COVID-19 Open Research Dataset Challenge (CORD-19)

ABSTRACT

The objective of the COVID-19 Open Research Dataset Challenge (CORD-19) is to develop a Question Answering (QA) Chatbot. The task of the QA Chatbot would be to give 10 relevant answers from various research papers for a given question. For this challenge, Kaggle provided Cord-19 Document Embeddings, Target Tables, Meta Data and Document Parses for both Portable Document Format (PDF) file and PubMed Central (PMC) File. Furthermore, the National Institute of Standards and Technology (NIST) provided the Questions, Query and Narrative Text for the CORD-19 Challenge. The Meta Data contains the data of about 1 million research papers and the Document Parses of PDF Research Papers are around 400 thousand. The main features of Meta Data were Publish Time, Title, Abstract, CORD ID (a document ID of research papers) and PDF JSON Files (directory path to the Document Parses). Also, the main features of Document Parses were Abstract and Body Text. The Cord-19 Document Embeddings were Sentence Embeddings of all Research Papers with keys as CORD IDS given in Meta Data.

The Methodology of the project has four sections namely Data Filtration, Data Preprocessing, Models Utilized and Method Used. In the first section of the Methodology, the research papers were filtered based on their Publish Time, Abstract, Title, Number of Characters in a Sentence of Abstract and Body Text. Moreover, the given target table files were loaded and then the research papers were filtered based on the Publish Time of the target table research papers. Then in the Data Preprocessing Section, the Paragraphs of Abstract and Body Text was converted to Sentences and these Sentences were further converted to Tokens. After this, in the Models Utilized Section, Word2Vec and Doc2Vec Models were utilized to generate Sentence Embeddings for Abstract, Body Text, Question Text, Query Text and Narrative Text. There were 3 variations of Word2Vec, and Doc2Vec Model based on 3 Sentence Embeddings namely Question, Mean and Concatenated Sentence Embeddings. At Last, in the Method Used Section Cosine Similarity was used between Sentence Embeddings to find Relevant Answers.

Based on the Relevance Judgements File of 5 rounds Aggregate Relevance Score was calculated by summing all the Relevance Value of the 10 Most Relevant Answers of every Question Text. The Relevance Score for Concatenated Word2Vec Model was 730 which makes it the best model. While the Relevance Score of the second-best model (Question Word2Vec Model) was 635.

1 INTRODUCTION

The CORD-19 Challenge involves the creation of a Question Answering (QA) System or Chatbot which would be able to extract 10 Relevant answers for a particular Scientific Question. Furthermore, creating such a QA System is necessary so that the general population of the world can get aware of the Covid-19 disease. Moreover, the QA Chatbot would be immensely useful for the research community because any researcher would be able to check previous research done on the coronavirus instantly and this will accelerate the research being conducted on the coronavirus.

Kaggle which is a Machine Learning (ML) and Artificial Intelligence (AI) competition platform provided a Meta Data Comma-Separated Values (CSV) File which is shown in Figure 1. It contains information about Publish Time, Title, Abstract, CORD ID (a document ID of research papers) and PDF JSON Files (directory path to the Document Parses) of the Research Articles. The CORD IDs of about 1 million research papers have been given. The Publish Time of research Articles ranges from the year 1825 to 2022. Also, from Figure 1 we can observe that around 22% of Articles have no Abstract and 65% of Articles have no PDF Document Parses available. Furthermore, Kaggle also provided CORD-19 Document Embeddings which were Sentence Embeddings of all the Research Papers with CORD IDS as keys as shown in Figure 2. The dimension of these document embeddings is 768.

Other Files which were provided include Target Tables, Question File and Relevance Judgement File given by Kaggle and the National Institute of Standards and Technology (NIST) respectively. The Target Table file contains information on certain topics or questions extracted from several Research Papers as Shown in Figure 3. It also contains a short extract of the Article shown in Figure 3 under Excerpt Heading. The Question File contained Questions, Query (a shorter form of the Question) and Narrative (explanation of the question) which would be entered in the QA Chatbot to get relevant answers. At Last, the Relevance Judgements File contains Question ID, CORD ID and Relevance Value of the Document as shown in Figure 4 below. The Relevance Values were 0 – 'Irrelevant Answer', 1 – 'Partially Relevant Answer' and 2 'Completely Relevant Answer' given in the Relevance Judgements File.

