HIGH SCHOOL EDITION
Journal of Student Research
Volume 11 Issue 3 (2022)

Investigating a Key COVID-19 Question by Using Natural Language Processing on Scientific Publications

Devika Dua[1] and John Mapes[2#]

[1]Cedar Creek High School
[2]Louisiana Tech University
[#]Advisor

ABSTRACT

The COVID-19 pandemic has brought an unprecedented challenge to public health. Numerous scientific publications are published daily on COVID-19 to understand the unexplored facets of the disease. The sheer volume of these publications makes it daunting for researchers to quickly find information and evaluate data related to specific COVID-19 queries. Natural Language Processing (NLP), a form of artificial intelligence, assists in churning these huge piles of data with a sophisticated algorithmic approach. The purpose of this study is to investigate key COVID-19 question by using NLP on scientific publications. Using the T5 (Text-To-Text Transfer Transformer) model, we analyzed 740,000 journal abstracts for specific answers an important COVID-19 question. We performed qualitative observations, T-Tests (p-values and inferences), and accuracy metrics (Precision, Recall, and F1 score) to evaluate the models in this study. As the number of scientific publications increases, our proposed methodology provides an efficient mechanism for performing specific information retrieval for emerging questions, diseases, and related conditions, especially for underrepresented populations.

Introduction

The COVID-19 pandemic has impacted the world in unimaginable ways and has claimed thousands of lives. With the surprising onset of the virus, researchers have scrambled for information and produced many publications on the virus. However, due to the volume of scientific publications, finding answers to specific COVID-19 related questions has posed as a challenging task. Therefore, we propose a deep- learning natural language processing model to extract answers to inputted questions from the abstracts of COVID-19 related scientific publications.

Recent advances in Transfer Learning methods in NLP has revolutionized the field of deep learning [16]. We utilized the T-5 [8] transformer (Text-To-Text Transfer Transformer): a deep learning model that is able to perform text-to-text tasks. Transformers can overcome the limitations of other neural network learning algorithms, such as Convolutional Neural Networks and Recurrent Neural Networks. From the perspective of analyzing COVID-19 related information through many published journals both quickly and accurately, researchers can benefit from the choice of the T5 model as it claims to be the most efficacious model on more than twenty well established NLP tasks.

This project employs the T5 model to narrow exact answers to COVID-19 related questions from the scientific publications provided in the Kaggle COVID-19 *Open Research Dataset*, or CORD-19 dataset.

Related Research

NLP based ensemble learning methods have been previously used to screen scientific publications for abstractive summarization, text classification and extraction, sentiment analysis, and market intelligence [1]. Several papers have used transformers for answer extraction: Lou *et al.* explored Bart, T5, and PropheNet to develop summaries on COVID-19 related publications [2], and Oniani *et al.* used BERT, BioBERT, and USE to filter answers to COVID-19 related questions from the CORD-19 dataset and develop a public chatbot [3]. Many other papers have similarly approached answer extraction, especially with the BERT model. However, the use of the T5 model for answer extraction isn't as common.

Georgia Tech University participated in the COVID-19 Open Research Dataset Challenge [17] and developed a hand-crafted model to extract answers to COVID-19 related questions from the CORD-19 dataset [4]. Their model is used as the state-of-the-art basis of comparison with our model and are acknowledged as *GA Tech CORD-19* throughout our research.

Dataset and Preprocessing

We utilized the COVID-19 Open Research Dataset (CORD-19) from Kaggle [5]. CORD-19 is a growing collection of over 500,000 scholarly articles, including over 200,000 with full text on COVID-19, SARS- CoV-2, and related coronaviruses. CORD-19 has been downloaded over 200,000 times and contains papers especially from PubMed Central (PMC) and World Health Organization (WHO) in the fields of virology, immunology, molecular biology, and more [6].

Each paper has been split into different chunks, including authors, abstracts, body text, and references, saved as a JSON file under different keys [2]. After critically reviewing multiple publications, we determined that the data was incomplete and inconsistent. To achieve efficient analysis, we focused on the abstracts of those publications. At the preprocessing stage, we excluded the duplicated abstracts and the ones with missing data. Additionally, we narrowed the abstracts by checking for COVID-19 related phrases in them ("covid", "-cov-2", "cov2", and "ncov"). After the preprocessing stage, we had 181,000 journal abstracts, narrowed from initially 740,000 journal abstracts. A snapshot of the narrowed data is provided in Table 1.

