Fact Checking with Insufficient Evidence

Pepa Atanasova Jakob Grue Simonsen Christina Lioma Isabelle Augenstein
Department of Computer Science, University of Copenhagen, Denmark
{pepa, simonsen, c.lioma, augenstein}@di.ku.dk

Abstract

Automating the fact checking (FC) process relies on information obtained from external sources. In this work, we posit that it is crucial for FC models to make veracity predictions only when there is sufficient evidence and otherwise indicate when it is not enough. To this end, we are the first to study what information FC models consider sufficient by introducing a novel task and advancing it with three main contributions. First, we conduct an in-depth empirical analysis of the task with a new fluency-preserving method for omitting information from the evidence at the constituent and sentence level. We identify when models consider the remaining evidence (in)sufficient for FC, based on three trained models with different Transformer architectures and three FC datasets. Second, we ask annotators whether the omitted evidence was important for FC, resulting in a novel diagnostic dataset, *SufficientFacts*[1], for FC with omitted evidence. We find that models are least successful in detecting missing evidence when adverbial modifiers are omitted (21% accuracy), whereas it is easiest for omitted date modifiers (63% accuracy). Finally, we propose a novel data augmentation strategy for contrastive self-learning of missing evidence by employing the proposed omission method combined with tri-training. It improves performance for Evidence Sufficiency Prediction by up to 17.8 F_1 score, which in turn improves FC performance by up to 2.6 F_1 score.

1 Introduction

Computational fact checking approaches typically use deep learning models to predict the veracity of a claim given background knowledge (Thorne et al., 2018; Leippold and Diggelmann, 2020; Augenstein, 2021). However, the necessary evidence is not always available, either due to incomplete knowledge sources, or because the claim has newly emerged and the relevant facts are not documented yet. In such cases, FC models should indicate that the information available is insufficient to predict the label, as opposed to making a prediction informed by spurious correlations.

Prior work shows that FC models can sometimes predict the correct veracity based on just the claim, ignoring the evidence, and that they can overly rely on features such as the word overlap between the evidence and the claim (Schuster et al., 2019, 2021), leading to biased predictions. However, there are no previous studies on what evidence a FC model considers to be enough for predicting a veracity label. To this end, this work introduces the **novel task of Evidence Sufficiency Prediction illustrated in Figure 1, which we define as the task of identifying what information is sufficient for making a veracity prediction.** This task is related to FC and can operate on instances and models from FC datasets, but is focused on evaluating the capability of models to detect missing important information in the provided evidence for a claim. The latter is usually not evaluated explicitly in current FC benchmarks, where joint scores disregard a FC model's prediction when insufficient evidence is retrieved.

We study the new task by, first, conducting a thorough empirical analysis of what models consider to be sufficient evidence for FC. Second, we collect human annotations for the latter, which results in a novel diagnostic dataset, *SufficientFacts*, for FC with omitted evidence. Finally, we employ the method introduced for the empirical analysis to improve the performance of models on the new task of Evidence Sufficiency Prediction, and show that considering it a component task of FC significantly improves FC performance. For the **empirical analysis**, we propose a new fluency-preserving method that occludes portions of evidence, automatically removing constituents

[1] We make the *SufficientFacts* dataset and the code for the experiments publicly available both on https://huggingface.co/datasets/copenlu/sufficient_facts and https://github.com/copenlu/sufficient_facts.

Transactions of the Association for Computational Linguistics, vol. 10, pp. 746–763, 2022. https://doi.org/10.1162/tacl_a_00486
Action Editor: Mohit Bansal. Submission batch: 12/2021; Revision batch: 3/2022; Published 8/2022.

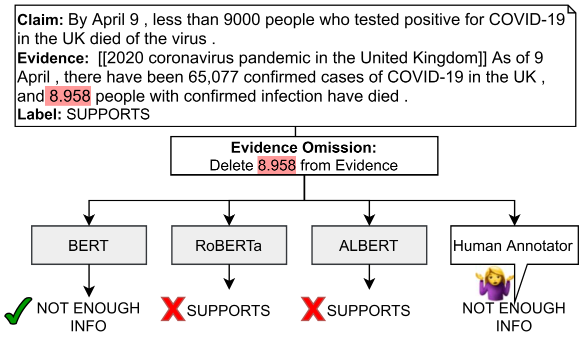

Claim: By April 9 , less than 9000 people who tested positive for COVID-19 in the UK died of the virus .
Evidence: [[2020 coronavirus pandemic in the United Kingdom]] As of 9 April , there have been 65,077 confirmed cases of COVID-19 in the UK , and 8.958 people with confirmed infection have died .
Label: SUPPORTS
Evidence Omission:
Delete 8.958 from Evidence
BERT
RoBERTa
ALBERT
Human Annotator
✓ NOT ENOUGH INFO
✗ SUPPORTS
✗ SUPPORTS
NOT ENOUGH INFO

Figure 1: An example from the VitaminC test set, where the number modifier has been omitted from the evidence. This results in there not being enough evidence for predicting its support for the claim as judged by human annotators, while two of the models still find the remaining evidence to be sufficient.

or entire sentences, to create incomplete evidence. We provide those as input to an ensemble of Transformer-based FC models to obtain instances on which FC models agree vs. disagree to have (in)sufficient information. We perform extensive experiments with three models—BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), and ALBERT (Lan et al., 2020)—and three textual FC datasets with different types of claims—FEVER (Thorne et al., 2018), HoVer (Jiang et al., 2020), and VitaminC (Schuster et al., 2021).

To compare model behavior with human rationales for Evidence Sufficiency Prediction, we ask annotators to indicate if the occluded evidence texts still provide enough information for a fact-check. This results in a **novel diagnostic test dataset, *SufficientFacts***, which contains information about the type of the omitted information, allowing for in-depth analyses of model behavior.

Finally, to improve model performance for detecting omitted important evidence and, in turn, FC, we propose to combine the proposed evidence omission method with tri-training (Zhou and Li, 2005), which utilizes the agreement of three different machine learning models to label unlabeled training instances (§5). This results in a **novel counterfactual data augmentation schema for learning of (in)sufficient information**. We find that the proposed approach is highly effective in improving model performance by up to 17.8 F_1 score on the newly introduced *SufficientFacts*. This also leads to improvements of up to 2.6 F_1 score on the standard FC test sets for the corresponding datasets.

2 Related Work

Here, we study when models trained on existing FC datasets find evidence with omitted important information to still be sufficient for veracity prediction. Such cases might be considered vulnerabilities of the models and can be due to models' faulty reasoning, learned biases, etc. Hence, our work is mainly related to studies exploring potential biases learned by FC models and the vulnerabilities of FC models to adversarial attacks. We further propose a method for evidence omission, which creates counterfactual instances, which is related to studies on input-level instance re-writing. We also use the proposed evidence omission method to collect counterfactually augmented data (CAD) and compare that to using the collected data in a contrastive learning (CL) loss to improve performance on Evidence Sufficiency Prediction and FC more generally. We thus discuss the relationship between our work and prior studies on CAD and CL. Finally, we compare our work based on deep learning models to FC performed against knowledge bases (KBs), where fact triples can also be missing.

Fact Checking Diagnostics. Previous work has exposed various biases of FC models. Although FEVER (Thorne et al., 2018) is one of the largest datasets for FC, Schuster et al. (2019) points out that models trained on it can verify a claim solely based on the text of the claim, without considering the evidence. To this end, Schuster et al. (2019) introduce a new diagnostic dataset, FeverSymmetric, of contrastively re-written claims and evidence. They show that the models fail to detect the contrastive changes in the text, leading to a drop of up to 57.46 F_1-score, compared with 85.85 F_1-score on the original FEVER development set. Furthermore, the claims in FEVER were manually written based on Wikipedia article sentences, and thus have a large token overlap between the evidence and the claim, especially for supporting evidence. Hence, Schuster et al. (2021) construct a new FC dataset, VitaminC, where they instruct the annotators to avoid using the same words as in the evidence. Ostrowski et al. (2021) further create PolitiHop—a dataset for claim verification of naturally occurring claims with evidence composed of multiple hops over interconnected evidence chunks. They study how multi-hop vs. single inference architectures reason over the evidence sets in PolitiHop. In addition, several papers (Thorne et al., 2019; Niewinski et al., 2019; Hidey et al., 2020) explored the vulnerability of FC models to adversarial attacks, for example, by discovering universal trigger words that fool a model into wrongly changing its prediction (Atanasova et al., 2020). In contrast, we are interested in how much evidence is enough for veracity prediction, studying this with three different FC models trained on three different datasets by omitting information at the constituent and sentence levels and comparing it to human judgments.