cord_uid	title	abstract	publish_time	pdf_json_files
970836 unique values	**850367** unique values	[null] 22% [Figure: see text]. 0% Other (820915) 78%	2021 22% 2020 20% Other (612067) 58%	[null] 65% document_parses/p... 0% Other (373763) 35%
ug7v899j	Clinical features of culture-proven Mycoplasma pneumoniae infections at King Abdulaziz University Ho...	OBJECTIVE: This retrospective chart review describes the epidemiology and clinical features of 40 pa...	2001-07-04	document_parses/pdf_json/d1aafb70c066a2068b02786f8929fd9c900897fb.json
02tnwd4m	Nitric oxide: a pro-inflammatory mediator in lung disease?	Inflammatory diseases of the respiratory tract are commonly associated with elevated production of n...	2000-08-15	document_parses/pdf_json/6b0567729c2143a66d737eb0a2f63f2dce2e5a7d.json
ejv2xln0	Surfactant protein-D and pulmonary host defense	Surfactant protein-D (SP-D) participates in the innate response to inhaled microorganisms and organi...	2000-08-25	document_parses/pdf_json/06ced00a5fc04215949aa72528f2eeaae1d58927.json
2b73a28n	Role of endothelin-1 in lung disease	Endothelin-1 (ET-1) is a 21 amino acid peptide with diverse	2001-02-22	document_parses/pdf_json/348055649b6b8cf2b9a376498df9bf41f71

Figure 1: Meta Data File Provided by Kaggle

To solve the CORD-19 Challenge, the project was divided into 4 parts namely Data Filtration, Data Preprocessing, Models Utilized and Method Used. In the Data Filtration step, the large data was filtered using features such as Publish Time, Abstract, Title, Body Text and Number of Characters in a sentence of Abstract and Body Text. Furthermore, in the Data Preprocessing Step, the Paragraphs of Abstract and Body Text would be converted to Sentences. After this, these Sentences are further converted to Tokens. In the Model Utilized step, the Word2Vec and Doc2Vec Model will be used to generate Word Embeddings and Sentence Embeddings respectively. There were three variations of Word2Vec and Doc2Vec Model based on Sentence Embedding of Question, Query and Narrative present in Question File. The last step is Method Used, where we would utilize Cosine Similarity to calculate the angle or similarity between two sentence embeddings. This method would help in extracting 10 Relevant Answers.

The Relevance Judgements File was utilized in the evaluation of the Models and Methods used. The Relevance Value of the top 10 Relevant Answers would be summed up for each question presented in the Question File. This is called the Aggregate Relevance Score. The Model with the highest Relevance Score will be the best Model.

A ug7v899j	# -2.939983606338501	# -6.312200546264648	# -1.04590308666229...	# 5.164162635803223
970835 unique values	-11.3 9.15	-13.2 6.61	-11 11.2	-6.8 15.1
02tnwd4m	4.688465595245361	-6.284460067749023	1.2282459735870361	5.569900035858154
ejv2xln0	0.5599197745323181	-6.754056930541992	2.271944999694824	7.644495964050293
2b73a28n	2.3525402545928955	-3.9970741271972656	2.3997738361358643	4.59423828125
9785vg6d	-1.2909215688705444	-5.849863529205322	1.5575721263885498	7.989901065826416
zjufx4fo	2.721076250076294	-0.4310147762298584	2.17907977104187	6.504300117492676
5yhe786e	-3.5625340938568115	-5.939841270446777	1.782112956047058	6.769374370574951
8zchiykl	-0.00439921021461486 8	-4.3564133644104	3.906830310821533	3.204770565032959
8qnrcgnk	2.5693230628967285	-5.584765434265137	4.62706184387207	6.406898498535156
jg13scgo	-4.686872959136963	-5.048419952392578	-3.628264904022217	4.775545597076416
5tkvsudh	-0.0718144029378891	-4.873435974121094	3.813286304473877	4.047080039978027
6lvn10f4	-5.229452133178711	-2.4634315967559814	1.9291044473648071	4.653711795806885
tvxpckxo	-3.914104461669922	-7.028103828430176	0.3996717929840088	5.510443687438965
mcuixluu	-2.8642659187316895	-5.376745700836182	3.310087203979492	5.558078289031982
6iu1dtyl	2.987952709197998	-5.908342361450195	1.7718592882156372	3.484102725982666

