Testing the Generalization of Neural Language Models for COVID-19 Misinformation Detection

Jan Philip Wohle\textsuperscript{1} [0000–0002–2116–9767], Nischal Ashok\textsuperscript{2} [0000–0002–7022–5948], Terry Ruas\textsuperscript{1} [0000–0002–9440–780X], Norman Meuschke\textsuperscript{1} [0000–0003–4648–8198], Tirthankar Ghosal\textsuperscript{3} [0000–0002–2358–522X], and Bela Gipp\textsuperscript{1} [0000–0001–6522–3019]

\textsuperscript{1} University of Wuppertal, Rainer-Gruenter-Straße, 42119 Wuppertal, Germany
\textit{last@uni-wuppertal.de}

\textsuperscript{2} Indian Institute of Technology Patna, Bihar-801106, India
\textit{1801cs33@iitp.ac.in}

\textsuperscript{3} Charles University, Malostranské náměstí 25, 118 00 Praha, Czech Republic
\textit{ghosal@ufal.mff.cuni.cz}

Abstract. A drastic rise in potentially life-threatening misinformation has been a by-product of the COVID-19 pandemic. Computational support to identify false information within the massive body of data on the topic is crucial to prevent harm. Researchers proposed many methods for flagging online misinformation related to COVID-19. However, these methods predominantly target specific content types (e.g., news) or platforms (e.g., Twitter). The methods capabilities to generalize were largely unclear so far. We evaluate fifteen Transformer-based models on five COVID-19 misinformation datasets that include social media posts, news articles, and scientific papers to fill this gap. We show tokenizers and models tailored to COVID-19 data do not provide a significant advantage over general-purpose ones. Our study provides a realistic assessment of models for detecting COVID-19 misinformation. We expect that evaluating a broad spectrum of datasets and models will benefit future research in developing misinformation detection systems.

Keywords: COVID-19 · Transformers · Health · Social Media.

1 Introduction

The COVID-19 pandemic has claimed more than four million lives by the time of writing this paper, and the number of infections remains high\textsuperscript{4}. The behavior of individuals strongly affects the risk of infection. In turn, the quality of information individuals receive strongly influences their actions \cite{23,10}. The novelty and rapid global spread of the SARS-CoV-2 virus has also led to countless life-threatening incidences of misinformation spread on the topic. Controlling COVID-19 and combating possible future pandemics early, requires reducing misinformation and increasing the distribution of facts on the subject \cite{5,35}.

\textsuperscript{4}Equal contribution.

\textsuperscript{4}https://coronavirus.jhu.edu/map.html
Researchers worldwide collaborate on automating the detection of false information on COVID-19. The initiatives build collections of scientific papers, social media posts, and news articles to analyze their content, spread, source, and propagators [33,7,19].

Natural Language Processing (NLP) research has extensively studied options to automate the identification of fake news [27], primarily by applying recent language models. Researchers proposed adaptations of well-known Transformer models, such as COVID-Twitter-BERT [20], to identify false information on COVID-19 from specific sources. However, most prior studies analyze specific content types (e.g., news) or platforms (e.g., Twitter). These limitations prevent reliable conclusions regarding the generalization of the proposed language models.

To fill this gap, we apply 15 Transformer models to five COVID-19 misinformation tasks. We compare Transformer models optimized on COVID-19 datasets to state-of-the-art neural language models. We exhaustively apply models to different tasks to test their generalization on unknown sources. The code to reproduce our experiments, and the datasets used are publicly available.

2 Related Work

The same way word2vec [18] inspired many models in NLP [4,26,25], the excellent performance of BERT [8], a Transformer-based model [28], caused its numerous adaption for language tasks [34,6,29,30]. Domain-specific models build on top of Transformers typically outperform their baselines for related tasks [12]. For example, SciBERT [2] was pre-trained on scientific documents and typically outperforms BERT for scientific NLP tasks, such as determining document similarity [22].

