**COVID-19 Research paper classification using ML**

Bita Faraji, Sindhura Vemuri, Yash Choksi

Department of Information Systems, California State University

Los Angeles

Tel. 858-798-6036

e-mail: [bfaraji@calstatela.edu](mailto:bfaraji@calstatela.edu), [svemuri2@calstatela.edu](mailto:svemuri2@calstatela.edu), [ychoksi@calstatela.edu](mailto:ychoksi@calstatela.edu)

**Abstract:** The outbreak of Coronavirus disease 2019 (COVID-19), caused by severe acute respiratory syndrome (SARS) coronavirus 2 (SARS-CoV-2), has thus far killed over 30,000 people and infected over 4 million people in the world, resulting in catastrophe for humans. In response to the rapidly increasing number of publications on the emerging disease, this paper attempts to provide accurate information for medical professionals from all around the world using classification algorithms to identify them as useful or not useful. This is done by using recent advances in natural language processing and other AI techniques to generate new insights in support of the ongoing fight against this infectious disease. We used supervised and unsupervised ML approach for applying different techniques to apply and seen which works best for our modelling. Our main goal is to design a classification system which can be worked in most efficient way and can help medical community in future.

**1. Introduction**

In this pandemic situation of covid-19, many medical professionals and scientists are working towards developing vaccination. Because this is a new virus, there are still things we do not know, such as how severe the illness can be, how well it is transmitted between people, and other features of the virus, so it 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.

The freely available dataset is provided by the kaggle to the global research community to apply recent advances in natural language processing and other AI techniques to generate new insights in support of the ongoing fight against this infectious disease contains around more than 33000 published papers and articles in different journals from all around the world Every research paper is having few keywords and format of research paper is JSON. The dataset is mostly comprised of Covid-19 and SARS-COV-2 and some other related to corona virus.We have worked on project using the dataset as the following steps,

* We have changed the dataset format from JSON to spark data frame and worked on it
* Dataset size is about 8GB
* Dataset features are Paper\_id,DOI, Abstract, body\_text,authors,title,journal,Abstract\_summary,Abstract\_count,Body\_count,body\_unique\_words,Languages, Processed\_text.
* Dataset that we have taken from Kaggle is not clean we have cleaned it using JSON blob objects and PySpark tokenizers.
* Dataset doesn’t have any labels; we have used two clustering models K-means and bisecting k-means algorithms to create labels in the dataset.
* After labeling the dataset we have used supervised machine learning, in supervised ML we have used around 4 types of algorithms and 8 algorithms in total for the project.

-Logistic Regression

- Decision Tree Classifier

- Random Forest Classifier

- Naïve Bayes

**-** Cross-validation in Logistic Regression

**-** Decision Forest Classifier

- Bayes point match classifier

- Boosted Decision Tree

* We have used 3 platforms for this project

Oracle BDCE, Data bricks and Azure ML

In Oracle and Data bricks we used full dataset and in Azure ML we have used around 600 rows which is 1.9% of whole dataset.

Table 1. Technical Specifications.

|  |  |  |
| --- | --- | --- |
| **Azure ML** | **Data bricks** | **Oracle BDCE** |
| Free Workspace | Data bricks Subscription | Cluster 5.2 (includes Apache Spark 2.4 |
| 10GB storage | Cluster 5.2 (includes Apache Spark 2.4.0, Scala 2.11)  6GB Memory, 0.88 Cores, 1 DBU | 2.20 GHz Memory, 802 GB Storage |
| Single node | Python Version 3 | Python Version 3 |

* Furthermore, to achieve our goal, we have tried to provide the best and needful information to the medical professionals
* So, at the end of this project when any new paper is feed in our model; it can able to predict what is category of that paper.

**2. Related Work**

As we gathered from all around the different research papers and we came to know that all of them have done work but most of them was in simple single node work. Not much included on Big Data. One of the papers which we referred was in NLP and they have done classification on the basics on entropy of each and every word. Now entropy is how much weight and probability distribution with other words. As there are so many other words are already present over there, we need to think that entropy would never be always accurate, and it would be assumption that this word would occur. So, in that case our model is much better. There is one other paper in which is mentioned that by tricking length of bag of words we can able to make some changes in that and it could expedite the processing. Yes, processing is always a huge challenge and due to that even we faced so many challenges in that case but still it would not be efficient model and they used small bag of words while we used large bag of words of 10000.

So, by looking at all the papers which we referred we can say we exhausted more resources but it’s much better and efficient than any other approach.

