CORD19STS: COVID-19 Semantic Textual Similarity Dataset

Xiao Guo∗, Hengameh Mirzaalian∗, Ekraam Sabir,
Aysush Jaiswal and Wael Abd-Almageed
Email: xiaoguo, hengameh, esabir, ajaiswal, wamageed@isi.edu
Information Sciences Institute, 4676 Admiralty Way #1001, Marina Del Rey, CA 90292

Abstract—In order to combat the COVID-19 pandemic, society can benefit from various natural language processing applications, such as dialog medical diagnosis systems and information retrieval engines calibrated specifically for COVID-19. These applications rely on the ability to measure semantic textual similarity (STS), making STS a fundamental task that can benefit several downstream applications. However, existing STS datasets and models fail to translate their performance to a domain-specific environment such as COVID-19. To overcome this gap, we introduce CORD19STS dataset which includes 13,710 annotated sentence pairs collected from COVID-19 open research dataset (CORD-19) challenge. To be specific, we generated one million sentence pairs using different sampling strategies. We then used a finetuned BERT-like language model, which we call Sen-SCI-CORD19-BERT, to calculate the similarity scores between sentence pairs to provide a balanced dataset with respect to the different semantic similarity levels, which gives us a total of 32K sentence pairs. Each sentence pair was annotated by five Amazon Mechanical Turk (AMT) crowd workers, where the labels represent different semantic similarity levels between the sentence pairs (i.e. related, somewhat-related, and not-related). After employing a rigorous qualification tasks to verify collected annotations, our final CORD19STS dataset includes 13,710 sentence pairs.

1 INTRODUCTION

As a response to the worldwide COVID-19 pandemic, a massive amount of scientific literature related to coronavirus has emerged and continues to grow rapidly [1]. With the assistance of these documents, various natural language processing (NLP) applications can be developed to combat not only the ongoing COVID-19 pandemic but also to tackle the spread of infectious diseases in the future. Existing COVID-19 corpora can help conversational medical diagnosis systems [2], [16] to determine affected patients by matching textual symptom descriptions or signs to the most related pre-cached medical records. It enables intelligent information retrieval engines to filter out misleading, or malicious information spreading through social media [31], [36], [39]. It also benefits NLP applications on article categorizations [8], [23] into various sub-topics such as: disease precaution and disease evolution.

Measuring sentence textual similarity (STS) plays a key role in the above mentioned applications, and is well-explored mainly in NLP community [11], [29], [32], [33], [40]. These models are trained for STS over sentence pairs collected from non-scientific sources with annotations such as the ones summarized in Table 1. However, the use of these datasets for COVID-19-related NLP tasks might cause deteriorated performance due different data distribution. For example, STS benchmark (STSB) samples [12] could be extracted from general English language context, i.e. grounded image captions, indicating that it may fail to match the in-domain jargon for COVID-19 applications. Also, STS benchmarks proposed in the medical domain [37], [41] have scarce amount of labeled data, which limits their utility as a testbed for learning distributed representation. In this work, we introduce CORD19STS, a dataset for COVID-19 which consists of 13,710 annotated sentence pairs with semantic labels. The sentence pairs in our dataset were selected out of COVID-19 open research dataset (CORD-19) Challenge [1].

Specifically, the construction of CORD19STS contains data-collection and annotation. The data-collection is a two-stage process. First, we sample one million sentence pairs from CORD-19 by applying different sampling strategies to obtain diverse textual similarity pairs. Then, we employ a balance selection scheme to sample 32K sentence pairs out of the previous step, with the aim of balancing samples across different semantic similarity score intervals. In fact, our balance selection scheme is based on the computed similarity scores of the sentence pairs resulting from Sen-SCI-CORD19-BERT, which is the language model that combines many state-of-the-art efforts [1], [9], [34] for discerning textual semantic similarity between sentence pairs extracted from CORD-19. Data annotation, is done by qualified Amazon Mechanical Turk (AMT) crowd workers who pass pre- and post- qualification tests. We present detailed statistical analysis of our dataset.
2 Related work

Sentence Textual Similarity The standard NLP approach to obtain meaningful semantic sentence-level features is to learn sentence-level embeddings for the general purpose in the first place, then apply the model to a downstream, low-complexity task, such as fine-tuning the model over STS dataset for regressing semantic labels to transfer and inject STS knowledge to the model [29], [40], [43].

