Automatic Text Summaration of COVID-19 Scientific Research Topics Using Pre-trained Model from HuggingFace®

Sakdipat Ontoum [1] and Jonathan H. Chan [1]

[1]Affiliation not available

October 30, 2023

Abstract

By identifying and extracting relevant information from articles, automated text summarizing helps the scientific and medical sectors. Automatic text summarization is a way of compressing text documents so that users may find important information in the original text in less time. We will first review some new works in the field of summarizing that use deep learning approaches, and then we will explain the "COVID-19" summarization research papers. The ease with which a reader can grasp written text is referred to as the readability test. The substance of text determines its readability in natural language processing. We constructed word clouds using the abstract's most commonly used text. By looking at those three measurements, we can determine the mean of "ROUGE-1", "ROUGE-2", and "ROUGE-L". As a consequence, "Distilbart-mnli-12-6" and "GPT2-large" are outperform than other.

Automatic Text Summaration of COVID-19 Scientific Research Topics Using Pre-trained Model from HuggingFace®

Sakdipat Ontoum* Jonathan H. Chan*
*Computer Science Program, School of Information Technology
King Mongkut's University of Technology Thonburi
Bangmod, Thung-Khru, Bangkok, Thailand

***Abstract*—By identifying and extracting relevant information from articles, automated text summarizing helps the scientific and medical sectors. Automatic text summarization is a way of compressing text documents so that users may find important information in the original text in less time. We will first review some new works in the field of summarizing that use deep learning approaches, and then we will explain the "COVID-19" summarization research papers. The ease with which a reader can grasp written text is referred to as the readability test. The substance of text determines its readability in natural language processing. We constructed word clouds using the abstract's most commonly used text. By looking at those three measurements, we can determine the mean of "ROUGE-1", "ROUGE-2", and "ROUGE-L". As a consequence, "Distilbart-mnli-12-6" and "GPT2-large" are outperform than other.**

***Index Terms*—automatic summarization, COVID-19, CORD-19, Hugging Face®, "Latent Dirichlet allocation (LDA)","Recall-Oriented Understudy for Gisting Evaluation (ROUGE)"**

I. Introduction

In December 2019, a severe acute respiratory syndrome coronavirus2 outbreak expanded to Wuhan, Hubei Province, China, and subsequently to the rest of China and the world. The sickness was dubbed "Coronavirus Disease 2019 (COVID-19)" by the "World Health Organization (WHO)" in February 2020 [4]. The "WHO" classified "COVID-19" a pandemic after 200,000 confirmed cases and 8,000 fatalities in more than 160 countries [5]. The hospital capacity was abruptly reduced in a short period of time due to the admission of "COVID-19" patients. This unfavorable circumstance, which was placed on countries unexpectedly, spurred experts to begin their study on the virus in order to uncover mechanisms of transmission, preventative techniques, new treatments, and vaccines. In a circumstance where every minute counts in saving the lives of hundreds of people, reviewing and digesting scientific publications takes time [6].

"Natural language processing (NLP)" is a field of computer science, artificial intelligence, and linguistics concerned with computer interactions with human (natural) language [1]. Natural languages are significant tools for communicating with each other that they can learn from their environment. Natural languages and forms of communication are used to represent emotions and knowledge and to transmit our responses to other people and their surroundings [1]. "Natural language processing (NLP)" is a set of methods for extracting grammatical structure and meaning from input in order to execute a meaningful task [1].

Automatic summarization has gained widespread popularity in "Natural Language Processing (NLP)" due to its potential for a variety of information access applications [2]. The act of reducing a piece of text into a shorter version that contains just the most significant information is known as summarization. There are two sorts of original summarizing methods: *extractive* and *abstractive*. *Extractive* techniques generate summaries primarily from passages (typically full sentences) extracted straight from the source text, but *abstractive* approaches, such as a human-written abstract, may generate fresh words and phrases not found in the original text [3]. However, additional talents required for high-quality summarizing, such as paraphrase, generalization, or assimilation of real-world knowledge, can only be attained in a *abstractive* framework [3].

Furthermore, the automatic text summarization assists the research and medical communities by recognizing and extracting important information from papers. Automatic text summarization is a method of creating a compressed version of text documents that allows readers to discover relevant information in the original text in less time [7]. The resulting summary should be brief and include key information from the text [8].

In this paper, we will first discuss some recent studies in the field of summarization utilizing deep learning methods, and then we will explain the "COVID-19" research articles in the field of summarization. The remainder of this paper is structured as follows: Section 2 discusses literature review; Section 3 discusses research methodology; Section 4 discusses results and discussion; Section 5 concludes; Section 6 appendices.

