COVID-19 Candidate Treatments, a Data Analytics Approach

Gerry Wolfe
Department of Engineering & Computer Science
Syracuse University
Bismarck, ND
gwolfe@syr.edu

Ashraf Elnashar
Department of Engineering & Computer Science
Syracuse University
Irvine, CA
aselnash@syr.edu

Will Schreiber
Department of Engineering & Computer Science
Syracuse University
Chicago, IL
wschreib@syr.edu

Izzat Alsmadi
Department of Computing and Cyber Security
Texas A&M,
San Antonio, TX
ialsmadi@tamusa.edu

***Abstract*— COVID-19, short for "coronavirus disease 2019", has majorly affected millions of people worldwide. In the U.S. alone as of the end of this week (June 1, 2020), there have been 1,790,191 total cases, with 104,383 deaths. There have been 6,166,978 cases in the entire world, with 372,037 deaths, these are just the reported cases. Our focus in this research is in evaluating a repository of research papers to extract knowledge related to COVID-19 and possible treatments. Driven by the COVID-19 Open Research Dataset Challenge from Kaggle, we focused on a subset of that, COVID-19 Pulmonary Risks Literature Clustering. The second dataset we are using is from the Maryland Transportation Institute (MTI). The data is broken up into four categories: (1) Mobility and Social Distancing, (2) COVID and Health, (3) Economic Impact, and (4) Vulnerable Population. The data is extracted from NPR, ESRI, the COVID tracking project, CDC, and several other sources. MTI has been the source of several papers regarding mobility impact, social distancing, stay-at-home orders, and non-pharmaceutical interventions.**

Keywords*— *COVID-19, Coronavirus, Risk Factors, Pulmonary Disease

I. Introduction

COVID-19 is the World Health Organization's official name for the disease caused by this newly identified coronavirus. Coronaviruses are a widespread cause of colds and other upper respiratory infections. The most up-to-date information is available from the World Health Organization, the US Centers for Disease Control and Prevention, Johns Hopkins University, and Maryland Transportation Institute ("MTI").

It has spread so rapidly and to so many countries that the World Health Organization has declared it a pandemic (a term indicating that it has affected a large population, region, country, or continent).

This project focuses on two aspects of analyzing the data surrounding COVID-19, specifically, text analysis of literature relating to COVID-19 research from more than 49,000 articles and the data from the MTI, which is summarized on its website as can be linked above.

The two datasets are quite different. The first dataset is unsupervised while the second dataset is supervised. What we were able to accomplish was bridging the gap between the two datasets. Precisely, we translated our unsupervised, unstructured text dataset into a supervised, structured dataset that allowed us to classify against a target.

II. Goals and Research Questions

A. Research Questions

Given a large amount of literature and the rapid spread of COVID-19, the literature has not been effectively organized. This project is meant to help organize the literature related to pulmonary diseases and their effect on COVID-19.

We will approach this project by modeling the following three areas.

(1) This paper intends to cluster the articles surrounding related pulmonary diseases to help researchers better determine possible trends, and opportunities of research. Specifically, rather than use a word-based feature, we used synonym based features structured as arrays to help us better determine possible trends, etc. This allowed us to better focus on quantifying this data so we can help determine the important factors in the tidal wave of articles. By using this 'out of box' synonym array as a feature, we uncovered some trends related to pulmonary diseases.

(2) We were also interested in some dynamics related to human mobility and their impact on COVID-19 cases. We used MTI dataset related to transportation and social distancing.

(3) Lastly, we wanted to take our unstructured text cluster and our structured data target and compare our accuracy results.

B. Approach

Fig. 1 below summarizes our overall project activities. Those replicate mostly standard data analytic activities in text-based data analytic projects.

