What Makes Medical Claims (Un)Verifiable?
Analyzing Entity and Relation Properties for Fact Verification

Amelie Wührl[1], **Yarik Menchaca Resendiz**[1], **Lara Grimminger**[1], and **Roman Klinger**[1,2]

[1]Institut für Maschinelle Sprachverarbeitung, University of Stuttgart, Germany
[2]Fundamentals of Natural Language Processing, University of Bamberg, Germany

```
{amelie.wuehrl, yarik.menchaca-resendiz,
     lara.grimminger}@ims.uni-stuttgart.de
            roman.klinger@uni-bamberg.de
```

Abstract

Biomedical claim verification fails if no evidence can be discovered. In these cases, the fact-checking verdict remains unknown and the claim is unverifiable. To improve upon this, we have to understand if there are any claim properties that impact its verifiability. In this work we assume that entities and relations define the core variables in a biomedical claim's anatomy and analyze if their properties help us to differentiate verifiable from unverifiable claims. In a study with trained annotation experts we prompt them to find evidence for biomedical claims, and observe how they refine search queries for their evidence search. This leads to the first corpus for scientific fact verification annotated with subject–relation–object triplets, evidence documents, and fact-checking verdicts (the BEAR-FACT corpus). We find (1) that discovering evidence for negated claims (e.g., X–does-not-cause–Y) is particularly challenging. Further, we see that annotators process queries mostly by adding constraints to the search and by normalizing entities to canonical names. (2) We compare our in-house annotations with a small crowdsourcing setting where we employ medical experts and laypeople. We find that domain expertise does not have a substantial effect on the reliability of annotations. Finally, (3) we demonstrate that it is possible to reliably estimate the success of evidence retrieval purely from the claim text ($.82F_1$), whereas identifying unverifiable claims proves more challenging ($.27F_1$). The dataset is available at http://www.ims.uni-stuttgart.de/data/bioclaim.

1 Introduction

Verifying scientific claims in user-generated content is sometimes unsuccessful, because no supporting or refuting evidence can be found. In these cases, the claim remains unverifiable. Previous work shows that both evidence retrieval and claim verification, the two core steps in automatic fact verification, are more robust for concisely formulated claims that have been extracted from noisy context compared to verifying user-generated claims directly (Sundriyal et al., 2022; Wuehrl et al., 2023).

Based on these findings, we hypothesize that breaking down claims into smaller units increases our understanding which properties impact verifiability. This knowledge is key to improving fact-checking (FC) systems. To this end, we assume that biomedical entities, e.g., mentions of treatments or medical conditions, and the relations between them (*causes*, *is–a–side-effect* etc.) make up the core variables that define the claim's anatomy. We analyze if these variables are connected to the claim's verifiability, i.e., that there are reoccurring patterns with respect to which types of claims tend to be SUPPORTED, REFUTED, or UNVERIFIABLE.

While disciplines like argument mining have an evolved understanding of claim properties (Boland et al., 2022), biomedical claims are poorly understood. The data resources to perform analyses do not exist yet: there is no corpus which is annotated both with (i) biomedical entities in relation that constitute claims and (ii) evidence and the veracity label it leads to. To create such a resource, we perform (1a) an annotation study of medical tweets in which we observe carefully trained in-house annotator's behavior to find evidence for entity-centered claims, (1b) a statistical analysis of the connection between entities/relations and the successful evidence retrieval, (2) a comparison of the in-house annotators' performance to crowdsourcing, in which we task laypeople and medical experts to verify the same claims. Finally, we (3) compare the performance of a fine-tuned transformer model to estimate the checkability of a claim.

We contribute BEAR-FACT, a novel Twitter[1] dataset for biomedical fact verification. It consists of 1,448 fact-checked claims, evidence documents and structured entity/relation information. We answer the following research questions:

---

[1]Twitter is now called X.

RQ1a Which properties, i.e., entity-relation patterns, make a claim (un)verifiable?

RQ1b How can we use medical entities in the claims as meaningful search queries for evidence discovery?

RQ2 What is the impact of the annotation setting, i.e., domain knowledge and crowdsourcing on annotation quality?

RQ3 Can we predict the verifiability, i.e., the likelihood that evidence for a claim exists, purely from the claim?

For RQ1a (§3.1), we find that entity–relation patterns in claims are connected to verifiability. Claims conveying a positive relation (e.g., *cause–of*) are more successfully fact-checked and more frequently SUPPORTED compared to their negative counterpart (*not–cause–of*). For RQ1b (§3.2), we find that study participants predominantly change an entity-centric predefined query by reformulating entity realizations to canonical names. For RQ2 (§4), we observe that domain expertise does not have a substantial effect on the reliability of fact-checking annotations. Finally, the hypothesis in RQ3 holds to some degree (§5): Fine-tuning a RoBERTa model to differentiate between verifiable and unverifiable claims is reliable for the verifiable class (.82 F_1). Detecting unverifiable claims is more challenging (.27 F_1).

2 Annotation study

We design an annotation study with two goals: (1) To construct a resource that enables us to explore claim properties, i.e., the role of entities and relations in the fact verification process. (2) To observe how fact-checkers modify entity-based search queries during the evidence retrieval process.

We construct a dataset with two annotation layers: (a) fact verification annotation, i.e., claims checked against evidence, and (b) structured knowledge, i.e., entity and relations. We build our dataset on top of BEAR (Wührl and Klinger, 2022b), a corpus of English tweets annotated with biomedical entities and relations.

2.1 Data

We identify relevant claims from BEAR for further annotation. The tweets have to:

Contain a claim. To identify tweets that convey a claim we employ a pretrained claim detection model (Wührl and Klinger, 2021) and only keep instances with claims.

Contain at least one medical relation. We use the documents for which at least one medical relation was annotated.

Out of 2,100 documents from BEAR this filtering leaves us with 646 claim-containing documents. To extract claims, the tweets undergo two steps:

Claim extraction. We identify individual claims within the tweets by extracting an entity–relation–entity triplet from the input documents based on the entity–relation annotation.

Manual filtering. To ensure data quality, we remove 166 claims that are incorrectly extracted from the tweet's context, repetitions within the same tweet, contain the relation "somehow related to", or are off-topic (e.g., discuss treatments of animals)[2].