Table 1: A snapshot of the data, including the publication date, authors, title, and an excerpt from the abstract.

#	Publication Date	Authors	Title	Excerpt from Abstract
1	2020-03-18	Xue, et al.	Is a 14-day quarantine period optimal for effectively controlling coronavirus disease 2019 (COVID-19)?	"We are questioning if the current-inferred median incubation time is representative for the whole COVID-19 population…"
2	2020-04-18	Rahimi, et al.	Challenges of managing the asymptomatic carriers of SARS-CoV-2	"Besides hospitalized cases, many individuals are likely asymptomatic but potentially carry the virus…"

3	2020-01-24	ɔnathan, et al.	Novel coronavirus 2019-nCoV: early estimation of epidemiological parameters and epidemic predictions	"Using a transmission model, we estimate a basic reproductive number of 3.11…"
…180,000				

Methodology

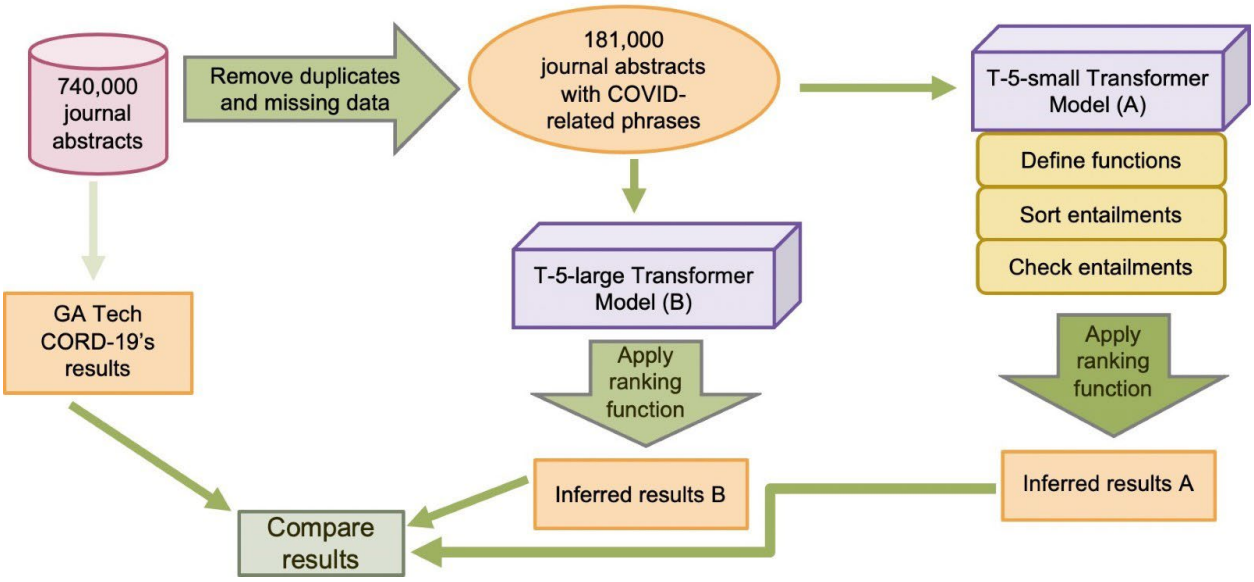

Figure 1. Overall Methodology

1. T5 Model

The T5 (Text-To-Text Transfer Transformer) model is a unified framework that converts all text-based language problems into a text-to-text format [6]. Every considered task, including translation, question answering, and classification, is cast as feeding the T5 model text as input and training it to generate some target text [7]. The T5 model was mainly used because of its ability to perform text-to-text tasks. Other transformers, including BERT, XLNet, and m(multilingual)T5, were experimented upon. However, with BERT and XLNet, only numerical class values were derived. The multilingual transformer was also experimented –base, -small, and –large, but there was not enough memory space available. In future work, we plan to experiment mT5 on GPU-based computers.

