Scientific Claim Verification with VERT5ERINI

Ronak Pradeep, Xueguang Ma, Rodrigo Nogueira, and Jimmy Lin

David R. Cheriton School of Computer Science
University of Waterloo

Abstract

This work describes the adaptation of a pretrained sequence-to-sequence model to the task of scientific claim verification in the biomedical domain. We propose a system called VERT5ERINI that exploits T5 for abstract retrieval, sentence selection, and label prediction, which are three critical sub-tasks of claim verification. We evaluate our pipeline on SCIFACT, a newly curated dataset that requires models to not just predict the veracity of claims but also provide relevant sentences from a corpus of scientific literature that support the prediction. Empirically, our system outperforms a strong baseline in each of the three sub-tasks. We further show VERT5ERINI's ability to generalize to two new datasets of COVID-19 claims using evidence from the CORD-19 corpus.

1 Introduction

The popularity of social media and other means of disseminating content, combined with automated algorithms that create "echo chamber" effects, has increased the proliferation of misinformation online. This has led to increased attention in the community on building better fact verification systems. Until recently, most fact verification datasets were constrained to domains such as Wikipedia, discussion blogs, and social media (Thorne et al., 2018; Hanselowski et al., 2019).

In the current environment, amidst the COVID-19 pandemic and the unease that comes with insufficient insight about the virus, there has been a sharp increase in curiosity among the general public toward scientific knowledge. While such curiosity is always appreciated, this has inadvertently resulted in a large spike of scientific facts being misrepresented, often to push a personal or political agenda, inducing ineffective and frequently even harmful policies and behaviours.

To mitigate this issue, Wadden et al. (2020) introduced the task of scientific claim verification, where systems need to evaluate the veracity of a claim against a scientific corpus. To facilitate this, they introduced the SCIFACT dataset that consists of scientific claims accompanied with abstracts that either support or refute the claim. The dataset also provides a set of rationale sentences for each claim that is necessary and sufficient to conclude its veracity. In addition, the authors provide VERISCI, a baseline for this task that takes inspiration from previous state-of-the-art systems (DeYoung et al., 2020) for the FEVER claim verification dataset (Thorne et al., 2018). This pipeline retrieves relevant abstracts by TF-IDF similarity, uses a BERT-based model (Devlin et al., 2019) to select rationale sentences, and finally labels each abstract as either SUPPORTS, NOINFO, or REFUTES with respect to the claim.

Despite the success of BERT for tasks like passage-level (Nogueira et al., 2019), document-level (Dai and Callan, 2019; MacAvaney et al., 2019; Akkalyoncu Yilmaz et al., 2019) and sentence-level (Soleimani et al., 2019) retrieval, there is evidence that ranking with sequence-to-sequence models can achieve even better effectiveness, particularly in zero-shot scenarios or with limited training data (Nogueira et al., 2020; Pradeep et al., 2021). This was further demonstrated in the TREC-COVID challenge (Roberts et al., 2020) where one of the best performing systems used sequence-to-sequence models for retrieval (Zhang et al., 2020; Pradeep et al., 2021). Similar trends are noted in CovidQA (Tang et al., 2020), a question answering dataset for COVID-19, where zero-shot sequence-to-sequence models outperformed other baselines.

Hence, we propose VERT5ERINI, where all three steps—abstract retrieval, sentence selection, and label prediction exploit T5 (Raffel et al., 2020),

Proceedings of the 12th International Workshop on Health Text Mining and Information Analysis, pages 94–103
April 19, 2021.

a powerful sequence-to-sequence language model. VERT5ERINI outperforms the VERISCI baseline on the SCIFACT tasks by a large margin and advances the state of the art for the task of Scientific Claim Verification. We also demonstrate the effectiveness of our system in verifying two different sets of COVID-19 claims with no additional training or hyperparameter tuning.

2 Task

In the SCIFACT task (Wadden et al., 2020), systems are provided with a scientific claim q and a corpus of abstracts $\mathcal{C}$ and tasked to return:

- A set of evidence abstracts $\hat{\mathcal{E}}(q)$.
- A label $\hat{y}(q, a)$ that maps claim q and abstract a to one of $\{\text{SUPPORTS}, \text{REFUTES}, \text{NOINFO}\}$.
- A set of rationale sentences $\hat{S}(q, a)$ when $\hat{y}(q, a) \in \{\text{SUPPORTS}, \text{REFUTES}\}$.

