information to include any practical training on those terms, so it is better to ignore them. The vector size, size of the embedding, in CVW2V is 200. Higher dimensionality increases the quality of

word embedding, but after reaching some point, the benefit would decrease [20]. Larger size values need more data from training but can lead to better (more precise) models. Usually, fair vector size values are in the tens to hundreds.

For the model type, CVW2V uses a skip-gram model for the training with negative sampling loss. With negative sampling, we randomly select just a small number of "negative" words (let's say n = 15) for which to update the weights. In this context, a "negative" word is one for which the neural network should output a 0. Negative sampling speeds up the training by decreasing the computational burden and also increases the efficiency of the resulting word vectors by having each training sample only modify a small percentage of the weights, rather than all of them [29]. According to Tshitoyan's result, it was found that skip-gram with negative sampling loss (n = 15) performed best [23]. With respect to Google code [25], the author indicated that CBOW is faster, while skip-gram is slower, but it does a better job for infrequent words because Skip-Gram seeks the prediction of the context given a word instead of the prediction of a word given its context like CBOW.

The window size parameter in CVW2V is 8, which relates to the maximum distance within a sentence between the actual and the predicted word, or how many words before or after a given word would be included as context words of the given word. The initial learning rate was 0.01 and it would drop linearly to 0.0001 as training progresses in 30 epochs. The typical range for the learning rate is between 1 and 10^{-6}. The learning rate aims to decide what proportion of the observed error should be used to change weights. The model will oscillate by setting the learning rate too high and will converge too slowly to a solution by setting it too low. A learning rate of 0.01 usually works with most networks [26]. To optimize the computation, we used a hierarchical SoftMax training algorithm, which uses a binary tree to represent all words in the vocabulary [27].

4 Results and Discussion

4.1 Verification

To verify CVW2V results, some general known words and phrases related to COVID-19 vocabulary such as SARS, vaccines, wearing a mask, dry cough were chosen as seeds for the model to test the relatedness between these words and their similarities depending on human judgment as shown in Table 1. For the first word, SARS, the top similar words are sarscov, sarscov1, sars_coronavirus, middle_east_respiratory_syndrome, sarscov2, severe_acute_respiratory_syndrome, mers_sars, corona_virus, infection, and pandemic. The first eight similarities are similar names for COVID-19, whereas infection and pandemic are more general references related to COVID-19. Similar words for vaccines are vaccine_candidates, inactivated_vaccines, mrna_vaccines, immunization, live_attenuated, viral_vector, proteinbased_vaccines, subunit_vaccines, rabies_vaccine, and hpv_vaccine. Most of these are types of vaccines. Wearing a mask, as a protective means for COVID-19, has washing_hands, avoid_touching_face, wearing_gloves, frequent_handwashing, maintaining_social_distance, staying_home, face_covering, personal_hygiene, avoiding_crowds, and cough_etiquette as the top 10 similarities and it is obvious that all of these are also protective measures against COVID-19. As a

symptom of COVID-19 we chose dry cough, resulting in the top 10 similarities of shortness_breath, fever_cough, dyspnoea, chills, runny_nose, chest_tightness, sore_throat, difficulty_breathing, myalgias, and diarrhea_abdominal_pain. Thus the similarities found are also symptoms of COVID-19.

4.2 Similarities for Some Vaccines

Vaccines are being developed with different technologies, some well-known and others completely new for human vaccines, such as peptide and nucleic acid technologies.

Whole Virus Vaccine. These are live-attenuated and inactivated vaccines. Live-attenuated vaccines use a weakened (or attenuated) pathogen that causes disease and that creates a strong and long-lasting immune response. Because these vaccines are so close to the real natural infection, only one or a maximum of two doses are needed to give lifetime immunity against the pathogen. They are used to protect against rotavirus, yellow fever, and chickenpox [34]. The inactivated vaccines use the dead version of the pathogen. Therefore, these vaccines do not provide immunity protection as strong as activated vaccines, so several doses are needed over time to get lasting immunity. They are used to immunize against hepatitis a, flu, polio, and rabies [34]. Sinopharm and CoronaVac Are examples of the inactivated virus vaccines for COVID-19.

Sinopharm. Also known as BBIBP-CorV and developed by Sinopharm, the National Pharmaceutical Group Corporation (CNPGC) in China, with 86% efficacy against the COVID-19 infection [36]. As the most similar words for Sinopharm, CVW2V found sinovac, coronavac, inactivated_vaccine_candidate, ad5ncov, sputnik_v, covaxin, codagenix, adenovirusvectored, adults_aged_18–59_years, and wibp_vaccine (see Fig. 1).