Instance Re-Writing. The above studies mainly perform re-writing or insertion operations for FC evidence. Here, we employ causal interventions on the evidence by omission to study when information is (in)sufficient for a model's prediction. Elazar et al. (2021) also use causal interventions that estimate the importance of a property by removing it from a representation. By comparison, even though text-level causal interventions are more intricate due to the discrete nature of text, we perform them on the text itself, by following linguistic rules for optional constituents to preserve the semantics and the fluency of the text. Thorne and Vlachos (2021) perform re-writing of claims by masking and then correcting separate words. They thus generate claims supported by the evidence, particularly for claims not supported before the factual correction. In a similar vein, Wright et al. (2022) decompose long, scientific claims into shorter, atomic claims. They then generate negative instances for those by masking single words in claims and replacing them with antonyms retrieved from a scientific knowledge base. In contrast, we perform omissions of evidence information at the sentence and constituent levels and for the new task of Evidence Sufficiency Prediction.

Contrastive Learning (CL) and Counterfactual Data Augmentation (CAD). Most existing work of CL in NLP employs contrastive self-learning for model pre-training (Rethmeier and Augenstein, 2021). Contrary to this, Rethmeier and Augenstein (2022) propose for CL to be performed jointly with the supervised objective. We follow the latter to improve the performance of FC models in detecting when important information is missing from the evidence, by using the original evidence texts paired with evidence texts with omitted information as contrastive data points. We perform contrastive self-training jointly with the supervised objective, as we use the contrastive loss as an unsupervised training for Evidence Sufficiency Prediction. In contrast, using it for pre-training followed by supervised training could lead to the models forgetting the information learned during pre-training, which is needed to improve the performance on *SufficientFacts*. An important factor for CL is the augmentation of negative and positive instances, which can be challenging due to the discrete nature of text. Related work explores augmentation through back-translation (Sennrich et al., 2016), masked word substitution with an LM (Wu et al., 2019), graph neighborhood sampling (Ostendorff et al., 2022), mix-up (Chen et al., 2020), or a combination thereof (Qu et al., 2021). In a similar vein, automated approaches for CAD in NLP include paraphrasing (Iyyer et al., 2018) and controlled (Madaan et al., 2021) text generation, which do not necessarily change the target label of an instance. CAD is found to improve model robustness to data artifacts (Kaushik et al., 2020; Teney et al., 2020) and to perform better out of domain (Samory et al., 2021). In contrast, we use evidence omission combined with tri-training for contrastive negative evidence mining (§5).

Knowledge-Base Fact Checking. A relevant line of work conducts FC against KBs by finding fact triple chains that are (in)consistent with the claim (Kim and Choi, 2021). Discovering such missing triples could also be used to detect insufficient evidence information. As KBs can contain an incomplete set of fact triples, related work completes KBs from unstructured textual data on the Web (Distiawan et al., 2019) or with graph embedding techniques (Kim et al., 2018). This work uses machine learning models that use textual evidence as input instead of performing an intermediate step of completing a knowledge base with needed fact triples.

3 Datasets

We employ three fact checking datasets (see Table 1) and use the gold evidence documents, that is, we do not perform document or sentence retrieval (apart from for the ablation experiment in Section 6.4). Thus, we avoid potential enforced biases for the veracity prediction models if they had to learn to predict the correct support of the evidence for the claim given wrong evidence sentences. Hence, each of the three fact

Dataset/Size	Example
FEVER 145,449 train 999,999 dev 999,999 test	**Label**: REFUTES (∈ {SUPPORTS, REFUTES, NOT ENOUGH INFO}) **Claim**: Sindh borders Indian states and is in India. **Evidence**: [Sindh] Sindh is home to a large portion of Pakistan's industrial sector and contains two of Pakistan's commercial seaports – Port Bin Qasim and the Karachi Port.
Vitamin C 370,653 train 63,054 dev 55,197 test	**Label**: SUPPORTS (∈ {SUPPORTS, REFUTES, NOT ENOUGH INFO}) **Claim**: Westlife sold more than 1 m. video albums and made over 23.5 m. sales in the UK. **Evidence**: [Westlife] According to the British Phonographic Industry (BPI), Westlife has been certified for 13 m. albums, 1.3 m. video albums, and 9.8 m. singles, with a total of more than 24 m. combined sales in the UK.
HoVer 18,171 train 1818 dev 4,000 test	**Label**: NOT SUPPORTED (∈ {SUPPORTS, NOT SUPPORTS=(REFUTES+NOT ENOUGH INFO)} **Claim**: Reason Is Treason is the second single release from a British rock band that are not from England. The band known for the early 90's album Novelty are not from England either. **Evidence**: [Kasabian] Kasabian are an English rock band formed in Leicester in 1997. [Jawbox] Jawbox was an American alternative rock band from Washington, D.C., United States. [Reason Is Treason] ''Reason Is Treason'' is the second single release from British rock band Kasabian. [Novelty (album)] Novelty is an album from the early 90's by Jawbox.

Table 1: Sizes and examples instances for the studied fact checking datasets (see §3).

checking datasets $D = \{(x_i, y_i)|x_i = (c_i, e_i), i \in [1, |D|]\}$ consists of instances with input x_i and veracity labels y_i. The input comprises a claim c_i and gold evidence e_i. The veracity label $y_i \in$ {0=SUPPORTS, 1=REFUTES, 2=NEI} for FEVER and VitamiC, and $y_i \in$ {0=SUPPORTING, 1= NOT SUPPORTING} for HoVer.

FEVER (Thorne et al., 2018) contains claim-evidence pairs, where the evidence consists of sentences from Wikipedia pages, and the claims are written manually based on the content of those Wikipedia pages. Note that 87% of the claims have evidence consisting of one sentence. The dataset has a high ratio of token overlap between the claim and the evidence, where the overlap is naturally higher for claims that are supporting (69%), than refuting (59%) and NEI (54%) claims. The high overlap ratio can create a bias for learning from token overlap, which can further prevent generalisation, as also noted in related work (Schuster et al., 2021).

Vitamin C (Schuster et al., 2021) is a collection of sentences from Wikipedia containing factual edits. For each factual edit, annotators construct a claim that is SUPPORTED and one that is REFUTED with the old and the new version of the evidence. When the factual edit introduces/removes facts from the evidence, claims are constructed so that there is NOT ENOUGH INFORMATION (NEI) to support them. Due to its contrastive nature and reduced claim-evidence overlap, the authors demonstrate that models trained on the dataset gain a 10% accuracy improvement on adversarial fact verification.

HoVer (Jiang et al., 2020) is designed to collect claims that need several hops over Wikipedia evidence sentences to verify a claim. The evidence contains between two and four sentences from different Wikipedia articles. As the test dataset is blind and we use the gold evidence, we use the development set for testing purposes and randomly select 10% of the training dataset for development.