Given the ground truth label $y(q, a)$, the set of gold abstracts $\mathcal{E}(q)$, and the set of gold rationales $\mathcal{R}(q, a)$ (each gold rationale is a set of sentences), the predictions are evaluated in two ways:

- **Abstract-level evaluation**, where systems are judged on whether they can identify abstracts that support or refute the claim. First, $a \in \hat{\mathcal{E}}(q)$ is *correctly labelled* if both $a \in \mathcal{E}(q)$ and $\hat{y}(q, a) = y(q, a)$. Second, it is *correctly rationalized*, if in addition, $\exists R \in \mathcal{R}(q, a)$ such that $R \subseteq \hat{S}(q, a)$.[1] These evaluations are referred to as $\text{Abstract}_{\text{Label-Only}}$ and $\text{Abstract}_{\text{Label+Rationale}}$, respectively.
- **Sentence-level evaluation**, where systems are evaluated on whether they can identify sentences sufficient to justify the abstract-level predictions. First, $\hat{s} \in \hat{S}(q, a)$ is *correctly selected* if $\exists R \in \mathcal{R}(q, a)$ such that both $\hat{s} \in R$ and $R \subseteq \hat{S}(q, a)$. Second, it is *correctly labelled*, if in addition, $\hat{y}(q, a) = y(q, a)$. These evaluations are referred to as $\text{Sentence}_{\text{Selection-Only}}$ and $\text{Sentence}_{\text{Selection+Label}}$, respectively.

Specifically, SCIFACT uses a corpus of 5,183 abstracts. Abstracts that support or refute each claim are annotated with rationale sentences (see Table 3 for examples). The label distribution is provided in Table 1. There are 1,409 claims, 809 of which are part of the training set and the rest are split equally across the development and test sets. Although the test set is balanced with 100 claims for each class (SUPPORTS, NOINFO, and REFUTES), it is clear that the training and development sets have significant class imbalance. This, coupled with the small dataset size, highlights the importance of zero- or few-shot systems for this task.

Set	SUPPORTS	NOINFO	REFUTES	Total
Train	332	304	173	809
Dev	124	112	64	300
Test	100	100	100	300

Table 1: SCIFACT label distribution.

Claim Set	SUPPORTS	REFUTES	Total
COVID-19 SCIFACT	-	-	36
COVID-19 Scientific	41	101	142

Table 2: COVID-19 claims.

To show that our system is able to verify claims related to COVID-19 by identifying evidence from the much larger CORD-19 corpus,[2] we evaluate VERT5ERINI in a zero-shot setting on two other datasets:

COVID-19 SCIFACT (Wadden et al., 2020) is a set of 36 COVID-related claims curated by a medical student. In this set, the same claim can sometimes be both supported and refuted by different abstracts, a scenario not observed in the main SCIFACT task. Two examples in this set are shown in Table 4.

COVID-19 Scientific (Lee et al., 2020) contains 142 claims (label distribution in Table 2) gathered by collecting COVID-related scientific truths and myths from sources like the U.S. Centers for Disease Control and Prevention (CDC), MedicalNewsToday, and the World Health Organization (WHO). Unlike the other two datasets, COVID-19 Scientific only provides a single label $y(q) \in \{\text{SUPPORTS}, \text{REFUTES}\}$ for a claim. According to the authors, during the construction of the dataset, claims that were unverifiable according to the CDC or the WHO were mapped to REFUTES. Hence, we make the following modifications to VERT5ERINI:

1. If $\hat{y}(q, a) = \text{NOINFO}$, then $\hat{y}(q, a)$ is modified to REFUTES.

[1] In SCIFACT's abstract-level evaluation, it is required that $|\hat{S}(q, a)| \leq 3$.

[2] We use the 2020-06-17 dump of CORD-19, which contains 192,459 abstracts, about 40 times as many abstracts as those in SCIFACT.

Claim	Label	Evidence
ALDH1 expression is associated with poorer prognosis in breast cancer.	SUPPORTS	Application of stem cell biology to breast cancer research has been limited by the lack of simple methods for identification and isolation of normal and malignant stem cells. . . . **In a series of 577 breast carcinomas, expression of ALDH1 detected by immunostaining correlated with poor prognosis.** . . .
CX3CR1 on the Th2 cells impairs T cell survival	REFUTES	Allergic asthma is a T helper type 2 (T(H)2)-dominated disease of the lung. . . . **We found that CX3CR1 signaling promoted T(H)2 survival in the inflamed lungs, and injection of B cell leukemia/lymphoma-2 protein (BCl-2)-transduced CX3CR1-deficient T(H)2 cells into CX3CR1-deficient mice restored asthma.** . . .
Arterioles have a larger lumen diameter than venules.	NOINFO	N/A

Table 3: Three SCIFACT claims, their labels, and their corresponding evidence (rationale highlighted in **bold**) if available. VERT5ERINI correctly predicts these labels and retrieves the matching evidence.

