RerrFact: Reduced Evidence Retrieval Representations for Scientific Claim Verification

Ashish **Rana**[1,*], Deepanshu **Khanna**[2,*], Tirthankar **Ghosal**[3], Muskaan **Singh**[3], Harpreet **Singh**[4] and Prashant Singh **Rana**[4]

[1]*School of Business Informatics and Mathematics, University of Mannheim, Germany*

[2]*Department of Electronics & Communication Engineering, Thapar Institute of Engineering & Technology, India*

[3]*Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University, Czech Republic*

[4]*Department of Computer Science & Engineering, Thapar Institute of Engineering & Technology, India*

Abstract

Exponential growth in digital information outlets and the race to publish has made scientific misinformation more prevalent than ever. However, the task to fact-verify a given scientific claim is not straightforward even for researchers. Scientific claim verification requires in-depth knowledge and great labor from domain experts to substantiate supporting and refuting evidence from credible scientific sources. The SciFact dataset and corresponding task provide a benchmarking leaderboard to the community to develop automatic scientific claim verification systems via extracting and assimilating relevant evidence rationales from source abstracts. In this work, we propose a modular approach that sequentially carries out binary classification for every prediction subtask as in the SciFact leaderboard. Our simple classifier-based approach uses reduced abstract representations to retrieve relevant abstracts. These are further used to train the relevant rationale-selection model. Finally, we carry out two-step stance predictions that first differentiate non-relevant rationales and then identify supporting or refuting rationales for a given claim. Experimentally, our system RerrFact with no fine-tuning, simple design, and a fraction of model parameters fairs competitively on the leaderboard against large-scale, modular, and joint modeling approaches. We make our codebase available at *https://github.com/ashishrana160796/RerrFact*.

1. Introduction

Misinformation is a modern day societal problem that has the potential to wreck havoc, especially with increasingly many people having an online footprint without adequate internet literacy. The problem grows intense when science gets associated with disinformation and provides a false sense of trustworthiness. Convincing statements derived from general public opinions like *"Ginger consumption in food reduces the risk of getting severely infected with COVID-19"* can effectively manipulate the masses. It is hard to verify such misleading statements from extensive scientific literature with appropriate reasoning even by providing relevant evidence. Also, it is a cumbersome task for experts to search for refuting or supporting argument rationales considering the amount of misinformation available on a plethora of outlets. Therefore, automatic fact-verification tools are essential, especially for scientific knowledge where the given system must understand scientific knowledge, interpret numeric and statistical inferences.

Previously, the veracity verification task has been extensively studied, and many datasets are available on various use-cases [1, 2, 3]. The most relevant amongst them is the FEVER shared task [4], which evaluates the veracity of human-generated claims from Wikipedia data. For the FEVER task, there are two paradigms: one that take a three-step modular approach and the other which is joint prediction approach for evidence retrieval & stance prediction [5, 6]. Similarly, for the SciFact task these two paradigms have been used either with very large language models like VerT5erini for modular architecture [7] or *ARSJoint*, *JointParagraph* for merged subtask architecture [8, 9]. In contrast to these diametrically opposite paradigms, QMUL-SDS's [10] partial binding between the abstract retrieval and rational selection stages offers a promising direction, which is also the inspiration for our current work. Our experiments demonstrate that this partial interdependence successfully introduces a form of regularization, providing much-needed improvements over precision and recall for the evidence retrieval component in the concerned task. Therefore, we present a computationally and architecturally simple pipeline-driven design for it.