4 Evidence Omission

To study what types of information the evidence models consider important, we propose to conduct causal interventions for the evidence by omitting information from it. We hypothesize that removing information important for the model to predict the support of evidence for a claim will cause a change in its original prediction, leading to the model indicating that there is missing information. If the removed information is not important for the model though, removing it would not change the model's prediction. We then ask whether the information that is important for a model when predicting the support of the evidence text for a claim, is actually important as judged by human annotators. The human annotations allow for a systematic study of common model errors, that is, when the models still predict the correct label even if important evidence information has been

Type	L	Claim	Evidence
S	R	The Endless River is an album by a band formed in 1967.	[[The Endless River]] The Endless River is a studio album by Pink Floyd. [[Pink Floyd]] Pink Floyd were founded in 1965 by students . . .
PP	R	Uranium-235 was discovered by Arthur Jeffrey Dempster in 2005.	[[Uranium-235]] It was discovered in 1935 by Arthur Jeffrey Dempster.
NOUNM	S	Vedam is a drama film.	[[Vedam (film)]] Vedam is a 2010 Indian drama film written and directed by Radhakrishna Jagarlamudi . . .
ADJM	S	Christa McAuliffe taught social studies.	[[Christa McAuliffe]] She took a teaching position as a social studies teacher at Concord High School. . .
ADVM	S	Richard Rutowski heavily revised the screenplay for Natural Born Killers.	[[Natural Born Killers]] The film is based on an original screenplay that was heavily revised by writer David Veloz, associate producer Richard Rutowski . . .
NUMM	S	Being sentenced to federal prison is something that happened to Efraim Diveroli.	[[Efraim Diveroli]] Diveroli was sentenced to four years in federal prison .
DATEM	R	Colombiana was released 1st October 2001.	[[Colombiana]] Colombiana is a French action film from 1st October 2011 . . .
SBAR	R	North Vietnam existed from 1945 to 1978.	[[North Vietnam]] North Vietnam, was a state in Southeast Asia which existed from 1945 to 1976.

Table 2: Examples from the FEVER dataset of constituent types (§4.1) removed from the evidence for a claim with Label (L) one of SUPPORTS (S) or REFUTES (R).

removed and when they consider the information to be insufficient if unrelated evidence has been removed.

4.1 Evidence Omission Generation

We omit information from the evidence text at the sentence and constituent level. Particularly, we aim to remove information from the evidence such that it does not change its stance towards the claim from SUPPORTS to REFUTES, or vice-versa, while preserving the grammatical correctness and fluency of the evidence. Following studies of linguistic sentence structure (Burton-Roberts, 2016; Börjars and Burridge, 2019), illustrated with examples in Table 2, we collect prepositional phrases, modifiers, and other optional sentence constructs—that is, those constructs that can be removed from the sentence without impairing its grammatical correctness, and where the remaining text is semantically identical to the original one, except for the additional information from the removed construct (Garvin, 1958). We use the following optional sentence constructs:

Sentences (S). In FEVER and HoVer, the evidence can consist of more than one sentence. The separate sentences are supposed to contain information important for the fact check, which we further verify with manual annotations as explained in Section 4.2. VitaminC consists of single sentences only, and we thus only perform constituent-level omissions for it, as described next.

Prepositional Phrases (PP) are optional phrases that are not part of a Verb Phrase (VP), but are child nodes of the root sentence in the constituent tree (Brown et al., 1991). These usually function as adverbs of place and consist of more than one word.

Noun Modifiers (NOUNM) are optional elements of a phrase or clause structure (Huddleston and Pullum, 2005). NOUNM can be a single or a group of nouns that modify another noun.

Adjective Modifiers (ADJM) are a single or a group of adjectives that modify a noun.

Adverb Modifiers (ADVM) are a single or a group of adverbs that modify verbs, adjectives, or other adverbs and typically express manner, place, time, and so forth.

Number Modifiers (NUMM) are a single or a group of words denoting cardinality that quantify a noun phrase.

Date Modifiers (DATEM) are a single or a group of words that express temporal reference. To preserve fluency, from a date expression consisting of a day, a month, and a year, we omit either the date, the date and the month, or the year.

Subordinate Clauses (SBAR) are introduced by a subordinating conjunction. Subordinate clauses depend on the main clause and complement its meaning. SBARs can be adverb clauses, adjective clauses, and noun clauses.

For the omission process, we use two pre-trained models with high performance from the Spacy library[2]: a part-of-speech (PoS) tagger with an accuracy of 97.2 and a constituency parser (Kitaev and Klein, 2018) with an F_1-score of 96.3 on the revised WSJ test set (Bies et al., 2015). During the omission process, we use the PoS tags to find nouns, adjectives, adverbs, and numbers and use the constituency tags to select only the modifiers. Thus, we find the NOUNM, ADJM, ADVM, and NUMM constructs. We collect SBAR and PP constructs by finding their corresponding tags in the constituent dependency tree. Finally, for the date, we use two regular expressions that are common date templates used in Wikipedia articles (<month name, date, year> or <date, month name, year>) and remove parts from the templates that preserve the coherency (<date>, <year>, <month name and date>, or <year and date>).

Overall, in this work, we perform a study of insufficient evidence for FC by removing information from the gold evidence. As explained in Section 2, we perform causal interventions on the evidence by omission to study when information is (in)sufficient for a model's prediction. Replacement of words is another operation that can be applied to the evidence. We can, for example, replace different types of named entities with pronouns, and different parts of the speech with demonstrative pronouns to induce insufficient information. However, the replacement operation does not allow for direct causal conclusions as any change of a word with another could potentially lead to confounding factors of the newly introduced word and the model's predictions. Note that there are some pronouns used in the evidence when they refer to the person/object of the article. We do not treat such cases as insufficient information as the title of the page with the name of the person/object is always prepended to the sentence, which allows for coreference resolution. Finally, another possible operation is the insertion of new information, which would lead to insufficient evidence when performed on the claim. The latter, however, requires the insertion of text that preserves the grammatical correctness and meaning of the claim, which is hard to achieve in an automated way.

4.2 Manual Annotations

Models. We train three Transformer-based FC models: BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), and ALBERT (Lan et al., 2020). BERT is pre-trained with masked language modeling and next sentence prediction objectives on the Toronto Book Corpus (Kiros et al., 2015) and the English Wikipedia.[3] It is also the most widely used pre-trained Transformer model.[4] RoBERTa improves upon BERT by optimizing key hyper-parameters, and is trained without the next sentence prediction objective. RoBERTa is one of the top-performing models on the GLUE (Wang et al., 2018) and SuperGLUE (Wang et al., 2019) benchmarks composed of various NLP tasks. The latter also holds for ALBERT, another Transformer architecture that improves upon BERT. It does so with parameter-reduction techniques, which lower the memory consumption of the model. ALBERT also employs a self-supervised pre-training loss for inter-sentence coherence. The latter is found to be beneficial for tasks with multiple sentences, and Schuster et al. (2021) report improved FC robustness with it on VitaminC compared to BERT.

We train each model on the respective training splits of each dataset with the claim c and the gold evidence e as input to predict the gold veracity label y: $f(c, x) = \hat{y}$. We optimize the supervised cross-entropy loss:

$$\mathcal{L}^S = -\frac{1}{m}\sum_{j=1}^{m} y^j \cdot \log(\hat{y}^j) \qquad (1)$$

where m is the label space size.

We then use an ensemble of these three different Transformer-based FC models to collect predictions for our new task Evidence Sufficiency Prediction, as we want to find instances with omitted information that are more broadly applicable (e.g., those on which the models agree). The (dis)-agreements between the models also allow us to study the differences between them in detecting omitted information. Transformer Language Models are pre-trained on large datasets, the veracity

[2] https://spacy.io/.

[3] https://en.wikipedia.org.

[4] https://huggingface.co/models.

of which can change over time (Schuster et al., 2021). This makes it important that the FC models take into account the facts in the given evidence. When provided with differences and similarities in the three FC models' predictions, future work could then also investigate the degree to which different Transformer-based FC models encode FC-relevant world knowledge they default to in their predictions.

Annotation Task. Next, we collect evidence with removed information as described above. We then use the models to find which of the omitted evidence they consider important, resulting in a prediction change to NEI. We consider instances from the original test splits of each of the datasets, where all models predicted the veracity correctly before the evidence omission was performed, as these are the cases where we can observe whether evidence omission causes the veracity prediction to change to NEI. We collect instances with omitted evidence information where the models: (1) agree that the evidence is still enough vs. (2) insufficient; and where they (3) disagree in their prediction. We collect a total of 400 instances at the sentence, and 600 instances at the constituent, level from the test splits of the corresponding datasets, distributed equally among the above three groups.