Figure 2: CORD-19 Document Embeddings Provided by Kaggle

	Date	Study	Study Link	Journal	Study Type	Factors	Influential	Excerpt	Measure of Evidence	Added on
0	16/05/2020	Optimal po	https://www.	Elsevier	Modeling	contact rate, quaran	Y	As well, Fig. 3 :	Countries: China; Cities: '	05/27/2020
1	15/05/2020	A model fo	https://arxiv.o	ArXiv	Modeling	lockdown, social con	Y	Comparing the	Timeline: February 15th	05/27/2020
2	13/05/2020	Modeling a	https://arxiv.o	ArXiv	Modeling	quarantine, lockdow	Y	Our study reve	Countries: India and seve	05/27/2020
3	9/05/2020	Impacts of	https://doi.or	J Popul Eco	Modeling	public health measu	Y	We then comp	Countries: China; Timelir	05/31/2020
4	8/05/2020	A multi-reg	https://doi.or	J Appl Mat	Modeling	protecting susceptib	Y	Figure 10 show	-	05/28/2020
5	6/05/2020	Now castin	http://medrx	MedRxiv	Modeling Study	case isolation	Y	Lockdown sho	Countries: India; Timelin	05/30/2020
6	5/05/2020	COVID-19 I	http://medrx	MedRxiv	Modeling Study	self isolation, social (	Y	The epidemic v	Countries: The UK; Timel	05/30/2020
7	1/05/2020	Modeling s	http://medrx	MedRxiv	Modeling Study	Increased antibody t	Y	However, if an	Countries: United States;	05/30/2020
8	1/05/2020	Modeling s	http://medrx	MedRxiv	Modeling Study	Social distancing	Y	Without any ir	Countries: United States;	05/30/2020
9	30/04/2020	COVID-19 C	http://medrx	MedRxiv	Modeling	lockdown, physical d	Y	Following a m	Countries: Nepal; City: K	05/28/2020
10	29/04/2020	Effectivene	http://medrx	MedRxiv	Modeling	limits placed on gath	Y	We estimated	Countries: UK; 40,162 pa	05/28/2020
11	29/04/2020	Detection a	http://medrx	MedRxiv	Modeling Study	case-isolation, gener	Y	Based on Euro	Countries: Argentina; Tin	05/30/2020
12	27/04/2020	Proactive s	https://www.	MedRxiv	Modeling Study	gathering bans; susp	Y	N/A; after the	Countries: China; Timelir	5/11/2020
13	22/04/2020	Modeling t	http://medrx	MedRxiv	Modeling Study	Case isolation, Public	Y	In Japan, durin	Countries: Japan; Timelir	05/28/2020
14	22/04/2020	Impact of S	https://www.	MedRxiv	Modeling Study	school closures coup	Y	ICU bed dema	Countries: Texas, United	5/11/2020
15	21/04/2020	Social inter	http://medrx	MedRxiv	Modeling Study	social distance, work	Y	-; We estimate	Countries: China, the Uni	05/28/2020
16	21/04/2020	Social inter	http://medrx	MedRxiv	Modeling Study	social distance, work	Y	-;Colombia wo	Countries: Columbia; Tir	05/28/2020
17	20/04/2020	Data Driver	https://arxiv.o	ArXiv	Modeling	case isolation, volun	Y	A combination	Countries: Ghana	05/31/2020
18	20/04/2020	Data Driver	https://arxiv.o	ArXiv	Modeling	telework, school clos	Y	The reference	Countries: Ghana	05/31/2020
19	15/04/2020	Interventio	https://doi.or	MedRxiv	Modeling Study	Strict quarantine, scl	Y	For the total a	Countries: Sweden; Time	05/28/2020
20	14/04/2020	A Stratified	https://doi.or	medRxiv	Modeling Study	schools and universit	Y	uncontained: e	Countries: Iran Timeline:	04/19/2020
21	14/04/2020	A Stratified	https://doi.or	medRxiv	Modeling Study	schools, offices, and	Y	uncontained: e	Countries: Iran Timeline:	04/19/2020
22	14/04/2020	Estimating	https://doi.or	MedRxiv	Modeling Study	community contact r	Y	With our parar	Countries: - ; Timeline: -	05/28/2020
23	11/04/2020	Evidence B	https://www.	Int J Surg	Modeling Study	Quarantine, workpla	Y	We ran simula	Countries: United States;	04/28/2020
24	6/04/2020	Sustainable	https://www.	medRxiv	Modeling Study	social protection (e.g	Y	R0 has been es	individuals : 10000000;	04/15/2020
25	3/04/2020	New measu	https://doi.or	Clin Infect I	Other	case isolation, 14-da	Y	These data sug	Countries: China: Cities: '	05/28/2020
26	2/04/2020	Only strict	https://doi.or	Euro Surveil	Other	time spent in public,	Y	Looking at the	Countries: Italy	05/28/2020
27	31/03/2020	Strong corr	https://arxiv.o	ArXiv	Modeling	social interaction rec	Y	For Italy and Fr	Countries: Brazil, China, I	05/28/2020
28	30/03/2020	General Mo	https://doi.or	medRxiv	Modeling Study	intercity travel and a	Y	if the level of a	Countries: Japan, USA; Ti	4/11/2020