Many models for COVID-19 misinformation detection employ domain-specific pre-training to improve their representation. COVID-Twitter-BERT [20] was pre-trained on 160M tweets and evaluated for sentiment analysis of tweets, e.g., about COVID vaccines. BioClinicalBERT [1] was trained into clinical narratives to incorporate linguistic characteristics from the biomedical and clinical domains.

Cui et al. [7] investigated the misinformation detection task by comparing traditional machine learning and deep learning techniques. Similarly, Zhou et al. [36] explored statistical learners, such as SVM, and neural networks to classify news as credible or not. The results of both studies show deep learning architectures as the most prominent alternatives for the respective datasets.

As papers on COVID-19 are recent, some contributions are only available as pre-trained models. COVID-BERT7 and COVID-SciBERT8 are pre-trained on the CORD-19 dataset and only available via the Huggingface API. Others, such as COVID-CQ [19] and CMU-MisCov19 [17] are used to either investigate intrinsic details (e.g., how dense misinformed communities are) or to explore the applicability of statistical techniques.

Although related works provide promising approaches to counter misinformation related to COVID-19, none of them explore multiple datasets. Research in many NLP

---

5 We collectively refer to fake news, disinfection, and misinformation as false information.

6 https://github.com/ag-gipp/iConference22_COVID_misinformation
7 https://tinyurl.com/86cpx6u2
8 https://tinyurl.com/9w24pc93
areas already uses diverse benchmarks to compare models [32,31]. To the best of our
knowledge, our study is the first to systematically test Transformer-based methods on
different data sources related to COVID-19.

3 Methodology

Models. Our study includes 15 Transformer-based models, which are detailed in Ap-
pendix A.2. We categorize the models into the following three groups:

General-Purpose Baselines. The first group consists of general-purpose Trans-
former models without domain-specific training, i.e., BERT [8], RoBERTa [16], BART [15],
DeBERTa [11]. These baselines show how vanilla Transformer-based models perform
on the COVID-19 misinformation detection task.

Intermediate Pre-Training. The second group contains models trained on spe-
cific content types and domains, i.e., SciBERT [2], BERTweet [21], and BioClinical-
BERT [1]. For example, SciBERT adapts BERT for scientific articles. These models
show the effect of intermediate pre-training on specific sources compared to general-
purpose training (e.g., whether BERTweet is superior to BERT for misinformation on
Twitter). Moreover, we compare the models in this group to language models optimized
using intermediate pre-training on COVID-19 data (third group).

COVID-19 Intermediate Pre-Training. The third group comprises models em-
ploying an intermediate pre-training stage on COVID-19 data. Due to task-specific pre-
training, we expect these models to achieve better results than the models in groups one
and two. We include a model that optimizes the pre-training objective on a large Twit-
ter corpus, i.e., CT-BERT [20], two models trained on the CORD-19 dataset (COVID-
BERT\textsuperscript{9} and COVID-SciBERT\textsuperscript{10}), and two popular models from the huggingface API
for which the intermediate pre-training sources are not released yet (ClinicalCOVID-
BERT\textsuperscript{11} and BioCOVID-BERT\textsuperscript{12}). We pre-train RoBERTa, BART, and DeBERTa on
the CORD-19 dataset to compare them to the models we used as general-purpose base-
lines.

Data. We compile an evaluation set from six popular datasets for detecting COVID-19
misinformation in social media, news articles, and scientific publications, i.e., CORD-
19 [33], CoAID [7], COVID-CQ [19], ReCOvery [36], CMU-MisCov19 [17], and
COVID19FN.\textsuperscript{13} Table 1 gives an overview of the datasets and Appendix A.1 presents
more details. For CORD-19, we only use abstracts in the dataset, as they provide an
adequate trade-off between size and information density. Additionally, less than 50%
of the articles in CORD-19 are available as full texts. For CoAID and ReCOvery, we
only extract news articles to reduce a bias towards Twitter posts in our evaluation. All
remaining datasets are used in their original composition.