We use the same partial interdependence pipeline design with *reduced evidence retrieval stage representations* for modeling our system RerrFact's subtask modules. We also align our efforts to maximize performance from each subtask performing binary classification instead of opting for approaches like external data fine-tuning, uti-
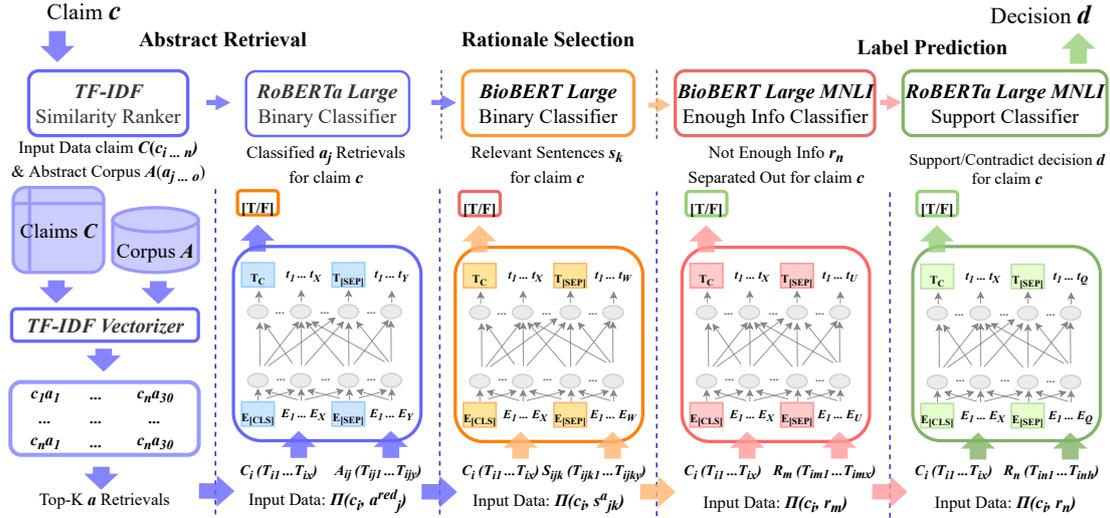
Figure 1: Architectural illustration of the RERRFACT scientific claim verification pipeline.

lizing extensive language models like T5, or using the joint learning architecture, etc. Here, we use the *reduced abstract representations* after the initial TF-IDF retrieval for the relevant abstract extraction subtask. After that, we use these retrieved abstracts for training the rationale selection model that adds a loose coupling effect between the two evidence retrieval subtasks. Finally, for stance prediction, we first segregate out {NoInfo} rationale instances and then predict stance for {Supports, Refutes} rationales. RERRFACT achieves the fourth rank in *SciFact* leaderboard by using language models of different BERT-variants, choosing the best performing one for each subtask. Our experimental results demonstrate the importance of this *loose coupling phenomenon* as we only stand after computationally expensive approaches that require much larger language models and optimization for various thresholding parameters for each subtask.

2. SciFact Dataset and Task Description

The SciFact dataset consists of a corpus with 5,183 relevant abstracts for 1,409 scientific claims [11]. These abstracts can either support or refute a claim with manually annotated rationales. Each claim has a unique single label, and no abstract has more than three rationales for a given claim. The natural claims derived from a paper and the papers cited in different paragraphs in it make the language modeling subtasks challenging especially due to added contextual scientific nuance.

For the SciFact task, one is given scientific claims $\mathscr{C}$ and a relevant abstract corpus $\mathscr{A}$ [11]. First, corresponding to a claim $c \in \mathscr{C}$, all unique abstracts $a \in \mathscr{A}$ are categorized as $y(c,a)$ in {Supports, Refutes, NoInfo}. Second, the sentence selection task functionally retrieves the relevant rationales $\{r_1(c,a), ..., r_m(c,a)\} \in \mathscr{R}$ for the given claim c for each abstract a. The performance of both these tasks is evaluated with *precision, recall, and F1* metrics for abstract and sentence-level tasks. Third, for the veracity verification task which is formulated as a stance prediction problem, labels {Supports, Refutes} are considered as positive labels, and {NoInfo} is taken as the negative label.

3. Methodology

We formulate each subtask for the SciFact task as a binary classification problem and create corresponding BERT representations for each sequence classifier. Figure 1 depicts the summarized view of the proposed RERRFACT system.