Sinovac is a biopharmaceutical company in China that focuses on the research, development, manufacturing, and commercialization of vaccines. This company developed the CoronaVac inactivated virus COVID-19 vaccine. Ad5-nCoV, or Convidicea, was the first novel coronavirus vaccine for COVID-19 in China. It is a viral vector-based vaccine. Both of these vaccines are under emergency use authorization status in China, Mexico, Pakistan, Hungary, and Chile [37]. Sputnik-V (or Gam-COVID-Vac) is a viral vector vaccine for COVID-19 developed by the Gamaleya Research Institute of Epidemiology and Microbiology in Russia. Registered on 11 August 2020 by the Russian Ministry of Health, followed by an interim analysis from the trial was published on February 2021, indicating 91.6% efficacy. Over a billion doses of that vaccine were requested by Russia, Argentina, Belarus, Hungary, Serbia and the United Arab Emirates for immediate distribution globally [38]. Covaxin is another inactivated virus-based COVID-19 vaccine being developed in India, by Bharat Biotech in collaboration with the Indian Council of Medical Research. After the second dose, it showed 81% intermediate effectiveness in preventing COVID-19 in those that had not previously been infected [39]. Codagenix is a clinical-stage synthetic biology company that uses programs to recode the genomes of viruses in order to create live-attenuated vaccines or viruses that can be used to prevent viral infections. It generated COVI-VAC, a single-dose, intranasal, live-attenuated

Table 1. The top 10 first similarities for SARS, vaccines,wearing a mask, and dry cough.

SARS	Vaccines	Wearing mask	Dry cough
sarscov	vaccine_candidates	washing_hands	shortness_breath
sarscov1	inactivated_vaccines	avoid_touching_face	fever_cough
sars_coronavirus	mrna_vaccines	wearing_gloves	Dyspnoea
middle_east_respiratory_syndrome	immunization	frequent_handwashing	Chills
sarscov2	live_attenuated	maintaining_social_distance	runny_nose
severe_acute_respiratory_syndrome	viral_vector	staying_home	chest_tightness
mers_sars	Proteinbased_vaccines	face_covering	sore_throat
corona_virus	subunit_vaccines	personal_hygiene	difficulty_breathing
infection	rabies_vaccine	avoiding_crowds	Myalgias
pandemic	hpv_vaccine	cough_etiquette	diarrhea_abdominal_pain

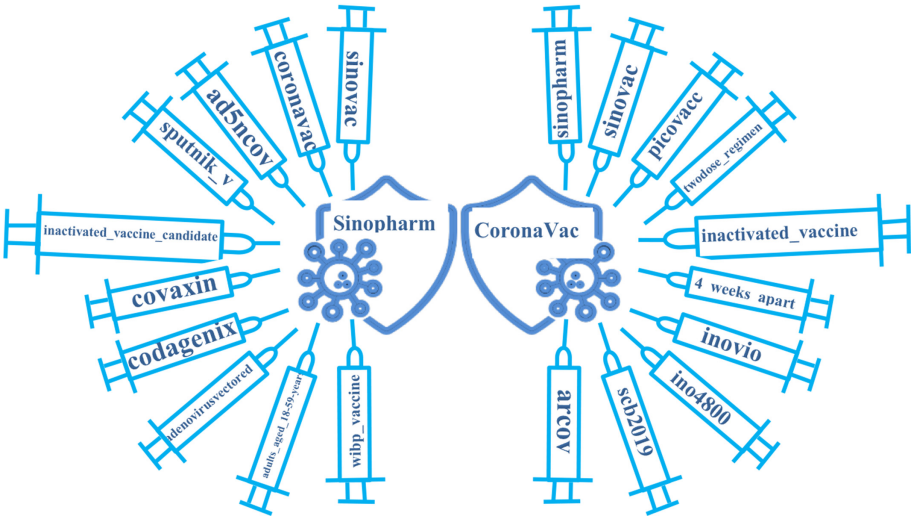

Fig. 1. The similar words for the inactivated virus vaccines for COVID-19, Sinopharm and CoronaVac

vaccine against COVID-19. In Phase I, COVI-VAC showed good results in safety and protection after just one dose in the gold standard animal model and now it is being evaluated in the clinical phase. It is designed to produce immunity against all SARS-CoV-2 proteins to protect against a range of SARS-CoV-2 strains [41]. Adenovirus vectored is a vector based vaccine, that adds a gene for the coronavirus vaccine into a modified version of a chimpanzee adenovirus that can enter human cells but not replicate inside, so the vaccine can target the spike proteins that SARS-CoV-2 uses to enter human cells [42]. The wibp_vaccine is one of two inactivated virus COVID-19 vaccines developed by Sinopharm [36].