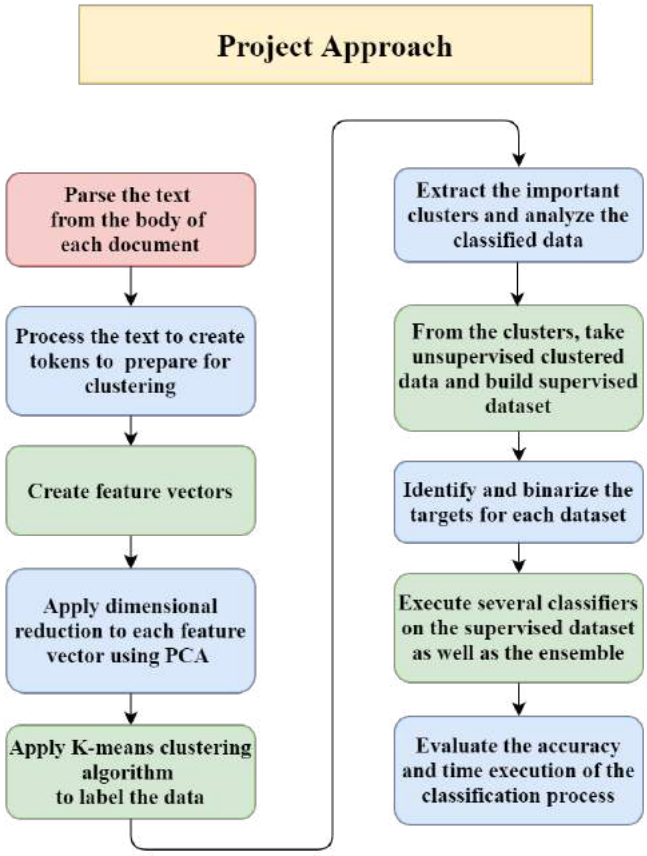

Figure 1 - COVID-19 overall data analytic activities

Our project includes a large input dataset of hundreds of documents. Throughout this process, the data is processed and stored in temporary memory files to ensure consistent data, fast iterations, and frequent testing with differing parameters.

C. Datasets Description

1. COVID-19 Open Research Dataset: In response to the COVID-19 pandemic, the White House and a coalition of leading research groups have prepared the COVID-19 Open Research Dataset (CORD-19). CORD-19 is a resource of over 51,000 scholarly articles, including over 40,000 with full text, about COVID-19, SARS-CoV-2, and related earlier coronaviruses. This freely available dataset is provided to the global research community to apply recent advances in natural language processing and other Artificial Intelligence (A.I.) techniques to extract new insights in support of the ongoing fight against this infectious disease. There is a growing urgency for these approaches because of the rapid acceleration in new coronavirus literature, making it difficult for the medical research community to keep up.

2. MTI Dataset: Researchers at the University of Maryland (UMD) are exploring how social distancing and stay-at-home orders are affecting mobility and travel behavior and the spread of the coronavirus. With privacy-protected data from mobile devices, government agencies, health care systems, and other sources, they are also studying the multifaceted impact of COVID-19 on our mobility, health, economy, and society.

III. Literature Review

This section provides summaries of nine subject matter related articles to build a framework of understanding for this paper.

X. Wang[10] focused on the concept that natural language processing requires tokenized words and phrases to evaluate and predict data outcomes. The CORD dataset has very specialized terms and phrases that are difficult for standard word embedding libraries to tokenize. Even SciSpacy (a popular NLP dictionary) has difficulty recognizing all the tokens that should be created. CORD-NER is an extensible NLP dictionary that can be used to take greater advantage of the CORD journals. As the CORD-19 dataset is updated, the CORD-NER dictionary will follow suit as well as having the ability to add additional entity types.

R. Tang[9] proposed a proof-of-concept providing a Style Question Answer Dataset (SQuAD), where the input is a query with scientific articles. The proof of concept is to provide the answer within the document with a percentage associated with its accuracy. They used CovidQA, which has 124 question-article pairs.

Our analysis makes it possible to build our own SQuAD based on risk factors to help researchers quickly ascertain the best articles to review, depending on the query.

For example, we could ask, "what is the incubation period of the virus?" or "what do we know about viral shedding in urine?"

T. Huang[14] focused on introducing COVID-19 Research Aspect Dataset (CODA-19), a human-annotated dataset. CODA-19 was created by 248 crowd workers from the Amazon Mechanical Turk program. The purpose of this study was to determine if non-experts could accurately label crowdsourcing labels on scientific publications related to COVID-19. They found that the crowd's accuracy in labeling was 82.2% compared with 85% accuracy for experts. The conclusion is that crowds can be rapidly employed to help research COVID-19 related issues, providing short-term support before machine learning models can be implemented.

H. Kroll[13] took two biomedical named entity recognition libraries, namely TaggerOne and GNormPlus, and created a new dictionary to analyze the CORD-19 dataset. The articles' metadata were then tagged with the entity types that best matched the article.

For our dataset, we tagged the article's JSON with tags related to risk factors rather than entity types. As an example, a journal article may be tagged with "smoking, diabetes, marijuana." From there, the tags can be counted, compiled, analyzed, and evaluated to determine further trends and other research. This would help researchers because they could count on the tags.

V. Kieuvongngam[11] used pre-trained NLP models, specifically Bidirectional Encoder Representations from Transformers (BERT) and OpenAI GPT-2, to summarize text

and generate abstracts. These are both trained using Wikipedia and other very large datasets of texts.

We would like to generate a different sort of abstracts, abstracts that focus on risk factors irrespective of the article's actual abstract.

A. Esteva[15] created a retriever-ranker semantic search engine designed to handle complex queries specifically for the CORD-19 literature. This project intended to return accurate and precise results from querying the CORD-19 dataset. Specifically, they used a Siamese-BERT encoder along with a combination of TF-IDF and BM25 vectorizers. They then used a ranker made up of a multi-hop question-answering module along with a multi-paragraph abstractive summarizer. The details of these encoder/vectorizers are not important for this summary. Instead, what is important is that the authors built an algorithm that has better semantic results, which provide researchers and health workers more accurate, trustworthy, and timely results.