3.1. Abstract Retrieval

Here, we retrieve relevant abstracts from corpus $\{a_1, ..., a_j\} \in \mathscr{A}$ for claims $c \in \mathscr{C}$. First, we calculate the TF-IDF similarity of each claim c_i with all abstracts a_o in $\in \mathscr{A}$ and restrict to top-K ($K = 30$) similar abstracts. Second, we create *reduced abstract representations (a^{red}_j)* from these abstracts which is given by $a^{red}_j=\{title, s_1, s_{n/2}, s_n\}$. These are empirically the most meaningful representations for RoBERTa large language model [12], which we use for

Abstract Classification Approach	F1-score
Total Abstract (a^{total}_j)	72.25
Diff-Size Abstracts, Five Sentences (a^{diff-5}_j)	74.41
Diff-Size Abstracts, Three Sentences (a^{diff-3}_j)	68.63
RERRFACT's Reduced Abstract (a^{red}_j)	**79.67**

Table 1
F1-score performances on *dev* set for different comparative abstract representations.

Sentence Selection Approach	F1-score
BioBERT large Oracle Retrieval	67.63
Oracle Retrieval + No Evidence & Cited	65.47
Oracle + No Evidence Cited + (-3)*TF-IDF	62.23
RERRFACT's Loose Coupling	**69.57**

Table 2
F1-score performances on *dev* set for different comparative sentence selection approaches.

Stance Prediction Approach	F1-score
BioBERT-MNLI *(Multiclass)*	74.09
RoBERTA-Large-MNLI *(Multiclass)*	76.58
RERRFACT's NoINFO *(Binary)*	**87.14**
RERRFACT's SUPPORTS/REFUTES *(Binary)*	**82.67**
RERRFACT Classifier *(Two-Step Binary)*	**85.23**

Table 3
F1-score performances on *dev* set for different comparative stance prediction approaches.

binary classification with input sequence $< c_i, [SEP], a^{red}_j>$ for obtaining all the relevant abstracts.

Additionally, we obtain the above-stated representation logic by permuting different combinations of abstract sentences. For all retrieval approaches, we append the title with different lengths of abstract. Keeping the language model architecture constant, for the baseline approach, we first feed the complete abstract a^{total}_j with the title into the model. But while appending the whole abstract due to the limitation of BERT models to take maximum 512 tokens as input on an average, our inputs get truncated, which possibly results in some information loss.

In the second approach, we divide our abstracts into different groups based on their sizes {*small ($\leq 8 * s_k$), medium ($>8 * s_k$ & $\leq 14 * s_k$), large ($>14 * s_k$ & $\leq 24 * s_k$), extra-large ($>24 * s_k$ & $\leq L_{max} * s_k$)*}, and for each group of abstracts formed, we consider the top five relative index positions of the most frequently occurring sentences for each group and sequentially append those five sentences after the title (a^{diff-5}_j) as our new input sequence to fine-tune our language model. Also, we follow the same methodology but limit our sentences to only top-three sentences appended after the title (a^{diff-3}_j) for observing performance and computational trade-off variations on smaller representations.

The results from Table 1 demonstrate our final reduced retrieval representations outperforming other representations with its best F1-score. Our manual analysis into workings of these representations shows that the a^{red}_j={*title, $s_1, s_{n/2}, s_n$*} method captures qualitatively best portions of the introduction, methodology & conclusion on an average. More importantly, unlike other approaches, it avoids the abstract's numeric & additional bulk information components, keeping the representations compact & precise.

3.2. Rationale Selection

In this subtask, relevant evidence rationales $\widehat{\mathcal{R}}(c,a)$ = {$r_1(c,a), ..., r_m(c,a)$} are retrieved, where each $r_1(c,a)$ comprises of {$s_1(c,a), ..., s_k(c,a)$}. We use all sentences from each retrieved abstract from the previous stage to fine-tune our pre-trained BioBERT large language model [13] with input sequence $<c_i, [SEP], s^a_k>$ and binary output

[T/F]. Binding the abstract retrieval module to the rationale selection module while model training helps in improving *co-reference* identification performance and gives special attention only to claim relevant data.

Also, we further analyze different training mechanisms for the sentence selection subtask. First, we train our baselines only by using oracle retrieved abstract. Further, as a new variation, we add negative label sentences for claims with no supporting/refuting evidence but only respective *cited_doc_id* in the abstract corpus. Second, we decide to add more negative samples by adding top-three falsely retrieved abstracts from initial TF-IDF similarity retrieval. Finally, we try our loose-coupling approach by binding training to classified abstracts only. The results from Table 3 demonstrate the importance of the binding mechanism & emphasize that adding negative samples does not necessarily improve results.