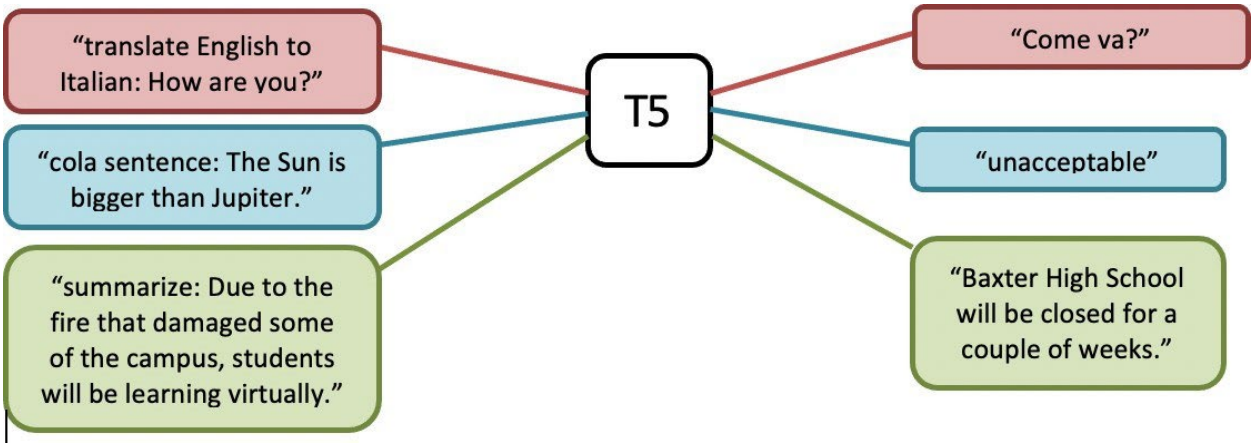

Figure 2: T5 model's tasks, illustrated

We obtained the T5 framework from HuggingFace [8]. Figure 2 demonstrates a series of examples that illustrates the tasks this text-to-text model can perform. For example, if we fed the task "translate English to Italian: How are you?" into the model, the response "come va?" will be returned.

1.1 T5-large

The T5-large model incorporates 770 million parameters with 24-layers, 1024-hidden-state, 4096 feed-forward hidden-state, and 16-heads [9].

1.2 T5-small

The T5-small model incorporates 60 million parameters with 6-layers, 512-hidden-state, 2048 feed-forward hidden-state, and 8 heads [9]. The larger transformer has a larger computational demand and more aspects of refinement.

2. Abstractive Extraction

We developed an algorithm that incorporates the tasks of the T5 model. Figure 3 illustrates the task training with a sample question. **Figure 3**: Task training for the question, "What is the range of the incubation period of COVID-19 in humans?" The incubation period is the number of days between infection and onset of symptoms [10].

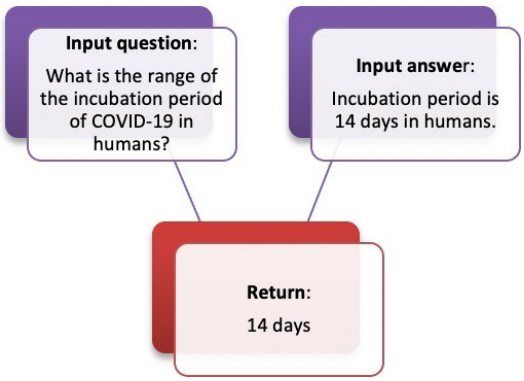

Figure 3: Task training for the question, "What is the range of the incubation period of COVID-19 in humans?"

2.1 String Matching

We distilled abstracts based on question-related phrases using a string-matching algorithm. For the question related to the incubation period of COVID-19, the phrases used were "shedding duration" and "contagious period". String matching finds patterns in the designated text that match the word or phrase inputted.

Table 2: Examples of questions and related phrases that may be inputted into the algorithm.

Question	**Related Phrases**
"What do we know about the seasonality of transmission?"	"seasonal", "transmission", "winter", "summer", "fall", "autumn", "spring"
"What is the effectiveness of personal protective equipment (PPE) against COVID-19?"	"protective", "clothing"
…26 other questions	

2.2 Entailments

We then applied the transformer's entailment task to yield binary answers (true/false). The goal is to predict whether a premise implies ("entailment") a hypothesis (in this case, our "input answer") [7].