CoronaVac. Is an inactivated virus COVID-19 vaccine developed by the Chinese company Sinovac Biotech. It has been in Phase III clinical trials in several countries. It does not need to be frozen, and can be transported and refrigerated at 2–8 °C (36–46 °F), as traditional flu vaccines are kept [37]. When it was used as a seed in our model, the top similar words Were sinopharm, inactivated_vaccine, sinovac, picovacc, twodose_regimen, 4_weeks_apart, inovio, ino4800, scb2019, and arcov (see Fig. 1).

Picovacc is the previous name for CoronaVac. twodose_regimen and 4_weeks_apart according to the treatment plan that specifies the amount and the schedule for the vaccination. Inovio, or ino4800, is a DNA vaccine INO-4700 against MERS CoV that is currently in preparation to initiate a Phase 2 vaccine trial [43]. The scb2019 subunit vaccine candidate is developed by Clover Pharmaceuticals [45]. Arcov is the first mRNA vaccine in China and may be ready in May for final stage trials overseas [44].

RNA or mRNA Vaccine. MRNA vaccine development uses a new technology, based on proteins to trigger an immune response. They have several benefits compared to other types of vaccines, including shorter manufacturing times and, because they do

not contain a live virus, no risk of causing disease in the person getting vaccinated. Moderna and pfizer-BioNTech are examples of mRNA vaccines for COVID-19. Under an Emergency Use Authorization (EUA) in December 2019, the FDA has authorized their emergency use to prevent or reduce the COVID-19 infection in individuals of age 16 years and older. We used these two vaccines as seeds to our model to search for similar candidate vaccines.

Moderna. The top similarities for Moderna are pfizer_biontech, mrna1273, Mrnabased_vaccine, encapsulated_mrna, Johnson_johnson, Astrazeneca, Curevac, Arcturus, Gsk, and Gx19 (see Fig. 2).

The mrna1273 similarity is another name for Moderna. CureVac COVID-19 is an mRNA COVID-19 vaccine candidate developed by CureVac N.V. and the Coalition for Epidemic Preparedness Innovations (CEPI). It is different from the current mRNA COVID-19 vaccines, Pfizer–BioNTech and Moderna. CureVac uses unmodified RNA, while the other two use nucleoside-modified RNA. It is currently in Phase III clinical trials, as of April 2021 [46]. Arcturus, or ARCT-02,1 is an mRNA COVID-19 vaccine candidate developed by Arcturus Therapeutics and currently in phase II [47]. Gsk, or VAT00002, is a protein subunit COVID-19 vaccine candidate in phase II and developed by Sanofi Pasteur and GSK [48]. The gx19 DNA COVID-19 vaccine candidate is developed by Genexine and in phase II status [49].