Claim	Label	Evidence
Hypertension and Diabetes are the most common comorbidities for COVID-19.	SUPPORTS	**Investigations reported that hypertension, diabetes, and cardiovascular diseases were the most prevalent comorbidities among the patients with coronavirus disease 2019 (COVID-19).** . . . The aim of this review was to summarize the current knowledge about the relationship between hypertension and COVID-19 and the role of hypertension on outcome in these patients.
The Secondary Attack rate of COVID-19 is 10.5% for household members/close contacts.	REFUTES	Background: As of April 2, 2020, the global reported number of COVID-19 cases has crossed over 1 million with more than 55,000 deaths. . . . **We estimated the household SAR to be 13.8% (95% CI: 11.1–17.0%) if household contacts are defined as all close relatives and 19.3% (95% CI: 15.5–23.9%) if household contacts only include those at the same residential address as the cases, assuming a mean incubation period of 4 days and a maximum infectious period of 13 days.**

Table 4: Two COVID-19 claims from SCIFACT, their *predicted* labels and their corresponding *predicted* evidence (rationale highlighted in **bold**).

Claim	Label	Evidence
Some people become infected by COVID-19 but don't develop any symptoms and don't feel unwell.	SUPPORTS	COVID-19 is an emerging infectious disease with widespread transmission of the coronavirus SARS-CoV-2 in the Netherlands. . . . **Others do not show any symptoms, but can still contribute to the transmission of the virus.** . . .
Young people will not get COVID-19.	REFUTES	Objective: To explore the epidemiological characteristics of COVID-19 associated with SARS-Cov-2 in Guizhou province, and to compare the differences in epidemiology with other provinces. . . . **Most of COVID-19 patients were 18-45 years old (52.27%).** . . . **CONCLUSION: Among the cases, most patients were young adults.**
Bill Gates caused the infection of COVID-19.	REFUTES	N/A

Table 5: Three COVID-19 Scientific claims from Lee et al. (2020), their *predicted* labels and their corresponding *predicted* evidence (rationale highlighted in **bold**). Note that if the system cannot find any supporting evidence for a claim, it is considered refuted.

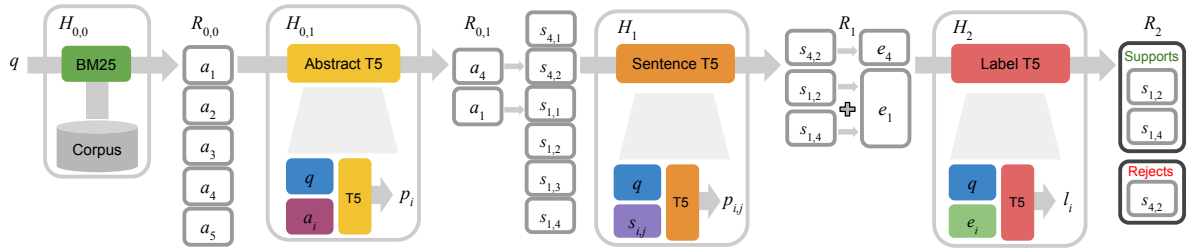

Figure 1: Illustration of the VERT5ERINI pipeline. In stage $H_{0,0}$, given a query q, the top $k_{0,0}$ ($= 5$ in the figure) candidate documents $R_{0,0}$ are retrieved using BM25. In stage $H_{0,1}$, Abstract T5 produces a relevance score p_i for each pair of query q and candidate $a_i \in R_{0,0}$. The top $k_{0,1}$ ($= 2$ in the figure) candidates with respect to these relevance scores are expanded to sentence-level granularity and passed to stage H_1, in which Sentence T5 computes a relevance score $p_{i,j}$ for each pair of query q and candidate sentence $s_{i,j}$. Sentences for a particular abstract a_i scoring above a threshold form its set of rationale sentences e_i and each set along with the query q are passed to stage H_2, in which Label T5 predicts the label.

2. $\hat{y}(q) = \max_{a \in \hat{\mathcal{E}}(q)} \hat{y}(q, a)$.

3. If $|\bigcup_{a \in \hat{\mathcal{E}}(q)} \hat{S}(q, a)| = 0$, i.e., the set of all evidence sentences across the abstracts is empty, then $\hat{y}(q) =$ REFUTES.

Three examples from this set are shown in Table 5. As one can imagine, it would be impossible to find any discussion of outlandish claims like "Bill Gates caused the infection of COVID-19" in a corpus of biomedical literature and hence VERT5ERINI maps them to REFUTES.

3 Methods

Our proposed system, VERT5ERINI (see Figure 1), has three major components:

1. H_0**: Abstract Retrieval** — which given claim q retrieves the top-k abstracts from corpus $\mathcal{C}$.

2. H_1**: Sentence Selection** — which given claim q and one of the top-k abstracts a, selects sentences from a that form $\hat{S}(q, a)$.