2.3 Ranking Algorithm

We developed a ranking approach using the concept of multiple-choice answers for randomized ranking. Figure 4 illustrates this process of ranking. This project's full code, including that of this ranking approach, is available on Kaggle and can be accessed through reference [11].

```
def multichoice(answers):
    joinedanswers = joinanswers(answers)
    GA_Tech_Question_1 = "How long are individuals contagious?"
    task = "question: %s answers: %s" % (GA_Tech_Question_1, joinedanswers)
    result = task_and_reply(task)
    return result
answers = ["Incubation period is 5 days in humans.",
           "Oranges are a great source of vitamin C and fiber.",
           "Bananas are delicious, but apples and oranges are better.",
           "Flamingos are my favorite animal, but dogs are awesome too."]
print(joinanswers(answers))
multichoice(answers)
```

Figure 4: Python code for the multiple-choice ranking approach.

Results

We adopted both qualitative and quantitative evaluation metrics to evaluate the efficiency of the T5-small, T5-large, and GA Tech CORD-19's model:

- Observations
- T-Tests (P-Values and Inferences)
- Precision, Recall, and F1 Scores

Throughout the presented results, the question "What is the range of the incubation period of COVID-19 in humans?" is primarily analyzed.

Observational Comparison of Transformers

Below are excerpts of the top results derived from both the T5-small and T5-large model. There was a total of 38 results from the T5-small model and 47 results from the T5-large model.

T5-small:
["6 days (range 1 to 13 days), "Auc, duration of viral shedding, and epithelial cells infected", "6 days", "18- 32 days", "144 days", "25-27", "65 and 0 62 g/kg/d"]
T5-large:
["iqr 18.5-41.0", "1 to 13 days", "52 days", "1-13 days", "144 days", "25-27", "0 to 4 weeks"]

The comparison between the values of the T5-small and T5-large model is not statistically significant (P- value: 0.1373). However, many key observations of the T5-large model (compared to the T5-small model) can be made:

- More results (in quantity)
- Results with the phrase "iqr" (incubation period) along with the value
- Less non-numerical results

T-Tests for Comparison with GA Tech CORD-19

A t-test allows us to compare the average values of two data sets and determine if they came from the same population [12]. We performed t-tests on our results versus those of GA Tech CORD-19, our results versus Centers for Disease Control and Prevention (CDC)'s value of the incubation period of COVID-19, and GA Tech CORD-19's results versus those of CDC. CDC's mean value for the incubation period of COVID-19 is 4-5 days (4.5 used for this study) [13]. Our mean value is 32.7117 and GA Tech CORD-19's mean value is 6.7971. Based off these values, the P- values were calculated for each set of comparison, as demonstrated in Table 3. Based off the analysis of the derived p-values, no comparison of two different results (e.g., our results vs. GA Tech CORD-19) agreed.

Table 3: An analysis of our results, those of GA Tech CORD-19, and those of CDC. "Statistically significant" implied difference in values.

	P-Value	Inference
Our results vs. GA Tech CORD- 19	0.0098	Very statistically significant
Our results vs. CDC	0.0002	Extremely statistically significant
GA Tech CORD-19 vs. CDC	0.0106	Statistically significant

Precision, Recall, and F1 Score for Comparison with GA Tech CORD-19's Work

We first established two confusion matrices - as presented in Figures 5 and 6 - from which we derived the Precision, Recall, and F1 score of both our results and those of GA Tech CORD-19. Confusion matrices evaluate the accuracy of a classification [14]. In order to determine the threshold number, we doubled the CDC's value of the average incubation period, from 4.5 days to 9 days. A number of standard deviations from the CDC's mean incubation period could have been employed.