**3. Background**

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 63,000 scholarly articles, including over 51,000 with full text, about COVID-19, SARS-CoV-2, and related coronaviruses. This freely available dataset is provided to the global research community to apply recent advances in natural language processing and other AI techniques to generate 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.

Allen institute has done phenomenal work for this project that they thought that ML can solve this approach and they give us chance to improve this as ML modelling. This whole project is in partnership of White House, Microsoft and Facebook as in partnership.

There is list of tasks and questions are assigned to this topic and they want us to think in direction of the idea that keep question starts from classifying journals and then there is some to the point questions are also present in the website. We try to answer the very basic question that weather any particular document is weather covid-19 related document or not. We try to make different models on it to predict which model would be best for this type of problems. As problems does not end there and every problem is specific to some algorithms while other relates to some other model.

They also posted some key words on the website regarding what are key words should we consider as our covid-19 identification papers and which are not in the metadata repository. The task list ranges from what are risk factors in covid-19, which factors primarily increase chance of this disease, what have been already published about no pharmaceutical and non-official treatments and finally what is diagnosis and surveillance in this scenario.

**4. Methodology**

In this project data we found was in the form of highly unstructured and in JSON format which is not very good sign for predictive modelling. As we have to convert it to pandas or PySpark data frames we have to do lot of processing and how to clean them is very big and huge challenge. We gathered data from Kaggle, and data size is almost 8 GB. Because of that we converted data to panda’s data frame using blob module of PySpark. Then we created many columns and extracted text from JSON files.

Figure 1. Flow chart.

After getting all the data normal text data can-not be processed in ML algorithms and those algorithms will be needed only numerical values to fir the data values. There is one good feature in Azure ML they call it as feature hashing while in PySpark we have in built feature name tokenizing and hashing. After creating this feature, we came to point that our whole dataset was comprised of 10000 bag of words and due to that those each row was converted to 10000 values.

By keeping that in mind we created many columns as paper id, publication, body, abstract and many more. As we created this module, we wanted to try to make it more useful but as we checked with feature importance and heat map and some co-relation materials, we came to know that only body text we can able to use for our predictive modelling. As our dataset is having no model this is typical unsupervised machine learning modelling. And as we kept that in mind, we have to create first labels to make any classification or any other ML algorithm to work on it.

Now as we do not have labels, we used 2 clustering algorithms to assign labels to them and we used K-Means and K-Means bisecting algorithm and then we checked the accuracy and Euclidian distance for them too. As we know as the Euclidian distance lesser that model is much better. So, after using these 2 models our distance was lesser for Bisecting K-Means and for that reason we used that algorithm for labelling our dataset. After labelling our dataset we used supervised ML algorithms and we used 4 supervised algorithms.

The 4 supervised algorithms which we used are Logistic regression, Decision Tree, Decision tree boosted, Naïve Bayes and Random Forest Classifier. As these 4 algorithms are important, they all come with different limitation and strength. As we trained our algorithm the best from all of them was Random Forest Classifier. As we tuned its hyper parameters and noted its performance, we found that it would be very good for us to use it. Then we used cross-validation on Logistic Regression, and it gave us best than all the other models we used by far. And with the ROC value of 0.96. We would like to show you the comparison on all the platforms which we performed our different algorithms. The table below is having details of running time and how much is the accuracy in all of the platforms.

Table 2. AUC values and run time on each of the platforms.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| AUC | K-Means | Bisecting K-Means | Logistic Regression | Decision Forest | Naïve Bayes | Random Forest Classifier & Boosted Decision Forest |
| Oracle | 0.289 | 0.232 | 0.99 | 0.89 | 0.38 | 0.824 |
| 2 min | 2 min | 3 min | 8 min | 5 min | 8 min |
| Databricks | 0.305 | 0.299 | 0.899 | 0.8303 | 0.134 | 0.9622 |
| 7 min | 8 min | 8 min | 13 min | 10 min | 14 min |
| Azure ML | 0.479 | - | 0.951 | 0.8 | 0.37 | 0.977 |
| 1.5 min | - | 1 min | 5 min | 3 min | 17 min |

Table 3. Comparing Decision Forest and Logistic Regression based on precision, recall, F1 score, accuracy, and AUC.

|  |  |  |
| --- | --- | --- |
|  | Decision Forest | Logistic Regression |
| Precision | 0.913 | 0.917 |
| Recall | 0.943 | 0.951 |
| F1 Score | 0.928 | 0.94 |
| Accuracy | 0.9 | 0.917 |
| AUC | 0.951 | 0.965 |