We used the spaCy NLP library, which gave us vectorization and thus acted as a 'retriever'. We used a similarity score to act as our 'ranker.'

X. Guo[12] worked on semantic textual similarity. Current Semantic Textual Similarity (STS) datasets do not provide great performance for domain-specific environments. The authors here built a new STS dataset to help bridge that gap. Here's how they did it. They took 1,000,000 sentence pairs and then created their training set using BERT to calculate similarity scores. They ended up with 32,000 labeled sentence pairs, which they labeled through the use of Amazon Mechanical Turk workers. This group focused on sentences where we focused on words.

J. Homolak[16] focused on data sharing of COVID-19 related information. He found that the scientific community did not quickly opt to share data related to COVID-19.

P. Resnik[17] built curated topic models. He used a bottom-up approach by employing spaCy to tokenize and identify words and phrases to find "meaningful semantic units." But rather than focus solely on machine learning, they used subject matter experts to help better understand the model, which is a top-down approach. By using both approaches, they concluded that there should be no less than 100 topic models that provide enough fine-grained topics in the model.

We did something similar by using spaCy to tokenize and then built clusters based on K-means, which we would need to label manually with "human-in-the-loop" interaction.

Author	Hypothesis	Current State
Wang[10]	CORD-NER is a useful extensible word embedding library.	General word embedding libraries are not extensible.
Tang[9]	They proposed a proof-of-concept for a Q&A dataset using one specifically for COVID-19.	Q&A in this context has not yet been thoroughly tested.
Huang[14]	Crowdsourced labeling is similarly accurate to expert knowledge.	Experts are better able to label training data.
Kroll[13]	They combined two biomedical named entity recognition libraries to create their own dictionary to analyze the CORD-19 dataset.	Libraries are not currently able to accurately analyze the CORD-19 dataset.
Kieuvongngam [11]	It is possible to use NLP models to construct written abstracts.	Domain experts are currently needed to write abstracts.
Esteva[15]	They created a retriever-ranker semantic search engine designed to handle complex queries for CORD-19 literature with better semantic results.	Current models do not provide strong semantic results.
Guo[12]	They built a new semantic textual similarity (STS) dataset using BERT to calculate similarity scores.	Current STS datasets do not provide great performance for certain environments.
Homolak[16]	It is critically important to share data quickly on COVID-19.	The scientific community does not quickly share data.
Resnik[17]	They built curated topic models using spaCy to tokenize the words while also using subject matter experts.	Machine learning models alone do not provide the accuracy needed.

IV. EXPERIMENTS AND ANALYSIS

We used two datasets, (1) the COVID-19 dataset and (2) data from the MTI. To obtain the second dataset, crawling and scraping the site was necessary.

A. COVID-19 Dataset

We took the raw data and conducted a significant amount of cleaning and preprocessing, which included parsing, removing duplicates, and removing null values.

For the unstructured text literature data, namely the articles' text body, we removed noise, tokenized, and vectorized each word within each article to analyze the literature more accurately. Also, we removed any articles from the dataset that were not authored in the English language.

We transformed the object types from object types to more fundamental types such as string, int, and float as well as binarizing the target for further analysis. We have also done some minor feature engineering related to word count and unique words.

We needed to aggregate all the words and build a unique word count to help with tokenization. Fig. 2 shows the unique number of words across the entire dataset.

```
sn.distplot(df['body_unique_words'])
df_covid['body_unique_words'].describe()
```

```
count    32417.000000
mean      1426.309005
std       1185.091529
min          1.000000
25%        911.000000
50%       1243.000000
75%       1667.000000
max      38298.000000
Name: body_unique_words, dtype: float64
```

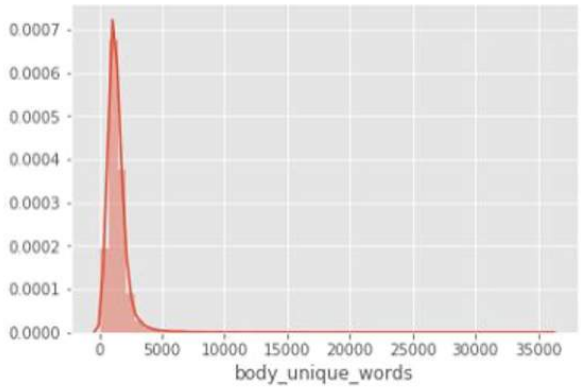

Figure 2 – Unique Word Count

A confusion matrix, shown in Fig. 3, was used to help determine which features were highly correlated and, thus, which features to continue to analyze.