We employ annotators on Amazon Mechanical Turk.[5] We first train potential annotators, presenting them with annotation guidelines and illustrative examples. We then select annotators using a qualification test with nine test annotations for our task. Each annotation had the cost of \$0.10, and annotators were paid \$10 on average per hour. The annotation task is to determine whether the evidence is still sufficient for predicting the label without the omitted information. If the remaining evidence is still sufficient, we ask them for the reason—whether this is because the removed evidence is repeated in the remaining text or because the removed evidence is not relevant to the veracity of the claim. Following the annotation guidelines for FEVER and HoVer, we ask the annotators not to use any world knowledge or knowledge they might have about the claim. For more details on the annotation task and the guidelines, we release the dataset with a detailed README file.

The final dataset *SufficientFacts* $= \{(x'_i, y'_i) | x'_i = (c_i, e'_i), i \in [1, |SufficientFacts|]\}$ consists of test instances x'_i with labels y'_i. All of the instances in *SufficientFacts* are a subset of the instances in the test datasets of FEVER, VitaminC, and HoVer with the following changes. The input x'_i comprises the original claim c_i and the evidence with omitted information e'_i. The tokens of e'_i are a subset of the tokens of the original gold evidence e_i of the instance. To re-iterate, the label of the originally selected instances is either SUPPORTS or REFUTES, that is, they have sufficient gold evidence information, where after omitting information from the evidence, the new label y'_i becomes either NEI if the majority of the annotators selected that important information was removed, and otherwise remains the original label – SUPPORTS and REFUTES for FEVER and VitamiC, or SUPPORTING for HoVer.

The resulting inter-annotator agreement (IAA) for *SufficientFacts* is 0.81 Fleiss' κ from three annotators. Due to the novelty of the introduced task of Evidence Sufficiency Prediction, we do not have direct points of comparison for IAA. However, we point as a reference the IAA reported for the related task of fact checking for the HoVer dataset (0.63 Fleiss' κ), and for the FEVER dataset (0.68 Fleiss' κ), where, for both datasets, the annotators were thoroughly trained and highly paid. The biggest challenges for our annotators, judging by their errors during the qualification test, were not to use common knowledge and assumptions in their annotations, and the general complexity of the task.

4.3 *SufficientFacts* Analysis

Overall Agreement with Annotators. The statistics of the resulting dataset, *SufficientFacts*, are presented in Table 3. We find that all three models agree that the remaining evidence is still sufficient (EI Agree) even when it has become insufficient after omitting information needed for verifying the claim (NEI) in 430 out of 1000 instances. We assume that these failures of all three models to detect missing information for FC point to the models making predictions based only on patterns observed in claims, or to the models defaulting to world knowledge encoded in the pre-trained Transformer models. We further find that when the models disagree about whether the remaining information is still sufficient (Disagree), they

[5] https://www.mturk.com/.

Dataset	Model Pred	EI_I	EI_R	NEI
FEVER SENT	EI Agree	61	20	119
	NEI Agree	13	9	<u>178</u>
	Disagree	39	24	137
	Total	113	53	434
FEVER CONST	EI Agree	146	3	51
	NEI Agree	0	0	200
	Disagree	43	1	156
	Total	189	4	407
HoVer SENT	EI Agree	<u>32</u>	<u>12</u>	156
	NEI Agree	4	1	195
	Disagree	7	1	192
	Total	43	14	543
HoVer CONST	EI Agree	139	6	55
	NEI Agree	1	0	199
	Disagree	48	1	151
	Total	188	7	405
VitaminC CONST	EI Agree	**146**	**5**	49
	NEI Agree	0	0	**200**
	Disagree	13	0	187
	Total	159	5	436
Total	EI Agree	524	46	430
	NEI Agree	18	10	972
	Disagree	150	27	823
	Total	692	83	2225

Table 3: Statistics of *SufficientFacts* presenting the predictions of the models in the ensemble (Model Pred: Agree Enough Information (EI Agree), Agree Not Enough Information (NEI Agree), Disagree, and Total) vs human annotations of the same (EI – Irrelevant (EI_I), EI – Repeated (EI_R), NEI). We present sentence (SENT) and constituent omission (CONST) dataset splits separately. We embolden/underline results of the datasets for predictions where the three models agree (NEI Agree, EI Agree) and have the highest/lowest agreement with human annotations about EI_I, EI_R, and NEI predictions. We use light blue/dark blue to denote where lower/higher results are better.

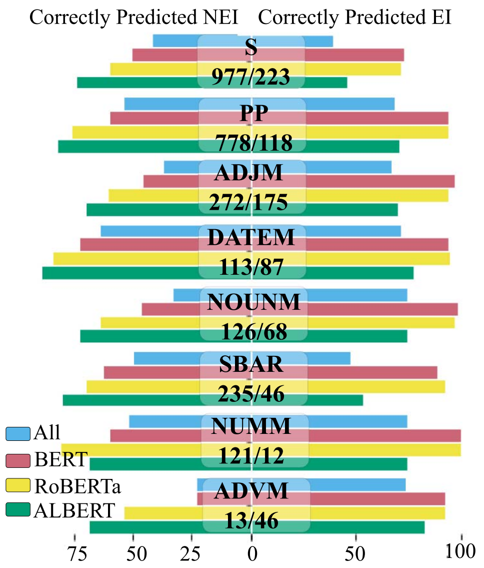
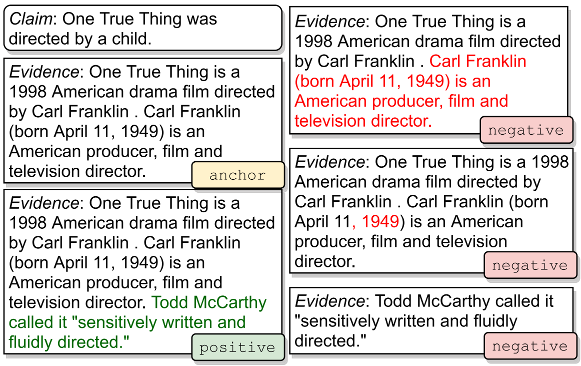


Figure 2: *SufficientFacts*: fine-grained analysis by type of removed evidence inftype 4.1) vs. proportion of correct predictions of NEI/EI instances. The proportion is computed for the separate models: BERT, RoBERTa, ALBERT, and for all three models agreeing on the correct NEI/EI label (All). The total number of NEI/EI instances of each type is provided under each of the types of removed evidence information. *A higher* proportion of correct predictions is *better*.

disagree mostly about instances where the omitted evidence information is needed for veracity prediction (NEI)—in 823 out of 1000 instances. By contrast, when the models agree that the remaining evidence is insufficient, they are correct in 972 out of 1000 of the instances.

Separate Dataset Agreement with Annotators. Looking at the separate datasets, it is the hardest for the models to identify missing evidence information needed for the fact check (EI Agree vs. NEI) for HoVer, particularly with sentence omissions, and the easiest for the VitaminC dataset with constituent omissions. We hypothesize that the latter is due to the HoVer dataset having more complex claims and requiring cross-sentence reasoning, whereas VitaminC contains contrastive instances which, during training, guide the models to identify the parts of the evidence needed for FC. Overall, the models fail to detect missing information more from sentences rather than from constituents. We hypothesize that this effect can be observed partly because models struggle to conduct multi-hop reasoning over them. Another possible reason for that is that the models could be better at verifying the type of information removed from a sentence constituent rather than from a sentence.

Performance by Omitted Evidence Type and Model. Figure 2 provides a fine-grained analysis of the performance of the models for different types of omitted constituents. We observe that it is the hardest to detect when the evidence is missing information for the prediction (Correctly Predicted NEI) that was removed from adverbial modifiers (ADVM), followed by subordinate clauses (SBAR). By contrast, it is easiest to detect missing information when it is a date modifier (DATEM), followed by number modifiers (NUMM). BERT has the lowest rate of



Claim: One True Thing was directed by a child.