3.3. Stance Prediction

In this subtask, we use the predicted rationales $\widehat{\mathcal{R}}(c,a)$ = {$r_1(c,a), ..., r_m(c,a)$} from the evidence retrieval stage to predict the veracity $\hat{y}(c,a)$ of the scientific claims $c \in \mathcal{C}$. We formulate this subtask as a two-stage binary classifier problem where the first classifier separates the rationales with $\hat{y}(c,a)$={NoINFO} with input sequence $<c_i, [SEP], r_m>$ and the second classifier predicts the stance $\hat{y}(c,a)$={SUPPORTS, REFUTES} with input representation $<c_i, [SEP], r_n>$. We choose the pre-trained BioBERT-MNLI language model for *Enough Information* detection and pre-trained RoBERTA-Large-MNLI for predicting *Claim Veracity*.

Further, we explore three-way classification by training the individual models of the RERRFACT veracity verification two-step module. We train our multiclass language model classifiers namely, BioBERT-MNLI & RoBERTa-Large-MNLI for directly predicting the {SUPPORTS, RE-

Models	Sentence-level						Abstract-level					
	Selection-only			Selection+Label			Label-Only			Label+Rationale		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
RERRFACT	**93.65**	**64.48**	**76.37**	**78.17**	**53.83**	**63.76**	79.17	54.55	64.59	**78.47**	54.07	**64.02**
ARSJoint	76.2	58.5	66.2	66.5	51.1	57.8	75.3	59.8	**66.7**	70.5	56.0	62.4
VERT5ERINI	64.81	57.37	60.87	60.8	53.83	57.1	65.07	**65.07**	65.07	61.72	**61.72**	61.72
ParagraphJoint	74.2	57.4	64.7	63.3	48.9	55.2	71.4	59.8	65.1	65.7	55.0	59.9
QMUL-SDS	80.75	58.47	67.83	72.08	52.19	60.54	**79.71**	52.63	63.40	76.81	50.72	61.10
VERISCI	54.3	43.4	48.3	48.5	38.8	43.1	56.4	48.3	52.1	54.2	46.4	50.0

Table 4
RERRFACT's performance on SCIFACT tasks on *dev* set.

Models	Sentence-level						Abstract-level					
	Selection-only			Selection+Label			Label-Only			Label+Rationale		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
VERT5ERINI	63.05	69.19	65.98	60.59	**66.49**	**63.40**	64.03	**72.97**	68.21	62.85	**71.62**	66.95
ARSJoint	79.53	**72.43**	**75.81**	66.17	60.27	63.08	72.22	64.41	68.10	69.70	62.16	65.71
RERRFACT	80.07	58.65	67.71	**73.43**	53.78	62.09	**82.89**	56.76	67.38	**81.58**	55.86	66.31
ParagraphJoint	79.86	63.24	70.59	68.94	54.59	60.94	75.81	63.51	**69.12**	73.66	61.71	**67.16**
QMUL-SDS	**81.58**	58.65	68.24	66.17	47.57	55.35	74.32	49.55	59.46	72.97	48.65	58.38
VERISCI	44.99	47.30	46.11	38.56	40.54	39.53	47.51	47.30	47.40	46.61	46.40	46.50

Table 5
RERRFACT's performance on SCIFACT tasks on *test* set.

FUTES, NOINFO} labels. The results in Table 2 demonstrates the advantage of using the two-step binary classification process in RERRFACT for the SCIFACT task. We attribute this performance increase to better prediction of REFUTES class, as multiclass classification models performed poorly for predicting this class due to its scarcity in the dataset. Hence, RERRFACT's two-step classification approach avoids false positive predictions of NOINFO class against the REFUTES class and improves on the claim refuting rationale prediction.