Table 4. Comparing Bayes Point Match and Decision Forest based on precision, recall, F1 score, accuracy, and AUC.

|  |  |  |
| --- | --- | --- |
|  | Bayes Point Match | Decision Forest Boosted |
| Precision | 0.884 | 0.922 |
| Recall | 0.992 | 0.959 |
| F1 Score | 0.935 | 0.94 |
| Accuracy | 0.906 | 0.917 |
| AUC | 0.856 | 0.977 |

Here as we can see Decision Forest boosted is best in all the cases in the small dataset, we applied this algorithm. As dataset size became larger and larger it always creates problem for us to manage it and computing time is also gets increases.

**5. Conclusion**

As world is falling apart because of covid-19 , many of the medical professionals are working day and night to provide us better tomorrow , they need covid-19 related data to know more about this virus and provide vaccination , so through this project we are trying to provide them the clean and filtered data related to covid-9 to them. Since the data is in unsupervised model we will not have any corressponding output variables, so our goal is to model the underlying structure of the data in order to learn more about the dataset. our dataset perform well with bisecting k-means which has euclidean distance of 0.232 that’s why we have choosed bisecting k-means clustering to label our dataset , after labelling it we ran four machine leraning algorithms and accuracy is best with cross validationin logistic regression and Accuracy is 0.9622, our model performs really well all the algorithms .we classified the different journal research paper so that we can able to help the research community ,and this model can be used for more number of datasets so in future any new research paper comes into market and if it feeds into our model, it will be able to predict what is the category of the paper . so finally by defining the category of the paper it will be really helpful for the researchers to filter out what they need means if that journal is related to covid-19 or not, also they will be able to search with any useful keywords for accurate data of what they needed. And as part of conclusions we also created some visuals which we could able to take out from this reesearch which wee done.

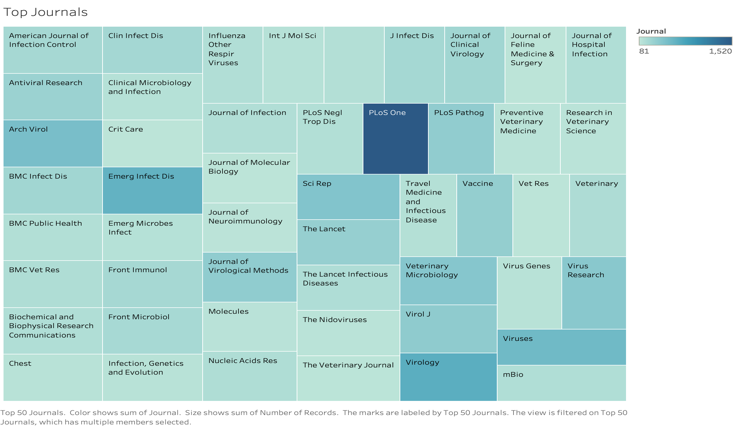


Figure 2. Top 50 jounals found in this dataset.

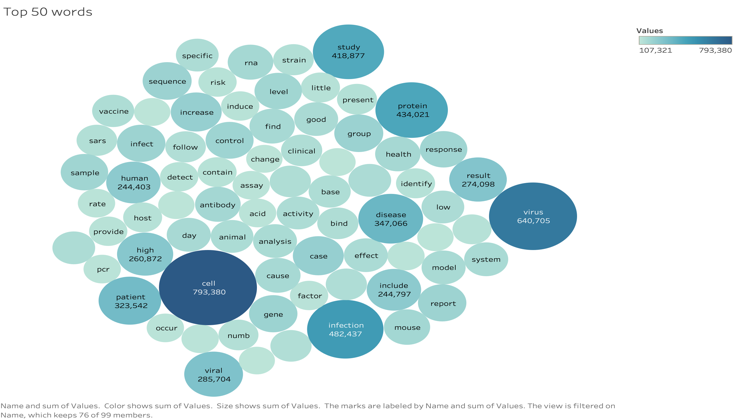


Figure 3. Top 50 words found from this research.

**6. GitHub Link**

1. <https://github.com/yashchks87/covid_19_nlp>

**7. Refrences**

[1]<https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge/>

[2]<https://www.kdnuggets.com/2019/05/guide-natural-language-processing-nlp.html>

[3]https://towardsdatascience.com/natural-language-processing-with-pyspark-and-spark-nlp-b5b29f8faba