Evidence: One True Thing is a 1998 American drama film directed by Carl Franklin . Carl Franklin (born April 11, 1949) is an American producer, film and television director.

anchor

Evidence: One True Thing is a 1998 American drama film directed by Carl Franklin . Carl Franklin (born April 11, 1949) is an American producer, film and television director. Todd McCarthy called it "sensitively written and fluidly directed."

positive

Evidence: One True Thing is a 1998 American drama film directed by Carl Franklin . Carl Franklin (born April 11, 1949) is an American producer, film and television director.

negative

Evidence: One True Thing is a 1998 American drama film directed by Carl Franklin . Carl Franklin (born April 11, 1949) is an American producer, film and television director.

negative

Evidence: Todd McCarthy called it "sensitively written and fluidly directed."

negative

Figure 3: Example of augmented contrastive instances for the original (anchor) instance. Red designates removed evidence information, where the models agree that the remaining evidence is not sufficient, producing a negative contrastive instance. Green designates an added distractor sentence, producing a positive instance. The distractor sentence, selected to have high overlap with the claim but with insufficient information, is used as another negative instance.

correctly detecting insufficient evidence from the three models, followed by RoBERTa, and ALBERT performs best. We conjecture that this is due to RoBERTa being an optimization of BERT, and due to ALBERT including pre-training with an inter-sentence coherence objective, which has been shown to make the model more robust for factual verification (Schuster et al., 2021). Even though ALBERT contains fewer parameters than BERT, it still detects better when the evidence is insufficient. Finally, we see a natural trade-off between correctly detecting sufficient and correctly detecting insufficient information. In particular, some models such as ALBERT have a higher number of correct predictions on instances without enough information (Figure 2, left). However, on instances with sufficient evidence information (Figure 2, right), ALBERT has the lowest number of correct predictions. In contrast, BERT has the worst performance on the NEI instances, but the best performance on EI instances.

5 Evidence Omission Detection

To improve the performance of models in recognizing when the evidence is not enough for verifying a claim, we experiment with CAD (§5.2) and a CL loss (§5.1). Both methods use contrastive data augmented with the proposed evidence omission method (§4.1) in combination with tri-training, as illustrated in Figure 3. We omit information from the original (anchor) evidence to collect potential negative instances with missing important evidence information compared to the original evidence (Figure 3, right). From the resulting candidates, we select as negative only those predicted as having insufficient information by the other two supervised models from the ensemble (§4) (e.g., RoBERTa and ALBERT predict NEI when we are training a model with a BERT Transformer architecture). We also collect positive instances that still have sufficient evidence information after applying a data augmentation operation. For each instance x_i, we find one distractor sentence from the document of the gold evidence that is the most similar to the claim by word overlap. We append the distractor sentence to the original evidence, which serves as a positive instance (Figure 3, left). Finally, we include only the distractor sentence as a negative instance as it does not have enough evidence contrasted both with the positive and the anchor instances. We conjecture that the latter would serve as a training signal for avoiding the bias for overlap between the claim and the evidence.

5.1 Contrastive Learning

We study self-supervised learning to train FC models that recognise when the evidence is not enough for verifying a claim. In particular, we propose to use self-supervised CL jointly with the supervised learning of the model to predict the support of the evidence for a claim. Given an anchor instance x_i, a positive instance x_i^+, and K^- negative instances $x_{i,k}^-$, $k \in [1, K^-]$, the objective of CL is to make the anchor and the positive instance closer in the representation space, and the anchor and the negative instances further apart. The anchor, positive, and negative instances are collected and/or augmented from the training splits of the corresponding datasets as described above. Each model, $g(x) = l(h(x)) = l(e) = \hat{y}$, uses 12 encoding layers to encode an input instance $h(x) = e$ and uses the encoding e of the last encoding layer to predict the veracity label with a linear layer: $l(e) = \hat{y}$. We encode the anchor, the positive, and the negative in stances with the corresponding model g, resulting in the anchor e_i, the positive e_i^+, and the negative $e_{i,j}^-$ representations, and minimise the following CL loss:

$$\mathcal{L}^{\text{CL}} = \log \sigma(s(e_i, e_i^+; \tau)) + \sum_{k=1}^{K^-} log\sigma(1 - s(e_i, e_{i,k}^-; \tau)) \tag{2}$$

where s is a similarity function between the representation of the two instances—cosine similarity in our case, τ is a temperature parameter subtracted from the cosine similarity (Ma and Collins, 2018), and K^- is the number of negatives. Note that the CL loss is the same as Noise Contrastive Estimation (Ma and Collins, 2018) expressed as a binary objective loss. The representation of each instance is obtained by mean pooling of the word representations of the instance in the last layer of the model M. We include the contrastive self-learning loss for those instances that are not annotated as NEI, as we cannot construct contrastive negative evidence with insufficient information for the instances that already do not have enough information for verification. Finally, the CL loss is optimised jointly with the supervised loss:

$$\mathcal{L}^S = -\frac{1}{m}\sum_{j=1}^{m} y^j \cdot \log(\hat{y}^j) \tag{3}$$

$$\mathcal{L} = \mathcal{L}^S + \mathcal{L}^{\text{CL}} \tag{4}$$

where $\hat{y}_i$ is the label prediction of model M, m the label space size, y_i is the gold label for instance x_i, $y_i \in \{$0=SUPPORTS, 1=REFUTES, 2=NEI$\}$ for FEVER and VitaminC, and $y_i \in \{$0=SUPPORTING, 1=NOT SUPPORTING$\}$ for HoVer.

5.2 Counterfactual Data Augmentation

We also experiment with counterfactually augmented evidence, using the negative and positive instances constructed as described above (§5 and Figure 3). As the models have high accuracy when they agree that a piece of evidence with omitted information is not sufficient (see agreement with human annotations in Table 3), we conjecture that the counterfactually augmented instances would serve as a good training signal for detecting (in)sufficient evidence information without incurring annotation costs for training data. The counterfactually augmented data is thus simply combined with the training instances of each dataset. In particular, we include in the training set the claim and the original evidence (anchor) with the corresponding gold label y_i. We include the positive instance—original evidence with distractor sentence appended to it, with the original gold label y_i. The negative instances, namely, those with insufficient evidence information, are included with a gold label y_i = NEI for FEVER and VitaminC, and y_i = NOT SUPPORTING for HoVer. Each model, $h(c, e) = \hat{y}$, receives as input the original claim c and the augmented or the original evidence e and predicts the veracity label $\hat{y}$. We optimize a supervised cross-entropy loss as per Equation 3.

5.3 Baseline Ensemble

We include a simple ensemble, consisting of the three models: BERT, RoBERTa, and ALBERT. Each ensemble contains only supervised models (§4.2), models trained with CAD (§5.2), or models trained with CL loss (§5.1). We employ majority voting, where the final prediction is the most common class among the predictions of the three models on an instance, defaulting to the class with the highest predicted probability if there is no most common class.

5.4 Experimental Details

All models are trained on the respective training splits of each dataset. We select the checkpoint with the highest macro F_1-score on the dev sets and provide results on the test sets. We note that for the newly introduced task Evidence Sufficiency Prediction, we have an annotated test dataset *SufficientFacts*, but no training dataset. The training is performed on the original training splits of the corresponding datasets, which have a different label distribution from the introduced diagnostic test set. Hence, it is possible that some of the instances in *SufficientFacts* are out of the original training distribution, which would make this diagnostic dataset of rather adversarial nature.

We select the learning rate $= 1e-5$ and the temperature parameters $\tau = 1.5$ by grid search over the performance on the dev sets from $[1e-5, 2e-5, 3e-5]$ and $[0, 0.5, 1, 1.5, 2]$, respectively. We use the batch sizes for corresponding models from prior work: 8 for HoVeR, 32 for FEVER, and 16 for VitaminC.