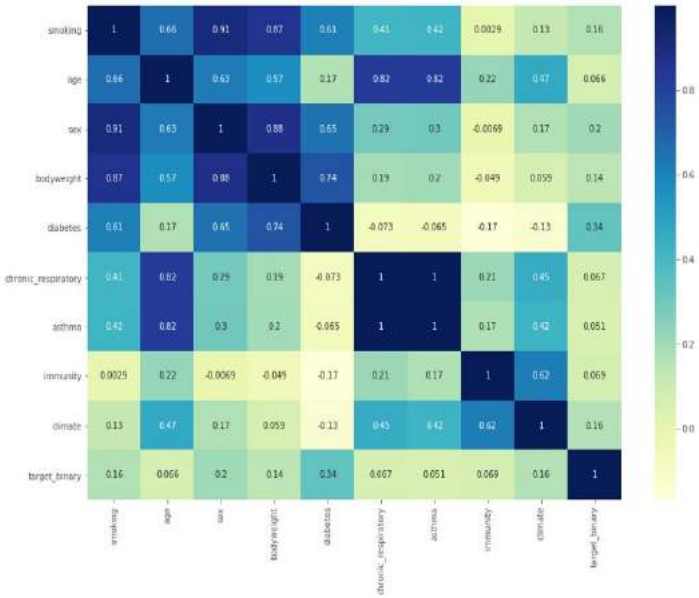

Figure 3 - Risk Factor Similarities Confusion Matrix

B. MTI Dataset

The confusion matrix in Fig. 4 for the MTI Dataset was somewhat easier to build since the data was structured.

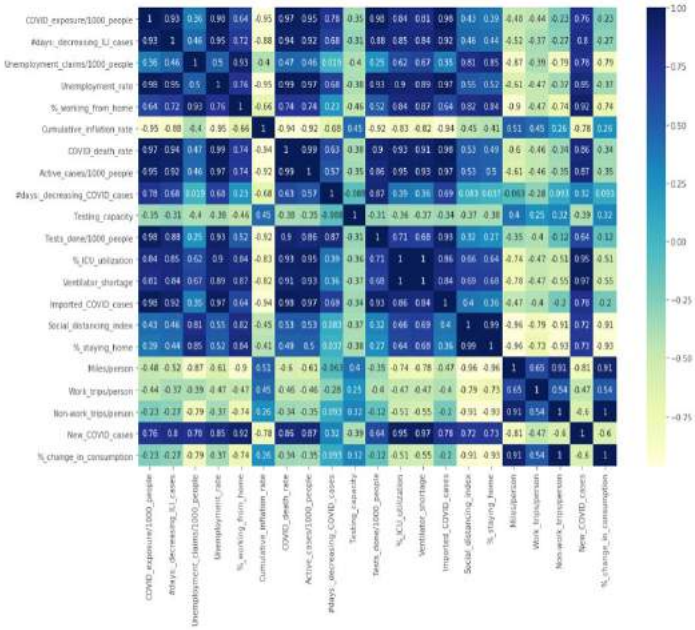

Figure 4 - Mobility and Travel Behavior Confusion Matrix

The confusion matrix for the MTI dataset was invaluable for our analysis. It helped us focus directly on the most important features.

1) Feature Selection, Reduction, and Extraction

We spent a significant amount of time in clustering, unsupervised models, supervised models, as well as extracting features. The process we went through was as follows. We took one article which contained approximately 18,000 words to build the feature selection process. We used this very small subset of data as a proof of concept because using larger amounts would have taken too much time to build the model.

We determined that 18,000 words was sufficient to run our process. For future consideration and follow-up submission, we would add the remaining 85,000+ articles to build upon.

2) Tokenization

We tokenized every word in the article using spaCy. spaCy is an API service that has pre-trained natural language processing models, including models related to scientific and medical terminology. Those models take into consideration vocabulary, semantics, and numerous other attributes. Tokenization of all the words contained in the medical journals is critical for our project.

After tokenization, we took already chosen risk factor features, for example, demographic and behavioral risk factors, age, weight, smoking, diabetic, chronic respiratory diseases, asthma, and immunity. We also analyzed generic risk factors.

We wanted to determine other highly correlative features that would help us make predictions. We used the spaCy models and built out other data frames.

We chose five features we thought were important from the data that we analyzed, and five more features were used based on the spaCy modeling. As you will be able to see from the graph

below, we uncovered a separate feature per feature that we could use.

The relationships we found were the similarities between the tokens. We calculated the maximum similarities and found interesting relationships. Fig. 5 shows the maximum token similarities. We noticed an interesting relationship between marijuana, coronavirus, cigar, and disease. We would not have thought that marijuana and coronavirus had a relationship nor that cigar and disease would have a relationship since it is widely accepted that that cigarette smoking creates more health problems than cigar smoking. This discovery deserves further attention, which is out of the scope of this project.

```
[16]: df_token_similarity['similarity_max'] = df_token_similarity.gr

[17]: index = df_token_similarity.groupby(['token1'])['similarity'].
      df_token_similarity[index]
```

[17]:

	token1	token2	similarity	similarity_max
10	smoking	finding	0.655741	0.655741
96	smoke	report	0.691356	0.691356
203	cigar	disease	0.721474	0.721474
508	nicotine	usa	0.649452	0.649452
532	cannabis	llc	0.588629	0.588629
547	marijuana	coronavirus	0.644733	0.644733

Figure 5 - Maximum Similarity Token to Token Ranking Table

As well as using the similarity method from spaCY, we also used PCA to reduce features. We used the tokenizer to transform the features, as well as to remove the stop words. We found some interesting information in our analysis. We found that there was an interesting relationship between cigar and heart, which had a similarity index of approximately 70%, and nicotine and t2, which had a similarity index of approximately 80%, as shown from the table above.