We correct grammatical errors in 346 of the automatically extracted claims to increase their readability. Table 9 shows an end-to-end example of the filtering process. These preprocessing steps lead to 1,532 claims to be fact-checked.

2.2 Annotation

2.2.1 Annotation Task

In the annotation study, annotators are tasked to verify claims against scientific evidence. For every claim, their task is to find an article which contains supporting or refuting evidence for the claim. They search for evidence using PubMed[3], a database for biomedical articles. Based on the evidence they find, the claims are assigned a fact-checking verdict. Claims can thus be labeled as follows:

SUPPORTED Evidence supports the claim.

PARTIALLY SUPPORTED Evidence partially support the claim, e.g., if evidence is more specific than the claim.

PARTIALLY REFUTED Evidence refutes the claim but is more specific than the claim.

REFUTED Evidence refutes the claim.

We provide the annotators with a starting query for the evidence search. The query is made up of the medical entities mentioned in the claim, connected by an AND operator. For example, a claim stating that "H2 blocker treats SpO2" has the starting query "(H2 blocker) AND (SpO2)".

Annotators provide the PubMed Identifier (PMID) of the respective article that they use to verify the claim along with the sentences that support or refute the claim. For a given claim, we

---

[2]Table 8 shows a description of each category.
[3]https://pubmed.ncbi.nlm.nih.gov/

instruct the annotators to go over the titles and abstracts of the first five search results for the pre-built query. After that, annotators refine the search query to fine-tune their search. If the refinement leads to evidence being discovered, we record their updated search query. If no evidence is discovered after three minutes, the claim is labeled as UNVERIFIABLE.[4] To understand why a particular claim appears to be unverifiable, annotators rate how confident they are that a continued search could uncover evidence. We refer to this as the 'evidence exists confidence'[5].

2.2.2 Evaluation

We evaluate the agreement for the two subtasks as follows: For the verdict assignment task we report the Cohen's κ score between annotators. For evidence retrieval, we gauge how often two annotators retrieve the same evidence document to verify a claim. We report the Jaccard similarity between the set of evidence documents which, given two sets, is calculated by dividing the size of the intersection of the two sets by the size of their union. Scores > 0 indicate that there is at least one shared member in the sets. We calculate the Jaccard similarity for each pair of evidence documents where both annotators assigned the same verdict. Since annotators may use the same evidence document to substantiate conflicting verdicts (e.g., A1 uses document 123 and assigns SUPPORTED, while A2 uses document 123 to REFUTED a claim), we also report the Jaccard score for conflicting verdicts.

2.2.3 Annotation Procedure

We set up the study using the online platform SoSci Survey[6]. We provide screenshots of the environment in the supplementary material[7]. We work with two in-house annotators (A1, A2) to label the claims. The annotators are male and female, aged 25 to 30, with backgrounds in computational linguistics. While they have no formal medical training, they are experienced annotators for biomedical social media data.

Both annotators label a test batch of 10 claims to evaluate our annotation guidelines. For assign-

claims	UN	REF	pREF	pSUP	SUP	total
#	447	60	38	224	679	1448
%	30.9	4.1	2.6	15.5	46.9	100

Table 1: Distribution of fact-checking verdicts.

ing the FC verdict, Cohen's κ is 1.0, indicating perfect agreement. For the evidence retrieval task, we calculate the Jaccard similarity between the evidence documents. The average similarity in all pairs where annotators assigned SUPPORTED is 0.29. Note that annotators do not assign REFUTED to any claim in the test batch, hence why we can not report a Jaccard score. For the SUP–SUP instances, only in 29 % of cases, annotators used the same evidence document to reach their verdict. This is noteworthy, as they are in perfect agreement about all verdicts. Our observation points to a key property of fact-checking evaluation: annotators may use different evidence documents to verify a claim. Low Jaccard scores therefore do not necessarily indicate low annotation quality.

In the main annotation phase, A1 and A2 research a total of 1,448 claims.[8] We refer to the dataset as BEAR-FACT.

2.3 Corpus statistics

For a better understanding of the resulting dataset, the following section provides corpus statistics.

Number of medical claims per tweets. BEAR-FACT consists of 1,448 claims in 572 tweets. We provide a a histogram of the number of claims per tweet in the Appendix (Fig. 3). We find that 201 out of the 572 tweets convey exactly one claim each. The tweet with the highest number of claims conveys 14 claims. Notably, the majority of tweets in BEAR-FACT expresses more than one claim.

Distribution of fact-checking verdicts. Table 1 provides an overview of how fact-checking verdicts are distributed in BEAR-FACT. The majority of claims are (PARTIALLY) SUPPORTED (62.4 %). (PARTIALLY) REFUTED has the smallest number of instances (6.7 %). Notably, 30.9 % of claims are unverifiable, meaning that there was no evidence to substantiate a verdict.[9]

---

[4]Note that this time limit only affects the refinement process, not the overall annotation process for an instance.

[5]Annotators rate this on a 5-point scale ranging from *I'm very confident relevant evidence exists* to *Very sure that there is no evidence out there which I could use to check the claim*.

[6]https://www.soscisurvey.de/

[7]The data and supplementary material is available at http://www.ims.uni-stuttgart.de/data/bioclaim.

[8]Due to limited resources, annotators could not complete the annotation for all 1,532 claims available in the dataset.

[9]From the annotators' comments we learn that 17 claims have been extracted incorrectly. This means that the claim we obtain from the entity-based claim extraction step (ref. to Sec. 2.1) does not accurately represent the claim expressed in the tweet, for example because the extracted claims omits relevant contextual information. As they are not checkable,
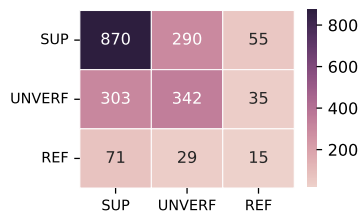
Figure 1: Pairwise co-occurrence of verdicts in BEAR-FACT tweets with more than one claim. (partially) SUP-PORTED and (partially) REFUTED are collapsed, resp.