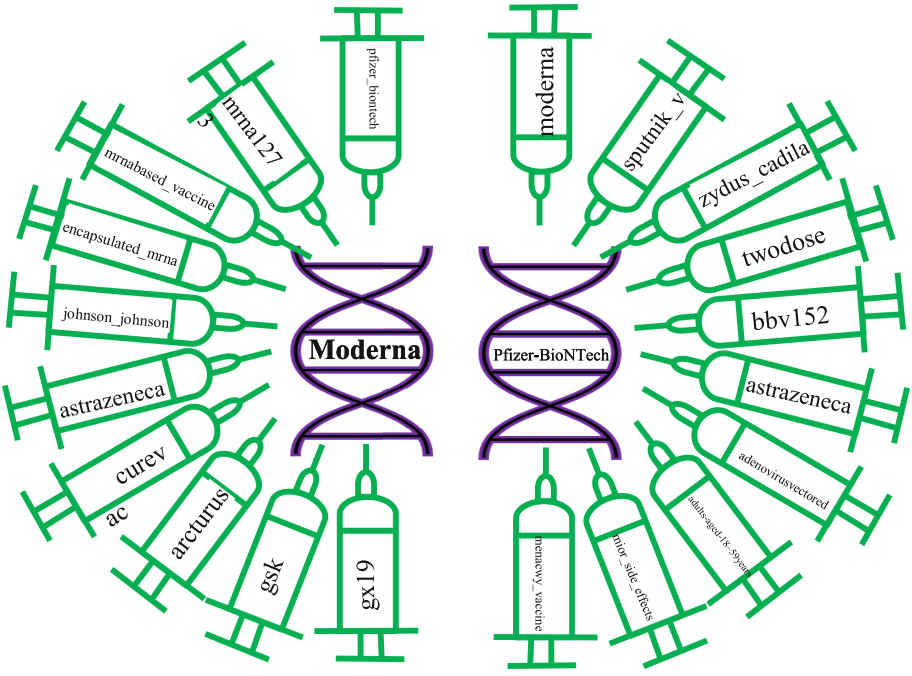

Fig. 2. The similar words for Moderna and Pfizer-BioNTech as mRNA vaccines for COVID-9.

Pfizer-BioNTech. The similar words found for pfizer-bioNTech are astrazeneca, moderna, sputnik_v, zydus_cadila, twodose, bbv152, adenovirusvectored, adults_aged_18–59_years, minor_side_effects, and menacwy_vaccine (see Fig. 2).

zydus_cadila is an Indian multinational pharmaceutical company that develops a DNA plasmid-based COVID-19 vaccine named ZyCoV-D, which is expected to get emergency use authorization in May or June [50]. BBV152, or Covaxin, is an inactivated virus-based COVID-19 vaccine. It was developed by the Indian Council of Medical Research and Bharat Biotech [51]. The MenACWY vaccine is a single injection given into the upper arm that protects against four meningococcal bacteria strains that cause meningitis and blood poisoning: A, C, W, and Y [52].

Viral Vector. Is a type of vaccine that uses a modified version of one virus as a vector. This vector is a DNA molecule used as a vehicle to artificially carry foreign genetic material into another cell. Even though it has two strange objects for the body, this type of vaccine does not cause infection with either the virus used as the vector or the source of the antigen. The genetic material it delivers does not integrate into a person's genome [40]. Johnson & Johnson's Janssen and Oxford–AstraZeneca are examples of viral vector-based COVID-19 vaccine.

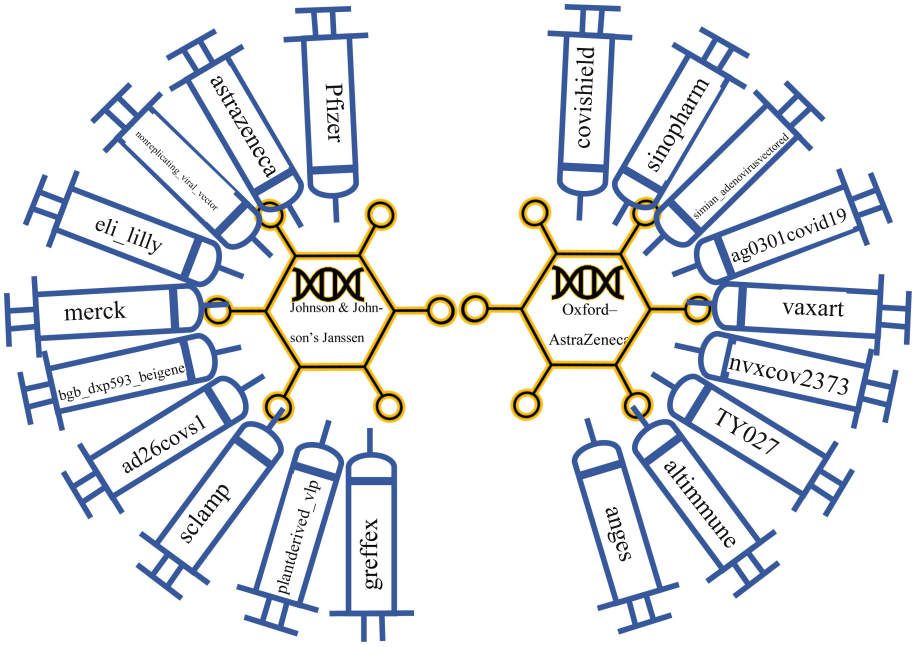

Fig. 3. The similar words for viral vector-based COVID-19 vaccine, Johnson & Johnson's Janssen and Oxford–AstraZeneca.