Figure 3: Target Table Provided by Kaggle

Question_ID	Round	CORD_ID	Relevance_Value
1	0.5	010vptx3	2
1	1.0	02f0opkr	1
1	1.0	04ftw7k9	0
1	1.0	05qglt1f	0
1	1.0	0604jed8	0
1	1.0	084o1dmp	0
1	0.5	0be4wta5	2
1	0.5	0hnh4n9e	1
1	1.0	0nh58odf	2
1	1.0	0pbjttv4	0
1	1.0	0xhho1sh	2
1	0.5	11edrkav	2
1	0.5	13jupb26	0
1	1.0	1585stal	0
1	1.0	1a8uevk8	0
1	1.0	1e28zj1d	0
1	0.5	1esupl4q	0

Figure 4: Relevance Judgements provided by NIST

query	question	narrative
coronavirus origin	what is the origin of COVID-19	seeking range of information about the SARS-CoV-2 virus's origin, including its evolution, animal source, and first transmission into humans
coronavirus response to weather changes	how does the coronavirus respond to changes in the weather	seeking range of information about the SARS-CoV-2 virus viability in different weather/climate conditions as well as information related to transmission of the virus in different climate conditions
coronavirus immunity	will SARS-CoV2 infected people develop immunity? Is cross protection possible?	seeking studies of immunity developed due to infection with SARS-CoV2 or cross protection gained due to infection with other coronavirus types
how do people die from the coronavirus	what causes death from Covid-19?	Studies looking at mechanisms of death from Covid-19.
animal models of COVID-19	what drugs have been active against SARS-CoV or SARS-CoV-2 in animal studies?	Papers that describe the results of testing drugs that bind to spike proteins of the virus or any other drugs in any animal models. Papers about SARS-CoV-2 infection in cell culture assays are also relevant.
coronavirus test rapid testing	what types of rapid testing for Covid-19 have been developed?	Looking for studies identifying ways to diagnose Covid-19 more rapidly.
serological tests for coronavirus	are there serological tests that detect antibodies to coronavirus?	Looking for assays that measure immune response to COVID-19 that will help determine past infection and subsequent possible immunity.
coronavirus under reporting	how has lack of testing availability led to underreporting of true incidence of Covid-19?	Looking for studies answering questions of impact of lack of complete testing for Covid-19 on incidence and prevalence of Covid-19.
coronavirus in Canada	how has COVID-19 affected Canada	seeking data related to infections (confirm, suspected, and projected) and health outcomes (symptoms, hospitalization, intensive care, mortality)
coronavirus social distancing impact	has social distancing had an impact on slowing the spread of COVID-19?	seeking specific information on studies that have measured COVID-19's transmission in one or more social distancing (or non-social distancing) approaches

Figure 5: Question File Provided by NIST

2 METHODOLOGY

This section is divided into four parts which are Data Filtration, Data Preprocessing, Models Utilized, and Method Used. The Pipeline of the Methodology is given in Figure 6.