The first step in STS is learning for the distributed sentence representation with fixed-length, which has been well-explored NLP task by many prior works. NLI offers a large-scale knowledge base which can be used as a supervised signal for learning universal sentence-level embedding [14], also it can be used with siamese and triplet network for more specific tasks [33]. Recently, it has been observed that large-scale multi-task learning helps model sentence-level embedding to gain generality and not be restricted to the single task and maintain high-standard performance over out-of-domain data [22], [28], [38], [40].

However, comparing to the development of distributed sentence-level embedding learning algorithms, relatively little attention has been made to bring the new knowledge base for the STS task. Currently, STSb [11] is the standard benchmark [40], which has been extracted from corpus of STS shared task (2012-2017) [3], [4], [5], [6], [7], consisting around 8K samples that have been extracted from a general context corpus [11], which we argue may fail offer insights for the COVID-19 domain. Moreover, two medical STS datasets have been proposed by [37], [41], that covers medical jargons and clinical notes, yet these two supervised task do not offer enough amount of labeled data (around 1K at most as Table 1), and such nature of low-resource may fail to serve as the benchmark for learning an effective distributed sentence features. Hence, we intend to share with the community CORD19STS, aiming to offer a new dataset for learning sentence semantics for COVID-19 consisting of 13,710 annotated samples.

COVID-19 Related Research/Datasets Very recently, many research efforts have been made as the response to the unprecedented COVID-19 epidemic. Chen et al. [13] provide multilingual COVID-19 twitter dataset to track COVID-19-related misinformation and unverified rumors and study the public attitude and reaction on COVID-19 epidemic; this data collection is ongoing in the foreseeable future and has inspired many other works [24], [26]. Apart from the social-media data collection, a coalition of leading research groups have prepared the COVID-19 Open Research Dataset (CORD-19) [1], which gathers papers and preprints historically and presently pertinent to COVID-19 from Semantic Scholar. CORD-19 serves as the sound research foundation for many following works including ours. Kleinberg et al. [18] provide dataset consists of 5,000 annotated texts with different emotion labels. A predictive modeling approach is employed to approximate the emotional responses of the texts. Huang et al. [19] collect segmented section annotation over 10,966 abstracts of CORD19 dataset, in which annotations provided by AMT users are highly-aligned with that of experts. COVID-Q [20] present annotations for 1,690 questions out of CORD19 dataset, where the labels determine the type of the questions asked within each text. They categorize questions to 15 classes and leverage BERT as their baseline. Moller et al. [21] collect 2,019 question-answer pairs annotated by volunteer biomedical experts. Both two works [20], [21] shows improved performance in QA related task. It is worth to mention that our CORD19STS dataset mainly focuses on offering information over sentence-level semantic meaning, which is orthogonal to the prior works in [18], [19], [20], [21].

3 Preliminary on CORD-19 Dataset

In this section, we provide a brief review of CORD-19 [1], which plays a significant role in the CORD19STS construction.

CORD-19 Dataset is collected out of 128,000 scholarly articles including over 59,000 full texts on COVID-19, SARS-CoV-2, and related coronaviruses. It integrates papers from several openly accessible sources, with the aim of bridging different research communities around the same scientific cause. The dataset can be downloaded from kaggle website and it contains JSON files of the full text of the PDFs in a structured format. General statistics of CORD19 dataset such as number of documents per category, and number of sentences are provided in Table 2.

Given the collected abstract and body-texts of the JSON files of CORD-19 dataset, we applied different preprocessing steps including lower casing the text, removing special characters, citation removal, and sentence tokenization.