Johnson & Johnson's Janssen. Is a one dose viral-based COVID-19 vaccine developed by Janssen Vaccines company. We used Johnson & Johnson's Janssen as a

seed to our model. pfizer, astrazeneca, nonreplicating_viral_vector, eli_lilly, merck, bgb_dxp593_beigene, ad26covs1, sclamp, plantderived_vlp, and greffex are some similar words for this vaccine (see Fig. 3).

eli_lilly is an antibody-based treatment for COVID-19 with an Emergency Use Authorization from the Food and Drug Administration [53]. Merck is a pharmaceutical company working on developing an oral medication for COVID-19. It is in phase II clinical trial with promising early findings [54]. BGB-DXP593 is a neutralizing antibody, or protein-based therapy, under investigation for SARS-CoV-2 in a clinical trial for participants with mild-to-moderate COVID-19 by BeiGene [55]; ad26covs1 is another name for Johnson & Johnson's Janssen vaccine. Sclamp, or V451, is a subunit_based COVID-19 vaccine developed by the University of Queensland and the Australian pharmaceutical company CSL Limited, and it is still in Phase I [56]. The plantderived_vlp vaccine is a different type of COVID-19 vaccine. Instead of using a viral vector, it uses living plants as bioreactors to produce non-infectious versions of viruses, called Virus-like Particles, or VLPs. This type of vaccine was developed by the Canada-based biopharma Medicago and GlaxoSmithKline. It is currently in Phase 3 clinical trials [57]. Greffex is a vaccine based on an adenovirus viral vector, currently in Pre-Clinical phase and developed by Greffex, a genetic engineering company [58].

Oxford–AstraZeneca. Is a viral vector-based COVID-19 vaccine developed by Oxford University and AstraZeneca. According to CVW2V, covishield, sinopharm, anges, ag0301covid19, vaxart, nvxcov2373, simian_adenovirusvectored', TY027, altimmune are interesting similar words for Oxford–AstraZeneca (see Fig. 3).

The word covishield is another name for Oxford–AstraZeneca. Anges, or AG0302-COVID-19, is a DNA candidate vaccine developed by AnGes Inc. in Japan and currently in Phase II/III [60]. Vaxart is a nonreplicating viral vector oral vaccine platform for COVID-19 in Phase I [59]; nvxcov2373 is a protein-based COVID-19 vaccine [35]. The simian_adenovirusvectored vaccine is used as a vector in the vector-based vaccine instead of the human adenovirus, because of its advantages over the human adenovirus. TY027 is a treatment for patients with COVID-19 to slow the progression of the disease and accelerate recovery, in addition to its potential to provide temporary protection against infection with SARS-CoV-2 [61]. Altimmune is a single dose viral vector COVID-19 vaccine that triggers a broad immune response – neutralizing IgG, mucosal IgA, and T-cells and it is still in Phase I [62].

5 Conclusions and Future Work

To stop the pandemic, wearing masks and social distancing are required but not enough; we need to use all prevention tools, especially vaccines. COVID-19 vaccines, like other vaccines, work with our immune system, to prepare our body to fight the virus if we are exposed. As an addition to our work in [63] on drug discovery using Word2vec for COVID-19 treatment, we trained an unsupervised word embedding model (CVW2V) to search for potential candidate vaccines in the COVID-19 literature according to their similarities with authorized or most promising existing ones. These studies rely on the

property that words sharing similar surrounding words are semantically close. After verifying CVW2V with the terms SARS, vaccines, wearing a mask, and dry cough as general words and their similarities, CVW2V was provided with currently authorized vaccine names such as Sinopharm and CoronaVac as inactivated whole virus-based vaccines, Moderna and Pfizer-BioNTech as mRNA vaccines, and Johnson & Johnson's Janssen and Oxford–AstraZeneca as viral vector-based COVID-19 vaccine. As a result, without any explicit insertion knowledge about COVID-19 or vaccines, or any supervision on the words or objects and their meanings, the CVW2V model was able to suggest several potential target vaccines for clinical investigation and support other ones in different clinical experimental phases according to their high similarity with currently authorized and most promising vaccines.

In future work, we plan on applying this approach in different areas, such as the materials science field. Since the data types may not be the same, the word embedding parameters need to be tuned. Different combinations of parameter values for the word embedding training could be changed to improve the approach, such as negative sampling distribution, number of epochs, window size [64], minimum word count, and vector size [65].

Acknowledgments. The authors would like to thank Dr. Alvis Fong and Dr. Pnina Ari-Gur for their valuable suggestions in the development of this work. Furthermore, we thank the anonymous reviewers for their valuable feedback and suggestions.