For the claims that are (partially) SUPPORTED by the literature we can infer that the information conveyed in them are most likely to be true. However, 37.6 % of instances are either UNVERIFIABLE or (partially) REFUTED, meaning there is no point of reference about that statement in the literature or they convey false information. This finding emphasizes the importance of fact-checking for biomedical claims for social media.

Verdict co-occurrence. For tweets with more than one claim, we analyze if medical claims co-occur with specific other types of claims with respect to their fact-checking verdict. In other words, if a claim in a tweet is SUPPORTED, REFUTED or UNVERIFIABLE, how common is it for the other claims in that tweet to have the same verdict? For all tweets with more than one claim, we form all possible pairs of claims in one tweet to obtain the verdict co-occurrences. We choose pairs to be able to handle the varying amount of claims per tweet. Subsequently, we visualize the pairwise co-occurrence of their fact-checking verdicts in Figure 1. We observe two major patterns: In the diagonal, we see that (PARTIALLY) SUPPORTED claims most frequently co-occur with claims of the same verdict, followed by pairings with UNVERIFIABLE claims. Further, (PARTIALLY) REFUTED claims do not show this pattern. Pairs of (PARTIALLY) RE-FUTED claims are very infrequent; it is more common for such claims to occur with (PARTIALLY) SUPPORTED or UNVERIFIABLE claims.

Our analysis indicates that medical tweets have a tendency to convey claims with mixed veracity levels emphasizing the importance of a fine-grained approach to fact verification.

3 Claim characteristics & evidence discovery

3.1 Which properties make claims (un)verifiable? (RQ1a)

We hypothesize that entity–relation properties in a claim are connected to its verifiability. To investigate this, we explore the following questions:
Which claim relations are (un)verifiable? To investigate if there are specific relation types in claims that tend to be SUPPORTED, REFUTED or UNVERIFIABLE, we analyze the distribution of verdicts across each relation type. Figure 2a shows the results. Each row in the heatmap represents a relation class, each column represents a fact-checking verdict. Each cell depicts the percentage of claims that express the relation and verdict.

We observe that for the positive relations, e.g., *cause–of*, *prevents* or *positive–influence–on* the distribution is dominated by the overall verdict distribution within BEAR-FACT, meaning the majority of claims fall into the SUPPORTED category, followed by UNVERIFIABLE and REFUTED. However, for their negative counterparts, e.g., *not–cause–of*, *does–not–prevent* etc., we observe the opposite. The dominant verdict class is REFUTED or UNVERI-FIABLE. The exception is the pair *positive/negative influence on* where the distribution of the negative relation is very similar to its positive counterpart.

To test if the distribution of positive and negative variants of a relation are in fact different, i.e., there is no relation between the two distributions, we compute the chi-square statistic[10] for relation classes with > 5 instances (*(not–)cause–of*, *(does–not–)treat*, *pos./neg.–influence–on*). The results show that for two out of three relation pairs, i.e., *(not–)cause–of* and *(does–not–)treat*) the distributions are in fact unrelated ($p < 0.05$).
Which claim entities are (un)verifiable? Figures 2b and 2c illustrate the distribution of entity types across each fact-checking verdict. From Fig. 2b, we observe that the vast majority of claims use a medical condition (*medC*) or treatment mention (*treat_drug*, *treat_therapy*) as their subject. The distribution across the verdicts for the majority of entity classes mirrors the distributing of verdicts in the overall dataset. In Fig. 2c, we observe that claim objects are almost exclusively mentions of medical conditions. This is most pronounced

---

[10]We use the Scipy implementation of chi-squared: https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.chi2_contingency.html

we assign those instances to the UNVERIFIABLE class.

in (PARTIALLY)REFUTED claims where 88 out of 98 (PARTIALLY)REFUTED claims make a claim. Across all verdicts, environmental factors are almost exclusively claim objects. When considering the entity distributions, note that to a certain degree the relation type dictates the subject and object entity type in a triplet.

3.2 Evidence discovery: Are medical entities meaningful search queries? (RQ1b)

We investigate **how we can use medical entities in the claims as meaningful search queries for evidence discovery**. We hypothesize that medical entities are key to connect claims to evidence. Further, we explore how search queries are refined during the evidence retrieval process. Thus, we analyze our annotators' search strategies.

Are medical entities meaningful search queries? Recall that annotators start their search with pre-built search queries that consist of the medical entities from the claim connected by AND operators (see Sec. 2.2.1). Out of the 1,001 claims in our dataset for which annotators found supporting or refuting evidence, 757 claims could be verified with the results from this original search query. In 244 cases (24.4 %) the annotators had to refine the search query, and subsequently discovered an evidence document. This shows that medical entities are an appropriate starting point for evidence search.[11]

How are queries refined to discover suitable evidence? We aim to understand how search queries are refined and analyze the query refinement strategies that lead to evidence being discovered. To that extend, we define seven types of refinements:
1. Generalizing search terms
2. Specifying search terms
3. Normalizing brand names by replacing brand names with the respective active ingredient
4. Normalizing informal language
5. Resolving abbreviations
6. Adding relation between search terms
7. Other

Table 2 shows examples and the number of search terms per refinement type class in a sub-sample of 50 claims. In those claims, annotators refined a query and subsequently discovered an evidence document. We count a total of 56 refinement operations in our sample. Normalizing informal and colloquial terms is the most frequent operation (in 18 out of 50 instances), followed by using a more general search term (11) and resolving abbreviations (10). Adding a relational term to the query is used the least (2).

With the exception of Strategy 6 (Adding relation), all other strategies we observe are operations to normalize the query terms. Considering the style gap between the social media claims and the scientific evidence documents, this finding intuitively makes sense. It also indicates that entity normalization or linking has the potential to improve automatic evidence retrieval methods.

Why do claims remain unverifiable? 447 claims are labeled as UNVERIFIABLE, because the query refinement was not successful and no relevant evidence could be discovered. To understand why a particular claim could not be checked, we analyze the annotators' estimate that evidence exists. For the majority of cases (54.1 %), annotators state that they cannot judge if evidence could exist. However, approx. 20 % of the unverifiable claims, the annotators are confident that evidence exists (*pretty confident*: 15.9%, *very confident: 4.7%*). For the remaining 25% of UNVERIFIABLE claims, annotators are either *pretty* (15.2%) or *very* confident (8.9%) that no evidence exists (Ex. 1/2 in Table 10 in Appendix). Those claims are ambiguous or general in which case it makes sense that discovering evidence is unlikely. For the claims with high confidence about the existence of evidence, consider Examples 3 and 4 in Table 10.