One of the issues with the literature is that we are dealing with unsupervised models.

We used the MTI dataset (with permission), and it was helpful because we wanted to test out feature selection using a supervised model. Luckily, we had built out a supervised model and, specifically, the national dataset contained in the MTI dataset.

We chose the highest 90% of the uncorrelated selected features, and we were able to achieve greater than 90-95% performance.

For this supervised model, our target was New_cases/1000_people. We feature transformed the target by binarizing the target using dataFrame.cut(). The correlation table is shown below in Fig. 6.

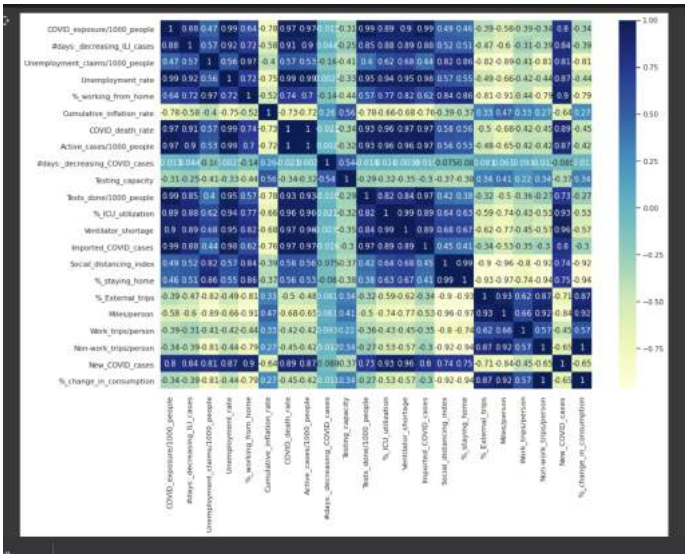

Figure 6 - Correlation Table

V. Training Methods

A. COVID-19 Dataset

We used the following training methods for our unstructured dataset.

1. NLP via the spaCy library
2. K-means

Once we had our clusters, we transformed our unstructured dataset into a structured dataset and applied all the methods described below.

We used the following training methods for our structured dataset: Perceptron, Support Vector Machines, Decision Tree, Random Forest, and Ensemble methods.

For the unstructured data, we used K-means. We needed somewhere between 7-12 clusters based on the elbow method, as shown in Fig. 7.

```
K = range(2, 20)
X_line = [K[0], K[-1]]
Y_line = [distortions_X1[0], distortions_X1[-1]]

# Plot the elbow
plt.plot(K, distortions_X1, 'b-')
plt.plot(X_line, Y_line, 'r')
plt.xlabel('k')
plt.ylabel('Distortion')
plt.title('The Elbow Method showing the optimal k BODY TEXT')
plt.show()
```

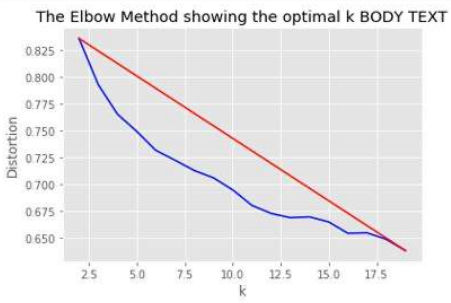

Figure 7 - Optimal K for K Means Based on Elbow Method

The Decision Tree classifier shown in Fig. 8 consistently provided the greatest accuracy score.

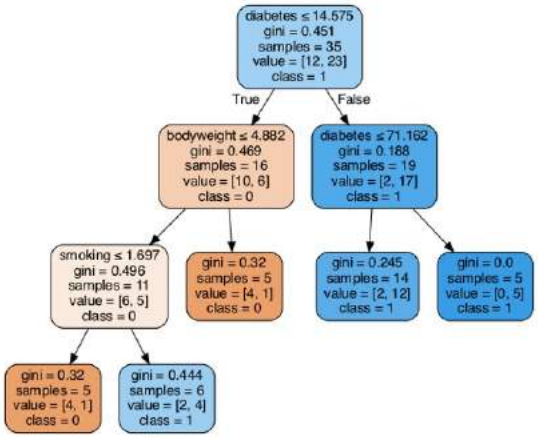

Figure 8 - Decision Tree for COVID-19 Dataset

As Fig. 9 shows, the Decision Tree's ROC stayed above the 50% line, making the Decision Tree valuable in our analysis.