3. H_2**: Label Prediction** — which given claim q and the rationale sentences $\hat{S}(q, a)$, predicts the final label $\hat{y}(q, a)$.

3.1 H_0: Abstract Retrieval

Given a scientific claim q and a corpus $\mathcal{C}$ of scientific abstracts, H_0 is tasked with retrieving the top-k abstracts from $\mathcal{C}$. We propose both a single-stage and a two-stage abstract retrieval pipeline.

In both cases, the first stage $H_{0,0}$ involves treating the query as a "bag of words" for ranking abstracts from the corpus using the BM25 scoring function (Robertson et al., 1994). Our implementation uses the Anserini IR toolkit (Yang et al., 2017, 2018),[3] which is built on the popular open-source Lucene search engine, and its Pyserini Python interface (Akkalyoncu Yilmaz et al., 2020; Lin et al., 2021) to support simple keyword search capabilities on the corpus. The output of this stage is a list of k_0 candidate abstracts.

The second abstract reranking stage, $H_{0,1}$, is tasked to estimate a score p quantifying how relevant a candidate abstract a is to a query q. In this stage, the abstracts retrieved in $H_{0,0}$ are reranked by a pointwise reranker, which we call monoT5. Our reranker is based on Nogueira et al. (2020), which uses T5 (Raffel et al., 2020), a sequence-to-sequence model pretrained with a similar masked language modeling objective as BERT. In this model, all target tasks are cast as sequence-to-sequence tasks. We adapt the approach to abstract reranking by using the following input sequence:

Query: q Document: a Relevant:

The model is fine-tuned to produce the words "true" or "false" depending on whether the abstract is relevant or not to the query. That is, "true" and "false" are the "target words" (i.e., ground truth predictions in the sequence-to-sequence transformation).

[3] http://anserini.io/

Since a considerable number ($\approx 15\%$) of SciFact abstracts are longer than the context limit of T5 (512 tokens), we first segment each abstract (on average 9 sentences) into spans by applying a sliding window of 6 sentences with a stride of 3.

In order to fine-tune monoT5 on abstract reranking in SciFact, we use all cited abstracts in the training set as positive examples. For each claim, we select negative examples by randomly selecting a non-ground truth abstract among the top-10 BM25 ranked candidates. We train on this set with a batch size of 128 for 200 steps, which corresponds to approximately 5 epochs.

At inference time, we first compute probabilities for each query–segment pair (in a reranking setting) by applying a softmax only on the logits of the "true" and "false" tokens. We then obtain the relevance score of the document as the highest probability assigned to the "true" token among all segments. The top-k_0 abstracts, R_0, with respect to these scores are then selected.

We run inference with three different monoT5[4] settings for abstract reranking: (1) fine-tuned on the MS MARCO passage dataset (Bajaj et al., 2016); (2) fine-tuned on MS MARCO and then fine-tuned again on the medical subset of MS MARCO (MacAvaney et al., 2020); and (3) fine-tuned on MS MARCO and then fine-tuned again on SciFact.

We choose to "pre–fine-tune" relevance classifiers on MS MARCO passages as it has been shown to help in various other tasks (Akkalyoncu Yilmaz et al., 2019; Zhang et al., 2020; Nogueira et al., 2020; Pradeep et al., 2021). Similarly, MacAvaney et al. (2020) demonstrated that fine-tuning the classifiers on the medical subset of MS MARCO helps with biomedical-domain relevance ranking.

3.2 H_1: Sentence Selection

In this stage, the goal is to select rationale sentences $\hat{S}(q, a)$ from each abstract a for each of the top-k abstracts retrieved $\hat{\mathcal{E}}(q)$. We use T5 for this task also. The following input sequence is used:

$$\text{Query: } q \quad \text{Document: } s \quad \text{Relevant:}$$

where s is a sentence in the abstract a.

We fine-tune a monoT5 (pre–fine-tuned on MS MARCO passage) on SciFact's gold rationales as positive examples and sentences randomly sampled from $\mathcal{E}(q)$ as negatives. We train on this set of sentences with a batch size of 128 for 2500 steps.

During inference, similar to abstract ranking, we compute a probability of the sentence being relevant based on the logits of the "true" and "false" tokens. Finally, we filter out all sentences whose "true" probability is below the threshold of 0.999 to obtain $\hat{S}(q, a)$.