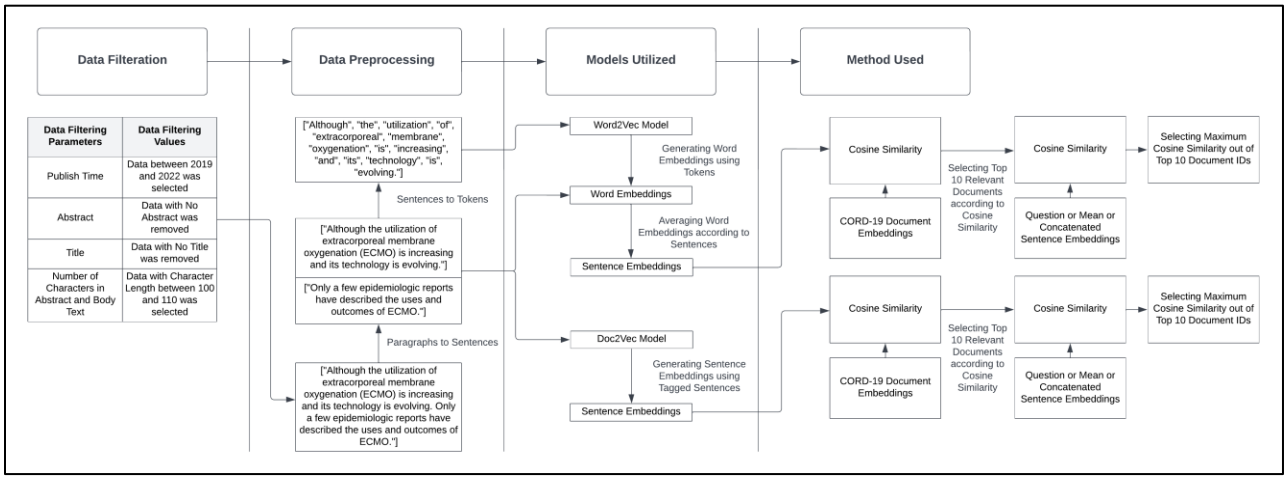

Figure 6: Pipeline of the Methodology Used

2.1 Data Filtration

The Document Parses of PDF Research Papers required approximately 40 Gigabytes (GB) of Random Access Memory (RAM). However, our system had only 16 GB of RAM. Therefore, data filtration was one of the most significant sections of this project. For this Publish Time, Title, Abstract and PDF JSON Files (Location of PDF Document Parses directory) of Meta Data File were utilized to filter the number of Research Papers. Moreover, the Number of Characters in the Abstract and Body Text given in the PDF Document Parses of the Research Papers were also utilized in Data Filtration.

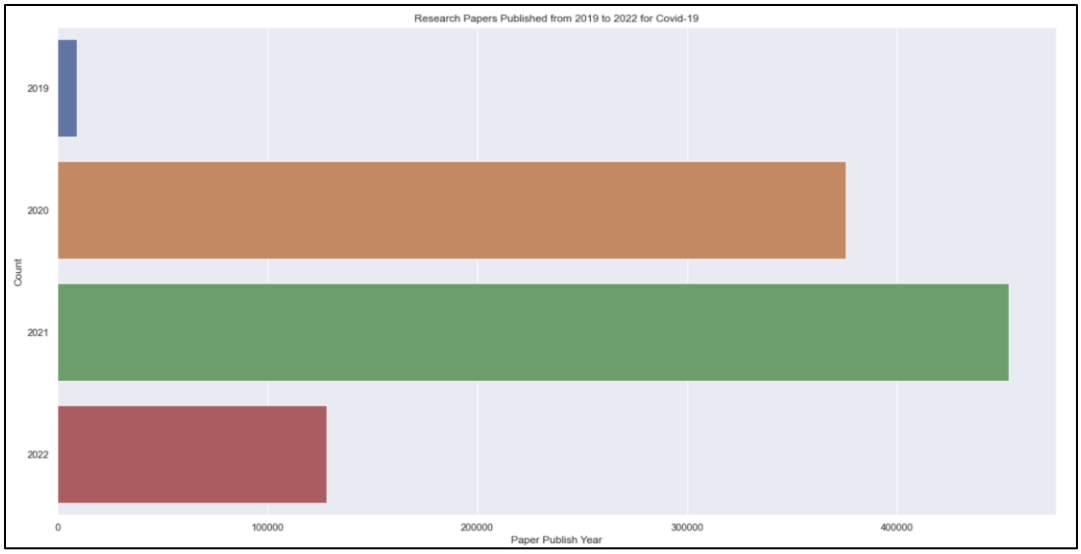

Figure 7: Publish Time of Research Papers which were selected

The Research Papers with Publish Time ranging from 2019 to 2022 were selected because Covid-19 started spreading in 2019. The number of Research Papers filtered using Publish Time is shown in Figure 7. Furthermore, the Research

Papers with No Abstract and No Title were also removed as most of the summarized information is present in the Abstract, therefore, Articles without it had to be removed. Furthermore, Research Articles without any PDF JSON Files were removed because the Body Text of the Research Article contains information about a topic in detail. The Statistics of Research Papers with and without an Abstract are shown in Figure 8 and Research Articles with and without PDF JSON Files are shown in Figure 9.