4. Experiment and Results

In our experiments, we analyze the performance of various language models in a standalone manner for each subtask and attempt multiple permutation settings for our system RERRFACT as shown in Tables 1, 2 and 3. Table 4 and Table 5 report the performance of our best language models in RERRFACT for each subtask in *SciFact* against the top leaderboard systems on both *dev* and *test* sets. For evaluation and reporting performance on the *dev* set, all language models for each subtask are trained only on the *train* set. Table 4 shows the evaluation results against the *dev* set having 300 claims. And for evaluation against the *test* set predictions, we train our models on the *train* set additionally combined with 75% of the *dev* set and validate our model results over the remaining 25% of the *dev* set. Table 5 reports the RERRFACT system's capabilities in terms of F_1 scores against 300 claims of the *test* set.

In the *abstract retrieval* subtask, we empirically observe that the *reduced abstract representations* substantially increase our retrieval performance, leading to a performance boost across all metrics in SCIFACT. This model is trained with batch size one for ten epochs. We achieve an *F1-score* of 79.67% against the *dev* set, which is higher than reported QMUL-SDS's *F1-score* of 74.15% but lower than VERT5ERINI's 89.95% *F1-score.* Second, for the *rationale selection* subtask, the BioBERT-large language model attains a higher recall score in the SCIFACT metrics because of the loose binding between the two subtasks for evidence retrieval as part of RERRFACT's system design. Though our *F1-score* performance for sentence selection was 69.57% which is again less than VERT5ERINI's *F1-score* of 76.14%, our performance on *dev* set supersedes all the systems, including the T5 language models of VERT5ERINI. Based on our analysis of predictions from abstract and sentence selection subtasks, this performance boost largely attributes to the regularization effect created by loosely binding the two evidence retrieval stages leading to highly accurate sentence predictions for the retrieved abstracts.

For the final *stance prediction* subtask, we train both our models in the two-step approach for 30 epochs with batch size 1. First, the {NOINFO} detector language model that eliminates evidence based on their unrelatedness to the scientific claim, achieves *F1-score* of 87.14%. The second stance predictor model for evidence that either supports or refutes the claim, achieves an *F1-score* of 82.67%. These two-step binary classifiers for neutral and support/refute evidence classification helps in achieving significant relative performance improvements on the *dev* set, as shown in Table 4's label prediction metrics. Also, from Table 5, we observe that RERRFACT's performance takes a relatively large dip in terms of prediction capabilities because of the relatively lower abilities to detect

Scientific Claim *(Reasoning Type, Frequency %)*	Wrongly Labeled Evidence *(Stance Gold Label)*
1/2000 in UK have abnormal PrP positivity. **(Numeric, 27.7%)**	*...indicating an overall prevalence of 493 per million population (95% confidence interval 282 to 801 per million)...* {SUPPORT}
Hypothalamic glutamate neurotransmission is crucial to energy balance. **(Directionality, 37.9%)**	*...secondary to impaired fasting-induced increases in the glucose-raising pancreatic hormone glucagon and...* {SUPPORT}
Breast cancer development is determined exclusively by genetic factors. **(Causal Effect, 34.4%)**	*...women who developed breast cancer... established environmental risk factors...alcohol consumption).* {CONTRADICT}

Table 6
Reasoning categories where RERRFACT fails to predict correct labels.

true negatives for each subtask and wrong predictions on scientifically exhaustive rationales.

5. Analysis

Our manual analysis shows that RERRFACT's increase in performance can be attributed to its ability to process *scientific background knowledge* and *co-references* more accurately. First, the reduced abstract representations help in qualitatively improving the *co-references* inference capabilities. Second, the dynamic biological pre-trained embeddings in classifier models help in increasing the *scientific background knowledge.* Additionally, by coupling the sentence selection module's training with retrieved abstract sentences as input, we add a form of regularization that increases generalization for rationale extraction subtask while keeping our sentence selection model compact. But, our system still fails to comprehend concepts like *quantitative directionality*, *numerical reasoning*, and *causal effects*. This we further demonstrate by examples in Table 6 alongside their corresponding error-occurring frequency in *dev* set over 29 misclassified *claim-rationale* pairs.