II. Literature Review

A. *"DeepMINE"*

"DeepMINE" is a system suggested by Joshi et al. [6]. This method is divided into two parts: "Mine Article" and "Article Summarization". The user inputs the required keywords in the first section, and the system retrieves related articles and links by searching in the title of the articles given by CORD-19. The second component uses deep learning and Natural Language Processing to summarize an input article (NLP).

B. *"SummaRuNNer"*

"Nallapati et al". proposed "SummaRuNNer", a recurrent neural network sequence model based on extractive summarization, in 2017 [18]. A "two-layer bidirectional GRU" is used in the proposed model. The first layer computes word hidden representations consecutively, whereas the second layer computes sentence hidden representations depending on the prior layer's word representations.

In a "logistic layer", each sentence is encountered sequentially, and a binary judgement is taken as to whether the sentence belongs to the summary.

III. Research Methodology

A. *Data Understanding*

1) ***Dataset:*** In response to the "COVID-19" epidemic, the "White House" and "a collaboration of premier academic groups" created the "COVID-19" Open Research Dataset (CORD-19)" [10]. "CORD-19" is a database including over 500,000 research publications on "COVID-19", "SARS-CoV-2", and other coronaviruses, including over 200,000 with full text [9][10]. This publicly available dataset is being made available to the global research community in order for them to leverage current advances in natural language processing and other AI technologies to generate new insights in support of the ongoing fight against this fatal sickness [9][10]. There is a rising need for these methodologies because to the great pace of new coronavirus literature, which makes it difficult for the medical research community to keep up [9][10].

2) ***Readability Tests:*** The ease with which a reader can grasp a written material is referred to as readability. The readability of text in natural language processing is determined by its content. It focuses on the words we pick and how we arrange them in phrases and paragraphs so that readers may understand them [13]. Higher results in the Flesch reading-ease test indicate easier-to-read content, while lower numbers suggest more difficult-to-read sections [13]. The "Flesch reading-ease score (FRES)" test formula as Eqn. 1 is

$$206.835 - 1.015(\frac{\text{total words}}{\text{total sentences}}) - 84.6(\frac{\text{total syllables}}{\text{total words}}) \quad (1)$$

Where:

90-100 :	Very Easy
80-89 :	Easy
70-79 :	Fairly Easy
60-69 :	Standard
50-59 :	Fairly Difficult
30-49 :	Difficult
0-29 :	Very Confusing

In Fig.1, "Flesch Reading Ease" calculation on abstracts shows that the mean value was about 30.2 with normal distribution curve bell shape for all data, implying that the abstracts are a little difficult to read and grasp.

B. *Data Preparation*

1) ***Stopwords Removal:*** the process of transforming data into something that a computer can understand. Filtering out worthless data is a common type of pre-processing. In natural language processing, stop words are worthless words [11]. Therefore, A stop word is a regularly used term (for example, "the," "a," "an," or "in") that a search engine has been configured to ignore, both while indexing items for searching and retrieving them as the result of a search query [11].

2) ***Bigram Words:*** the two-element sequence from a string of tokens, which are commonly letters, syllables, or words [12]. For simple statistical analysis of text, the frequency distribution of bigrams in a string is widely utilized [12].

3) ***Lemmatization:*** The abstract of covid-19 scientific paper have many kind of word that not specify for analysing the frequency distribution of text. For instance, punctuation word, abd Math Symbol. Therefore, there are keeping only noun, adjective, verb, and adverb. Then lemmatize word to original words (Root Words).

C. *Exploratory Data Analysis (EDA)*

1) ***Word Clouds of Abstract:*** The most text frequently of the abstract, we created word clouds. We were created by those method from data preparation or data pre-processing of the texts. The size of each term in the word clouds is then proportional to its frequency of appearance in the top words on each abstract. As the result, the most of the words that show in Fig.2, there is about covid, patient, outbreak, hospital, and stemi.

2) ***"Latent Dirichlet allocation (LDA)":*** The crucial step of "EDA" for determining the most often occurring topic modeling analysis. There are two types of statistical distributions: those based on the distributional hypothesis and those based on statistical distributions. As a result, "LDA" will assist us in mapping each document in the corpus. Then there is a collection of subjects that encompass a significant portion of the words in the paper, as illustrated in Fig.3 and Fig.4 For example, "intertopic distance map" and "Top-30 Most Relevant Terms for Topic 1".

As a result, when we adjust relevance metric with lambda is corresponding with one. The topic 1 is about study about "COVID-19", the case of infection that high risk areas, and what kind of clinical to "COVID-19" symptomatic threament.