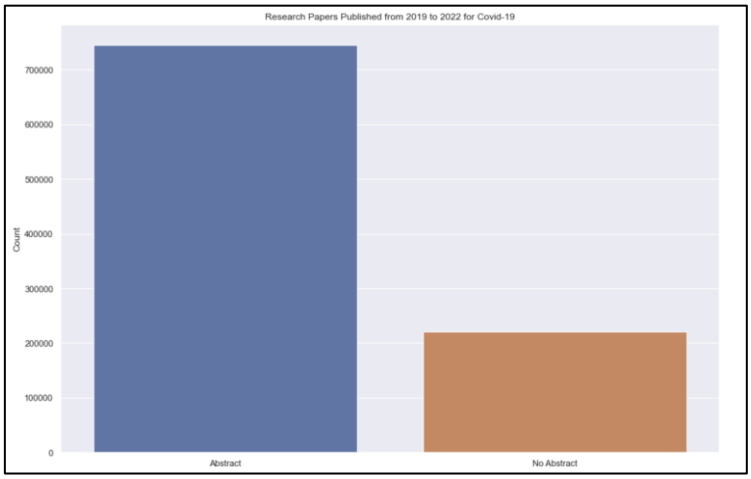

Figure 8: Number of Research Papers with and without Abstract

The Body Text and Abstract extracted from the PDF Document Parses were filtered based on the Number of Characters in their Sentences. The Abstract and Body Text sentences with Character Lengths between 100 and 110 were selected because a sentence generally contains approximately 12 to 16 words which equate to around 100 to 110 characters. Moreover, the frequency of sentences with character lengths between 100 and 110 was in the order of 10^4 shown in Figure 10 due to which this range for Character Length was selected.

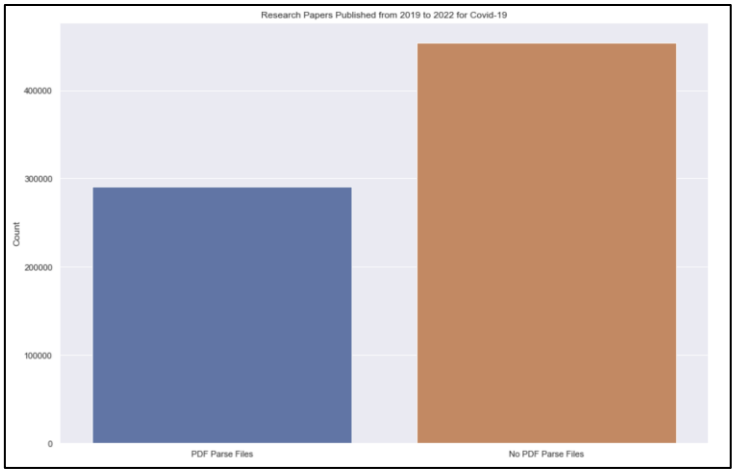

Figure 9: Number of Research Papers with and without PDF JSON or Parse Files

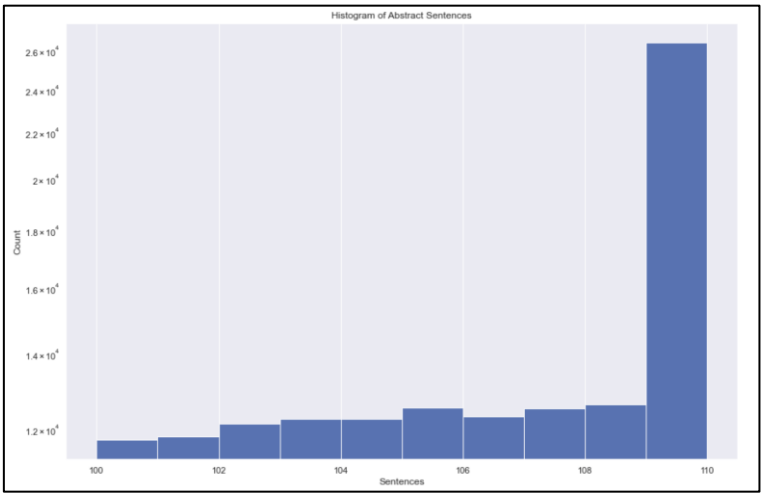

Figure 10: Number of Research Papers with Abstract Character Length between 100 and 110

2.2 Data Preprocessing

The Data Preprocessing step was necessary for the Model Utilized section as the Word2Vec model requires Tokens and the Doc2Vec model requires Tagged Sentences. Therefore, to accomplish this Paragraphs of Abstract and Body Text were converted to Sentences and these Sentences were further converted to tokens which are explained below in detail.

2.2.1Paragraph to Sentences

The Paragraphs of Abstract and Body Text were converted to sentences with the help of Punkt Sentence Tokenizer in the NLTK library. Each of these sentences was provided with its corresponding CORD ID with the help of the TaggedDocument() function present in the Gensim Library. These Tagged Sentences were fed to Doc2Vec Model for generating Sentence Embeddings by training the model [1]. This is shown in Figure 6.