We use CORD-19 to extend the pre-training of general-purpose models and all other
datasets to evaluate the models for a downstream task. CORD-19 consists of scientific

\textsuperscript{9}https://tinyurl.com/86cpx6u2
\textsuperscript{10}https://tinyurl.com/9w24pc93
\textsuperscript{11}https://tinyurl.com/kebysw
\textsuperscript{12}https://tinyurl.com/4xx9vdkm
\textsuperscript{13}https://tinyurl.com/4ne9vtzu
articles, while the other datasets primarily contain news articles and Twitter content. We chose different domains for training and evaluation to test the models’ generalization capabilities and avoid overlaps between the datasets.

Table 1. An overview of the COVID-related datasets. CORD-19 has no specific Task or Label as it provides a general collection of documents. Details on the labels are given in Memon et al. [17].

| Corpus          | Corpus | Task                          | Domain                                | Source(s)                          | Label(s)           |
|-----------------|--------|-------------------------------|---------------------------------------|-------------------------------------|--------------------|
| CORD-19 ([33])  | 497,906| Scientific articles           | CZI, PMC, BioRxiv, MedRxiv            | [true, false]                      |
| CoAID ([7])     | 302,926| Misinformation detection      | Healthcare misinformation             | Twitter, news, social media         | [true, false]      |
| ReCOVery ([16]) | 142,849| Credibility classification    | Low information credibility           | Twitter, news                      | [reliable, unreliable] |
| COVID-CQ ([19]) | 14,374 | Efficacy of treatments        | Drug treatment                        | Twitter                            | [neutral, against, for] |
| CMU-MisCov19 ([17]) | 4,573   | Communities detection        | Misinformed communities               | Twitter                            | [17 labels†]       |
| COVID19FN (2020)| 2,800  | Misinformation detection      | Misinformation in news                | Poynter                             | [true, false]      |

4 Experiments

Overview. Our study includes three experiments. The first experiment tests how static word embeddings and frozen contextual embeddings perform compared to fine-tuned language models. The second experiment studies whether tokenizers specifically tailored to a COVID-19 vocabulary are superior to general-purpose ones. The third experiment evaluates and compares all 15 Transformer models on the five evaluation datasets.

Training & Evaluation. To compare general-purpose baselines and COVID-19 intermediate pre-trained models, we perform pre-training on the CORD-19 dataset for three models (RoBERTa, BART, and DeBERTa) and use pre-trained configurations for the remaining models (BERT, SciBERT, BioCOVID-BERT, ClinicalCOVID-BERT). We then fine-tune all models for each of the five test tasks (COVID-CQ, CoAID, ReCOVery, CMU-MisCov19, and COVID19FN). We use a split of 80% and 20% of the documents in a dataset for training and testing, respectively. This split is the most common configuration for the tested datasets [36,19] and is comparable to other studies [7]. We use 10% of the train dataset as a hold-out validation set.

5 Results & Discussion

Static and Frozen Embeddings. Table 2 compares the classification results of a baseline composed of BiLSTM and GlobalVectors (GloVe) to the frozen embeddings of three Transformer models for the COVID-CQ dataset. The results show no significant difference between GloVe and the frozen models. However, fine-tuning the same three models end-to-end generally increases their performance. Therefore, we choose to fine-tune neural language models for the classification of COVID-19 misinformation.

14 General-purpose refers to the tokenizers released with the pre-trained models.
Tokenizer Ablation. Table 3 shows the results on COVID-CQ for the best configuration of the models using a standard \textsuperscript{15} tokenizer for pre-training and fine-tuning. We expected adjusting the tokenizer to the CORD-19 dataset would improve the results, as it adds valuable tokens to the vocabulary, which are often not present in standard tokenizers. However, using specialized tokenizers decreased the performance. The content in CORD-19 originates from the scientific domain. We hypothesize tweets lack similar token relations, which causes the performance drop on the COVID-CQ dataset. Therefore, we use the standard tokenizer for our full evaluation experiments.