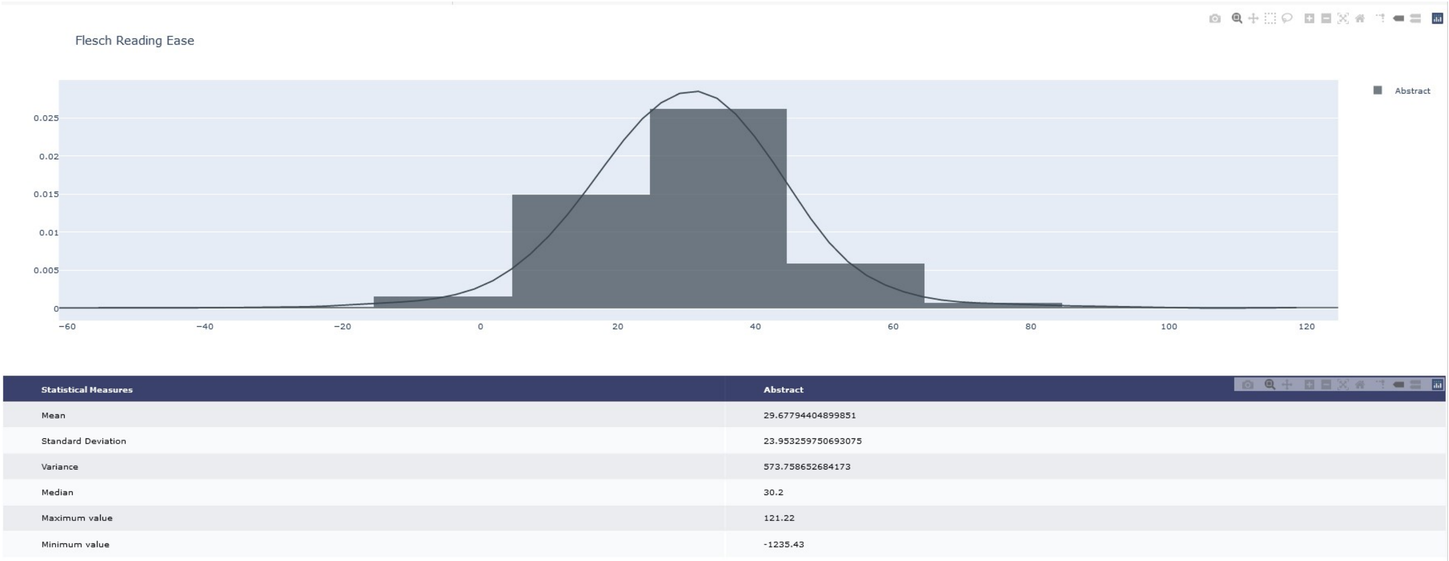
3) ***Topic Distribution Plot:*** The distribution of dominant Topics in each document as Fig.5. Moreover, the distribution of topic weightage as Fig.6. As a result, the topic 6 which about patient, study, and covid are both highest number of documents.

4) ***Perplexity and Coherence Score:*** The coherence score is used to examine the quality of the learnt subjects. Perplexity is a measurement of how well a probability distribution or probability model predicts a sample. As the result, the perplexity of of corpus that measured with topic modeling is -7.828 , and the coherence score is 0.436.

D. *Modeling*

The Hugging Face®, a company that provides the open-source of NLP technologies. For instance, language model, the library that imperative to trained the data with specific problem. As our text summarization of "CORD-19", we utilized four pre-trained models.

1) ***"Distilbart-mnli-12-6":*** the model that low usage environment.The technique of this model using "No Teacher Distillation technique" [15]. There is default model for proposing text summarization. This is a very simple and effective technique, since the performance are succinct dropping, and trade-offs [16].



Statistical Measures	Abstract
Mean	29.67794404899851
Standard Deviation	23.953259750693075
Variance	573.758652684173
Median	30.2
Maximum value	121.22
Minimum value	-1235.43

