International Journal of Social Science and Economic Research

ISSN: 2455-8834

Volume:07, Issue:07 "July 2022"

A NOVEL MACHINE LEARNING APPROACH TO ACCUMULATE KNOWLEDGE CONCERNING COVID-19 OVER TIME

Nihal Boina, Jai Agarwal, Taruna Agrawal, John Leddo, Kinnari Chaubal, Deepika Ravi, Mason Elkas, SudarsanNallabola, SmaranPasupuleti, Riya Pasupuleti, Dhanush Kundur, Dillon Michlena, Avyakt Gaur, SathvikRedrouthu, Poorvaja Gopinath, Vaneesha Gupta, HridaySainathuni, Nate Levkov, SudhitSangela, Saranya Ganne, NikhitRachapudi, CharanNarra, Ananya Dandemraju, Aneesh Sreedhara, SakethChintalapati, Parth Patel, Jishnu Patel, Manohar Nookala, SharanyaChilukuri, RithvikDandemraju, Vihaan Cherukuri, JagadeepramMaddipatla, Satvik Matta, Surya Vallamkonda, Shashank Varma, Rohan Suri, Manav Sabharwal, Swetha Rajan, HavishRallabandi, Vikram Rudraju, SonithSunku, Ranveer Kataria, PrishaGattu, Samarth Jain and Nithya Jayakaran

MyEdMaster, LLC, 13750 Sunrise Valley Drive, Herndon, VA, United States of America

John Leddo is the director of research at MyEdmaster.

DOI: 10.46609/IJSSER.2022.v07i07.018 URL: https://doi.org/10.46609/IJSSER.2022.v07i07.018

Received: 3 July 2022 / Accepted: 15 July 2022 / Published: 28 July 2022

ABSTRACT

COVID 19, commonly known as coronavirus is a cold-like virus that can spread through droplets from an infected person's coughs, sneezing, and/or breathing. Symptoms of COVID 19 include cough, fatigue, fever, shortness of breath, body aches, headaches, loss of taste and smell, and more. Some severe symptoms include acute pain in the chest area, pale skin, trouble breathing. COVID 19 is said to have originated in Wuhan, China as the first infections were discovered there. Many methods have been developed To give citizens updated statistics on rising issues, providing easy access to global information. Knowledge graphs are one of many ways professionals and regular citizens of the global network get to know the effects of COVID-19 in their society. This is one of many ways the ANNs (Artificial Neural Networks) takes place in promoting the analysis of knowledge graphs to everyone. Part of the ANN includes how a person's mental physique operates. For example, the input goes through a theoretical process of analysis to give the most accurate output. In this case, the effects of COVID even after students

ISSN: 2455-8834

Volume:07, Issue:07 "July 2022"

are back to school can be seen as a continuity in the environment. ANNs and general graphs play of great importance in this scenario.

Keywords: General artificial intelligence, Covid-19, deep learning

Introduction

In the present day, COVID-19 has affected the lives of people resulting in more than 450 million infections and over 6 million deaths around the world. Depending on the severity, COVID can cause diarrhea, fatigue, headaches, new loss of taste or smell, body or muscle aches, coughing, fever chills, nausea, and other such symptoms [1]. The COVID-19 pandemic clearly affected the human population on a large scale during the three years of its discovery. Social distancing, quarantining, and virtual learning changed the mental processing for not only patients and students, but also adults who lost habit to their regular day to day tasks before the pandemic started [2]. Parts of this include writing habits, productivity, and stress levels due to the lack of physical interaction [3].

Although now much is known about COVID-19, during the time COVID-19 was present there were numerous cases of scientific inaccuracies and misinformation [4]. For example, while it is now known that COVID-19 is airborne, this was not thought to be true by the general population during the past few years. On a broader level, in the scientific world there exists many instances of misinformation. A major issue of science is how to take into account differing sources and research papers to bring about a holistic knowledge-based view of the current state.

Artificial intelligence has been demonstrated to complete complex, language-based tasks before, an example being OpenAI's GPT-2's high accuracy in predicting the next word in 40GB of Internet text. One of the primary objectives of Artificial Intelligence is to make Artificial General Intelligence. Artificial General Intelligence, or AGI for short, is a hypothetical intelligent software that can learn any intellectual task a human can [5]. One of the most complex and important human actions is the consolidation of knowledge. Through the use of knowledge graphs, a structured tree-based representation of facts, relationships between various objects can be noted.