4 What is the impact of the annotation setup?

As described in the previous section, we employ in-house annotators to create BEAR-FACT. Crowdsourcing is, however, a viable alternative to collect fact-checking annotations (Martel et al., 2023; Mohr et al., 2022). Therefore, to understand how the annotation setting, i.e., in-house annotation vs. crowdsourcing, impacts our task, we investigate **RQ2a: How reliably do untrained crowdworkers verify biomedical claims?** and **RQ2b What is the impact of domain knowledge** (i.e., biomedical expertise) **in the crowdsourcing setting?**

4.1 Experimental setting

To study this we require crowd annotations from people with and without biomedical expertise. For

---

[11]Note that PubMed's internal article ranking contributes to evidence discovery. The properties of this ranking need to be taken into account when designing systems that do not rely on PubMed.

Figure 2

(a) Relation types.

	UNVER	REF	SUP
cause of	207 (35%)	25 (4%)	352 (60%)
not cause of	13 (45%)	10 (34%)	6 (21%)
prevents	12 (15%)	5 (6%)	65 (79%)
does not prevent	3 (38%)	2 (25%)	3 (38%)
treats	69 (25%)	12 (4%)	190 (70%)
does not treat	7 (26%)	14 (52%)	6 (22%)
has symptom	14 (22%)	0 (0%)	50 (78%)
side effect of	14 (27%)	4 (8%)	33 (65%)
prescribed for	0 (0%)	0 (0%)	4 (100%)
is contraindicated drug	2 (50%)	0 (0%)	2 (50%)
may diagnose	3 (17%)	1 (6%)	14 (78%)
may not diagnose	5 (100%)	0 (0%)	0 (0%)
pos influence on	41 (27%)	14 (9%)	99 (64%)
neg influence on	32 (33%)	7 (7%)	58 (60%)
pos interaction	10 (59%)	0 (0%)	7 (41%)
neg interaction	3 (60%)	1 (20%)	1 (20%)
is similar to	4 (36%)	0 (0%)	7 (64%)
other	8 (47%)	3 (18%)	6 (35%)

(b) Subject entities.

	UNVER	REF	SUP
medC	146 (32%)	13 (3%)	298 (65%)
treat_drug	131 (31%)	33 (8%)	261 (61%)
treat_therapy	72 (26%)	27 (10%)	179 (64%)
diag	9 (39%)	1 (4%)	13 (57%)
biochem_proc	3 (38%)	0 (0%)	5 (62%)
biochem_subst	17 (42%)	4 (10%)	19 (48%)
env_diet	13 (21%)	8 (13%)	40 (66%)
env_geo-cli	3 (100%)	0 (0%)	0 (0%)
env_habit	11 (48%)	4 (17%)	8 (35%)
env_pollut	2 (67%)	0 (0%)	1 (33%)
env_socio-eco	14 (32%)	2 (5%)	28 (64%)
pathogen	3 (12%)	1 (4%)	22 (85%)
qol	0 (0%)	0 (0%)	0 (0%)
other	23 (40%)	5 (9%)	29 (51%)

(c) Object entities.

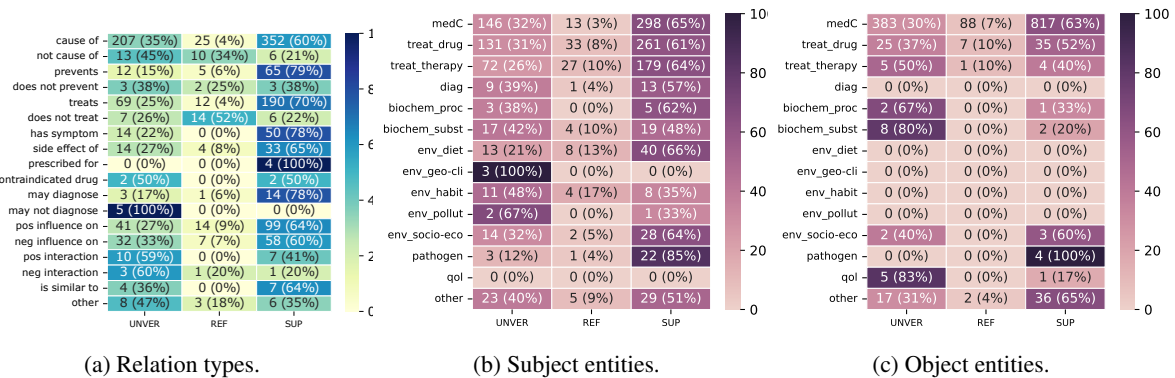|	UNVER	REF	SUP
medC	383 (30%)	88 (7%)	817 (63%)
treat_drug	25 (37%)	7 (10%)	35 (52%)
treat_therapy	5 (50%)	1 (10%)	4 (40%)
diag	0 (0%)	0 (0%)	0 (0%)
biochem_proc	2 (67%)	0 (0%)	1 (33%)
biochem_subst	8 (80%)	0 (0%)	2 (20%)
env_diet	0 (0%)	0 (0%)	0 (0%)
env_geo-cli	0 (0%)	0 (0%)	0 (0%)
env_habit	0 (0%)	0 (0%)	0 (0%)
env_pollut	0 (0%)	0 (0%)	0 (0%)
env_socio-eco	2 (40%)	0 (0%)	3 (60%)
pathogen	0 (0%)	0 (0%)	4 (100%)
qol	5 (83%)	0 (0%)	1 (17%)
other	17 (31%)	2 (4%)	36 (65%)

Figure 2: Verdict distribution across claim relation and entity types. Color coding is based on the percentage of verdicts per relation/entity class. We collapse (partially) SUPPORTED and (partially) REFUTED instances into one group, respectively.