References.

1. Usama, M., et al.: Unsupervised machine learning for networking: techniques, applications and research challenges. IEEE Access. **7**, 65579–65615 (2019)
2. Yu, L.-C., et al.: Refining word embeddings using intensity scores for sentiment analysis. IEEE/ACM Trans. Audio Speech Lang. Process. **26**(3), 671–681 (2018)
3. Manning, C.D., Raghavan, P., Schütze, H.: Introduction to Information Retrieval, Cambridge University Press, vol. 39 (2008)
4. Chen, W., et al.: Distributed feature representations for dependency parsing. IEEE/ACM Trans. Audio, Speech Lang. Process. **23**(3), 451–460 (2015)
5. Ouchi, H., et al.: Transition-based dependency parsing exploiting supertags. IEEE/ACM Trans. Audio, Speech, Lang. Process. **24**(11), 2059–2068 (2016)
6. Shen, M., et al.: Dependency parse reranking with rich subtree features. IEEE/ACM Trans. Audio Speech Lang. Process. **22**(7), 1208–1218 (2014)
7. Zhou, G., et al.: Learning the multilingual translation representations for question retrieval in community question answering via non-negative matrix factorization. IEEE/ACM Trans. Audio Speech Lang. Process. **24**(7), 1305–1314 (2016)
8. Hao, Y., et al.: An end-to-end model for question answering over knowledge base with cross-attention combining global knowledge. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). (2017)
9. Zhang, B., et al.: A context-aware recurrent encoder for neural machine translation. IEEE/ACM Trans. Audio Speech Lang. Process. **25**(12), 2424–2432 (2017)
10. Chen, K., et al.: A neural approach to source dependence based context model for statistical machine translation. IEEE/ACM Trans. Audio Speech Lang. Process. **26**(2), 266–280 (2018)

11. Sun, F., et al.: Learning word representations by jointly modeling syntagmatic and paradigmatic relations. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers) (2015)
12. Lai, S., et al.: How to generate a good word embedding. IEEE Intell. Syst. **31**(6), 5–14 (2016)
13. Yin, W., Schütze, H.: Discriminative phrase embedding for paraphrase identification. In: Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (2015)
14. Li, Y., Yang, T.: Word Embedding for Understanding Natural Language: a Survey. Studies in Big Data, pp. 83–104 (2017)
15. Mellet, J., Pepper, M.S.: A COVID-19 vaccine: big strides come with big challenges. Vaccines. 9(1), 39 (2021)
16. Beck, B.R., et al.: Predicting commercially available antiviral drugs that may act on the novel coronavirus (SARS-CoV-2) through a drug-target interaction deep learning model. Comput. Struct. Biotechnol. J. **18**, 784–790 (2020)
17. Zhavoronkov, A., et al.: Deep learning enables rapid identification of potent DDR1 kinase inhibitors. Nat. Biotechnol. **37**(9), 1038–1040 (2019)
18. Keshavarzi Arshadi, A. et al.: Artificial Intelligence for COVID-19 Drug Discovery and Vaccine Development. Frontiers in Artificial Intelligence. 3, (2020).
19. Center for Biologics Evaluation and Research: Vaccine Development – 101. https://www.fda.gov/vaccines-blood-biologics/development-approval-process-cber/vaccine-development-101
20. Mikolov, T., Chen, K., Corrado, G., et al.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)
21. ElsevierDev: ElsevierDev/elsapy. https://github.com/ElsevierDev/elsapy. Accessed 22 Apr 2021
22. gensim: topic modelling for humans. https://radimrehurek.com/gensim_3.8.3/index.html. Accessed 22 Apr 2021
23. Tshitoyan, V., et al.: Unsupervised word embeddings capture latent knowledge from materials science literature. Nature **571**(7763), 95–98 (2019)
24. Yang, Z., et al.: An in silico deep learning approach to multi-epitope vaccine design: a SARS-CoV-2 case study. Sci. Rep. **11**, 1 (2021)
25. "Google Code Archive - Long-term storage for Google Code Project Hosting". code.google.com. Retrieved 22 October 2020.
26. Strandqvist, W.: Neural Networks for Part-of-Speech Tagging (Dissertation) (2016). http://urn.kb.se/resolve?urn=urn:nbn:se:liu:diva-129296
27. Rong, X.: Word2vec Parameter Learning Explained. ArXiv Preprint ArXiv:1411.2738 (2014)
28. Fast, E., Altman, R.B., Chen, B.: Potential t-cell and b-cell epitopes of 2019-ncov. (2020). Goldberg, Y. and Levy, O., "word2vec Explained: deriving Mikolov et al.'s negative-sampling word-embedding method", arXiv e-prints (2014)
29. Chen, B., et al.: Predicting HLA class II antigen presentation through integrated deep learning. Nat. Biotechnol. **37**, 1332–1343 (2019)
30. Jurtz, V., Paul, S., Andreatta, M., Marcatili, P., Peters, B., Nielsen, M.: NetMHCpan-4.0: Improved Peptide–MHC Class I Interaction Predictions Integrating Eluted Ligand and Peptide Binding Affinity Data. J. Immunology **199**, 3360–3368 (2017)
31. Crossman, L.C.: Leveraging Deep Learning to Simulate Coronavirus Spike proteins has the potential to predict future Zoonotic sequences (2020)
32. Abbasi, B.A., Saraf, D., Sharma, T., Sinha, R., Singh, S., Gupta, P., Sood, S., Gupta, A., rawal, kamal: Identification of vaccine targets & design of vaccine against sars-cov-2 coronavirus using computational and deep learning-based approaches. (2020).