A knowledge base is any structure that stores complex structured and unstructured information in a computer system. The fundamental difference between a knowledge base and a database is that entries in a database are opaque tokens, whereas entries in a knowledge base are interrelated [6]. This is why knowledge bases are commonly referred to as gray boxes, whereas databases are referred to as black boxes. One way to construct a knowledge base is with knowledge graphs to

ISSN: 2455-8834

Volume:07, Issue:07 "July 2022"

represent the relationship between various topics. A knowledge graph represents a network of real-world entities in a node format, with diverse relationships between each related node. However, knowledge graphs -like any form of knowledge structure- are limited in certain areas.

One of the largest drawbacks to knowledge graphs is that these knowledge structures scale enormously, with increasing time and storage complexity. As a knowledge graph is grown, more nodes and connections are traversed, and more nodes and connections are added, so there is the possibility for there to be a dynamic programming component to ensure algorithmic optimization.

Deep learning is an important branch of machine learning, and has had numerous applications leading to feats that reach and sometimes exceed human performance in many different areas. One popular type of machine learning model is K-means clustering. K-means clustering is a method of separating graphical data points into k clusters with respect to the nearest mean. K-means is used to find groups that have not been explicitly labeled within a particular dataset, and for that reason is known as an unsupervised machine learning algorithm.

Artificial intelligence (AI) involves the training and testing of a computational model, or several models. An artificial neural network, ANN, consists of neurons sorted in layers that hold activation numbers, weights, and biases that are tweaked through a process called backpropagation in order to fit the training data and then tested through another set of data. Neural networks are a fundamental part of AI and are inspired by the way biological brains work [7]. Neural networks typically require large amounts of data and training parameters. Because neural networks are efficient at modeling highly complex data between numerous parameters, they provide a possible means to consolidate knowledge from a specified source.

An important concept in understanding the way ANNs work is the black box. The basic idea is that while a neural network can approximate any function, studying the actual values of each node in a neural network will not give any insight on the structure of the function being identified. In other words, neural networks are great at matching a dataset or specific problem, but not in understanding why it's working the way that it is. For this reason, a neural network in a knowledge-based system would have to be trained on outputting abstractions in order to automate the acquisition of knowledge.

One of the largest knowledge-base efforts to date is CYC. The CYC project began in 1984, and the purpose of this project was to build a knowledge base containing a significant percentage of all "common-sense" knowledge of a human being. For the past 40 years, researchers at CYC

ISSN: 2455-8834

Volume:07, Issue:07 "July 2022"

worked to hardcode millions of symbolic assertions and rules in order to map out all commonsense knowledge. Originally, it was projected that only a million assertions were needed in order to accomplish this hefty goal, but as time went on the researchers at CYC found out that they were wrong by a factor of more than 100. CYC is still ongoing, with hundreds of thousands of human hours put into hardcoding all of this information [6]. However, in hindsight it appears that CYC was and is mistaken for one fundamental idea: that human knowledge and common sense is constantly evolving. Thus, it makes more sense to make a dynamic, morphing knowledge base that can potentially automate itself rather than making a solid, set-in-stone knowledge base that is expected to last forever.

In Wang et al, researchers were able to make a generalized machine learning framework to extract knowledge trees from specific sentences [8]. The researchers were able to do this by using the popular pre-trained NLP models BERT & GPT-⅔ to recognize the head, tail, and relation of a particular line. Then, a 1 node to 1 node connected "knowledge segment" was formed. Dynamic programming was then utilized to put this one segment in terms with an existing knowledge base by looking at the other nodes and relations. There are certain limitations to this approach: 1) there is no actual machine learning used to generate the line segments, and there is no incentive for the model to actually learn how to produce knowledge trees in a more optimized/better/larger manner, 2) there is no system put into place for combining whole knowledge trees with other knowledge trees, and 3) there is no machine learning framework for optimizing knowledge tree construction.

In Tseng & Lin, researchers were able to use Machine Learning to extract knowledge trees for computer science education from stackoverflow [9]. This article provides a potential machine learning framework to use in order to extract knowledge from a particular website. Through using a K-means clustering algorithm alongside a hierarchical sort, the researchers were able to construct knowledge bases to a limited extent from stack overflow pages. These researchers concluded that an array of current machine learning methods alongside an optimized algorithm has much to offer in extracting knowledge.