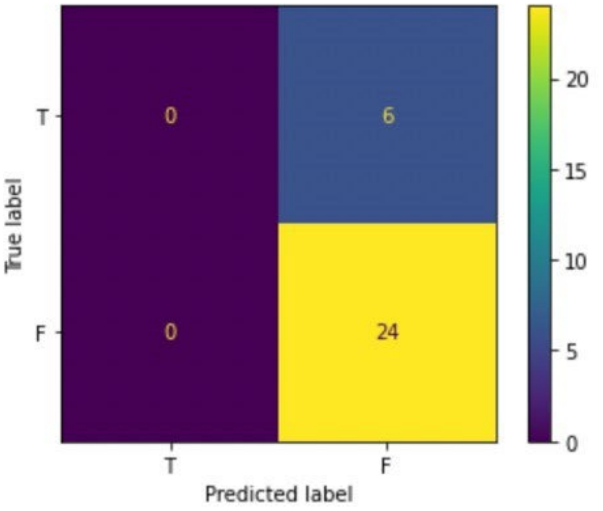

Figure 5: Confusion matrix for our results

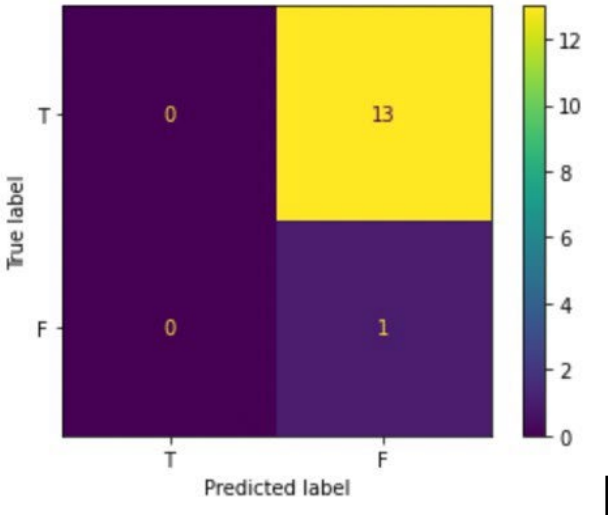

Figure 6: Confusion matrix for GA Tech CORD-19's results

Precision, Recall, and F1 score are threshold-based metrics, using a qualitative understanding of error [15].

Table 4: Quantitative evaluation metrics for our results and those of GA Tech CORD-19.

	Our results	GA Tech CORD-19's results
Precision	**0.8**	**.07**
Recall	**1.0**	**1.0**
F1 Score	**.89**	**.13**

Based on Table 4, the higher precision of our model relates to the low false positive rate, and the higher F1 score of our model relates to its higher accuracy.

Analysis of Special Populations

Based on the P-values collected from the T-Tests performed upon our results, those of GA Tech CORD- 19, and that of CDC, we inferred that our higher values refer to special populations, as presented in Table 5. The difference between our values with those of GA Tech CORD-19 (Table 6) can be attributed to differently sorted data due to our algorithm and transformer's ability to analyze complex language.

Table 5: Results of the mean incubation period from our model and the special population(s) each refers to.

Our results	Special Populations
6 days	Taiwan; N/A
20 days	Hospitalized patients in Wuhan, China
20 days	Severely ill patients
9.57 days	China; Children
3.78 days	France & Switzerland; critically ill patients
34 days	Wuhan; hospitalized conditions

Table 6: A snapshot of GA Tech CORD-19's results for the mean incubation period.

GA Tech CORD-19's Results					
16.5 days	6.15 days	6.15 days	5 days	6.4 days	6.4 days

Many other questions can be easily inputted into our model, but due to the timing of this study, thorough analysis was not possible.

Conclusion and Future Work

Transformers offer a significant learning tool to extract answers to COVID-19 questions from scientific abstracts. The proposed approach has the ability to discover answers accurately and efficiently. Our results demonstrate the specificity of our findings of special populations previously not discovered. The scientific investigation of COVID-19 continues to grow, and methods, such as the one proposed, perform specific information retrieval for emerging questions, diseases, and conditions.

In future work, we will apply and analyze our model on many other COVID-19 related questions, such as "What is the role of the environment in transmission of COVID-19?". Additionally, not only can our model apply to COVID-19, but it can also be applied on datasets related to many different diseases and conditions.

Specific phrases from conversations in medical settings (clinics, hospitals) can be used as question-related phrases in our model. This can refine results that are more specific to clinicians, physicians, and other medical professionals' specific interests. Moreover, we can create a searching tool similar to Google search that can implement our model and a wide variety of disease/condition-related datasets. This resource can be available to clinicians, physicians, and other medical professionals as well.