Fig. 1. Flesch Reading Ease Calculation

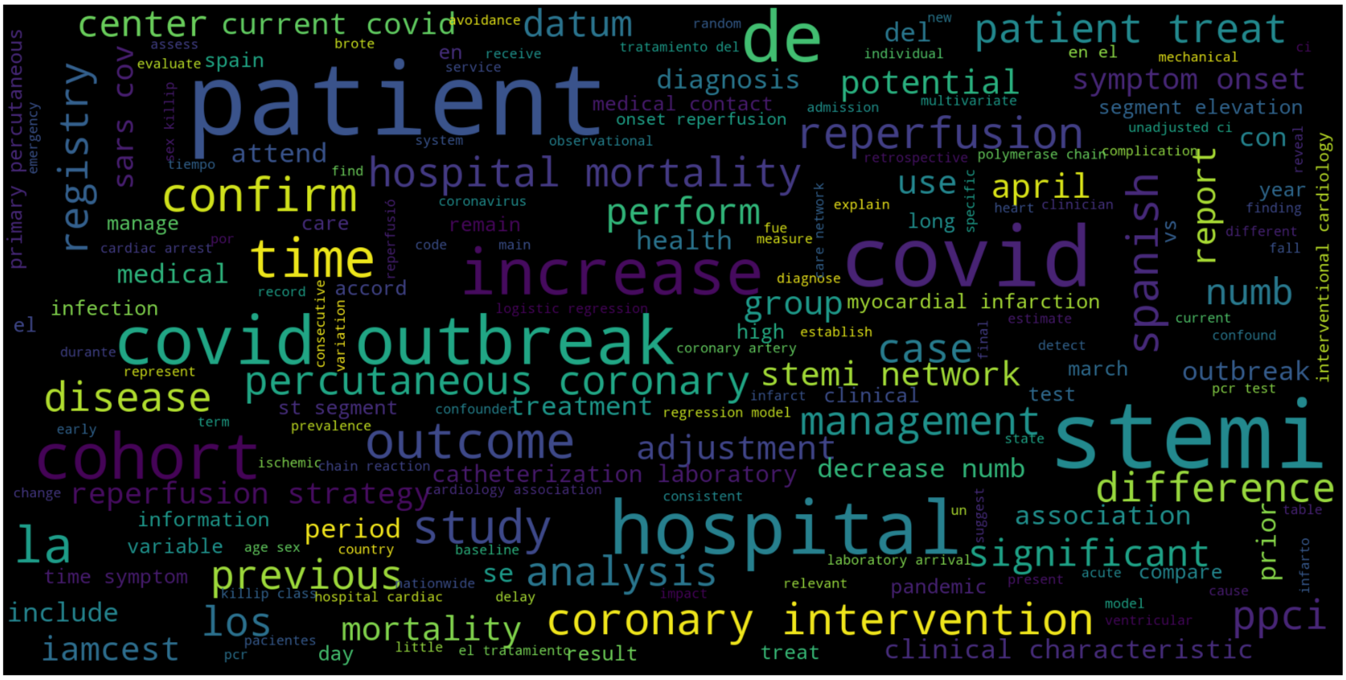


Fig. 2. The most text frequently of the abstract.

2) “Bigbird-pegasus-large-arxiv”: the model is utilizing “sparse-attention” that cam make model have much longer sequences [17][18]. Also, the model is checkpoint from “RoBRRTa”. So, there is compatible with various task involving long document such as text summarization on scientific abstract paper [20].

3) “GPT2-large”: the distinguished model that embedding of words with absolutely in term of position. there is prudent to pad the inputs from the right to the left with alternatively. “GPT-2” is kind of “casual language modeling (CLM)”. Therefore, in term of predicting the token in sequence is forceful.

4) “T5-large”: the model that encode-decode the multi-task, and converted the format with text-to-text. There is works well on thought process of task by prepending a different prefix to the input corresponding.

IV. Results and Discussion

The evaluation of automatic summarization, we utilized the “Recall-Oriented Understudy for Gisting Evaluation (ROUGE)” metric. there are There is a comparison of an autonomously generated summary or translation to a reference or group of references (human-generated) summary or translation [15]. The formula of ROUGE-N is shown in below as Eqn.2 [15].

Most of “ROUGE-N” that utilized for comparing the performance on each model to summarize the text [16][17]. There are have “ROUGE-1”, they would be monitoring the match-rate of unigrams between our model output and the reference. Bigrams