33. Schnabel, T., Labutov, I., Mimno, D., Joachims, T.: Evaluation methods for unsupervised word embeddings. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (2015)
34. The different types of COVID-19 vaccines. https://www.who.int/news-room/feature-stories/detail/the-race-for-a-covid-19-vaccine-explained
35. Pollet, J., Chen, W.-H., Strych, U.: Recombinant protein vaccines, a proven approach against coronavirus pandemics. Adv. Drug Deliv. Rev. **170**, 71–82 (2021)
36. Wee, S.-lee, Qin, A.: China Approves Covid-19 Vaccine as It Moves to Inoculate Millions, https://www.nytimes.com/2020/12/30/business/china-vaccine.html.
37. Corum, J., Zimmer, C.: How the Sinovac Vaccine Works. https://www.nytimes.com/interactive/2020/health/sinovac-covid-19-vaccine.html
38. Logunov, D.Y., et al.: Safety and efficacy of an rAd26 and rAd5 vector-based heterologous prime-boost COVID-19 vaccine: an interim analysis of a randomised controlled phase 3 trial in Russia. The Lancet. **397**, 671–681 (2021)
39. Bharat Biotech-Vaccines & Bio-Therapeutics Manufacturer in India. https://www.bharatbiotech.com/covaxin.html. Accessed 22 Apr 2021
40. Understanding Viral Vector COVID-19 Vaccines. https://www.cdc.gov/coronavirus/2019-ncov/vaccines/different-vaccines/viralvector.html?CDC_AA_refVal=https%3A%2F%2Fwww.cdc.gov%2Fvaccines%2Fcovid-19%2Fhcp%2Fviral-vector-vaccine-basics.html. Accessed 22 Apr 2021
41. Codagenix Home. https://codagenix.com/. Accessed 27 Apr 2021
42. Dutta, D.S.S.: What are Adenovirus-Based Vaccines? https://www.news-medical.net/health/What-are-Adenovirus-Based-Vaccines.aspx
43. Modjarrad, K.,et al.: Safety and immunogenicity of an anti-Middle East respiratory syndrome coronavirus DNA vaccine: a phase 1, open-label, single-arm, dose-escalation trial. Lancet Infectious Diseases **19**, 1013–1022 (2019)
44. China's mRNA COVID-19 vaccine may start late-stage trial in May - state media. https://www.reuters.com/business/healthcare-pharmaceuticals/chinas-mrna-covid-19-vaccine-may-start-late-stage-trial-may-state-media-2021-04-13/. Accessed 27 Apr 2021
45. COVID-19 S-Trimer (SCB-2019) Vaccine. https://www.precisionvaccinations.com/vaccines/covid-19-s-trimer-scb-2019-vaccine. Accessed 28 Apr 2021
46. Celonic and CureVac Announce Agreement to Manufacture over 100 Million Doses of CureVac's COVID-19 Vaccine Candidate, CVnCoV. https://www.curevac.com/en/2021/03/30/celonic-and-curevac-announce-agreement-to-manufacture-over-100-million-doses-of-curevacs-covid-19-vaccine-candidate-cvncov/. Accessed 28 Apr 2021
47. Ascending Dose Study of Investigational SARS-CoV-2 Vaccine ARCT-021 in Healthy Adult Subjects – Full Text View. https://www.clinicaltrials.gov/ct2/show/NCT04480957. Accessed 2 May 2021
48. Study of Recombinant Protein Vaccine Formulations Against COVID-19 in Healthy Adults 18 Years of Age and Older - Full Text View. https://www.clinicaltrials.gov/ct2/show/NCT04537208. Accessed 2 May 2021
49. Philippidis, A.: Genexine - GX-19. https://www.genengnews.com/covid-19-candidates/genexine-gx-19/. Accessed 2 May 2021
50. Dey, A., Chozhavel Rajanathan, T.M., Chandra, H., Pericherla, H.P.R., Kumar, S., Choonia, H.S., Bajpai, M., Singh, A.K., Sinha, A., Saini, G., Dalal, P., Vandriwala, S., Raheem, M.A., Divate, R.D., Navlani, N.L., Sharma, V., Parikh, A., Prasath, S., Rao, S., Maithal, K.: Immunogenic Potential of DNA Vaccine candidate, ZyCoV-D against SARS-CoV-2 in Animal Models. (2021).
51. Ella, R., et al.: Safety and immunogenicity of an inactivated SARS-CoV-2 vaccine, BBV152: a double-blind, randomised, phase 1 trial. Lancet. Infect. Dis **21**, 637–646 (2021)