6 Results and Discussion

6.1 Supervised Model Performance

We start by discussing the performance of models trained on the supervised splits of the corresponding datasets to predict labels for claims based on the newly created dataset *SufficientFacts* for Evidence Sufficiency Prediction, presented in

Dataset	Model	Veracity Pred. / Orig.Test				Evidence Sufficiency Pred. / Suff.Facts			
		BERT	RoBERTa	ALBERT	Ens.	BERT	RoBERTa	ALBERT	Ens.
FEVER	Supervised	87.16	88.69	86.67	88.81	59.51	59.10	63.00	61.36
	+ CL	87.62	88.81	86.62	89.02	65.79	67.98	**70.83**	**69.90**
	+ CAD	**87.86**	**89.23**	**87.31**	**89.14**	**67.18**	**69.58**	68.56	69.25
HoVer	Supervised	80.75	83.37	76.88	82.73	58.15	64.81	66.28	65.88
	+ CL	81.82	83.38	77.62	83.08	74.91	75.41	72.83	78.05
	+ CAD	**81.87**	**83.65**	**79.44**	**83.65**	**74.98**	**77.14**	**76.12**	**79.07**
VitaminC	Supervised	82.26	84.98	83.38	86.01	58.51	69.07	66.57	66.76
	+ CL	83.00	**85.54**	83.48	**86.22**	62.34	72.18	68.13	70.42
	+ CAD	**83.56**	85.65	**83.82**	86.14	**72.93**	**75.79**	**75.13**	**78.60**

Table 4: Macro F_1-score test performance of models and an ensemble (Ens.) (§5.3) trained on the supervised training splits of each dataset (Supervised), and in addition with the contrastive objective (+CL) (§5.1) and the counterfactually augmented data (+CAD) (§5.2). Results are the average of three different seed runs. The highest results for a test dataset and a model are in bold, and the overall highest result of a model for a test dataset are additionally underlined.

Table 4. Recall that the instances in *SufficientFacts* had correct predictions from all models before the evidence omission was performed (§4.2), that is, the performance of the models on the instances in *SufficientFacts* had 100 F_1-score before the evidence omission. Hence, the omission of information from the evidence results in a performance decrease from 100 to 58 F_1-score (BERT model for the HoVer dataset)—a decrease of up to 42 F_1-score. Out of the three FC models, BERT has the lowest performance on *SufficientFacts*, and ALBERT has the highest. The latter corroborates that ALBERT is a more robust model for fact verification, as explained in more detail in Section 4.2.

Further, we observe the worst performance on *SufficientFacts* for the HoVer dataset—down to 58 F_1-score—followed by FEVER, and with the best performance on VitaminC. We suggest that the contrastive nature of the instances in VitaminC that contain factual edits of the evidence, changing the support of the evidence for the claim, as described in Section 3, can indeed provide a better learning signal for the models about which parts of the evidence are important for verifying the claim.

6.2 CL and Augmented Model Performance

Including a CL loss or CAD results in improvements for all models and datasets on *SufficientFacts* by up to 17.2 F_1-score. Note that the proposed technique does not incur additional annotation costs for training data for Evidence Sufficiency Prediction. This corroborates that our proposed evidence omission approach combined with tri-training improves the recognition of (in)sufficient evidence. This, in turn, improves the performance on the original test sets by up to 3.6 F_1-score. Comparing the CL loss with counterfactually augmented data, we see that CAD improves the model performance in more cases on *SufficientFacts*, except for ALBERT for the FEVER dataset. This could be because the augmented data uses raw labels obtained with tri-learning, while the CL loss only drives apart the negative instances from the anchor in the representation space.

Finally, we compare the performance of CAD and CL loss that rely on the agreement predictions of the supervised models with the simple majority voting ensembles (§5.3). Single models trained with CAD and CL loss still outperform the ensembles of the supervised models. A majority voting classifier from the models trained with CAD and CL loss improves the performance on the original and *SufficientFacts* sets even further.

6.3 Comparison to Related Work

We further compare the performance of our models to existing systems on the used datasets (see Table 5). Note that we are particularly interested in veracity prediction to study what evidence models consider as sufficient for factuality prediction.

Dataset	Model	F_1
FEVER	DA *(Thorne et al., 2018)*	83.84
	RoBERTa Supervised	88.69
	+ CL	88.68
	+ Augmented	**89.23**
HoVer	BERT *(Jiang et al., 2020)*	*81.20*
	BERT Supervised	80.75
	+ CL	81.82
	+ Augmented	**81.87**
VitaminC	ALBERT *(Schuster et al., 2021)*	82.76
	ALBERT Supervised	83.38
	+ CL	83.48
	+ Augmented	**83.82**

Table 5: Macro F_1-score on the original test set compared to baseline (FEVER) and SOTA (HoVer, VitaminC) oracle results. Highest results for a dataset are in bold.

Thus, in the base setting, we do not conduct evidence retrieval, as typically performed for the HoVer and FEVER datasets, but train models using gold evidence (oracle). For FEVER, existing systems report results on both tasks, hence we can only compare to the veracity prediction results with oracle evidence available in the FEVER dataset paper with a Decomposable Attention (DA) model (Parikh et al., 2016). For HoVer and VitaminC, the presented results are also from the dataset papers of models trained with oracle evidence. As there are no other reported results on these datasets, they also represent the state-of-the-art for these two datasets. To compare to them, we pick those of our models with the same Transformer architecture as used in the respective dataset papers, and the best-performing model architecture for FEVER. Note that we use the same training setting as in related work (§5.4) for all models and datasets. We find that our supervised models are close in performance to prior reported results. Furthermore, including counterfactual data augmentation and contrastive learning leads to improvements over prior results for all three datasets, by up to 2.6 F_1-score.

Model	BERT	RoBERTa	ALBERT	Ens.
FEVER				
Supervised	82.18	81.88	85.03	84.24
+ CL	87.63	93.53	**95.18**	**91.60**
+ CAD	**89.50**	**94.73**	90.89	90.95
HoVer				
Supervised	97.27	78.64	97.65	88.57
+ CL	99.58	**99.71**	**99.45**	**99.98**
+ CAD	**99.65**	98.52	99.30	99.97
VitaminC				
Supervised	69.99	80.36	**80.69**	78.33
+ CL	75.77	79.32	78.95	78.90
+ CAD	**80.71**	**82.69**	75.69	**80.78**

Table 6: Accuracy of models trained on the supervised training splits of each dataset (Supervised), the contrastive objective in addition to training with Supervised (+CL), and the counterfactually augmented data (+CAD). The models are evaluated on the task of Evidence Sufficiency Prediction on datasets with extracted unrelated evidence information (§6.4).

6.4 Incorrect Evidence

So far, we studied model performance on instances with omitted information from the gold evidence. We now probe how well the models detect missing information given retrieved incorrect evidence, which does not contain sufficient information. The latter is possible in real-world scenarios. The evidence we feed to the fact checking model depends on the preceding evidence retrieval step, which can retrieve gold evidence with varying performance. While the fact checking model is possibly trained on gold evidence to avoid learning spurious correlations, we want to evaluate its capability to recognize when the retrieval system has discovered incorrect evidence as well. Note that current FC benchmarks do not consider the prediction of a veracity model if the correct evidence is not retrieved. However, in realistic situations, we do not know whether the evidence is correct, and FC models would still provide a veracity for a claim. Hence, we further study the performance of models on incorrect evidence. For each instance in the original test splits, we retrieve incorrect evidence by selecting the closest evidence of another claim in the dataset by word overlap between the claim and the evidence candidates. We then use the retrieved instead of the original evidence. This results in a test set of claims with incorrect evidence of the same size as the original test split.

Table 6 reports results on the test datasets incorrect evidence. As all instances in the dataset have the new gold label of NEI, we report accuracy, which corresponds to the ratio of the instances with a predicted NEI label. We find that the performance of the models is improved by as much as 27 accuracy points after training with CAD or CL, which is another indication for the effectiveness of the proposed training methods. We also find that CAD again brings larger performance gains than CL, except for HoVer, where the two approaches achieve very similar accuracy scores.

The extended evaluation of incorrect evidence is an important complement to the study of missing evidence. However, the two are not necessarily directly comparable. First, in Table 4, the two test datasets—the Original Test and SufficientFacts—both have instances with and without sufficient evidence. The extended study on incorrect evidence in this section only has instances that do not have sufficient evidence. This also results in our use of different measures to report results: accuracy in Table 6, which is the percentage of detected incorrectly retrieved evidence, and macro F1-score in Table 4, which combines the performance on up to three classes in a balanced way.