Full Evaluation. Table 4 reports the results of our full evaluation. All results are statistically significant using bootstrap and permutation tests ($p < .05$) \cite{bootstrap}. General-purpose baselines achieved the best result for two of the five datasets (BART on CoAID and BART on COVID19FN). For two datasets (ReCOVery and COVID-CQ), a model we pre-trained on CORD-19 data (COVID-RoBERTa) performed best. CT-BERT achieved the best result on CMU-MisCov19, an expected outcome as the datasets consist only of Twitter content. BERTweet, which was also trained on Twitter data, does not achieve better results than general-purpose baselines. We expected a minor drop in performance for BERTweet compared to CT-BERT as the former was not trained on COVID-19 vocabulary, but better a performance than general-purpose models as BERTweet was trained mainly on Twitter data.

Table 2. F1-Macro scores of neural language models and a baseline (BiLSTM+GloVe) for the COVID-CQ dataset. The static and frozen models use a stacked BiLSTM; fine-tuned models were pre-trained on the CORD-19 dataset and fine-tuned for the task.

| Type   | Models      | F1-Macro |
|--------|-------------|----------|
| static | GloVe       | .71      |
| frozen | BERT        | .72      |
| frozen | RoBERTa     | .70      |
| frozen | SciBERT     | .68      |
| fine-tuned | BERT     | .75      |
| fine-tuned | RoBERTa | .80      |
| fine-tuned | SciBERT | .76      |

Table 3. F1-Macro scores of BART and RoBERTa on the COVID-CQ dataset using different Pre-Training and Fine-Tuning Tokenizers. All models were pre-trained on the CORD-19 dataset.

| Models     | PT Tok. | FT Tok. | F1-Macro |
|------------|---------|---------|----------|
| RoBERTa Standard | Standard | .78     |
| RoBERTa COVID | Standard | .73     |
| RoBERTa COVID COVID | Standard | .72     |
| BART Standard | Standard | .77     |
| BART COVID Standard | Standard | .73     |
| BART COVID COVID | Standard | .70     |

All models achieved low scores for CMU-MisCov19, making it the most challenging dataset in our evaluation. The best results were obtained for CoAID. Overall, general-purpose baselines achieved comparable results to COVID-19 intermediate pre-trained models for all datasets. For example, the best mean result for the dataset ReCOVery was achieved by COVID-RoBERTa (F1=.91, std=.03) while the general-purpose model BART (F1=.90, std=.01) was only .01 score points worse. We observe similar results for COVID-CQ, where the best model COVID-RoBERTa (F1=.78, std=.03) has a

\textsuperscript{15}Pre-Trained tokenizer provided by HuggingFace
### Table 4.

| IT CA Model          | CMU-MisCov19 | CoAID | ReCOVery | COVID19FN | COVID-CQ |
|----------------------|--------------|-------|----------|-----------|----------|
| - - BERT             | .54±.03      | .93±.01 | .78±.02  | .65±.01   | .76±.01  |
| - - RoBERTa          | .53±.03      | .95±.01 | .81±.01  | .73±.02   | .64±.01  |
| - - BART             | .49±.03      | .96±.01 | .90±.01  | .83±.01   | .75±.01  |
| - - DeBERTa          | .52±.04      | .95±.01 | .75±.01  | .67±.03   | .63±.02  |
| ✓ - SciBERT          | .46±.02      | .95±.01 | .77±.01  | .76±.01   | .61±.02  |
| ✓ - BioClinicalBERT  | .48±.02      | .89±.01 | .82±.01  | .81±.02   | .63±.02  |
| ✓ - BERTweet         | .51±.03      | .88±.02 | .84±.03  | .65±.01   | .75±.01  |
| ✓ ✓ CT-BERT *        | **.58±.04**  | .94±.01 | .80±.02  | .40±.01   | .63±.02  |
| ✓ ? ClinicalCOVID-BERT | .50±.03    | .93±.01 | .82±.01  | .78±.02   | .74±.02  |
| ✓ ? BioCOVID-BERT *  | .45±.01      | .91±.02 | .81±.01  | .68±.03   | .63±.02  |
| ✓ ✓ COVID-BERT †     | .46±.02      | .94±.01 | .85±.03  | .73±.02   | .72±.01  |
| ✓ ✓ COVID-SciBERT †  | .34±.02      | .92±.03 | .78±.03  | .67±.02   | .76±.03  |
| ✓ ✓ COVID-RoBERTa    | .40±.04      | .92±.04 | **.91±.03** | .67±.03   | **.78±.03** |
| ✓ ✓ COVID-BART       | .33±.01      | .91±.06 | .89±.03  | .66±.05   | .77±.07  |
| ✓ ✓ COVID-DeBERTa    | .30±.01      | .92±.05 | .89±.02  | .81±.03   | .73±.01  |

We conclude that pre-training language models on COVID data before fine-tuning on a misinformation task did not generally provide an advantage for the tested datasets in this paper.