3.3 H_2: Label Prediction

Given the claim q, an abstract a and their corresponding set of rationale sentences $\hat{S}(q, a)$, H_2 is tasked to predict a label $\hat{y}(q, a) \in \{\text{SUPPORTS}, \text{NOINFO}, \text{REFUTES}\}$. Yet again, we use T5 for this task with the input sequence:

$$\text{hypothesis: } q \quad \text{sentence1: } s_1 \quad \cdots \quad \text{sentence}z\text{: } s_z$$

where $s_1, \cdots, s_z$ are the rationale sentences in $\hat{S}(q, a)$. The target sequence is one of "true", "weak", or "false" tokens corresponding to the labels SUPPORTS, NOINFO, or REFUTES, respectively. Note that a feature of this approach is feeding a collection of sentences into the model at once, as opposed to the perhaps more obvious approach of performing per-sentence independent label prediction. This requires the model to process longer input sequences, but allows predictions to incorporate evidence from multiple sources. In the parlance of learning to rank in the context of information retrieval, this would be called a "listwise" approach (Li, 2011).

SUPPORTS and REFUTES training examples are selected from evidence sets of cited abstracts for each claim. The sentences in each evidence set are concatenated with the claim in the above input sequence template as a single example for the corresponding label. The NOINFO examples are selected by concatenating one or two randomly-selected non-rationale sentences from each of the cited abstracts across all labels. Here, we fine-tune a fresh T5-3B (that was just pretrained on the mixture task) and not a pre–fine-tuned monoT5 since there is no natural transfer from the relevance ranking task. We use a batch size of 128 and select the best checkpoint after $[200, 400, 600, 800, 1000]$ steps based on the development set scores.

During inference, the token with the highest probability is assigned the label $\hat{y}(q, a)$ for abstract $a \in \mathcal{E}(q)$.

[4] All models are T5-3B.

Method	R@3	R@5
Oracle	97.61	100.00
TF-IDF	69.38	75.60
BM25	79.90	84.69
T5 (MS MARCO)	86.12	**89.95**
T5 (MS MARCO MED)	85.65	89.00
T5 (SciFact)	**86.60**	89.40

Table 6: Comparison of abstract retrieval methods on the development set of SciFact.

Method	P	R	F_1
RoBERTa-large	73.71	70.49	72.07
T5	**79.29**	**73.22**	**76.14**

Table 7: Comparison of different sentence selection methods on SciFact's development set.

Method	Label	P	R	F_1
RoBERTa-large	Supports	92.56	81.16	86.49
	NoInfo	74.82	**92.86**	82.87
	Refutes	77.05	66.20	71.21
T5	Supports	**93.13**	**88.41**	**90.71**
	NoInfo	**85.25**	**92.86**	**88.89**
	Refutes	**86.76**	**83.10**	**84.89**

Table 8: Comparison of different label prediction models on SciFact's development set.

4 Results

4.1 Baselines

For the SciFact and COVID-19 SciFact end-to-end tasks, the baseline system used is VeriSci (Wadden et al., 2020). It has an abstract retrieval module that uses TF-IDF, a sentence selection module trained on SciFact, and a label prediction module trained on FEVER + SciFact. For the abstract retrieval module, the authors report the best full-pipeline development set scores by retrieving the top three documents.

For the COVID-19 Scientific task, we compare with the following two baselines established by Lee et al. (2020):

- LiarMisinfo (Lee et al., 2020) uses a BERT-large (Devlin et al., 2019) label prediction model fine-tuned on LIAR-PolitiFact (Wang, 2017), a set of 12.8k claims collected from PolitiFact. It is worth noting that LIAR-PolitiFact does not contain any claims related to COVID-19.

- LM Debunker (Lee et al., 2020) uses GPT-2 (Radford et al., 2019) to determine the perplexity of the claim given evidence sentences. Claims with a perplexity score higher than a threshold are labeled Refutes while the others are labeled Supports.

The sentence selection module in both baselines employ TF-IDF followed by some rule-based evidence filtering to select the top three sentences for each claim. LiarMisinfo represents a zero-shot model where no fine-tuning is performed on the COVID-19 Scientific set. LM Debunker, on the other hand, first partitions the data into a validation and a test set. The validation set is used to tune the perplexity threshold for the model following which evaluation is performed on the test set.

4.2 Abstract Retrieval

Table 6 reports recall at rank three (R@3) and rank five (R@5) for abstract retrieval. The oracle (first row) shows that most claims from the development set have fewer than three relevant abstracts and all have fewer than five. For comparison, we show the effectiveness of the TF-IDF method used by Wadden et al. (2020).

We find that using BM25 results in an effectiveness improvement of around 10 points in comparison to the TF-IDF baseline. Using T5 to rerank the top-20 abstracts retrieved from BM25 results in a 17-point improvement over the baseline.

However, results show almost no difference in effectiveness whether T5 was fine-tuned on SciFact or on MS MARCO MED. This might be due to the relatively small size of the SciFact dataset and the fact that MS MARCO MED data is not entirely relevant to the target task. Hence, we use T5 fine-tuned only on the full MS MARCO dataset (i.e., no further fine-tuning) in the end-to-end pipeline experiments (Section 4.5).