However, it is worth addressing the high performance of the models on the irrelevant evidence dataset. We employ evidence that has word overlap with the claim, but is not necessarily semantically similar to the claim. If the models were to only rely on features of the claim or on surface word overlap between the claim and the evidence, the models would have low performance on the irrelevant evidence dataset. We train models to avoid such spurious correlations with CAD and CL loss, which make discovering missing evidence information in irrelevant evidence easy, leading to the observed high performance in Table 6.

6.5 Error Analysis

Lastly, we conduct an error analysis on the newly introduced *SufficientFacts* to understand whether known biases in models trained on FC datasets (§2) also affect predictions on *SufficientFacts*.

Claim-Only Prediction. Schuster et al. (2019) found that FC models often learn spurious correlations and can predict the correct label even when no evidence is provided, as they learn only features of the claim. We investigate whether it is also among the reasons for incorrect predictions of the models on the *SufficientFacts* dataset. We compute the percentage of instances in *SufficientFacts* where the models do not predict when provided with evidence. We find that for the HoVer dataset, the supervised BERT model does not predict an NEI label for 36% of the instances in *SufficientFacts*, whereas the respective number for RoBERTa is 23% and 14% for ALBERT. This indicates that supervised models trained on HoVer learn claim-only features for some instances. After training the models with CAD (§5.2) and CL loss (§5.1), fewer than 1% of instances from *SufficientFacts* are predicted as having enough information by each of thee models when given only the claim. This indicates that training with CAD and CL loss decreases the claim-only bias for the HoVer dataset. For FEVER and VitaminC, we find a lower percentage of instances (fewer than 4%) in the corresponding *SufficientFacts* splits that the supervised models predict as having enough information when given only the claim. We hypothesises that this is due to the larger amount of training data in both datasets and due to the contrastive nature of VitaminC, which requires the models to learn features from the evidence as well. The percentage is again decreased after training with CAD and CL (fewer than 1%). Finally, we find that the instances that are still not detected as having insufficient evidence after training with CAD/CL loss are those that the model could have gained world knowledge about during pre-training. One example of such a claim is given in Table 7, row 3.

1. *Claim:* Unison (Celine Dion album) was originally released by Atlantic Records.
Evidence: [Unison (Celine Dion album)] The album was originally released on 2 April 1990.
Dataset: FEVER, *Model:* BERT *Gold:* NEI, *Sup.:* SUPPORTS, *+CAD:* NEI, *+CL:* NEI

2. *Claim:* Jean-Jacques Dessalines was born on October 2nd, 2017.
Evidence: [Jean-Jacques Dessalines] He defeated a French army at the Battle of Vertiéres.
Dataset: FEVER, *Model:* RoBERTa, *Gold:* NEI, *Sup.:* SUPPORTS, *+CAD:* NEI, *+CL:* SUPPORTS

3. *Claim:* The Times is a website. *Evidence:* N/A
Dataset: FEVER, *Model:* RoBERTa, *Gold:* NEI, *Sup.:*REFUTES, *+CAD:* REFUTES, *+CL:* REFUTES

4. *Claim:* The Bragg–Gray cavity theory was developed by Louis Harold Gray, William Lawrence Bragg, and a man knighted in the year 1920.
Evidence: [William Henry Bragg] He was knighted in 1920.
Dataset: HoVer, Model: RoBERTa, Gold: NEI, supervised: SUPPORTS, +CAD: SUPPORTS, +CL: SUPPORTS

Table 7: Example model predictions before (Sup.) and after including CAD/CL loss training.

Claim-Evidence Overlap. Schuster et al. (2021) also find that FC models are biased in predicting the SUPPORT class when the overlap between the claim and the evidence is high. We conjecture that this is another possible reason that the instances in *SufficientFacts* are hard for the models to distinguish as having missing important evidence information, as their evidence still has a high overlap with the claim. To probe this, we compute the average overlap between the claim and the evidence, disregarding stop words, of instances in the *SufficientFacts* that are predicted as having insufficient information by the supervised models and by the models trained with CAD and CL loss. For FEVER and HoVer, the instances predicted as NEI by the supervised models have low overlap with the claim that increases after training with CAD and CL loss (61% to 68% for HoVer and 63% to 65% for FEVER). An example instance where the evidence has high overlap with the claim and is predicted as NEI only after training with CAD and CL loss can be found in Table 7, row 1. The latter is an indication that training with CAD and CL loss also reduces the overlap bias of FC models. We do not observe a change in the overlap ratio for VitaminC, where we assume that training with contrastive instances already prevents learning biases, including the overlap bias.

Spurious Patterns. Finally, we investigate whether the models learn other spurious patterns that could lead to low results on *SufficientFacts*. We already observed that for some instances, the supervised models predict that the evidence is not sufficient after removing irrelevant information (Table 3), which is one indication of learned spurious patterns. Further, when removing important information, the supervised models still predict the same label for some instances, as they rely on other parts of the input, which might not be important. Table 7 shows one example where the supervised models did not recognise that the evidence is missing important information (row 1), but after training with CAD or CL loss, it was detected as NEI. However, there are still possible spurious correlations that the models learn even after training with CAD or CL loss, for example, the example in row 4. Another such example is in row 3, where even after training with CAD and CL loss, the models still find the claim without any provided evidence sufficient for predicting a refuted claim. As this example relies on knowledge of common facts, we assume that the models rely on knowledge obtained during pre-training or fine-tuning instead. Finally, we find that CAD can prevent the model from learning spurious correlations more than the CL loss. This leads to more instances having the correct prediction only after training with CAD, as in the example in row 2.

7 Conclusion

We propose a new task related to fact checking, namely, detecting when evidence with omitted information is (in)sufficient. To this end, we conducted an in-depth empirical analysis with a newly introduced fluency-preserving method for omitting evidence information. We compared what Transformer-based models and humans find to be sufficient information for FC, resulting in a novel dataset, *SufficientFacts*. Finally, we showed that the proposed evidence omission method can be used for collecting contrastive examples for CL and CAD, which improved the performance of the studied models on the Evidence Sufficiency Prediction task and on veracity prediction.

The resulting models could be applied to detect emergent false claims, which gain popularity before any reputable source can refute them, as our proposed models can indicate when the provided input is insufficient for making a decision and whether to provide the user with the veracity prediction. Such models could also be used for detecting knowledge or evidence gaps that need to be filled to refute or support popular claims. Another possible future research direction would be to build FC models that indicate the particular part of the claim that they are missing supporting evidence for. Moreover, our proposed analysis and methods could be applied to other knowledge-intensive tasks, such as question answering.

Acknowledgments

The research documented in this paper has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement no. 801199. Isabelle Augenstein's research is further partially funded by a DFF Sapere Aude research leader grant. The authors would like to thank the anonymous reviewers and action editors for their helpful comments and suggestions.

References

Pepa Atanasova, Dustin Wright, and Isabelle Augenstein. 2020. Generating label cohesive and well-formed adversarial claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3168–3177, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.256

Isabelle Augenstein. 2021. *Towards Explainable Fact Checking*. Dr. Scient. thesis, University of Copenhagen, Faculty of Science.

Ann Bies, Justin Mott, and Colin Warner. 2015. English news text treebank: Penn treebank revised. https://doi.org/10.35111/m5b6-4m82

Kersti Börjars and Kate Burridge. 2019. *Introducing English Grammar*. Routledge. https://doi.org/10.4324/9780429023293

Keith Brown, Jim Miller, and James Edward Miller. 1991. *Syntax: A Linguistic Introduction to Sentence Structure*. Psychology Press.