### 6 Conclusion & Future Work

This study empirically evaluated 15 Transformer models for five COVID-19 misinformation tasks. Our analysis shows domain-specific models and tokenizers do not generally perform better in the classification of misinformation. We conclude that the vocabulary related to COVID-19 and possibly text-patterns about COVID-19 do not have a significant effect on the models’ ability to classify misinformation.

The main limitation of our study is the non-standardized pre-training of models due to the models’ diversity. To reliably detect misinformation across content types and platforms, researchers need access to diverse data. We see this study as an initial step to compile a benchmark for COVID-19 data similar to widely adopted natural language understanding benchmarks (e.g., GLUE, SuperGLUE) which enable an evaluation across diverse sets of misinformation domains, sources, and tasks.

Controlling the current and future pandemics requires reliable detection of false information propagated through many streams and having different unique features. This study is a first step for researchers and policymakers to devise and deploy systems that reliably flag misinformation related to COVID-19 from a broad spectrum of sources.
The usefulness of NLP models increases significantly if they are applicable to multiple tasks [31]. We anticipate future NLP technologies for detecting misinformation need to adopt the trend of evaluating on several benchmark datasets. This work provides a first milestone in evaluating general model capabilities and questioning the advantage of domain-specific model pre-training.

Although COVID-19 accelerated the propagation of misinformation and disinformation, these problems are not unique to the current pandemic. The effects of COVID-19 misinformation and disinformation on elections, ethical biases, and the portrayal of ethical groups [3] can have similar or even more severe consequences on society than misinformation related to COVID-19. Therefore, identifying false information streams across domains will remain a challenging problem, and identifying which models can generalize for many sources is crucial.