4 Construction of CORD19STS

In this section, we provide detailed information of our data-collection process including: selection of one million

| Dataset       | #Pairs | Domain       | Annotators  | Labels                                                                 |
|---------------|--------|--------------|-------------|----------------------------------------------------------------------|
| BIOSSES [37]  | 100    | Biomedical   | Medical experts | Similarity scores from 0 (no relation) to 4 (equivalent).                |
| MedSTS [41]   | 1,068  | Clinical     | Medical experts | Semantic similarity scores from 0 to 5 (low to high similarity)          |
| MedNLI [35]   | 14,049 | Clinical     | Radiologists  | Entailment, Contradicts, Neutral                                      |
| SNLI [10]     | 570,000| Grounded Human Writing | AMT users | Entailment, Contradicts, Neutral                                      |
| STSb [12]     | 8628   | Misc         | AMT users    | Similarity scores from 0 to 5                                         |
| CORD19STS     | 13,710 | COVID-19     | AMT users    | Related, Somewhat-related, Not-related                                 |
... elevated isd in the ancestral member of the gene pair might have already been there at this moment of gene birth, or it might have subsequently evolved, representing two subhypotheses within the overall hypothesis of constraint. the overlapping gene pairs that we observe are those that have been retained. if either member of the overlapping pair was born with low isd, then constraint makes it difficult to adapt to a changing environment, and that pair is less likely to be retained. ... how the relative ages of the genes were classified for 47 out of the 92 overlapping gene pairs for which sequence data were available. within each shaded region, each gene pair is counted within only one of the subregions shown. ...

... while these ancestral sequences are not available, as a proxy we use homologous sequences from basal lineages whose most recent common ancestry predate the origin of the overlap. For simplicity, we refer to these sequences as “pre-overlapping” to distinguish them ...

4.1 Sentence-Pair Sampling Strategies

We employ four different sampling strategies in our pipeline, in order to generate sentence pairs with diverse underlying semantic similarity patterns, i.e. covering different levels of semantic similarities that can be inherently related to the text spanning distance, as follows: (S1) selecting two consecutive sentences within a same paragraph of a document, (S2) selecting pairs of sentences from same paragraphs, (S3) selecting sentence pairs from different paragraphs of a same document, and (S4) selecting sentence pairs from two different documents.

Examples of the four different strategies (S1-S4) are shown in Fig. 1, in which we can observe a meaningful correlation between the physical text spanning distance and semantic similarity level. The anchor sentence contains "gene pair might have already been there" (in orange color) can be expressed as "gene pairs that we can observe", contained in the sentence sampled by S1. Such correlation gradually degrades over sentences collected by S2 and S3, which merely contains overlapping words such as "gene pair" and "ancestral" but hardly express the meaning of the anchor sentence. Finally, it can be seen that the sentence sampled by S4 does not have much semantic correlation with the anchor sentence.

Applying the above mentioned sampling strategies, we collect one million sentence pairs from CORD-19. Although, the initial intuition of applying them was to create sentence pairs reflecting different levels of similarities, i.e. having high to low levels of sentence pair similarities while moving from strategy S1 to S4, surprisingly there still exist many sentence pairs with low level of similarities comparing to high level ones even applying the first two strategies (S1 and S2).

4.2 STS Computation using Our Proposed Sen-SCI-CORD19-BERT Model

Sen-SCI-CORD19-BERT is our proposed language model for measuring textual similarity between sentence pairs of STSCORD19 dataset. To be specific, SCIBERT [9] is a pretrained BERT model over 1.14M papers from semantic scholar, and proves its effectiveness over a suite of tasks and datasets from scientific domains, such as Named Entity Recognition (NER) and Relation Classification (REL) [17], [25], [27]. Due to the fact that CORD-19 has been established over a large amount of scientific articles and pre-prints from semantic scholar, we employ SCIBERT for modeling the textual information. More formally, we fine-tune SCI-BERT over preprocessed CORD-19 text via Masked Language Modeling following [15]. We call the resulting language model SCI-CORD19-BERT.

Meanwhile, Sentence-BERT (SBERT) [34] is the state-of-the-art method for efficiently retrieving sentence pairs with
textual semantic similarity, measured as the cosine distance between the two fixed-dimension sentence embeddings. In fact, SBERT incorporates a pretrained language model followed by a pooling mechanism, into a Siamese network architecture to compute the similarity scores, and the training procedure is orthogonal to the word-level language models such as BERT [15] and RoBERTa [30]. Therefore, we combine the SCI-CORD19-BERT with the Siamese architecture as Sen-SCI-CORD19-BERT, for obtaining the semantic similarity level for CORD-19 sentence pairs, as shown in Figure 3.