In this paper, a system of various machine learning components that can reliably output and update knowledge graphs is needed for this key process to work. In order to do this, several ANN & GNN-based generators were used in order to filter raw text into commands to then be used to construct and update knowledge graphs as a whole using a dynamic-programming based method.

ISSN: 2455-8834

Volume:07, Issue:07 "July 2022"

Software components

The software utilizes python and imports in order to function, as listed below:

1. Pandas
2. Tensorflow
3. Numpy
4. Scikit-learn
5. Spacy
6. NLTK

Technology

The software starts out by first accepting user input as to what date to search information up to. Then, using Selenium and searching only for information relating to COVID-19 up to that point of time, the software is able to simulate and train as if it was on that given date. Then, the software starts out by querying a series of pre-set questions. From here, the websites suggested by the Google search browser accessed through Selenium are used as raw text to be put through a series of NLP filters and processors. Then, the resulting text is used with a combination of dynamic programming and K-means clustering to construct a series of node-to-node connections and eventually an initial knowledge base. Then, the software goes through each connection made in the initial knowledge base and verifies each statement with the open-source web browser (given date constraints). Specific websites that were given a higher pre-built level of "trust" were those ending with the *.org, .edu, and .gov*. Upon verifying each node-to-node connection, more connections are made through additional processing and sorting using DP. The software is able to learn and become better at putting together knowledge bases by being given explicit data identifying overall summaries/brief descriptions and preidentified misinformation.

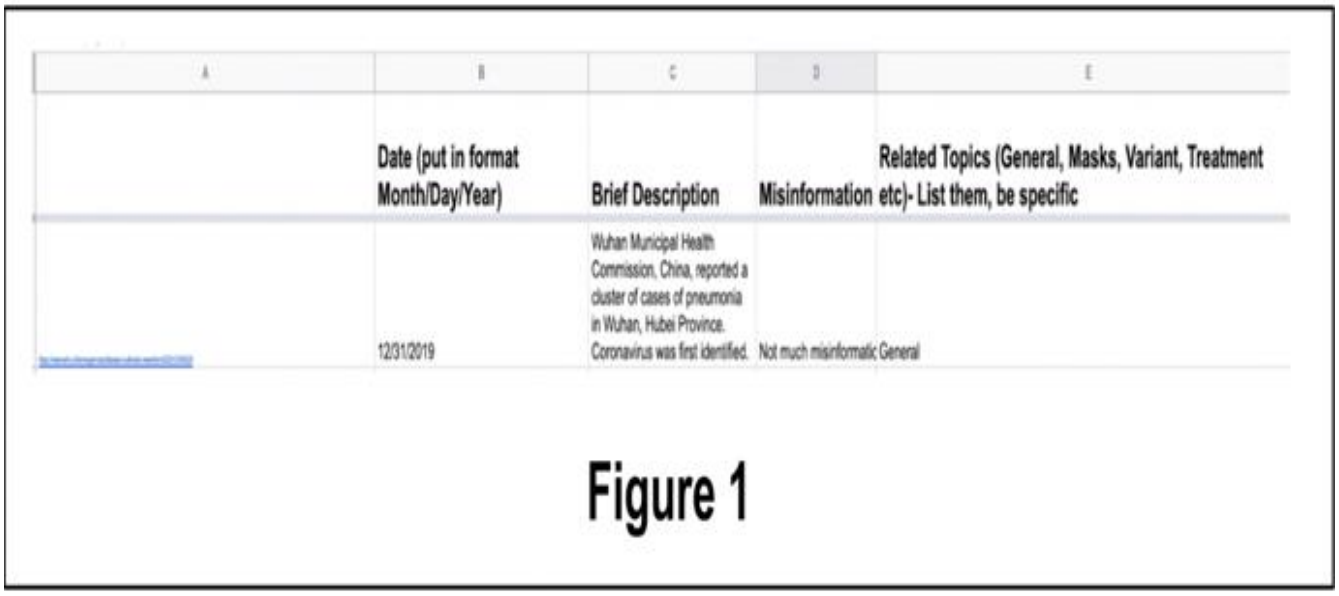
Dataset

For the dataset portion of this study, an original source had to be compiled for the model to train on. This dataset consists of various websites linked from various points of time of Covid-19's existence and the relevance of each source listed. Through doing this, a constructive dataset was able to be made consisting of 300 entries. For each entry, there were 5 components: a date for the source, a website link to the source, raw text from the source, a key summary for the source, and

ISSN: 2455-8834

Volume:07, Issue:07 "July 2022"

additional relevance into how this source matters. An example for this dataset structure is shown in Figure 1.