References
1. Alsentzer, E., Murphy, J.R., Boag, W., Weng, W.H., Jin, D., Naumann, T., McDermott, M.B.A.: Publicly Available Clinical BERT Embeddings. arXiv:1904.03323 [cs] (Jun 2019), http://arxiv.org/abs/1904.03323
2. Beltagy, I., Lo, K., Cohan, A.: SciBERT: A Pretrained Language Model for Scientific Text. 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). pp. 3613–3618. Association for Computational Linguistics, Hong Kong, China (2019). https://doi.org/10/ggcgtm
3. Benkler, Y., Farris, R., Roberts, H.: Network Propaganda, vol. 1. Oxford University Press (Oct 2018). https://doi.org/10.1093/oso/9780190923624.001.0001
4. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching Word Vectors with Subword Information. Transactions of the Association for Computational Linguistics 5, 135–146 (Dec 2017). https://doi.org/10/gfw9cs
5. Cinelli, M., Quattrociocchi, W., Galeazzi, A., Valensise, C.M., Brugnoli, E., Schmidt, A.L., Zola, P., Zollo, F., Scala, A.: The COVID-19 social media infodemic. Scientific Reports 10(1), 16598 (Dec 2020). https://doi.org/10.1038/s41598-020-73510-5
6. Clark, K., Luong, M.T., Le, Q.V., Manning, C.D.: ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. arXiv:2003.10555 [cs] (Mar 2020), http://arxiv.org/abs/2003.10555
7. Cui, L., Lee, D.: CoAID: COVID-19 Healthcare Misinformation Dataset. arXiv:2006.00885 [cs] (Aug 2020), http://arxiv.org/abs/2006.00885
8. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805 [cs] (May 2019), http://arxiv.org/abs/1810.04805
9. Dror, R., Baumer, G., Shlomov, S., Reichart, R.: The Hitchhiker’s Guide to Testing Statistical Significance in Natural Language Processing. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 1383–1392. Association for Computational Linguistics, Melbourne, Australia (Jul 2018). https://doi.org/10.18653/v1/P18-1128
10. Hale, T., Angrist, N., Goldszmidt, R., Kira, B., Petherick, A., Phillips, T., Webster, S., Cameron-Blake, E., Hallas, L., Majundar, S., Tatlow, H.: A Global Panel Database of Pandemic Policies (Oxford COVID-19 Government Response Tracker). Nature Human Behaviour 5(4), 529–538 (Apr 2021). https://doi.org/10.1038/s41562-021-01079-8
11. He, P., Liu, X., Gao, J., Chen, W.: DeBERTa: Decoding-enhanced BERT with Disentangled Attention. arXiv:2006.03654 [cs] (Jan 2021), http://arxiv.org/abs/2006.03654
12. Howard, J., Ruder, S.: Universal Language Model Fine-tuning for Text Classification. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 328–339. Association for Computational Linguistics, Melbourne, Australia (Jul 2018). https://doi.org/10.18653/v1/P18-1031
13. Johnson, A.E., Pollard, T.J., Shen, L., Lehman, L.H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L.A., Mark, R.G.: MIMIC-III, a freely accessible critical care database. Scientific Data 3, 160035 (May 2016). https://doi.org/10.1038/sdata.2016.35
14. Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C.H., Kang, J.: BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics pp. 1–7 (Sep 2019). https://doi.org/10.1093/bioinformatics/btz682
15. Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., Zettlemoyer, L.: BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 7871–7880. Association for Computational Linguistics, Online (Jul 2020). https://doi.org/10.18653/v1/2020.acl-main.703
16. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv:1907.11692 [cs] (Jul 2019), http://arxiv.org/abs/1907.11692
17. Memon, S.A., Carley, K.M.: Characterizing COVID-19 Misinformation Communities Using a Novel Twitter Dataset. arXiv:2008.00791 [cs] (Sep 2020), http://arxiv.org/abs/2008.00791
18. Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J.: Distributed Representations of Words and Phrases and their Compositionality. arXiv:1310.4546 [cs, stat] (Oct 2013), http://arxiv.org/abs/1310.45464
19. Mutlu, E.C., Oghaz, T., Jasser, J., Tutunculer, E., Rajabi, A., Tayebi, A., Ozmen, O., Garibay, I.: A stance data set on polarized conversations on Twitter about the efficacy of hydroxychloroquine as a treatment for COVID-19. Data in Brief 33, 106401 (2020). https://doi.org/10.1016/j.dib.2020.106401
20. Müller, M., Salathé, M., Kummervold, P.E.: COVID-Twitter-BERT: A Natural Language Processing Model to Analyse COVID-19 Content on Twitter. arXiv:2005.07503 [cs] (May 2020), http://arxiv.org/abs/2005.07503
21. Nguyen, D.Q., Vu, T., Tuan Nguyen, A.: BERTweet: A pre-trained language model for English Tweets. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. pp. 9–14. Association for Computational Linguistics, Online (2020). https://doi.org/10.18653/v1/2020.emnlp-demos.2
22. Ostendorff, M., Ruas, T., Blume, T., Gipp, B., Rehm, G.: Aspect-based Document Similarity for Research Papers. In: Proceedings of the 28th International Conference on Computational Linguistics. pp. 6194–6206. International Committee on Computational Linguistics, Barcelona, Spain (Online) (2020). https://doi.org/10.18653/v1/2020.coling-main.545
23. Pennycook, G., McPhetres, J., Zhang, Y., Lu, J.G., Rand, D.G.: Fighting COVID-19 Misinformation on Social Media: Experimental Evidence for a Scalable Accuracy-Nudge Intervention. Psychological Science 31(7), 770–780 (Jun 2020). https://doi.org/10.1177/0956797620930954
24. Press, O., Smith, N.A., Lewis, M.: Shortformer: Better Language Modeling using Shorter Inputs. arXiv:2012.15832 [cs] (Dec 2020), http://arxiv.org/abs/2012.15832
25. Ruas, T., Ferreira, C.H.P., Grosky, W., de França, F.O., de Medeiros, D.M.R.: Enhanced Word Embeddings Using Multi-Semantic Representation through Lexical Chains. Information Sciences 532, 16–32 (Sep 2020). https://doi.org/10.1016/j.ins.2020.04.048
26. Ruas, T., Grosky, W., Aizawa, A.: Multi-sense embeddings through a word sense disambiguation process. Expert Systems with Applications 136, 288–303 (Dec 2019). https://doi.org/10.1016/j.eswa.2019.06.026