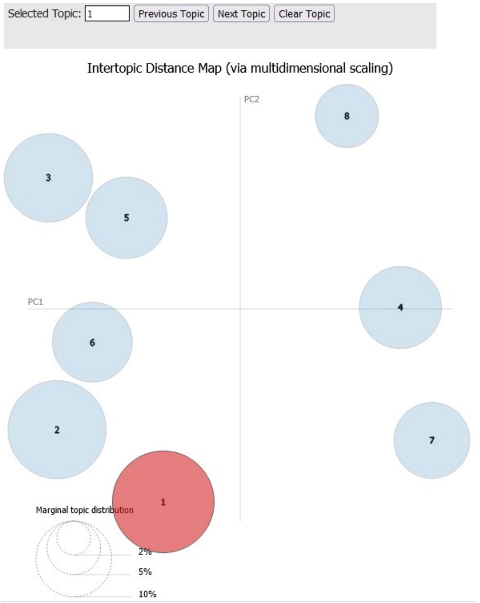
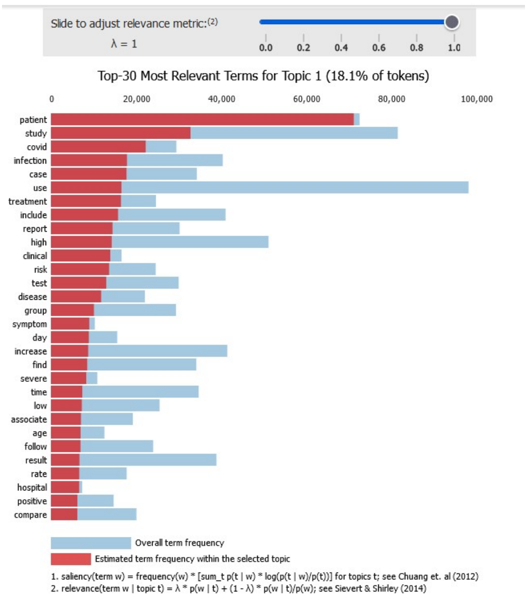

Selected Topic: 1
Previous Topic
Next Topic
Clear Topic
Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
8
3
5
4
6
2
7
1
Marginal topic distribution
2%
5%
10%

Fig. 3. Intertopic Distance Map



1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Fig. 4. Top-30 Most Relevant Terms for Topic 1

and trigrams would be used by “ROUGE-2” and “ROUGE-3”, respectively [16][17].

In this case, we utilized the F-score, Precision, and Recall metric on each “ROUGE-N”. Because of this, there is can make known more of the result in accurately. In Fig.7 shows that “Distilbart-mnli-12-6” and “GPT2-large” are outperform than other pre-trained model. If we calculated mean of “ROUGE-1”, “ROUGE-2”, and “ROUGE-L” by looking those three metrics. As a result, “Distilbart-mnli-12-6” have 0.556 ROUGE-F-Score, 0.90 ROUGE-Precision, 0.40 ROUGE-Recall.“GPT2-large” have 0.551 ROUGE-F-Score, 0.99 ROUGE-Precision, 0.34 ROUGE-Recall. Those value is come from selecting most one topic that generate with “LDA model”.

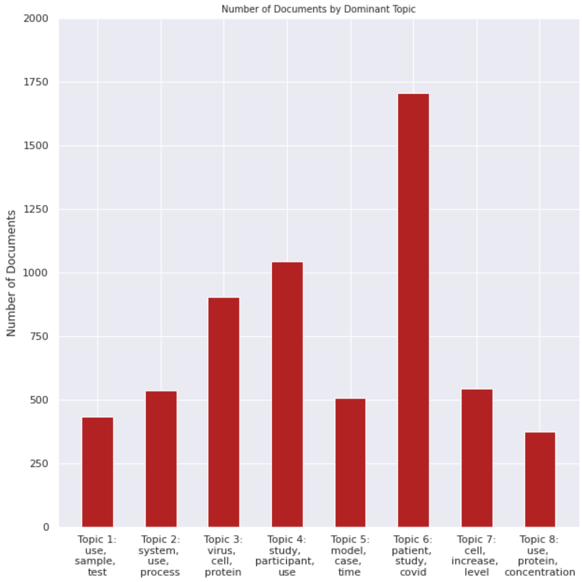


Fig. 5. The distribution of dominant Topics in each document.

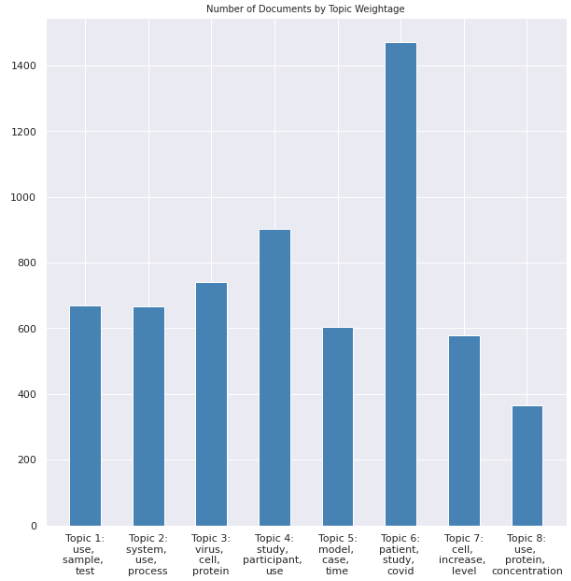


Fig. 6. the distribution of topic weightage.