6. Conclusion

In this work, our proposed system RERRFACT demonstrates that reduced evidence retrieval representations and loosely binding the evidence retrieval stages for flexible regularization lead to better and concise retrieved rationale sentences. Additionally, combined with RERRFACT's two-step stance prediction approach, it outperforms all the other veracity verification systems on the SCIFACT *dev* set. Also, for RERRFACT, the performance especially takes a relatively high dip on the *test* set, which can be attributed to a high false-positive rate on the *test* set & also that SCIFACT metric penalizations requiring more regularized predictions for each subtask. Our proposed system RERRFACT ranks 4th on the SCIFACT leaderboard, with 62.09% F1-score for the Sentence+Label prediction module, while the top-performing system has an F1-score of 67.21%. As future work, we would systematically im-

prove upon these limitations and further explore novel premise assimilation architectures to create qualitatively improved veracity verification systems.

References

[1] J. DeYoung, S. Jain, N. F. Rajani, E. Lehman, C. Xiong, R. Socher, B. C. Wallace, ERASER: A benchmark to evaluate rationalized NLP models, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 4443–4458. URL: https://aclanthology.org/2020.acl-main.408. doi:10.18653/v1/2020.acl-main.408.

[2] W. Ferreira, A. Vlachos, Emergent: a novel data-set for stance classification, in: Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: Human language technologies, 2016, pp. 1163–1168.

[3] A. Vlachos, S. Riedel, Fact checking: Task definition and dataset construction, in: Proceedings of the ACL 2014 workshop on language technologies and computational social science, 2014, pp. 18–22.

[4] J. Thorne, A. Vlachos, O. Cocarascu, C. Christodoulopoulos, A. Mittal, The fact extraction and VERification (FEVER) shared task, in: Proceedings of the First Workshop on Fact Extraction and VERification (FEVER), Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 1–9. URL: https://aclanthology.org/W18-5501. doi:10.18653/v1/W18-5501.

[5] Y. Nie, H. Chen, M. Bansal, Combining fact extraction and verification with neural semantic matching networks, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 33, 2019, pp. 6859–6866.

[6] S. Chen, D. Khashabi, W. Yin, C. Callison-Burch, D. Roth, Seeing things from a different angle:discovering diverse perspectives about claims, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Associ-

ation for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 542–557. URL: https://aclanthology.org/N19-1053. doi:10.18653/v1/N19-1053.

[7] R. Pradeep, X. Ma, R. Nogueira, J. Lin, Scientific claim verification with VerT5erini, in: Proceedings of the 12th International Workshop on Health Text Mining and Information Analysis, Association for Computational Linguistics, online, 2021, pp. 94–103. URL: https://aclanthology.org/2021.louhi-1.11.

[8] X. Li, G. A. Burns, N. Peng, A paragraph-level multi-task learning model for scientific fact-verification, in: A. P. B. Veyseh, F. Dernoncourt, T. H. Nguyen, W. Chang, L. A. Celi (Eds.), Proceedings of the Workshop on Scientific Document Understanding co-located with 35th AAAI Conference on Artificial Inteligence, SDU@AAAI 2021, Virtual Event, February 9, 2021, volume 2831 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2021. URL: http://ceur-ws.org/Vol-2831/paper8.pdf.

[9] Z. Zhang, J. Li, F. Fukumoto, Y. Ye, Abstract, rationale, stance: A joint model for scientific claim verification, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2021, pp. 3580–3586. URL: https://aclanthology.org/2021.emnlp-main.290.

[10] X. Zeng, A. Zubiaga, QMUL-SDS at SCIVER: Step-by-step binary classification for scientific claim verification, in: Proceedings of the Second Workshop on Scholarly Document Processing, Association for Computational Linguistics, Online, 2021, pp. 116–123. URL: https://aclanthology.org/2021.sdp-1.15. doi:10.18653/v1/2021.sdp-1.15.

[11] D. Wadden, S. Lin, K. Lo, L. L. Wang, M. van Zuylen, A. Cohan, H. Hajishirzi, Fact or fiction: Verifying scientific claims, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online, 2020, pp. 7534–7550. URL: https://aclanthology.org/2020.emnlp-main.609. doi:10.18653/v1/2020.emnlp-main.609.

[12] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Ro{bert}a: A robustly optimized {bert} pretraining approach, 2020. URL: https://openreview.net/forum?id=SyxS0T4tvS.

[13] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, J. Kang, Biobert: a pre-trained biomedical language representation model for biomedical text mining, Bioinformatics 36 (2020) 1234–1240.