27. Shu, K., Sliva, A., Wang, S., Tang, J., Liu, H.: Fake News Detection on Social Media: A Data Mining Perspective. ACM SIGKDD Explorations Newsletter 19(1), 22–36 (Sep 2017). https://doi.org/10.1145/3137597.3137600

28. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. p. 6000–6010. NIPS’17, Curran Associates Inc., Red Hook, NY, USA (2017), https://arxiv.org/abs/1706.03762

29. Wahle, J.P., Ruas, T., Foltynek, T., Meuschke, N., Gipp, B.: Identifying Machine-Paraphrased Plagiarism. In: Proceedings of the iConference (February 2022)

30. Wahle, J.P., Ruas, T., Meuschke, N., Gipp, B.: Are Neural Language Models Good Plagiarists? A Benchmark for Neural Paraphrase Detection. In: Proceedings of the ACM/IEEE Joint Conference on Digital Libraries (JCDL). IEEE, Washington, USA (Sep 2021)

31. Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., Levy, O., Bowman, S.: SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems. In: Wallach, H., Larochelle, H., Beygelzimer, A., d’Alché-Buc, F., Fox, E., Garnett, R. (eds.) Advances in Neural Information Processing Systems 32, pp. 3266–3280. Curran Associates, Inc. (2019), https://arxiv.org/abs/1905.00537

32. Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., Bowman, S.R.: GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. arXiv:1804.07461 [cs] (Feb 2019), https://arxiv.org/abs/1804.0746

33. Wang, L.L., Lo, K., Chandrasekhar, Y., Reas, R., Yang, J., Burdick, D., Eide, D., Funk, K., Katsis, Y., Kinney, R., Li, Y., Liu, Z., Merrill, W., Mooney, P., Murdick, D., Rishi, D., Sheehan, J., Shen, Z., Stilson, B., Wade, A., Wang, K., Wang, N.X.R., Wilhelm, C., Xie, B., Raymond, D., Weld, D.S., Etzioni, O., Kohlmeier, S.: CORD-19: The COVID-19 Open Research Dataset. arXiv:2004.10706 [cs] (Jul 2020), http://arxiv.org/abs/2004.10706

34. Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., Le, Q.V.: XLNet: Generalized Autoregressive Pretraining for Language Understanding. arXiv:1906.08237 [cs] (Jun 2019), https://arxiv.org/abs/1904.0746

35. Zarocostas, J.: How to fight an infodemic. The Lancet 395(10225), 676 (Feb 2020). https://doi.org/10.1016/S0140-6736(20)30461-X

36. Zhou, X., Mulay, A., Ferrara, E., Zafarani, R.: ReCOVery: A Multimodal Repository for COVID-19 News Credibility Research, p. 3205–3212. Association for Computing Machinery, New York, NY, USA (2020). https://doi.org/10.1145/3340531.3412880
A Appendix

A.1 Dataset Details

COVID-19 Open Research Dataset (CORD-19) [33] is the largest open source dataset about COVID-19 and coronavirus-related research (e.g., SARS, MERS). CORD-19 is composed of more than 280K scholarly articles from PubMed, bioRxiv, medRxiv, and other resources maintained by the WHO. We use this dataset to extend the general pre-training from selected neural language models (cf. Section 3) into the COVID-specific vocabulary and features.

Covid-19 heAlthcare mIsinformation Dataset (CoAID) [7] focuses on healthcare misinformation, including fake news on websites, user engagement, and social media. CoAID is composed of 5 216 news articles, 296 752 related user engagements, and 958 posts about COVID-19, which are broadly categorized under the labels true and false.