Id	Refinement type	#refined	Original term/query	Refined term/query
1	Generalizing search term	11	(Vit C 500mg)	(Vitamin C)
2	Specifying search term	8	(delta)	(delta) AND (Covid)
3	Normalizing brand name	4	(Tecentriq)	(Atezolizumab)
4	Normalizing informal terms	18	(Rona)	(corona)
5	Resolving abbreviations	10	(IBS)	(irritable bowel syndrome)
6	Adding relation	2	(fever) AND (Ivermectin)	(fever) side effect of (Ivermectin)
7	other	3	(elevated Uric acid levels)	(Uric acid) AND (lower)

Table 2: Number of refined search terms per refinement type across a sample of 50 successfully refined search queries along with examples.

the general crowd (to which we refer as layCrowd), we recruit university students of a Master's Program in Computational Linguistics for a voluntary on-site study. For the crowdworkers with domain expertise (expCrowd), we recruit participants with a background in (bio)medicine on the online crowd-sourcing platform Prolific[12]. Each participant verifies ten claims. The study is hosted using Google Forms[13]. Crowdworkers on Prolific are reimbursed with £9 per hour. All partipants are instructed with the same guidelines as in-house annotators.

4.2 Results

4.2.1 Fact-checking verdicts

We obtain ten sets of annotations in the expCrowd setting and nine sets in the layCrowd setting[14]. Table 3 presents the agreement scores for the verdict assignment task.[15] We report the agreement among each group (layCrowd, expCrowd) as well as the

---

[12] https://prolific.com

[13] https://docs.google.com/forms/

[14] One annotator in the layCrowd setting dropped out.

[15] One participant in layCrowd only managed to work on 8 out of 10 claims in the scheduled time. We include their annotations for the completed claims. For the agreement, we calculate the κ metric for all pairs involving this annotator only with the completed claims.

Annotator 1	Annotator 2	κ
layCrowd		0.24
expCrowd		0.22
in-house	agg. layCrowd	0.30
in-house	agg. expCrowd	0.40
agg. expCrowd	agg. layCrowd	0.65

Table 3: Inter-annotator agreement for the verdict assignment task. We report (av.) pairwise Cohen's κ.

agreement between our in-house annotations and the aggregated label from each group. Assuming that for modeling purposes we would use an aggregation of the individual crowdworkers' labels, we first aggregate via majority voting before calculating the agreement between groups. Appendix A.4 shows details on the aggregation strategy.

Trained annotators compared to crowdworkers. The agreement among the individual annotators in layCrowd is $\kappa = 0.24$, indicating fair agreement. The result is similar for expCrowd ($\kappa = 0.22$). The in-house annotators of the full corpus showed perfect agreement in their training phase. These scores are not directly comparable, but this indicates that the task is more challenging in a crowd setting. The agreement between the two crowd settings (agg. ex-

		expert		
		SUP	UNV	REF
	SUP	3	0	0
lay	UNV	2	5	0
	REF	0	0	0

Table 4: Confusion matrix illustrating the verdicts assigned by agg. layCrowd vs. agg. expCrowd.

pCrowd, agg. layCrowd) is moderate ($\kappa = 0.65$).

Impact of biomedical expertise. To understand if domain expertise has an impact on the annotation, we compare the agreement scores for layCrowd and expCrowd. The agreement among the domain experts is slightly lower (a Δ of 0.02κ). This indicates that biomedical expertise does not have a substantial impact on the reliability of the results, in fact, their judgments are more varied compared to the general crowd. Note, however, that annotation quality in an anonymous online setting may also vary more compared to an on-site setting.

The agreement between the in-house annotators and the agg. expCrowd verdicts is higher than their agreement with the agg. layCrowd verdicts. We hypothesize that this may be an effect of their prior experience in annotating biomedical data.

Finally, we visualize the verdict assignments in a confusion matrix in Table 4 which shows the verdicts assigned by agg. layCrowd on the vertical and the agg. expCrowd verdicts on the horizontal axis. The diagonal represents the instances where both groups assigned the same verdict. We observe the strongest agreement in the UNVERIFIABLE instances. Notably, we observe that in two instances the agg. expCrowd is able to verify the claim (SUPPORTED), the lay crowd is not. We presume that this a result of their domain expertise.

4.2.2 Evidence retrieval

We want to understand how frequently annotators use the same evidence to SUPPORT or REFUTE a claim, while the verdict label is or is not the same. Table 5 reports the Jaccard similarities that measure the overlap of evidence documents. For each combination of verdicts (SUP–SUP, REF–REF, SUP–REF), we report the average Jaccard score as well as the number and percentage of instances within that group with a Jaccard score >0. The upper half of the table reports the results for layCrowd, the bottom half the results for expCrowd.

Across all verdict combinations, layCrowd annotators agree more on which PubMed document

they use to substantiate a verdict compared to the expCrowd annotators. For the SUP–SUP instances, in 48 % of evidence pairs, layCrowd annotators relied on at least one common document, while in expCrowd this is only the case for 33 % of evidence pairs. Apparently, experts chose the evidence more selectively and do not accept the first document that might be a fit. This is in line with the agreement scores for the verdict assignment: If annotators use more diverse evidence documents, we can also expect their verdicts to be more diverse. Generally, the agreement on evidence documents is the highest in REF–REF pairs compared to the other combinations of verdicts. This may be an effect of negative results being published more seldomly.

Annotators sometimes use the same evidence documents, but reach opposing verdicts. This happens more frequently in the layCrowd setting. For SUP–REF pairs, we observe an average Jaccard score of 0.38 in layCrowd and 0.3 in expCrowd. We hypothesize that this is an effect of domain knowledge as it might take biomedical expertise to interpret evidence correctly.

The relatively low Jaccard scores potentially also result from the annotators' evidence research strategy. In cases where the query returns multiple relevant evidence documents, one annotator may stop their research after discovering the first document, while the other continues, the Jaccard score is zero. In our setup, we do not explicitly instruct annotators to look at all evidence documents before assigning the verdict. Instead, we instruct them to return to the survey once they find suitable evidence. While we assume that they go over the results from top to bottom, we cannot control for that. Defining more fine-grained retrieval strategies could be an interesting task for future research. That being said, we do not strictly want to optimize for a high Jaccard score. Two annotators could reach the same verdict using different evidence documents, so we should always take into account both the agreement for the verdict assignment as well as the Jaccard score for evidence retrieval.