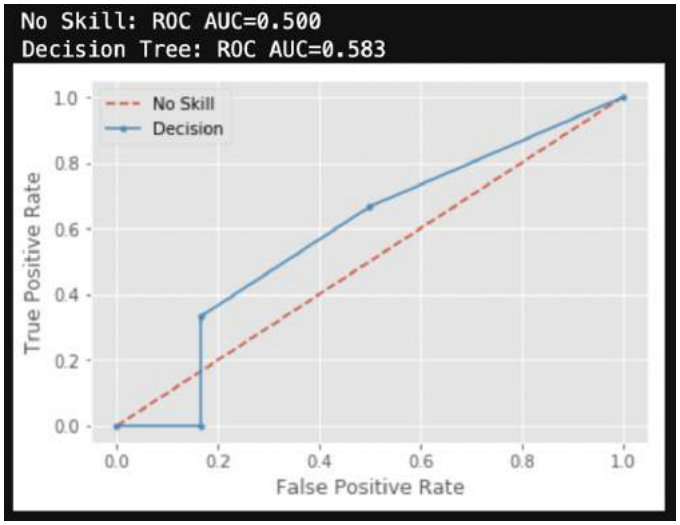

Figure 9 - Decision Tree ROC for COVID-19 Dataset

We ensembled the models. We attempted to weight each model within the ensemble. We found that a 2 to 1 weight with the Decision Tree taking the two and the other models taking the 1 to be our best of the numerous ensembled weights we attempted. It was notable that ensembling did not provide a better result than the Decision Tree. We also found that the Decision Tree classifier was, by far, the most accurate of all classifiers. This dataset simply preferred the Decision Tree. Is this some sort of bias? We are unsure but nonetheless found this interesting.

We divided our training and testing data using an 80/20 split. We tried different splits, but this split seemed to provide the most consistent high output.

B. MTI Dataset

The Decision Tree classifier in Fig. 10 also gave a very high accuracy result for the MTI Dataset.

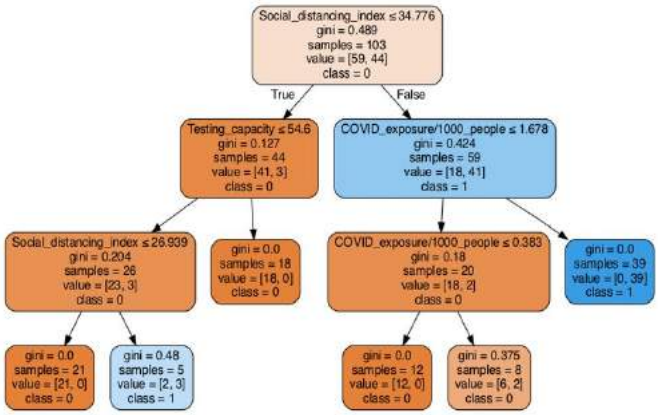

Figure 10 - Decision Tree for MTI Dataset

The Decision Tree classifier provided an almost perfect ROC/AUC curve, as shown in Fig. 11.

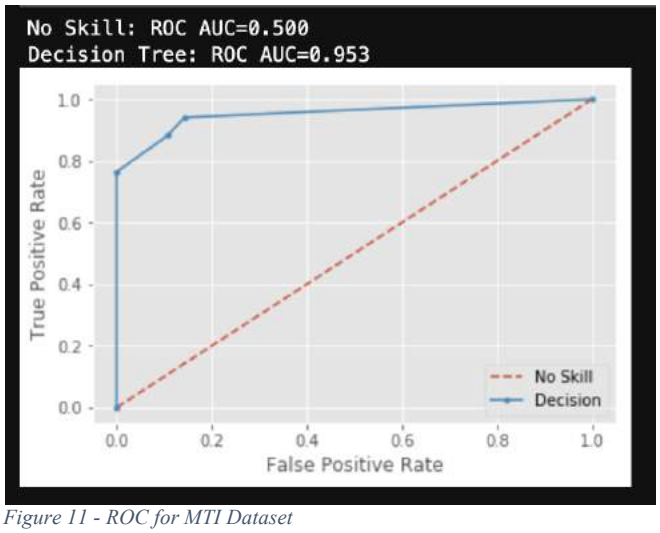

Figure 11 - ROC for MTI Dataset

C. Model Execution Time

Table 1 – Evaluation Execution Time (fastest time emboldened)

TRAINING METHOD	**TIME**
Decision Tree	**time: 0.002**
Logistic Regression	time: 0.03
Support Vector Machine	time: 0.09
Random Forest	time: 1.72
Perceptron	time: 0.08
Neural Network	time: 1.15
Ensemble	time: 1.05
K-means	time: 338.96

The most efficient prediction method used is the Decision Tree with two milliseconds of overall processing time.

The simplified model is the Decision Tree, which is emboldened in the table above. The simplified model was trained and predicted, as listed in the tables above.