![Cosine Similarity between U and V](image)

**Fig. 3.** The visualization over Sen-SCI-CORD19-BERT, which incorporates SCI-CORD19-BERT with the pooling mechanism into a siamese network.

### 4.3 Sampling-Balance Scheme

Given the one million sentence pairs generated out of the prior step, we attempt to balance them with respect to the different semantic similarity levels. To do so, we use our proposed Sen-SCI-CORD19-BERT model (Section 3) to compute similarity scores of the one million samples. We set different score intervals as \([-1, 0.25), (0.25, 0.5), (0.5, 0.75), (0.75, 1]\). For each interval, we sample 2000 sentence pairs per sampling strategy (8000 = 2000^4 samples per interval). As a result, we collect 32K (=4^8K) sentences in total and largely maintain the balance in terms of both underlying similarity patterns and computed similarity levels.

Distributions of the computed similarity scores of the one million sentence pairs are shown in Fig. 2. We address the problem of having unbalanced samples over different interval scores by applying the balance selection scheme, which helps to obtain a dataset balanced over similarity score levels as well as the strategies.

### 4.4 Amazon Mechanical Turk Annotation

Annotation of the sentence pairs of our dataset were provided by AMT crowd users. In our AMT API, the labels were set to three different levels: {related, somewhat-related, not-related}. To reduce potential subjectivity bias introduced to the annotations, each individual sentence pair was assigned to five different annotators. **Pre-Qualification:** Our AMT pipeline had a pre-test qualification part, by which, we first train AMT users by providing them examples of sentence pairs with their correct corresponding labels (Table 4), and then, ask them to attend to our pretest quiz including six sentence pairs. Users providing at least four correct labelings were allowed to attend to our data annotation.

**Post-qualification:** We also include some hidden qualification test samples at each AMT-HIT to verify the performance of the users. Very poor performance on those hidden samples leads to AMT-HIT rejection. We have to take this step because some annotators have tried to game the system by annotating without thinking leading to useless annotations.

**Fig. 4.** Illustrates the statistics of the AMT workers in terms of worker performance over the hidden qualification questions (a-axis) versus total number of the HITs made by
Fig. 4. Visualization of the statistics of the AMT workers in terms of: worker performance over the hidden qualification questions (x-axis); total number of the HITs made by each worker (y-axis); and the average time spent per HIT by the worker, which is encoded/proportional to the radii of the scatter plot. As a data cleaning step in our pipeline, we exclude HITs for which AMT users did not have a good performance over the hidden questions or the ones who did not spend enough time per HIT.

Fig. 5. Statistics of the final gold labels (R: related, SWR: somewhat-related, NR: not-related) per strategy (a-d) as well as overall (e). The NULL bar represents numbers of the sentence pairs which did not have at least three concurring annotations. Distribution of the textual similarity scores over different labels are shown (f-j).

5 Statistics of the Collected Annotations

Fig. 5 visualizes statistics of the final gold labels per strategy (Figures 5a-5d) as well as overall (Figure 5e). The NULL bar represents numbers of the sentence pairs which did not have at least three concurring annotations. Distribution of the textual similarity scores over different labels are shown in Figures 5f-5j.

5.1 Sen-SCI-CORD19-BERT Evaluation

Since we design our sampling method based on Sen-SCI-CORD19-BERT, it is important to validate its performance for modeling language. In this work, we evaluate Sen-SCI-CORD19-BERT based on two experiments, more details of which can be found at our project page1. Performance on STSb dataset. STSb dataset contains 8628 sentence pairs, and we follow the data partition defined in [11] which train, validation and test sets have 5749, 1500, 1379 sentence pairs respectively. In order to report a comprehensive model performance, we conduct experiment over supervised and unsupervised settings, following [34]. Specifically, in the unsupervised manner, models are trained over NLI dataset, the combination of SNLI [10] and Multi-Genre NLI [42]; and in the supervised manner, models are training on the NLI first, then STSb dataset.