4.3 Sentence Selection

Table 7 reports the precision, recall, and F_1 scores for the sentence selection task. We find that T5 (MS MARCO) fine-tuned on SciFact outperforms the RoBERTa-large baseline fine-tuned on SciFact used by Wadden et al. (2020). This result, together with results from Table 6, demonstrates the effectiveness of the T5 model at selecting evidence at various levels of granularity.

Label Only			
Method	P	R	F_1
(1) Oracle (VERISCI)	90.97	67.46	77.47
(2) Oracle (ours)	**92.70**	**78.95**	**85.27**
(3) VERISCI	55.31	47.37	51.03
(4) VERT5ERINI (BM25)	**70.88**	61.72	**65.98**
(5) VERT5ERINI (T5)	65.07	**65.07**	65.07
Label+Rationale			
Method	P	R	F_1
(6) Oracle (VERISCI)	85.16	63.16	72.53
(7) Oracle (ours)	**88.76**	**75.60**	**81.65**
(8) VERISCI	52.51	44.98	48.45
(9) VERT5ERINI (BM25)	**67.03**	58.37	**62.40**
(10) VERT5ERINI (T5)	61.72	**61.72**	61.72

Table 9: Full pipeline abstract-level effectiveness on SCIFACT's development set.

Selection Only			
Method	P	R	F_1
(1) Oracle (VERISCI)	79.41	59.02	67.71
(2) Oracle (ours)	**83.54**	**72.13**	**77.42**
(3) VERISCI	52.46	43.72	47.69
(4) VERT5ERINI (BM25)	**67.70**	53.83	59.97
(5) VERT5ERINI (T5)	64.81	**57.37**	**60.87**
Selection+Label			
Method	P	R	F_1
(6) Oracle (VERISCI)	71.32	53.01	60.82
(7) Oracle (ours)	**78.16**	**67.49**	**72.43**
(8) VERISCI	46.89	39.07	42.62
(9) VERT5ERINI (BM25)	**63.92**	50.82	56.62
(10) VERT5ERINI (T5)	60.80	**53.83**	**57.10**

Table 10: Full pipeline sentence-level effectiveness on SCIFACT's development set.

4.4 Label Prediction

In Table 8, we present label-wise precision, recall, and F_1 scores for the label prediction task. For SUPPORTS and REFUTES labels, the input to the model comprises gold rationales from cited abstracts. For NOINFO labels, recall that cited abstracts are available but no gold rationales exist. In this case, we pick the two most similar sentences according to TF-IDF from each abstract.

The results across all labels demonstrate that T5 fine-tuned on SCIFACT's label prediction task shows significant improvements over the baseline RoBERTa-large that was fine-tuned on FEVER followed by fine-tuning on SCIFACT's label prediction task. We believe some of this can be credited to T5's pretraining on a mixture of multiple tasks. Although this mixture does not include FEVER, the corpus contains various other NLI datasets, including MNLI (Williams et al., 2018) and QNLI (Rajpurkar et al., 2016).

4.5 Full Pipeline

In Tables 9 and 10, we report the precision, recall, and F_1 scores of abstract-level evaluation and sentence-level evaluation, respectively, for full pipeline systems.

Rows 1, 2, 6, 7 present the scores in the oracle abstract retrieval setting, where gold evidence abstracts are provided to systems. We see that our pipeline outperforms VERISCI by around 10 F_1 points at both the abstract and sentence level. The improvements are even larger in the $\text{Abstract}_{\text{Label+Rationale}}$ and $\text{Sentence}_{\text{Selection+Label}}$ evaluation settings (rows 6, 7 in Tables 9 and 10, respectively) which require more from systems in terms of sentence selection and label prediction.

In rows 3–5 and 8–10, we report scores in the full pipeline setting where systems are also required to retrieve relevant abstracts. We evaluate two full pipeline systems, one that uses BM25 alone and another that uses BM25 followed by T5 (MARCO) for abstract retrieval. Both these systems outperform the baseline system VERISCI by about 14 F_1 points. This comes as no surprise seeing that our models display notable improvements for each of the three sub-tasks.

Notice that in Table 6, using T5 (MARCO) brings large gains in terms of R@3 over the BM25 baseline. Yet, in the case of the full pipeline, with these two abstract retrieval methods, we only observe comparable effectiveness on the development set. We believe this might be linked to the relatively small size of the development set; below, we choose to probe the SCIFACT hidden test set with both configurations.