V. Conclusion

In this article, we provide an interpretable deep learning-based technique for summarizing “COVID-19” scientific research publications. This challenge is similar to a sentence regression approach. We employ any pre-trained model from The Hugging Face® to locate the most informative words.

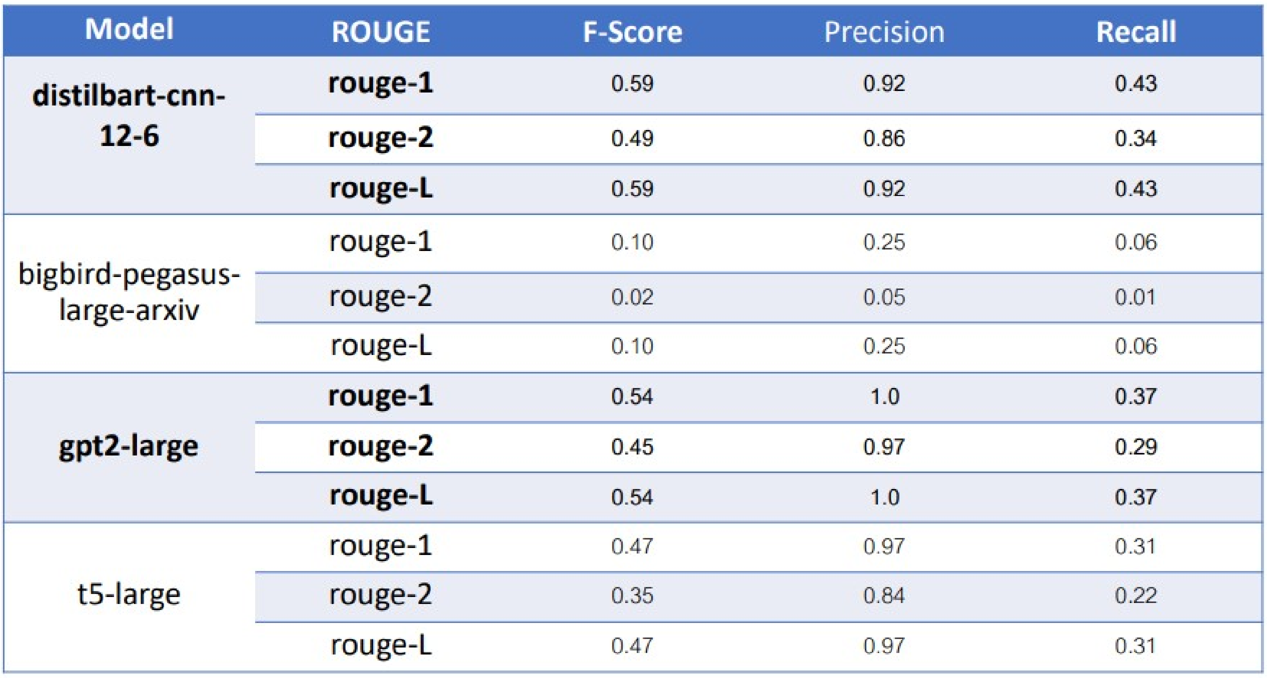
In the realm of summarization, we investigated the performance of several types of pretrained models. The best results were obtained by “Distilbart-mnli-12-6” and “GPT2-large”. We demonstrated that selecting better sentences to include in the summary is more suitable information from the sentences by taking both in the sentences into account.

In the future, We intend to utilize the “Generative Pre-trained Transformer 3 (GPT-3)” which is an autoregressive language model that leverages deep learning to generate human-like text

$$ROUGE-N=\frac{\sum_{S\in\{ReferenceSummaries\}}\sum_{gram_n\in S}Count_{match}(gram_n)}{\sum_{S\in\{ReferenceSummaries\}}\sum_{gram_n\in S}Count(gram_n)} \tag{2}$$

Where:

$$\begin{aligned} n&: \text{“the length of the n-gram, and } gram_n\text{”}\\ Count_{match}(gram_n)&: \text{“the maximum number of n-grams”}\\ -------------''-----------&: \text{“co-occurring in a candidate summary”}\\ -------------''-----------&: \text{“and a set of reference summaries.”} \end{aligned}$$