1. https://gitlab.vista.isi.edu/xiaoguo/cord_19/tree/master
As reported in Table 3, our proposed Sen-SCI-CORD19-BERT is comparable or slightly outperform the previous work, which indicates that our proposed language model still maintain the ability of measuring textual similarity between sentence pairs extracted from the general contexts, even Sci-CORD19-BERT has been finetuned over CORD-19 dataset for learning specific, in-domain language features.

| Model                     | Correlation Coeff. |
|---------------------------|---------------------|
| Unsupervised (trained on NLI data) | 77.03               |
| SBERT                     | 85.35               |
| Sen-SCI-CORD19-BERT       | 78.90               |
| Supervised (trained on NLI + STSB data) | 84.79               |
| SBERT                     | 85.35               |
| SRoBERTa                  | 85.79               |
| Sen-SCI-CORD19-BERT       | 85.60               |

The comparison on supervised and unsupervised STS.

### 6 Conclusions

In this work, we propose CORD19STS, a semantic textual similarity dataset for COVID-19, which solve the issue that previous STS dataset lacks the in-domain knowledge for the outbreak worldwide pandemic. In order to obtain a dataset which is balanced in terms of latent similarity patterns and different similarity levels, we use four different sentence-pair sampling strategies with balance sampling scheme that relies on our proposed language model, Sen-SCI-CORD19-BERT. Also, we design multiple qualification for AMT to guarantee high-quality annotation can be provided. To this end, we present detailed statistical analysis on the proposed CORD19STS, and evaluate our proposed language model.

### References

[1] Covid-19 open research dataset challenge (cord-19).
[2] Asma Ben Abacha and Pierre Zweigenbaum. Means: A medical question-answering system combining nlp techniques and semantic web technologies. Information processing & management, 51(5):570–594, 2015.
[3] Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weimei Guo, Inigo Lopez-Gazpio, Montse Maritzalar, Rada Mihalcea, et al. Semeval-2015 task 2: Semantic textual similarity, english, spanish and pilot on interpretability. In Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015), pages 232–263, 2015.
[4] Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weimei Guo, Rada Mihalcea, Germain Rigau, and Janice Wiebe. Semeval-2014 task 10: Multilingual semantic textual similarity. In Proceedings of the 8th international workshop on semantic evaluation (SemEval 2014), pages 81–91, 2014.
[5] Eneko Agirre, Carmen Banea, Daniel Cer, Mona Diab, Aitor Gonzalez Agirre, Rada Mihalcea, German Rigau Claramunt, and Janice Wiebe. Semeval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual focused evaluation. arXiv preprint arXiv:1705.02364, 2017.
[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018.
[7] Gunjan Dhole and Nilesh Uke. Nlp based retrieval of medical information for diagnosis of human diseases. Int J Renew Energy Technol, 3(10):243e8, 2014.
[8] Rezarta Islamaj Do˘gan, Robert Leaman, and Zhiyong Lu. Ncbi disease corpus: a resource for disease name recognition and concept normalization. Journal of biomedical informatics, 47(1–10), 2014.
[9] Inc. GitHub. Open source data. https://github.com/JerryWe03/CODA-19, 2020.
[10] Inc. GitHub. Open source data. https://github.com/Inc. GitHub. Open source data. https://github.com/JerryWe03/CODA-19. 2020.
[11] Inc. GitHub. Open source data. https://github.com/Inc. GitHub. Open source data. https://github.com/JerryWe03/CODA-19. 2020.
[12] Xiaoguo Xu and Jongmoo Choi. Human motion prediction via learning local structure representations and temporal dependencies. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 33, pages 2580–2587, 2019.
[13] Peter Jackson and Isabelle Moulinier. Natural language processing for online applications: Text retrieval, extraction and categorization, volume 5. John Benjamins Publishing, 2007.
[14] Bennett Kleinberg, Isabelle van der Vegt, and Maximilian Mozes. Measuring emotions in the covid-19 real world worry dataset. arXiv preprint arXiv:2004.04225, 2020.
[15] Jens Kringlem, Sonny Kim Kjaerulf, Soren Brunak, Ole Lund, Tudor I Oprea, and Olivier Taboureau. Chempred-3.0: a global chemical biology diseases mapping. Database, 2016, 2016.
[16] Sharon Levy and William Yang Wang. Cross-lingual transfer learning for covid-19 outbreak alignment. 2020.
[17] Thomas C Wiegers, and Zhiyong Lu. Biocreative v cdr task corpus: a resource for disease name recognition and concept normalization. Journal of biomedical informatics, 47:1–10, 2014.
[18] Inc. GitHub. Open source data. https://github.com/Inc. GitHub. Open source data. https://github.com/Inc. GitHub. Open source data. https://github.com/JerryWe03/CODA-19. 2020.
[19] Inc. GitHub. Open source data. https://github.com/Inc. GitHub. Open source data. https://github.com/JerryWe03/CODA-19. 2020.
[20] Inc. GitHub. Open source data. https://github.com/Inc. GitHub. Open source data. https://github.com/JerryWe03/CODA-19. 2020.
[21] Inc. GitHub. Open source data. https://github.com/Inc. GitHub. Open source data. https://github.com/JerryWe03/CODA-19. 2020.
[22] Inc. GitHub. Open source data. https://github.com/Inc. GitHub. Open source data. https://github.com/JerryWe03/CODA-19. 2020.
[28] Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. Improving multi-task deep neural networks via knowledge distillation for natural language understanding. *arXiv preprint arXiv:1904.09482*, 2019.