5 Can we predict if claims are unverifiable?

Moving to the modeling perspective, we investigate if we **can predict verifiability, i.e., the likelihood that evidence for a claim exists, purely from the claim** (RQ3). A setup like that could allow us to adapt the manual evidence search

		SUP-SUP	REF-REF	SUP-REF
lay	avg. J	.31	.58	.38
	# J>0	40	10	27
	% J>0	.48	.77	.56
expert	avg. J	.19	.43.	.2
	# J>0	53	4	17
	% J>0	.33	.57	.3

Table 5: Avg. Jaccard similarity (avg. J), number and percentage of Jaccard scores > 0 per verdict combination for the lay and expert crowd. We collapse the partially SUPPORTED/REFUTED verdicts into their respective major class.

Id	claim	G	P
1	dantrolene treats malignant hyperthermia	V	V
2	inappropriate cleaning methods is/are the cause of outbreaks	U	U
3	NAC is contraindicated drug to pregnancy	U	V
4	fear causes a weakened immune system	V	U

Table 6: Claims along with gold and predicted verifiablitiy labels (U: UNVERIFIABLE, V: VERIFIABLE). G: gold, P: Prediction.

Class	Recall	Precision	F1
VERIFIABLE	.93	.74	.82
UNVERIFIABLE	.18	.52	.27
Macro	.56	.63	.54
Weighted	.71	.67	.66

Table 7: Precision, Recall, F_1 of the verifiability class.

procedure by giving experts more time for such claims. To investigate this, we train a model that differentiates UNVERIFIABLE from verifiable (SUPPORTED/REFUTED) claims.

Experimental setting. To train the models, we define two classes: UNVERIFIABLE and VERIFIABLE ((PARTIALLY) SUPPORTED, (PARTIALLY) REFUTED). We train a classifier on top of RoBERTa (Conneau et al., 2020) for 15 epochs with default parameters on an Nvidia RTX A6000, using a 80/20 train-test split of the BEAR-FACT data. The input for the classifier is the claim phrase.

Results. Table 7 shows the results. They indicate that it is possible to reliably identify claims that are verifiable (.82F_1), whereas identifying unverifiable claims proves more challenging (.27F_1). We further look into the connection between the model performance and annotators' 'evidence exists confidence' (see Sec. 2.2.1). We hypothesize that this rating may be correlated with the predicted labels. We find, however, that the point biserial correlation[16] that measures the correlation between the binary predicted labels and the continuous confidence scores is limited (0.22).

Qualitative Analysis. To expand our analysis of (un)verifiable claims to the modeling perspective, we conduct a manual error analysis. We find that VERIFIABLE claims frequently include technical terms and medical terminology making the entities and relation more specific (see Ex. 1, Table 6). UNVERIFIABLE claims on the other hand include more general vocabulary, potentially making them more difficult to verify (see Ex. 2). Based on this introspection, we hypothesize that the model picks up on the varying specificity levels of the medical terminology and overgeneralizes this property

---

[16]https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.pointbiserialr.html

in the incorrectly classified instances (Ex. 3, 4). This aligns with Sec. 3.2 where we find that entity normalization is key to verify claims.

6 Related Work

Scientific biomedical fact-checking focuses on verifying scientific claims against evidence sources. Typically, this is modeled in two steps: evidence retrieval, i.e., discovering relevant sources, and claim verification, the task of assigning a verdict to a claim based on the evidence. This can also be modeled jointly. Vladika and Matthes (2023) provide a comprehensive overview.

The focus in the area has been on concise, sometimes synthetic claims (Wadden et al., 2022; Kotonya and Toni, 2020). More recently, user-generated medical content has received more attention (Zuo et al., 2022; Vladika et al., 2023; Saakyan et al., 2021, i.a.). This type of content is challenging (Kim et al., 2021) and complex (Sarrouti et al., 2021), posing the question of which units of information should serve as the input to FC systems. Nevertheless currently it is standard to either process full sentences or even paragraphs (Mohr et al., 2022). Alternatively claims are atomic by design, e.g., claims in SCIFACT (Wadden et al., 2022). Recent work shows that both evidence retrieval and claim verification are more robust for concisely formulated claims. Apart from that, **properties of biomedical claims** and their impact on fact-checking are poorly understood.

Outside biomedical fact-checking, i.e., in argument mining and argument theory, there exists a more developed understanding of claim properties.

In argument mining claims have been categorized according to their function, i.e., epistemic vs. practical vs. moral claims (Lippi and Torroni, 2016), their semantic type (Hidey et al., 2017; Egawa et al., 2019; Jo et al., 2020, i.a.) or studied with respect to their conceptualization across domains (Daxenberger et al., 2017; Boland et al., 2022).

For fact-checking, some datasets categorize their claims into groups such as numerical claims or position statements (Francis and Full Fact, 2016), but the inherent structural properties of claims are not understood. Structured knowledge has been proposed to represent claims in scientific discourse (Magnusson and Friedman, 2021), and as a method to detect (Yuan and Yu, 2019) and extract health-related claims (Wührl and Klinger, 2022a).

Two strands of research are related to our task of estimating a claim's (un)verifiablity: studies that explore stylistic properties of claims to detect misinformation (Rashkin et al., 2017; Schuster et al., 2020, i.a.) and Atanasova et al. (2022) who detect when evidence with omitted information is (in)sufficient to reach a fact-checking verdict.