Model	ROUGE	F-Score	Precision	Recall
distilbart-cnn-12-6	**rouge-1**	0.59	0.92	0.43
	rouge-2	0.49	0.86	0.34
	rouge-L	0.59	0.92	0.43
bigbird-pegasus-large-arxiv	rouge-1	0.10	0.25	0.06
	rouge-2	0.02	0.05	0.01
	rouge-L	0.10	0.25	0.06
gpt2-large	**rouge-1**	0.54	1.0	0.37
	rouge-2	0.45	0.97	0.29
	rouge-L	0.54	1.0	0.37
t5-large	rouge-1	0.47	0.97	0.31
	rouge-2	0.35	0.84	0.22
	rouge-L	0.47	0.97	0.31

Fig. 7. “Distilbart-mnli-12-6” and “GPT2-large” are outperform than other pre-trained model by calculated ROUGE.”

[19].

REFERENCES

[1] Reshamwala, Alpa Mishra, Dhirendra Pawar, Prajakta. (2013). REVIEW ON NATURAL LANGUAGE PROCESSING. IRACST – Engineering Science and Technology: An International Journal (ESTIJ). 3. 113-116.

[2] Narayan, S., Cohen, S. B., Lapata, M. (2018). Ranking sentences for extractive summarization with reinforcement learning. arXiv preprint arXiv:1802.08636.

[3] See, A., Liu, P. J., Manning, C. D. (2017). Get to the point: Summarization with pointer-generator networks. arXiv preprint arXiv:1704.04368.

[4] Zu, Z. Y., Jiang, M. D., Xu, P. P., Chen, W., Ni, Q. Q., Lu, G. M., Zhang, L. J. (2020). Coronavirus Disease 2019 (COVID-19): A Perspective from China. Radiology, 296(2), E15–E25. https://doi.org/10.1148/radiol.2020200490

[5] Spinelli, A., Pellino, G. (2020). COVID-19 pandemic: perspectives on an unfolding crisis. British Journal of Surgery, 107(7), 785–787. https://doi.org/10.1002/bjs.11627

[6] Joshi, B., Bakarola, V., Shah, P., Krishnamurthy, R. (2020). deepMINE - Natural Language Processing based Automatic Literature Mining and Research Summarization for Early-Stage Comprehension in Pandemic Situations specifically for COVID-19. DeepMINE - Natural Language Processing Based Automatic Literature Mining and Research Summarization for Early-Stage Comprehension in Pandemic Situations Specifically for COVID-19. Published. https://doi.org/10.1101/2020.03.30.014555

[7] Yousefi-Azar, M., Hamey, L. (2017). Text summarization using unsupervised deep learning. Expert Systems with Applications, 68, 93–105. https://doi.org/10.1016/j.eswa.2016.10.017

[8] PadmaPriya, G., Duraiswamy, K. (2014). AN APPROACH FOR TEXT SUMMARIZATION USING DEEP LEARNING ALGORITHM. Journal of Computer Science, 10(1), 1–9. https://doi.org/10.3844/jcssp.2014.1.9

[9] Wang, L.L., Lo, K., Chandrasekhar, Y., Reas, R., Yang, J., Eide, D., Funk, K., Kinney, R.M., Liu, Z., Merrill, W.C., Mooney, P., Murdick, D.A., Rishi, D., Sheehan, J., Shen, Z., Stilson, B., Wade, A.D., Wang, K., Wilhelm, C., Xie, B., Raymond, D.A., Weld, D.S., Etzioni, O., Kohlmeier, S. (2020). CORD-19: The COVID-19 Open Research Dataset. ArXiv.

[10] COVID-19 Open Research Dataset Challenge (CORD-19). (2021, November 9). Kaggle. https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge

[11] GeeksforGeeks. (2021, May 31). Removing stop words with NLTK in Python. https://www.geeksforgeeks.org/removing-stop-words-nltk-python/

[12] Wikipedia contributors. (2021, November 1). Bigram. Wikipedia. https://en.wikipedia.org/wiki/Bigram

[13] Wikipedia contributors. (2021, December 4). Flesch–Kincaid readability tests. In Wikipedia, The Free Encyclopedia. Retrieved 15:45, December 21, 2021, from https://en.wikipedia.org/wiki/Flesch-Kincaid_readability_tests

[14] Topic Modeling: An Introduction. (2019, September 26). MonkeyLearn Blog. https://monkeylearn.com/blog/introduction-to-topic-modeling/

[15] Wikipedia contributors. (2021b, November 23). ROUGE (metric). Wikipedia. https://en.wikipedia.org/wiki/ROUGE_(metric)

[16] Lin, C. (2004). ROUGE: A Package for Automatic Evaluation of Summaries. ACL 2004.