[29] Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. Multi-task deep neural networks for natural language understanding. *arXiv preprint arXiv:1901.11504*, 2019.

[30] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized bert pretraining approach, 2019.

[31] Vahed Qazvinian, Emily Rosengren, Dragomir Radev, and Qiaozhu Mei. Rumor has it: Identifying misinformation in microblogs. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1589–1599, 2011.

[32] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training, 2018.

[33] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019.

[34] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China, November 2019. Association for Computational Linguistics.

[35] Alexey Romanov and Chaitanya Shivade. Lessons from natural language inference in the clinical domain. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1586–1596, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

[36] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. Fake news detection on social media: A data mining perspective. *ACM SIGKDD explorations newsletter*, 19(1):22–36, 2017.

[37] Gizem Sogancıoglu, Hakime Ozturk, and Arzucan Ozgur. Biosses: a semantic sentence similarity estimation system for the biomedical domain. *Bioinformatics*, 33(2017):i49–i585, 2017.

[38] Sandeep Subramaniam, Adam Trischler, Yoshua Bengio, and Christopher J Pal. Learning general purpose distributed sentence representations via large scale multi-task learning. *arXiv preprint arXiv:1804.00079*, 2018.

[39] Shaibib Syed-Abdul, Luis Fernandez-Luque, Wen-Shan Jian, Yu-Chuan Li, Steven Crain, Min-Huei Hsu, Yao-Chin Wang, Dorjsuren Khandregzen, Enkhzaya Chuluunbaatar, Phung Anh Nguyen, et al. Misleading health-related information promoted through video-based social media: anorexia on youtube. *Journal of medical Internet research*, 15(2):e30, 2013.

[40] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*, 2018.

[41] Yanshan Wang, Naveed Afzal, Sunyang Fu, Liwei Wang, Feichen Shen, Majid Rastegar-Mojarad, and Hongfang Liu. Medsts: a resource for clinical semantic textual similarity, language resources and evaluation. *Bioinformatics*, page 1–16, 2018.

[42] Adina Williams, Nikita Nangia, and Samuel R Bowman. A broad-coverage challenge corpus for sentence understanding through inference. *arXiv preprint arXiv:1704.05426*, 2017.

[43] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in neural information processing systems*, pages 5753–5763, 2019.