From Table 11, it is clear that in the hidden test set, both our systems outperform the baseline VERISCI, with evaluation aspects like Sentence+Label (rows 10–12) showing relative improvements of around 50%. Comparing with the corresponding conditions in Tables 9 and 10, we also see no indication of overfitting. We also note that abstract retrieval using the two-stage approach brings large gains here (rows 5, 11 vs. 6, 12) unlike in the development set. This shows that neural reranking, even though used in a zero-shot formu-

Label Only			
Method	P	R	F_1
(1) VERISCI	47.5	47.3	47.4
(2) VERT5ERINI (BM25)	63.1	60.8	61.9
(3) VERT5ERINI (T5)	**63.6**	**66.2**	**64.9**
Label+Rationale			
Method	P	R	F_1
(4) VERISCI	46.6	46.4	46.5
(5) VERT5ERINI (BM25)	60.3	58.1	59.2
(6) VERT5ERINI (T5)	**61.5**	**64.0**	**62.7**
Selection Only			
Method	P	R	F_1
(7) VERISCI	45.0	47.3	46.1
(8) VERT5ERINI (BM25)	64.9	58.9	61.8
(9) VERT5ERINI (T5)	**66.2**	**63.5**	**64.8**
Selection+Label			
Method	P	R	F_1
(10) VERISCI	38.6	40.5	39.5
(11) VERT5ERINI (BM25)	58.3	53.0	55.5
(12) VERT5ERINI (T5)	**60.0**	**57.6**	**58.8**

Table 11: Full pipeline effectiveness of VERT5ERINI on SCIFACT's test set.

lation, is critical to getting higher quality abstracts from the corpus $\mathcal{C}$, thereby improving effectiveness in later stages too.

4.6 Verification of COVID-19 Claims

Finally, we evaluate our most effective pipeline configuration, VERT5ERINI (T5), on the two sets of COVID-related claims. We do this in a zero-shot setting in that we do not fine-tune our model on either of these datasets.

In the COVID-19 SCIFACT set, for each claim q, we use VERT5ERINI (T5) to predict evidence abstracts, $\hat{\mathcal{E}}(q)$. A $(q, \hat{\mathcal{E}}(q))$ pair is considered *plausible* if at least half of the evidence abstracts in $\hat{\mathcal{E}}(q)$ are found to have reasonable rationales and labels. For 30 out of 36 claims, we find that VERT5ERINI (T5) provides plausible evidence abstracts. These claims have reasonable labels and evidence rationales selected successfully from evidence abstracts. This is in comparison to the 23 out of 36 claims for which VERISCI provides plausible evidence, demonstrating the effectiveness of our system in the zero-shot setting.

In the COVID-19 Scientific set, we compare the effectiveness of VERT5ERINI with that of two baselines considered by Lee et al. (2020). Table 12 reports the accuracy, the F_1-Macro, and the F_1-Binary scores on the test set. The F_1-Binary score

Method	Accuracy	F_1-Macro	F_1-Binary
LiarMisinfo	61.5	59.2	82.8
LM Debunker	75.4	69.8	83.1
VERT5ERINI (T5)	**78.2**	**73.2**	**83.8**

Table 12: Label prediction effectiveness on COVID-19 Scientific claims

corresponds to the F_1 score of the REFUTES label, since debunking misinformation is critical. Note that the LM Debunker baseline uses the average scores across four-fold cross-validation on the test set, unlike VERT5ERINI and LiarMisinfo. We observe that VERT5ERINI outperforms both baselines in a zero-shot setting, without any in-task tuning like the LM Debunker. The adaptability of VERT5ERINI to both these new tasks with no additional training makes a strong case for the effectiveness of our system.

5 Conclusions

In this paper, we introduced VERT5ERINI, a novel system for scientific claim verification that exploits a generation-based approach to abstract ranking, sentence selection, and claim verification. Such systems are of significance in this age of misinformation, amplified by the COVID-19 pandemic. Experiments show that our system outperforms the state of the art in the end-to-end task on the SCIFACT dataset. We note improvements in each of the three sub-tasks, demonstrating the importance of this sequence-to-sequence approach as well as zero-shot and few-shot transfer capabilities. Finally, we find that VERT5ERINI generalizes to two new COVID-19 related datasets with no tuning of parameters while maintaining high effectiveness.

Yet, there is still a large gap between our system and an oracle. Ideally, a system that performs scientific claim verification should possess additional attributes such as:

- Numerical reasoning — the ability to interpret statistical and numerical findings and ranges.
- Biomedical background — the ability to leverage knowledge about domain-specific lexical relationships.

Future work that incorporates such attributes might be critical towards building higher-quality scientific fact verification systems. We report progress, but there is much more work to be done.

Acknowledgements

This research was supported in part by the Canada First Research Excellence Fund, the Natural Sciences and Engineering Research Council (NSERC) of Canada, and the Waterloo–Huawei Joint Innovation Laboratory. Additionally, we would like to thank Google for computational resources in the form of Google Cloud credits.

References

Zeynep Akkalyoncu Yilmaz, Charles L. A. Clarke, and Jimmy Lin. 2020. A lightweight environment for learning experimental IR research practices. In *Proceedings of the 43rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2020)*, pages 2113–2116.