Noel Burton-Roberts. 2016. *Analysing Sentences: An Introduction to English Syntax*. Routledge. https://doi.org/10.4324/9781315646046

Jiaao Chen, Zichao Yang, and Diyi Yang. 2020. MixText: Linguistically-informed interpolation of hidden space for semi-supervised text classification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2147–2157, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.194

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. Association for Computational Linguistics. https://www.aclweb.org/anthology/N19-1423

Bayu Distiawan, Gerhard Weikum, Jianzhong Qi, and Rui Zhang. 2019. Neural relation extraction for knowledge base enrichment. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 229–240.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. 2021. Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Transactions of the Association for Computational Linguistics*, 9:160–175. https://doi.org/10.1162/tacl_a_00359

Paul L. Garvin. 1958. Syntactic units and operations. In *Proceedings of the 8th International Congress of Linguists at Oslo*, pages 58–59. De Gruyter Mouton.

Christopher Hidey, Tuhin Chakrabarty, Tariq Alhindi, Siddharth Varia, Kriste Krstovski, Mona Diab, and Smaranda Muresan. 2020. DeSePtion: Dual sequence prediction and adversarial examples for improved fact-checking. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8593–8606, Online. Association for Computational Linguistics. https://www.aclweb.org/anthology/2020.acl-main.761

Rodney Huddleston and Geoffrey Pullum. 2005. The Cambridge grammar of the English language. *Zeitschrift für Anglistik und Amerikanistik*, 53(2):193–194. https://doi.org/10.1515/zaa-2005-0209

Mohit Iyyer, John Wieting, Kevin Gimpel, and Luke Zettlemoyer. 2018. Adversarial example generation with syntactically controlled paraphrase networks. 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 1875–1885, New Orleans, Louisiana. Association for Computational Linguistics. https://doi.org/10.18653/v1/N18-1170

Yichen Jiang, Shikha Bordia, Zheng Zhong, Charles Dognin, Maneesh Singh, and Mohit Bansal. 2020. HoVer: A dataset for many-hop fact extraction and claim verification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3441–3460, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.findings-emnlp.309

Divyansh Kaushik, Eduard Hovy, and Zachary Lipton. 2020. Learning the difference that makes a difference with counterfactually-augmented data. In *International Conference on Learning Representations*. https://openreview.net/forum?id=Sklgs0NFvr

Ji-Seong Kim and Key-Sun Choi. 2021. Fact checking in knowledge graphs by logical consistency. http://www.semantic-web-journal.net/system/files/swj2721.pdf

Jiho Kim, Kijong Han, and Key-Sun Choi. 2018. KBCNN: A knowledge base completion model based on convolutional neural networks. In *Annual Conference on Human and Language Technology*, pages 465–469. Human and Language Technology.

Ryan Kiros, Yukun Zhu, Russ R. Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Skip-thought vectors. In *Advances in neural information processing systems*, pages 3294–3302. https://proceedings.neurips.cc/paper/2015/file/f442d33fa06832082290ad8544a8da27-Paper.pdf

Nikita Kitaev and Dan Klein. 2018. Constituency parsing with a self-attentive encoder. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2676–2686. https://doi.org/10.18653/v1/P18-1249

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A lite BERT for self-supervised learning of language representations. In *International Conference on Learning Representations*. https://openreview.net/forum?id=H1eA7AEtvS

Markus Leippold and Thomas Diggelmann. 2020. Climate-FEVER: A dataset for verification of real-world climate claims. In *NeurIPS 2020 Workshop on Tackling Climate Change with Machine Learning*. https://www.climatechange.ai/papers/neurips2020/67

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

Zhuang Ma and Michael Collins. 2018. Noise contrastive estimation and negative sampling for conditional models: Consistency and statistical efficiency. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3698–3707, Brussels, Belgium. Association for Computational Linguistics. https://www.aclweb.org/anthology/D18-1405

Nishtha Madaan, Inkit Padhi, Naveen Panwar, and Diptikalyan Saha. 2021. Generate your counterfactuals: Towards controlled counterfactual generation for text. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13516–13524. https://ojs.aaai.org/index.php/AAAI/article/view/17594

Piotr Niewinski, Maria Pszona, and Maria Janicka. 2019. GEM: Generative enhanced model for adversarial attacks. In *Proceedings of the Second Workshop on Fact Extraction and VERification (FEVER)*, pages 20–26, Hong Kong, China. Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-6604

Malte Ostendorff, Nils Rethmeier, Isabelle Augenstein, Bela Gipp, and Georg Rehm. 2022. Neighborhood contrastive learning for scientific document representations with citation embeddings. In *arXiv:2202.06671*.

Wojciech Ostrowski, Arnav Arora, Pepa Atanasova, and Isabelle Augenstein. 2021. Multi-hop fact checking of political claims. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 3892–3898. International Joint Conferences on Artificial Intelligence Organization. Main Track. https://doi.org/10.24963/ijcai.2021/536

Ankur Parikh, Oscar Täckström, Dipanjan Das, and Jakob Uszkoreit. 2016. A decomposable attention model for natural language inference. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2249–2255, Austin, Texas. Association for Computational Linguistics.

https://doi.org/10.18653/v1/D16-1244

Yanru Qu, Dinghan Shen, Yelong Shen, Sandra Sajeev, Weizhu Chen, and Jiawei Han. 2021. CoDA: Contrast-enhanced and diversity-promoting data augmentation for natural language understanding. In *International Conference on Learning Representations*. https://openreview.net/forum?id=Ozk9MrX1hvA

Nils Rethmeier and Isabelle Augenstein. 2021. A primer on contrastive pretraining in language processing: Methods, lessons learned and perspectives. *arXiv preprint arXiv:2102.12982*.

Nils Rethmeier and Isabelle Augenstein. 2022. Long-tail zero and few-shot learning via contrastive pretraining on and for small data. In *Proceedings of AAAI 2022 Workshop on Artificial Intelligence with Biased or Scarce Data (AIBSD 2022)*.

Mattia Samory, Indira Sen, Julian Kohne, Fabian Flöck, and Claudia Wagner. 2021. Call me sexist, but...: Revisiting sexism detection using psychological scales and adversarial samples. In *Proceedings of the Fifteenth International Conference on Web and Social Media*. AAAI Press. https://ojs.aaai.org/index.php/ICWSM/article/view/18085

Tal Schuster, Adam Fisch, and Regina Barzilay. 2021. Get your vitamin C! Robust fact verification with contrastive evidence. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 624–643, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.naacl-main.52

Tal Schuster, Darsh Shah, Yun Jie Serene Yeo, Daniel Roberto Filizzola Ortiz, Enrico Santus, and Regina Barzilay. 2019. Towards debiasing fact verification models. 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 3419–3425, Hong Kong, China. Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1341

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics. https://doi.org/10.18653/v1/P16-1009

Damien Teney, Ehsan Abbasnedjad, and Anton van den Hengel. 2020. Learning what makes a difference from counterfactual examples and gradient supervision. In *Computer Vision – ECCV 2020*, pages 580–599, Cham. Springer International Publishing. https://doi.org/10.1007/978-3-030-58607-2_34

James Thorne and Andreas Vlachos. 2021. Evidence-based factual error correction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3298–3309, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.acl-long.256

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. Association for Computational Linguistics. https://doi.org/10.18653/v1/N18-1074

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2019. Evaluating adversarial attacks against multiple fact verification systems. 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 2944–2953, Hong Kong, China. Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1292

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. SuperGLUE: A stickier benchmark for general-purpose language understanding systems. In

Advances in Neural Information Processing Systems, volume 32. Curran Associates, Inc.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics. https://doi.org/10.18653/v1/W18-5446

Dustin Wright, David Wadden, Kyle Lo, Bailey Kuehl, Arman Cohan, Isabelle Augenstein, and Lucy Lu Wang. 2022. Generating scientific claims for zero-shot scientific fact checking. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*. https://arxiv.org/abs/2203.12990

Xing Wu, Shangwen Lv, Liangjun Zang, Jizhong Han, and Songlin Hu. 2019. Conditional BERT contextual augmentation. In *Computational Science – ICCS 2019*, pages 84–95, Cham. Springer International Publishing. https://doi.org/10.1007/978-3-030-22747-0_7

Zhi-Hua Zhou and Ming Li. 2005. Tri-training: Exploiting unlabeled data using three classifiers. *IEEE Transactions on Knowledge and Data Engineering*, 17(11):1529–1541. https://doi.org/10.1109/TKDE.2005.186