7 Appendix
| Ex. Sentence-Pair | Label |
|-------------------|-------|
| #1.a **Text**#1: researchers believe that americans with diabetes, chronic lung disease, and cardiovascular disease all diseases linked to obesity are at higher cov risk  
**Text**#2: major risk factors for covid include obesity, diabetes, and cardiovascular diseases  
**Explanation:** both sentences are talking about a set of similar risk factors on covid | Related |
| #1.b **Text**#1: researchers believe that americans with diabetes, chronic lung disease, and cardiovascular disease all diseases linked to obesity are at higher cov risk  
**Text**#2: 30 minutes of moderate-intensity aerobic activity like dog walking helps lower blood pressure and improve overall cardiovascular health.  
**Explanation:** both sentences are related to cardiovascular diseases but not addressing the same topic | Somewhat-related |
| #1.c **Text**#1: researchers believe that americans with diabetes, chronic lung disease, and cardiovascular disease all diseases linked to obesity are at higher cov risk  
**Text**#2: this structural and immune characterization provides new insights into coronavirus spike stability determinants and eludes the immune landscape of viral spike proteins  
**Explanation:** it is not easy to find any mutual connection between the two sentences | Not-related |
| #2.a **Text**#1: covid - 19 does not care about your race, ethnicity, gender, age, or immigration status  
**Text**#2: it has been shown that demographic characteristics might have some correlation with sars  
**Explanation:** both sentences are talking about the effect of demographic features such as: race, ethnicity, gender, and age on covid/sars | Related |
| #2.b **Text**#1: covid - 19 does not care about your race, ethnicity, gender, age, or immigration status  
**Text**#2: information was collected on potential risk factors for SARS-CoV infection (such as having a chronic disease), personal hygiene (such as washing hands), and the use of masks  
**Explanation:** sars/covid risk factors from completely two different categories are mentioned within the two sentences | Somewhat-related |
| #2.c **Text**#1: covid - 19 does not care about your race, ethnicity, gender, age, or immigration status  
**Text**#2: we apply an intervention analysis to assess if this influenza season deviates from eectations  
**Explanation:** although the second sentence is related to influenza, which is a similar disease to covid, but it has nothing to do with the risk factors | Not-related |
| #3.a **Text**#1: the mask also protects you against an infected person sneezing and coughing airborne covid and prevent it from entering your nose or mouth  
**Text**#2: individual level studies have found that the use of face masks was protective for the acquisition and transmission of a range of respiratory viruses including sars cov - 1  
**Explanation:** both sentences are talking about the effectiveness of wearing mask on sars/covid disease | Related |
| #3.b **Text**#1: the mask also protects you against an infected person sneezing and coughing airborne covid and prevent it from entering your nose or mouth  
**Text**#2: disposable medical face masks are intended for a single use only  
**Explanation:** both sentences are related to mask; but one is on its effectiveness and the other one is on the importance of changing disposable masks | Somewhat-related |
| #3.c **Text**#1: the mask also protects you against an infected person sneezing and coughing airborne covid and prevent it from entering your nose or mouth  
**Text**#2: linear mixed models revealed that the presence of mandated bcg policies was associated with a significant flattening of the enential increase in both confirmed cases and deaths during the first - day period of country - wide outbreaks .  
**Explanation:** not easy to find any commonality between the two sentences | Not-related |
| #4.a **Text**#1: we aimed to develop a deep learning method that could extract covid - 19’s graphical features in order to provide a clinical diagnosis ahead of the pathogenic test , thus saving critical time for disease control  
**Text**#2: we established an auxiliary diagnostic tool based on artificial intelligent algorithm to diagnostic hyperlipemia and automatically predict the corresponding diagnostic markers using hematological parameters  
**Explanation:** both sentences are talking about applying machine learning techniques for disease predictions | Related |
| #4.b **Text**#1: we aimed to develop a deep learning method that could extract covid - 19’s graphical features in order to provide a clinical diagnosis ahead of the pathogenic test , thus saving critical time for disease control  
**Text**#2: our method is based on naive bayes, a machine learning algorithm which uses the observed frequencies in the training dataset to estimate the probability that a pair is linked given a set of covariates  
**Explanation:** both sentences have commonalities on applying machine learning | Somewhat-related |
| #4.c **Text**#1: we aimed to develop a deep learning method that could extract covid - 19’s graphical features in order to provide a clinical diagnosis ahead of the pathogenic test , thus saving critical time for disease control  
**Text**#2: we characterized importations timeline to assess the rapidity of isolation , and epidemiologically linked clusters to estimate the rate of detection  
**Explanation:** not easy to find any commonality between the two sentences | Not-related |