	Date (put in format Month/Day/Year)	Brief Description	Misinformation	Related Topics (General, Masks, Variant, Treatment etc)- List them, be specific
	12/31/2019	Wuhan Municipal Health Commission, China, reported a cluster of cases of pneumonia in Wuhan, Hubei Province. Coronavirus was first identified.	Not much misinformatic	General

Figure 1

Algorithm

A similar method to Folsonomy is used in the finding of the same repetitive words gathered from the websites. Node to node relationships within buzzwords are used to create a knowledge graph. Artificial Neural Networks (ANN) are also be used to create a knowledge graph that is based out of the dataset of the different types of websites that were gathered. The app redeems the information that is gathered from the 137 websites and gives the best overall thought out process as the output to the input which was given from the app's user. Of course, the question can only be answered if it is relevant to the medical sciences of COVID-19.

Specifically, an ANN and GNN are generally used for identifying specific textual information and generating knowledge graphs. Also, ADAM is used as an optimizer function and ReLU as the activation function, with various parameters changing with each iteration to achieve the best randomized testing results with eachtrained model.

ISSN: 2455-8834

Volume:07, Issue:07 "July 2022"

Results

Table 1: Cosine Similarity for Mod

Cosine Similarity
0.97

The table above depicts the averaged Cosine similarity for the model's predicted knowledge graph output vs. the expected.

Conclusion

Our results show that high testing accuracies were obtained by our software, with the model having a relatively high cosine similarity value of 0.97. The next step for this project is to do more finetuning, to fit a variety of cases, and then implement in real life. Specifically, the software must be tested on knowledge relating to medical ailments extending beyond primarily COVID-19 and the ability to encounter diseases that are completely unfamiliar.

References

[1]. De, Priyasha, et al. "Brief Review on Repurposed Drugs and Vaccines for Possible Treatment of COVID-19." European Journal of Pharmacology, vol. 898, May 2021, p. 173977. ScienceDirect, https://doi.org/10.1016/j.ejphar.2021.173977.

[2]. Meo, Sultan Ayoub, et al. "Impact of Lockdown on COVID-19 Prevalence and Mortality during 2020 Pandemic: Observational Analysis of 27 Countries." European Journal of Medical Research, vol. 25, no. 1, Nov. 2020, p. 56. BioMed Central, https://doi.org/10.1186/s40001-020-00456-9.

[3]. Wood, Heather. "New Insights into the Neurological Effects of COVID-19." Nature Reviews. Neurology, vol. 16, no. 8, 2020, p. 403. PubMed Central, https://doi.org/10.1038/s41582-020-0386-7.

ISSN: 2455-8834

Volume:07, Issue:07 "July 2022"

[4]. Xydakis, Michael S., et al. "Post-Viral Effects of COVID-19 in the Olfactory System and Their Implications." The Lancet Neurology, vol. 20, no. 9, Sept. 2021, pp. 753–61. ScienceDirect, https://doi.org/10.1016/S1474-4422(21)00182-4.

[5]. Hodson, H. (2019, March 1). DeepMind and Google: The battle to control artificial intelligence. The Economist. https://www.economist.com/1843/2019/03/01/deepmind-and-google-the-battle-to-control-artificial-intelligence

[6]. Lex Fridman. (2021, September 15). Douglas Lenat: Cyc and the Quest to Solve Common Sense Reasoning in AI | Lex Fridman Podcast #221. YouTube. https://www.youtube.com/watch?v=3wMKoSRbGVs&ab_channel=LexFridman

[7]. DeMuro, J. (2019, December 17). What is a neural network? TechRadar. https://www.techradar.com/news/what-is-a-neural-network

[8]. Wang, Chenguang, et al. "Language Models Are Open Knowledge Graphs." ArXiv:2010.11967 [Cs], Oct. 2020. arXiv.org, http://arxiv.org/abs/2010.11967.

[9]. Tseng, C.H., & Lin, J.R. (2020). A semi-hierarchical clustering method for constructing knowledge trees from stackoverflow. Journal of Information Science, 1-13.