52. Pizza, M., Bekkat-Berkani, R., Rappuoli, R.: Vaccines against Meningococcal Diseases. Microorganisms. **8**, 1521 (2020)
53. Commissioner, O.of the: Coronavirus (COVID-19) Update: FDA Authorizes Monoclonal Antibodies for Treatment of COVID-19. https://www.fda.gov/news-events/press-announcements/coronavirus-covid-19-update-fda-authorizes-monoclonal-antibodies-treatment-covid-19-0. Accessed 2 May 2021
54. Miller, K.: Merck Oral COVID-19 Drug Shows Promise in Early Trials. https://www.verywellhealth.com/merck-oral-covid-19-drug-clinical-trial-5115909. Accessed 2 May 2021
55. BGB-DXP593. https://go.drugbank.com/drugs/DB16357. Accessed 2 May 2021
56. A Study on the Safety, Tolerability and Immune Response of SARS-CoV-2 Sclamp (COVID-19) Vaccine in Healthy Adults - Full Text View. https://www.clinicaltrials.gov/ct2/show/NCT04495933
57. Dhama, K., et al.: Plant-based vaccines and antibodies to combat COVID-19: current status and prospects. Hum. Vaccin. Immunother. **16**, 2913–2920 (2020)
58. The Future of Genetic Engineering. https://www.greffex.com/. Accessed 2 May 2021
59. New Data from Vaxart Oral COVID-19 Vaccine Phase I Study Suggests Broad Cross-Reactivity against Other Coronaviruses. https://investors.vaxart.com/news-releases/news-release-details/new-data-vaxart-oral-covid-19-vaccine-phase-i-study-suggests. Accessed 2 May 2021
60. Phase II / III Study of COVID-19 DNA Vaccine (AG0302-COVID19) - Full Text View. https://www.clinicaltrials.gov/ct2/show/NCT04655625. Accessed 2 May 2021
61. Efficacy and Safety of TY027 a Treatment for COVID-19 in Humans. https://www.centerwatch.com/clinical-trials/listings/259289/efficacy-and-safety-of-ty027-a-treatment-for-covid-19-in-humans/. Accessed 2 May 2021
62. Safety and Immunogenicity of AdCOVID in Healthy Adults (COVID-19 Vaccine Study) - Full Text View. https://clinicaltrials.gov/ct2/show/NCT04679909. Accessed 4 May 2021
63. Gharaibeh, T., de Doncker, E.: Unsupervised Learning with Word Embeddings Captures Knowledge from COVID-19 Literature. CSCI 2020 (Dec. 2020), IEEE CPS, Accepted
64. Caselles-Dupré, H., Lesaint, F., Royo-Letelier, J.: Word2vec applied to recommendation. In: Proceedings of the 12th ACM Conference on Recommender Systems. (2018)
65. Yildiz, B., Tezgider, M.: Learning quality improved word embedding with assessment of hyperparameters. In: Schwardmann, U., et al. (eds.) Euro-Par 2019. LNCS, vol. 11997, pp. 506–518. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-48340-1_39