7 Conclusion

With this paper, we contribute a better understanding of what makes claims (un)verifiable. To this end we design a study that tasks annotators with varying levels of expertise to search for evidence and label claims that have entity/relation annotations. In an in-depth analysis of the resulting resource we find that claims with particular relations are more challenging to find evidence for and successfully assign fact-checking verdicts. *This leads to important future work, namely to focus on methods that are able to find evidence also for negated claims (X–does-not-cause–Y).* Through the study we also observe that some specific topics appear to be more challenging, including environmental factors. *We suggest that future work studies which data bases are promising sources to provide evidence for specific biomedical topics. Presumably, PubMed is not equally well suited across topics.*

We further analyze if the expertise level of annotators has an impact on the annotation quality and how far the annotations by various groups overlap. Evidently, domain expertise leads to more carefully selected evidence (which might also be more accurately selected), but not to a too large difference in verdicts. *This aspect requires further studies – how exactly does an evidence document need to relate to a claim to allow for a correct verdict? Does, perhaps, the inference procedure vary between annotators for a good reason? Following a perspectivist approach, the diversity in verdicts should be carefully investigated in future studies.*

Finally, we perform a modeling study to understand if we can develop a system supporting the annotation process. We find that a text classifier is surprisingly successful to predict if a claim is likely to have evidence – with a nearly perfect recall and a high precision. *Future work is required to study how such classifier can be used in an annotation setup. We assume that not restricting the time of annotators to find evidence for such claims would be an appropriate design decision.* Apart from that, future work should explore other modeling approaches such as few-shot prompting to explore the capabilities of LLMs for our task. With respect to both understanding claim and evidence properties as well as modeling, we further need to explore how to handle complex claims that may not follow the subject-relation-object structure we are investigating in this work.

8 Ethical Considerations

It lies in the nature of fact verification that annotators may be exposed to false medical information. We educate annotators about this possibility before they start the task. They can stop working on the task at any time. The resulting resource is first and foremost a dataset intended for analyses, designed to enable further research in modeling biomedical claim verification. It should not be the basis of in-production, fully automatic fact-checking systems. Further, biomedical research itself constantly evolves which means the evidence and verdicts in BEAR-FACT may be outdated.

It is important to point out that the fact-checking verdicts in the dataset are not to be taken out of context from the evidence document. While we can cautiously infer a veracity label for claims that are SUPPORTED or REFUTED, because we trust annotators to base their verdict on reliable evidence, there is no objective measure of truth in our task. In theory, there could be a number of reasonable evidence documents for a single claim, as medical research might produce multiple studies on the same topic.

9 Limitations

Our work studies claim properties based on real-world data. Therefore, the conclusions we draw from the analyses represent the underlying sample. The extend to which our findings generalize to other datasets should be the focus of future research. Such a manually annotated corpus can never represent the entirety of data in the real-world appropriately. While we believe that the corpus we created can be used to induce machine learning models, it might lead to biases and other unwanted confounding variables, given such limitations.

While we recruit a substantial number of annotators in the crowd experiments (§4) (9 and 10 crowd workers, respectively, who work on the same claims), the number of instances we study is comparably small. Particularly difficult or straightforward claims could have a stronger impact on annotation performance in the small study setting. It is important to contextualize our findings further. We see this as an opportunity for future work.

With respect to the evidence retrieval process, and the conclusions we draw from annotators not being able to discover evidence for a given claim, we have to consider that the PubMed knowledge base could be incomplete. Evidence could indeed exist, but not being indexed by PubMed. Further, the database does not guarantee that the documents within it are accurate or of reliable quality.

Even with thorough document filtering (§2.1), we cannot guarantee that tweets' authors always intent to make a claim as opposed to opinion statement. The latter usually are not considered to be verifiable (Merpert et al., 2018). However, following Toulmin (2003) who defines a claim as an "assertion that deserves our attention" (Toulmin, 2003) we argue that in the medical context, we need a to adopt a wide definition of what constitutes a claim. Any statement that conveys false medical information poses immediate harm and therefore deserves fact-checkers' attention.

Acknowledgments

References

Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. 2022. Fact checking with insufficient evidence. *Transactions of the Association for Computational Linguistics*, 10:746–763.

Katarina Boland, Pavlos Fafalios, Andon Tchechmedjiev, Stefan Dietze, and Konstantin Todorov. 2022. Beyond facts–a survey and conceptualisation of claims in online discourse analysis. *Semantic Web – Interoperability, Usability, Applicability*, 13(5):793–827.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Johannes Daxenberger, Steffen Eger, Ivan Habernal, Christian Stab, and Iryna Gurevych. 2017. What is the essence of a claim? cross-domain claim identification. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2055–2066, Copenhagen, Denmark. Association for Computational Linguistics.

Ryo Egawa, Gaku Morio, and Katsuhide Fujita. 2019. Annotating and analyzing semantic role of elementary units and relations in online persuasive arguments. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 422–428, Florence, Italy. Association for Computational Linguistics.

Diane Francis and Full Fact. 2016. Fast & furious fact check challenge. [Accessed on 2023-10-10].

Christopher Hidey, Elena Musi, Alyssa Hwang, Smaranda Muresan, and Kathy McKeown. 2017. Analyzing the semantic types of claims and premises in an online persuasive forum. In *Proceedings of the 4th Workshop on Argument Mining*, pages 11–21, Copenhagen, Denmark. Association for Computational Linguistics.

Yohan Jo, Elijah Mayfield, Chris Reed, and Eduard Hovy. 2020. Machine-aided annotation for fine-grained proposition types in argumentation. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1008–1018, Marseille, France. European Language Resources Association.

Byeongchang Kim, Hyunwoo Kim, Seokhee Hong, and Gunhee Kim. 2021. How robust are fact checking systems on colloquial claims? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1535–1548, Online. Association for Computational Linguistics.

Neema Kotonya and Francesca Toni. 2020. Explainable automated fact-checking for public health claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7740–7754, Online. Association for Computational Linguistics.

Marco Lippi and Paolo Torroni. 2016. Argument mining from speech: Detecting claims in political debates. *Proceedings of the AAAI Conference on Artificial Intelligence*, 30(1).

Ian Magnusson and Scott Friedman. 2021. Extracting fine-grained knowledge graphs of scientific claims: Dataset and transformer-based results. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4651–4658, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Cameron Martel, Jennifer Allen, Gordon Pennycook, and David G. Rand. 2023. Crowds can effectively identify misinformation at scale. *Perspectives on Psychological Science*. Advance online publication.

Ariel Merpert, Melina Furman, María Victoria Anauati, Laura Zommer, and Inés Taylor. 2018. Is that even checkable? an experimental study in identifying checkable statements in political discourse. *Communication Research Reports*, 35(1):48–57.