D. COVID-19

Ensemble learning is the process by which multiple models, such as classifiers or experts, are strategically generated and combined to solve a particular computational intelligence problem. With that in mind, we ensembled the models. We attempted to weight each model within the ensemble. We found that a 2 to 1 weight with the Decision Tree taking the two and the other models taking the one to be our best of the numerous ensembled weights we attempted. Interestingly, for the dataset, the ensembling did not provide a better result than the Decision Tree alone. This was a strange result but nonetheless was our finding.

E. MTI Dataset

Deep learning is an artificial intelligence function that imitates the human brain's workings in processing data and creating patterns for use in decision-making. Deep learning is a subset of machine learning in artificial intelligence (A.I.) with networks capable of unsupervised learning from unstructured or unlabeled data. This is also known as deep neural learning or deep neural network.

For the deep learning neural network, we had an accuracy of .98 after 35 epochs. We built the model using Tensorflow and Keras, using the Sequential class, creating a Sequential instance, adding three dense layers. The summary of our model is below.

```
model = tf.keras.Sequential()
model.add(tf.keras.layers.Dense(12, input_dim=7,
activation='relu'))
model.add(tf.keras.layers.Dense(8, activation='relu'))
model.add(tf.keras.layers.Dense(1, activation='sigmoid'))
model.summary()
```

Then we compiled the instantiated model using binary cross-entropy, the adam optimizer, and focusing on accuracy for our metrics.

```
model.compile(loss='binary_crossentropy', optimizer='adam',
metrics=['accuracy'])
```

Then we fit the model with the trained data and the target, fitting it over 35 epochs with a batch size of 10.

```
model.fit(X1, y1, epochs=35, batch_size=10)
```

Finally, the model was evaluated as shown below, which provided an accuracy of 98.65, which took approximately 1.65 seconds to fit and evaluate.

```
_, accuracy = model.evaluate(X1, y1)
```

F. A Comparison Study

We compared the results of Maksim to our results. Maksim took the COVID-19 unstructured data and built a highly robust visualization that clusters all the articles. We took a different approach in that we clustered the articles, rather than building a robust visualization. We used the unstructured data to build a structured dataset. From there, we analyzed further and plotted accordingly. In short, our goals were different.

Fig. 12 shows Maksim's visualization.

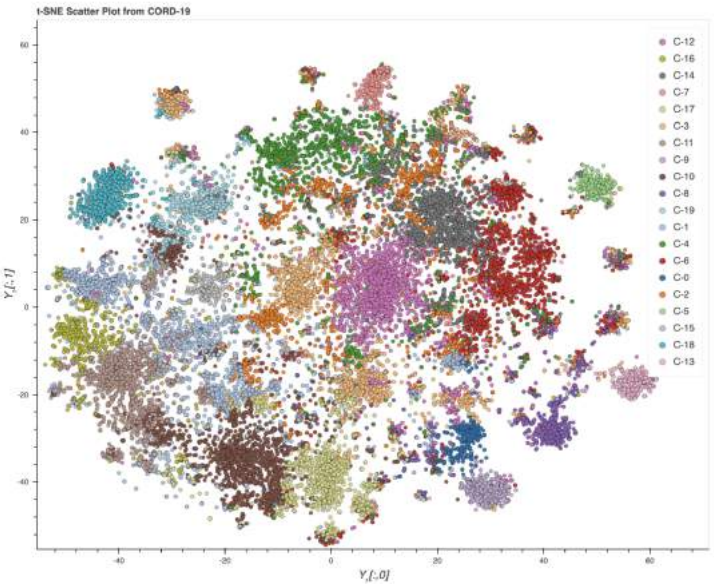

Figure 12 - Maksim's Visualization

Further results are shown in Fig. 13 and Fig. 14.

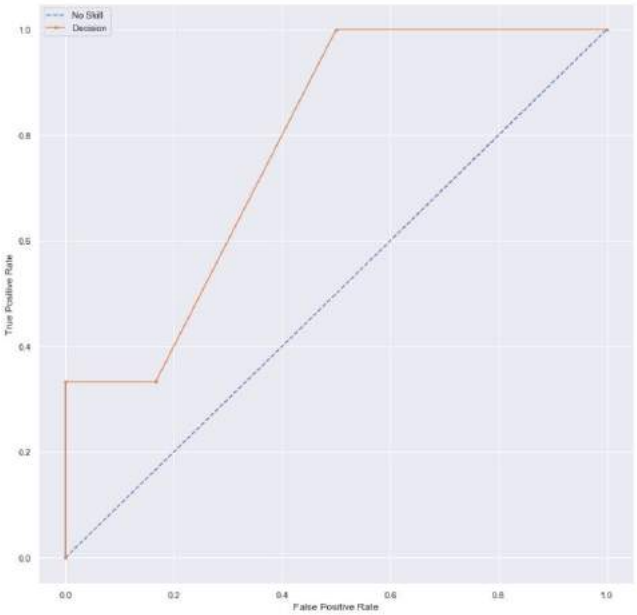

Figure 73 - TPR/FPR Chart

Figure 14 - CORD-19 Unstructured Clustering Visualization

VI. Summary and Conclusion

This project was challenging in many ways. There was an enormous amount of unstructured data. The unstructured data kept growing, and the schema of the data kept changing. This required, eventually, that we pickle the data.