Acknowledgements

This project was done as an independent research project, and most of the work was performed in the summer of 2021; the writing of this paper was extended into the fall of 2021. We wish to extend our special thanks to Dr. Brian Benson, who assisted in the discussion of applications of this project and future work.

References

1. Qin X, Liu J, Wang Y, Liu Y, Deng K, Ma Y, Zou K, Li L, Sun X. Natural language processing was effective in assisting rapid title and abstract screening when updating systematic reviews. J Clin Epidemiol. 2021 May; 133:121-129. doi: 10.1016/j.jclinepi.2021.01.010. Epub 2021 Jan 21. PMID: 33485929.

2. Lou Z, Zhang J. Abstractive Summarization on COVID-19 Publications. CS230 Deep Learning, Stanford University. Spring 2020.

3. Oniani D, Wang Y. A qualitative evaluation of language models on automatic question-answering for COVID-19. Association for Computing Machinery Digital Library. 21 September 2020.

4. Mlconsult. (2020, May 3). *Transmission, incubation and environment 2.0*. Kaggle. Retrieved November 14, 2021, from https://www.kaggle.com/mlconsult/transmission-incubation-and- environment-2-0.

5. COVID-19 Open Research Dataset (CORD-19), available for download at https://allenai.org/data/cord-19

6. Wang L, Lo K, Chandrasekhar Y, Reas R, Yang J, Eide D, Funk K, Kinney R, Liu Z, Merrill W, Mooney P, Murdick D, Rishi D, Sheehan J, Shen Z, Stilson B, Wade A, Wang, K, Wilhelm, C, Xie B, Raymond D, Weld D, Etzioni O, Kohlmeier S. CORD-19: The Covid-19 Open Research Dataset. National Institutes of Health. 22 April 2020. PMID: 32510522

7. Raffel C, Shazeer N, Roberts A, Lee K, Narang S, Matena M, Zhou Y, Li W, Liu P. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer.

8. T*5*. T5 - transformers 4.12.2 documentation. (n.d.). Retrieved November 14, 2021, from https://huggingface.co/transformers/model_doc/t5.html.

9. *Pretrained models*¶. Pretrained models - transformers 4.0.0 documentation. (n.d.). Retrieved November 14, 2021, from https://huggingface.co/transformers/v4.0.1/pretrained_models.htm

10. Pathak, N. (2021, September 30). *Coronavirus incubation period: How long and when most contagious*. WebMD. Retrieved November 14, 2021, from https://www.webmd.com/lung/coronavirus- incubation-period#1.

11. Devika Dua. (2021, November 13). *AMIA 2021 1fbc4b*. Kaggle. Retrieved November 14, 2021, from https://www.kaggle.com/devikadua/amia-2021-1fbc4b.

12. Hayes, A. (2021, November 13). *T-test definition*. Investopedia. Retrieved November 14, 2021, from https://www.investopedia.com/terms/t/t-test.asp.

13. Centers for Disease Control and Prevention. (2021, February 12). *Management of patients with confirmed 2019-ncov*. Centers for Disease Control and Prevention. Retrieved November 14, 2021, from https://www.cdc.gov/coronavirus/2019-ncov/hcp/clinical-guidance-management- patients.html.

14. *Sklearn.metrics.confusion_matrix*. scikit. (n.d.). Retrieved November 14, 2021, from https://scikit-learn.org/stable/modules/generated/sklearn.metrics.confusion_matrix.html.

15. Yacouby R, Axman D. Probabilistic Extension of Precision, Recall, and F1 Score for More Thorough Evaluation of Classification Models. Eval4NLP. 20 November 2020.

16. Azunre, P. (2021, August 11). *Recent advances in transfer learning for Natural Language Processing*. Medium. Retrieved November 14, 2021, from https://towardsdatascience.com/why- should-you-leverage-transfer-learning-14d08a60f616.

17. AI, A. I. F. (2021, November 9). *Covid-19 open research dataset challenge (cord-19)*. Kaggle. Retrieved November 15, 2021, from https://www.kaggle.com/allen-institute-for-ai/CORD-19- research-challenge.