[17] Briggs, J. (2021, December 18). Measure NLP Accuracy With ROUGE — Towards Data Science. Medium. https://towardsdatascience.com/the-ultimate-performance-metric-in-nlp-111df6c64460

[18] Nallapati, R., Zhai, F., Zhou, B. (2017, February). Summarunner: A recurrent neural network based sequence model for extractive summarization of documents. In Thirty-First AAAI Conference on Artificial Intelligence.

[19] Wikipedia contributors. (2021c, December 21). GPT-3. Wikipedia. https://en.wikipedia.org/wiki/GPT-3

VI. Appendix

In particular of text summarization that utilize "GPT2-large" to summarizing the abstract.

A. Patient

Title: "Impact of COVID-19 on ST-segment elevation myocardial infarction care. The Spanish experience "

Original: "introduction and objectives: the covid-19 outbreak has had an unclear impact on the treatment and outcomes of patients with st-segment elevation myocardial infarction (stemi). the aim of this study was to assess changes in stemi management during the covid-19 outbreak. methods: using a multicenter, nationwide, retrospective, observational registry of consecutive patients who were managed in 75 specific stemi care centers in spain, we compared patient and procedural characteristics and in-hospital outcomes in 2 different cohorts with 30-day follow-up according to whether the patients had been treated before or after covid-19. results: suspected stemi patients treated in stemi networks decreased by 27.6% and patients with confirmed stemi fell from 1305 to 1009 (22.7%). there were no differences in reperfusion strategy (¿ 94% treated with primary percutaneous coronary intervention in both cohorts). patients treated with primary percutaneous coronary intervention during the covid-19 outbreak had a longer ischemic time (233 [150-375] vs 200 [140-332] minutes, p ¡ .001) but showed no differences in the time from first medical contact to reperfusion. in-hospital mortality was higher during covid-19 (7.5% vs 5.1%; unadjusted or, 1.50; 95%ci, 1.07-2.11; p ¡ .001); this association remained after adjustment for confounders (risk-adjusted or, 1.88; 95%ci, 1.12-3.14; p = .017). in the 2020 cohort, there was a 6.3% incidence of confirmed sars-cov-2 infection during hospitalization. conclusions: the number of stemi patients treated during the current covid-19 outbreak fell vs the previous year and there was an increase in the median time from symptom onset to reperfusion and a significant 2-fold increase in the rate of in-hospital mortality. no changes in reperfusion strategy were detected, with primary percutaneous coronary intervention performed for the vast majority of patients. the co-existence of stemi and sars-cov-2 infection was relatively infrequent."

Summary: "introduction and objectives: the covid-19 outbreak has had an unclear impact on the treatment and outcomes of patients with st-segment elevation myocardial infarction (stemi). results: suspected stemi patients treated in stemi networks decreased by 27.6% and patients with confirmed stemi fell from 1305 to 1009 (22.7%). patients treated with primary percutaneous coronary intervention during the covid-19 outbreak had a longer ischemic time (233 [150-375] vs 200 [140-332] minutes, p ¡ .001) but showed no differences in the time from first medical contact to reperfusion."

B. Vaccine

Title: "Emerging roles of extracellular vesicles in COVID-19, a double-edged sword? "

Original: "the sudden outbreak of sars-cov-2-infected disease , initiated from wuhan, china, has rapidly grown into a global pandemic. emerging evidence has implicated extracellular vesicles (evs), a key intercellular communicator, in the pathogenesis and treatment of covid-19. in the pathogenesis of covid-19, cells that express ace2 and cd9 can transfer these viral receptors to other cells via evs, making recipient cells more susceptible for sars-cov-2 infection. once infected, cells release evs packaged with viral particles that further facilitate viral spreading and immune evasion, aggravating covid-19 and its complications. in contrast, evs derived from stem cells, especially mesenchymal stromal/stem cells, alleviate severe inflammation (cytokine storm) and repair damaged lung cells in covid-19 by delivery of anti-inflammatory molecules.therapeutic beneficial evs can also be engineered into drug delivery platforms or vaccines to fight against covid-19. therefore, evs from diverse sources exhibit distinct effects in regulating viral infection, immune response, and tissue damage/repair, 2 —"

Summary: "the sudden outbreak of sars-cov-2-infected disease , initiated from wuhan, china, has rapidly grown into a global pandemic. in contrast, evs derived from stem cells, especially mesenchymal stromal/stem cells, alleviate severe inflammation (cytokine storm) and repair damaged lung cells in covid-19 by delivery of anti-inflammatory molecules."