2.2.2Sentences to Tokens

The sentences of Abstract and Body Text received from Sentence Tokenizer as well as Question, Query and Narrative Text present in Question File are tokenized (by using the text_to_word_sequence() function in TensorFlow Library) to words for Training the Word2Vec Model to generate Word Embeddings [6]. This is shown in Figure 6.

2.3 Models Utilized

The most important aspect of any project is the selection of different models which can be trained and evaluated to get the best results. The two models selected for the CORD-19 challenge were Word2Vec Model and Doc2Vec Model which are explained below.

2.3.1*Word2Vec Model*

The Word2Vec Model comes in two variations that are Continuous Bag of Words (CBOW) and Continuous Skip-Gram (CSG). In the CSG Word2Vec Model, the context words are Predicted from a target word. Whereas, in CBOW Word2Vec Model the target words are predicted using the context words [3][7][8]. In the Project, the CBOW Word2Vec Model was

selected because it took less time to get trained which is ideal for a big dataset. The implementation of the Word2Vec model was provided by the Gensim Library. The CBOW Word2Vec Model was trained using the Tokens generated in the Data Preprocessing step shown in Figure 6. The CBOW Word2Vec Model generated Word Embeddings of each word (received from text_to_word_sequence() function) using the models.word2vec.wv() function. These Word Embeddings were averaged according to sentences provided by Punkt Tokenizer to generate Sentence Embeddings that would be utilized in Cosine Similarity. This is also displayed in Figure 6. The CBOW Word2Vec Model is shown in Figure 11.

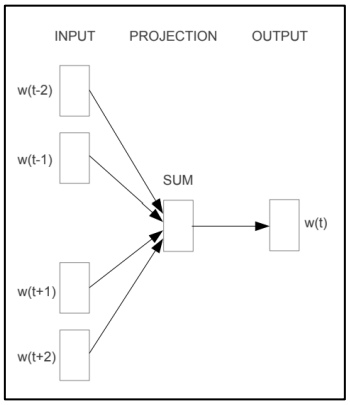

Figure 11: The Word2Vec CBOW Model Provided by Gensim [8]

2.3.2*Doc2Vec Model*

The Doc2Vec Model comes in two variations that are Distributed Bag of Words (DBOW) and Distributed Memory (DM). In the DBOW Doc2Vec Model, the Document Tags which were CORD IDs in this project were utilized to generate Sentence Embeddings. Whereas, in Distributed Memory Doc2Vec Model the Document Sentences are utilized to generate Sentence Embeddings [2][4]. The Documents were Tagged with the help of the TaggedDocument() function of Gensim Library. The DBOW Doc2Vec Model was utilized in this challenge because it took less time to get trained. The DBOW Doc2Vec Model used doc2vec.infer_vector() function to generate sentence embeddings for sentences which were provided by Punkt Tokenizer by tokenizing Abstract and Body Text of Document Parses and Question File. These Sentence Embeddings were further used in Cosine Similarity as shown in Figure 6. The DBOW Doc2Vec Model is shown in Figure 12 provided by Gensim Library.

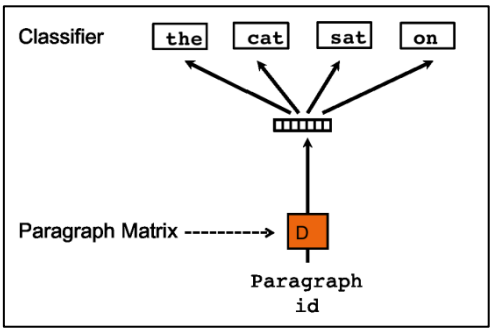

Figure 12: The Doc2Vec DBOW Model Provided by Gensim [2]

2.4 Method Used

To select the Top 10 Relevant Answers or Sentences Cosine Similarity was used as shown in Figure 6. Cosine Similarity is the angle or similarity between two vectors or sentence embeddings of an inner product space. The formula of Cosine Similarity is shown in Figure 13. The Scipy Library's 1 – cosine.distance() was utilized [5].