Zeynep Akkalyoncu Yilmaz, Wei Yang, Haotian Zhang, and Jimmy Lin. 2019. Cross-domain modeling of sentence-level evidence for document retrieval. 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 3481–3487.

Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, Mir Rosenberg, Xia Song, Alina Stoica, Saurabh Tiwary, and Tong Wang. 2016. MS MARCO: A human generated MAchine Reading COmprehension dataset. *arXiv:1611.09268*.

Zhuyun Dai and Jamie Callan. 2019. Context-aware sentence/passage term importance estimation for first stage retrieval. *arXiv:1910.10687*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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)*, pages 4171–4186, Minneapolis, Minnesota.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. 2020. ERASER: A benchmark to evaluate rationalized NLP models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4443–4458.

Andreas Hanselowski, Christian Stab, Claudia Schulz, Zile Li, and Iryna Gurevych. 2019. A richly annotated corpus for different tasks in automated fact-checking. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 493–503, Hong Kong, China.

Nayeon Lee, Yejin Bang, Andrea Madotto, and Pascale Fung. 2020. Misinformation has high perplexity. *arXiv:2006.04666*.

Hang Li. 2011. *Learning to Rank for Information Retrieval and Natural Language Processing*. Morgan & Claypool Publishers.

Jimmy Lin, Xueguang Ma, Sheng-Chieh Lin, Jheng-Hong Yang, Ronak Pradeep, and Rodrigo Nogueira. 2021. Pyserini: An easy-to-use Python toolkit to support replicable IR research with sparse and dense representations. *arXiv:2102.10073*.

Sean MacAvaney, Arman Cohan, and Nazli Goharian. 2020. SLEDGE: A simple yet effective baseline for COVID-19 scientific knowledge search. *arXiv:2005.02365*.

Sean MacAvaney, Andrew Yates, Arman Cohan, and Nazli Goharian. 2019. CEDR: Contextualized embeddings for document ranking. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1101–1104.

Rodrigo Nogueira, Zhiying Jiang, Ronak Pradeep, and Jimmy Lin. 2020. Document ranking with a pre-trained sequence-to-sequence model. In *Findings of EMNLP*.

Rodrigo Nogueira, Wei Yang, Jimmy Lin, and Kyunghyun Cho. 2019. Document expansion by query prediction. *arXiv:1904.08375*.

Ronak Pradeep, Rodrigo Nogueira, and Jimmy Lin. 2021. The expando-mono-duo design pattern for text ranking with pretrained sequence-to-sequence models. *arXiv:2101.05667*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Technical Report*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas.

Kirk Roberts, Tasmeer Alam, Steven Bedrick, Dina Demner-Fushman, Kyle Lo, Ian Soboroff, Ellen Voorhees, Lucy Lu Wang, and William R. Hersh. 2020. TREC-COVID: Rationale and structure of an information retrieval shared task for COVID-19. *Journal of the American Medical Informatics Association*.

Stephen E. Robertson, Steve Walker, Susan Jones, Micheline Hancock-Beaulieu, and Mike Gatford. 1994. Okapi at TREC-3. In *Proceedings of the 3rd Text REtrieval Conference (TREC-3)*, pages 109–126.

Amir Soleimani, Christof Monz, and Marcel Worring. 2019. BERT for Evidence Retrieval and Claim Verification. *arXiv:1910.02655*.

Raphael Tang, Rodrigo Nogueira, Edwin Zhang, Nikhil Gupta, Phuong Cam, Kyunghyun Cho, and Jimmy Lin. 2020. Rapidly bootstrapping a question answering dataset for COVID-19. *arXiv:2004.11339*.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: A large-scale dataset for Fact Extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7534–7550.

William Yang Wang. 2017. "Liar, Liar Pants on Fire": A new benchmark dataset for Fake News Detection. *arXiv:1705.00648*.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana.

Peilin Yang, Hui Fang, and Jimmy Lin. 2017. Anserini: Enabling the use of Lucene for information retrieval research. In *Proceedings of the 40th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2017)*, pages 1253–1256, Tokyo, Japan.

Peilin Yang, Hui Fang, and Jimmy Lin. 2018. Anserini: Reproducible ranking baselines using Lucene. *Journal of Data and Information Quality*, 10(4):Article 16.

Edwin Zhang, Nikhil Gupta, Raphael Tang, Xiao Han, Ronak Pradeep, Kuang Lu, Yue Zhang, Rodrigo Nogueira, Kyunghyun Cho, Hui Fang, and Jimmy Lin. 2020. Covidex: Neural ranking models and keyword search infrastructure for the COVID-19 Open Research Dataset. *arXiv:2007.07846*.