The project provided certain insights that would not have otherwise been apparent without the significant use of computational power and machine learning algorithms.

We also used the following non-exhaustive list of libraries: Tensorflow, Keras, Pandas, Numpy, Matplotlib, Pickle, spaCy, and several other libraries.

Each training shown in Tables 1 and 2 took less than 2 seconds, and most took milliseconds only, except for K-means, which took 338 seconds.

Substantively, we were able to extract structured output from an unsupervised dataset, which allowed us to accomplish the goal of determining the risk factors associated with COVID-19 based on the research articles.

The MTI dataset provided an opportunity to answer another goal related to the target of the number of COVID-19 cases per 1000 people.

The top three most important features for the first dataset were smoking, age, and asthma. The top three most important features for the MTI dataset were social distancing, COVID exposure, and testing capacity.

References

[1] Allen Institute For A.I. (2020, June 05). COVID-19 Open Research Dataset Challenge (CORD-19). Retrieved June 09, 2020, from https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge

[2] Industrial-Strength Natural Language Processing in Python. (2020). Retrieved from https://www.spacy.io

[3] CDC COVID Data Tracker. (2020, June 5). Retrieved from https://www.cdc.gov/coronavirus/2019-ncov/cases-updates/cases-in-us.html?CDC_AA_refVal=https://www.cdc.gov/coronavirus/2019-ncov/cases-in-us.html

[4] Coronavirus disease (COVID-19) Weekly Epidemiological Update and Weekly Operational Update (2020). Retrieved from https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports

[5] Maksim, E., Solovyev, N., Nicholas, C., & Raff, E. (2020, April 16). COVID-19 Literature Clustering. Retrieved June 09, 2020, from https://www.kaggle.com/maksimeren/covid-19-literature-clustering

[6] University of Maryland COVID-19 Impact Analysis Platform. (2020). Retrieved from https://data.covid.umd.edu/

[7] COVID-19 Dashboard by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (JHU) (2020 September 9). Retrieved from https://www.arcgis.com/apps/opsdashboard/index.html#/bda7594740fd40299423467b48e9ecf6

[8] tf.keras.preprocessing.text.Tokenizer : TensorFlow Core v2.3.0. (2020, September 12). Retrieved from https://www.tensorflow.org/api_docs/python/tf/keras/preprocessing/text/Tokenizer

[9] Tang, R., Nogueira, R., Zhang, E., Gupta, N., Cam, P., Cho, K., & Lin, J. (2020). Rapidly Bootstrapping a Question Answering Dataset for COVID-19. *arXiv preprint arXiv:2004.11339*.

[10] Wang, X., Song, X., Guan, Y., Li, B., & Han, J. (2020). Comprehensive named entity recognition on cord-19 with distant or weak supervision. *arXiv preprint arXiv:2003.12218*.

[11] Kieuvongngam, V., Tan, B., & Niu, Y. (2020). Automatic Text Summarization of COVID-19 Medical Research Articles using BERT and GPT-2. *arXiv preprint arXiv:2006.01997*.

[12] Guo, X., Mirzaalian, H., Sabir, E., Jaiswal, A., & Abd-Almageed, W. (2020). CORD19STS: COVID-19 Semantic Textual Similarity Dataset. *arXiv preprint arXiv:2007.02461*.

[13] Kroll, H., Pirklbauer, J., Ruthmann, J., & Balke, W. T. (2020). A Semantically Enriched Dataset based on Biomedical NER for the COVID19 Open Research Dataset Challenge. *arXiv preprint arXiv:2005.08823*.

[14] Huang, T. H. K., Huang, C. Y., Ding, C. K. C., Hsu, Y. C., & Giles, C. L. (2020). CODA-19: Reliably Annotating Research Aspects on 10,000+ CORD-19 Abstracts Using Non-Expert Crowd. *arXiv preprint arXiv:2005.02367*.

[15] Esteva, A., Kale, A., Paulus, R., Hashimoto, K., Yin, W., Radev, D., & Socher, R. (2020). Co-search: Covid-19 information retrieval with semantic search, question answering, and abstractive summarization. *arXiv preprint arXiv:2006.09595*.

[16] Homolak, J., Kodvanj, I., & Virag, D. (2020). Preliminary analysis of COVID-19 academic information patterns: A call for open science in the times of closed borders. *Scientometrics, 124*(3), 2687-2701. doi:10.1007/s11192-020-03587-2

[17] Resnik, P., Goodman, K. E., & Moran, M. (2020, June 30). Developing a Curated Topic Model for COVID-19 Medical Research Literature. ACL 2020 Workshop.