First, the Cosine Similarity is calculated between the CORD-19 Sentence Embeddings Provided by Kaggle and the Question, Mean (generated by averaging the sentence embeddings of Question, Query and Narrative in Question File provided by NIST) and Concatenated (generated by concatenating the sentence embeddings of Question, Query and Narrative in Question File provided by NIST) Sentence Embeddings which were created using the Question File provided by Question File. Then 10 unique CORD IDs with the highest Cosine Similarity were Selected and Sent to the Second stage. In the Second Stage, the Cosine Similarity between the Sentences of Abstract and Body Text of Document Parses and the Question, Mean and Concatenated Sentence Embedding. Then under 10 unique CORD IDs extracted in Stage one the sentences with maximum Cosine Similarity were retrieved by using GroupBy Function of Library Pandas and these were the Top 10 Relevant Answers.

$$\cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$

Figure 13: The formula of Cosine Similarity

3 RESULTS AND FINDINGS

The Relevance Score is calculated by utilizing the Relevance Value Provided in the Question File Provided by NIST. The Relevance Value of 10 CORD IDs (for each question) retrieved in the Second Stage of Cosine Similarity was summed up for all Questions. Therefore, if there were 10 Questions in the Question File then the Relevance Value of 100 Answers or Sentences was summed up to generate a Relevance Score.

Table 1: Relevance Score of different Word2Vec and Doc2Vec Models

MODELS	RELEVANCE SCORE
Word2Vec Question Model	635
Word2Vec Mean Model	557
Word2Vec Concatenated Model	730
Doc2Vec Question Model	600
Doc2Vec Mean Model	600
Doc2Vec Concatenated Model	611

The Relevance Value is of three types namely ‘0’ – Irrelevant Answer, ’1’ – Partially Relevant Answer and ‘3’ – Fully Relevant Answer. The higher the value of the Relevance Score the better the Model is. When observing Table 1 we realize that the best model is the Word2Vec Concatenated Model with a Relevance Score of 730 while the Second-Best Model is the Word2Vec Question Model with a Relevance Score of 635. The worst performing models had an averaged sentence embedding of Question, Query and Narrative i.e., Word2Vec and Doc2Vec Mean Models.

4 DISCUSSIONS AND CONCLUSIONS

Out of all six models, the best model was the Word2Vec Concatenated Model with Concatenated Sentence Embedding (of Question, Query and Narrative given in Question File) as it showed a Relevance Score of 730 which was the highest compared to other models. Whereas the Second-best Model Word2Vec Question Model with Question Sentence Embedding had a Relevance Score of 635. In Doc2Vec Models, the best Model was found to be Doc2Vec Concatenated Model with Concatenated Sentence Embedding.

Data Filtering became an essential unit of this project as the data of 10 million Research Papers was massive and difficult to process with a RAM of about 16GB.

Data Preprocessing played a significant role in the implementation of a QA Chatbot as both Sentences and Tokens were required by both Word2Vec and Doc2Vec Models to generate sentence Embeddings which would be utilized to calculate Cosine Similarity.

The COVID-19 Open Research Dataset Challenge (CORD-19) opens various conversations for future improvements. The Word and Sentence Embedding of each word and sentence respectively can be multiplied by their Inverse Document Frequency (IDF) Score to make less frequent words more important. Moreover, the Word and Sentence Embeddings of each word and sentence can also be weighted by using the BM (Best Matching) – 25 Ranking Algorithm. This would increase the accuracy of the Word2Vec or Doc2Vec Model. Furthermore, the FastText Model can also be used instead of Word2Vec or Doc2Vec Model because it uses character N-Grams. Also, a pre-trained sentence embedding Model can be used such as SentenceBERT which takes considers context while generating Sentence Embeddings.

REFERENCES

[1] NLTK. Punkt Sentence Tokenizer. Retrieved August 13, 2022 from https://www.nltk.org/_modules/nltk/tokenize/punkt.html

[2] Quoc V. Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. (May 2014). Retrieved August 13, 2022 from https://arxiv.org/abs/1405.4053v2

[3] Radim Řehůřek. 2022. Gensim: Topic modelling for humans. (May 2022). Retrieved August 13, 2022 from https://radimrehurek.com/gensim/models/word2vec.html

[4] Radim Řehůřek. 2022. Gensim: Topic modelling for humans. (May 2022). Retrieved August 13, 2022 from https://radimrehurek.com/gensim/models/doc2vec.html

[5] Scipy. Scipy.spatial.distance.cosine#. Retrieved August 13, 2022 from https://docs.scipy.org/doc/scipy/reference/generated/scipy.spatial.distance.cosine.html

[6] TensorFlow. Tf.keras.preprocessing.text.text_to_word_sequence : Tensorflow Core v2.9.1. Retrieved August 13, 2022 from https://www.tensorflow.org/api_docs/python/tf/keras/preprocessing/text/text_to_word_sequence

[7] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. (October 2013). Retrieved August 13, 2022 from https://arxiv.org/abs/1310.4546

[8] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. (September 2013). Retrieved August 13, 2022 from https://arxiv.org/abs/1301.3781