Twitter Stance Dataset (COVID-CQ) [19] is a dataset of user-generated Twitter content in the context of COVID-19. More than 14K tweets were processed and annotated regarding the use of Chloroquine and Hydroxychloroquine as a valid treatment or prevention against the coronavirus. COVID-CQ is composed of 14 374 tweets from 11 552 unique users labeled as neutral, against, or favor.

ReCOVery [36] explores the low credibility of information on COVID-19 (e.g., bleach can prevent COVID-19) by allowing a multimodal investigation of news and their spread on social media. The dataset is composed of 2 029 news articles on the coronavirus and 140 820 related tweets labeled as reliable or unreliable.

CMU-MisCov19 [17] is a Twitter dataset created by collecting posts from unknowingly misinformed users, users who actively spread misinformation, and users who disseminate facts or call out misinformation. CMU-MisCov19 is composed of 4 573 annotated tweets divided into 17 classes (e.g., conspiracy, fake cure, news, sarcasm). The high number of classes and their imbalanced distribution make CMU-MisCov19 a challenging dataset.

COVID19FN [20] is composed of approximately 2 800 news articles extracted mainly from Poynter categorized as either real or fake.

A.2 Model Details

General-Purpose Baselines. BERT [8] mainly captures general language characteristics using a bidirectional Masked Language Model (MLM) and Next Sentence Prediction (NSP) tasks. RoBERTa [16] improves BERT with additional data, compute budgets, and hyperparameter optimizations. RoBERTa also drops the NSP as it contributes little to the model representation. BART [15] optimizes an auto-regressive forward-product
and auto-encoding MLM objective simultaneously. DeBERTa [11] improves the attention mechanism using a disentanglement of content and position.

**Intermediate Pre-Trained**. SciBERT [2] optimizes the MLM for 1.14M randomly selected papers from Semantic Scholar\(^2\). BioClinicalBERT [1] specializes on 2M notes in the MIMIC-III database [13], a collection of disidentified clinical data. BERTweet [21] optimizes BERT on 850M tweets each containing between 10 and 64 tokens.

**COVID-19 Intermediate Pre-Trained** COVID-Twitter-BERT [20] (CT-BERT) uses a corpus of 160M tweets for domain-specific pre-training and evaluates the resulting model’s capabilities in sentiment analysis, such as for tweets about vaccines. BioClinicalBERT [1] fine-tunes BioBERT [14] into clinical narratives in the hope to incorporate linguistic characteristics from both the clinical and biomedical domains.

Cui et al. [7] propose CoAID and investigate the misinformation detection task by comparing traditional machine learning (e.g., logistic regression, random forest) and deep learning techniques (e.g., GRU). In a similar layout, Zhou et al. [36] compare traditional statistical learners, such as SVM and neural networks (e.g., CNN), to classify news as credible or not. In both studies, the results show deep learning architectures as the most prominent options.

### A.3 Evaluation Details

**Pre-Training**. We use the data from the abstracts of the CORD-19 dataset for pre-training. For pre-processing the CORD-19 abstract data, we consider only alphanumerical characters. We use a sequence length of 128 tokens, which reduces training time while being competitive to longer sequence lengths when fine-tuning [24]. We mask words randomly with a probability of .15, a common configuration for Transformers [8,11], and perform the MLM with the following remaining parameters: a batch size of 16 for all the base models, and eight for the large models, the Adam Optimizer \((\alpha = 2e-5, \beta_1 = .9, \beta_2 = .999, \epsilon = 1e-8)\), and a maximum of five epochs. All experiments were performed on a single NVIDIA GeForce GTX 1080 Ti GPU with 11 GB of memory.

**Fine-Tuning**. The classification model applies a randomly initialized fully-connected layer to the aggregate representation of the underlying Transformer (e.g., [CLS] for BERT) with dropout \((p = .1)\) to learn the annotated target classes with cross-entropy loss for five epochs and with a sequence length of 200 tokens. We use the same configuration of the optimizer as in pre-training.

\(^2\)https://www.semanticscholar.org/