Isabelle Mohr, Amelie Wührl, and Roman Klinger. 2022. CoVERT: A corpus of fact-checked biomedical COVID-19 tweets. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 244–257, Marseille, France. European Language Resources Association.

Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. 2017. Truth of varying shades: Analyzing language in fake news and political fact-checking. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2931–2937, Copenhagen, Denmark. Association for Computational Linguistics.

Arkadiy Saakyan, Tuhin Chakrabarty, and Smaranda Muresan. 2021. COVID-fact: Fact extraction and verification of real-world claims on COVID-19 pandemic. 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 2116–2129, Online. Association for Computational Linguistics.

Mourad Sarrouti, Asma Ben Abacha, Yassine Mrabet, and Dina Demner-Fushman. 2021. Evidence-based fact-checking of health-related claims. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3499–3512, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tal Schuster, Roei Schuster, Darsh J. Shah, and Regina Barzilay. 2020. The limitations of stylometry for detecting machine-generated fake news. *Computational Linguistics*, 46(2):499–510.

Megha Sundriyal, Atharva Kulkarni, Vaibhav Pulastya, Md. Shad Akhtar, and Tanmoy Chakraborty. 2022. Empowering the fact-checkers! automatic identification of claim spans on Twitter. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7701–7715, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Stephen E. Toulmin. 2003. *The Uses of Argument*, 2 edition. Cambridge University Press.

Juraj Vladika and Florian Matthes. 2023. Scientific fact-checking: A survey of resources and approaches. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6215–6230, Toronto, Canada. Association for Computational Linguistics.

Juraj Vladika, Phillip Schneider, and Florian Matthes. 2023. Healthfc: A dataset of health claims for evidence-based medical fact-checking. *arXiv preprint arXiv:2309.08503*.

David Wadden, Kyle Lo, Bailey Kuehl, Arman Cohan, Iz Beltagy, Lucy Lu Wang, and Hannaneh Hajishirzi. 2022. SciFact-open: Towards open-domain scientific claim verification. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4719–4734, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Amelie Wuehrl, Lara Grimminger, and Roman Klinger. 2023. An entity-based claim extraction pipeline for real-world biomedical fact-checking. In *Proceedings of the Sixth Fact Extraction and VERification Workshop (FEVER)*, pages 29–37, Dubrovnik, Croatia. Association for Computational Linguistics.

Amelie Wührl and Roman Klinger. 2021. Claim detection in biomedical Twitter posts. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 131–142, Online. Association for Computational Linguistics.

Amelie Wührl and Roman Klinger. 2022a. Entity-based claim representation improves fact-checking of medical content in tweets. In *Proceedings of the 9th Workshop on Argument Mining*, pages 187–198, Online and in Gyeongju, Republic of Korea. International Conference on Computational Linguistics.

Amelie Wührl and Roman Klinger. 2022b. Recovering patient journeys: A corpus of biomedical entities and relations on Twitter (BEAR). In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4439–4450, Marseille, France. European Language Resources Association.

Shi Yuan and Bei Yu. 2019. Hclaime: A tool for identifying health claims in health news headlines. *Information Processing & Management*, 56(4):1220–1233.

Chaoyuan Zuo, Kritik Mathur, Dhruv Kela, Noushin Salek Faramarzi, and Ritwik Banerjee. 2022. Beyond belief: a cross-genre study on perception and validation of health information online. *International Journal of Data Science and Analytics*, pages 1–16.

A Appendix

A.1 Data Processing

Table 8 shows the filtering criteria for removing irrelevant claims before the fact-checking annotation.

Table 9 shows the filtering process described in Sec. 2.1 using an example.

A.2 Annotation study

A.2.1 Study disclaimer

Crowd annotators obtain the following description and task disclaimers when starting the study:

"**Purpose of this Study** We want to understand how people check if a biomedical statement is true or not and how they find evidence to judge those statements. **Your Task** You will be presented with a Twitter post which contains a biomedical claim. You will fact-check this claim by searching for relevant evidence on PubMed, a database of biomedical articles. The study should take you about 1 hour to complete. Please be aware that you might be looking at claims which convey false biomedical information. You can stop the study at any time. Note that you won't be paid in this case. **Data Collection** The data we collect will not contain any personal information. It will be used for researching automatic fact-checking and made publicly available in an anonymized form. We will write a scientific paper about this study which can include anonymized examples from the collected data."

A.2.2 Annotator compensation

In-house annotators are compensated with 12,52 € per hour. The crowdworkers on Prolific are compensated with £9 per hour, which corresponds to the recommended amount on the platform.

A.2.3 Annotator screening on Prolific

We add the following screeners to recruit participants for the Prolific study:

- Fluent languages: English

- Highest education level completed: Technical/community college, Undergraduate degree (BA/BSc/other), Graduate degree
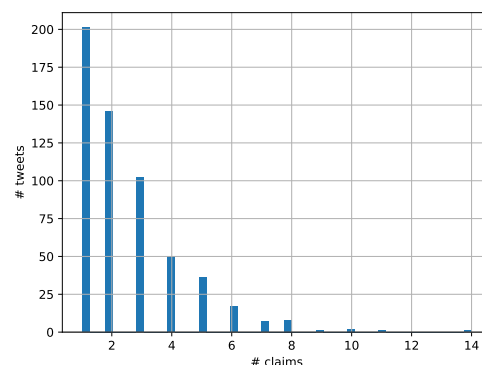

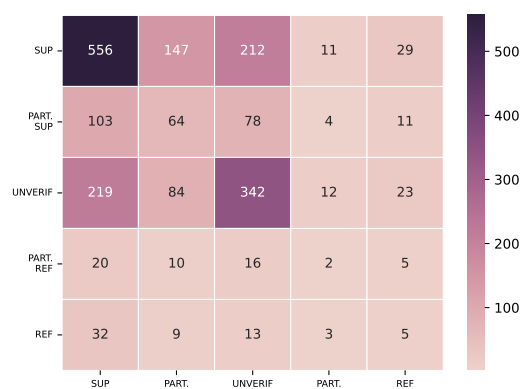Figure 3: Number of claims per tweet in BEAR-FACT.

Figure 4: Pairwise co-occurrence of fact-checking verdicts in BEAR-FACT tweets with more than one claim.

(MA/MSc/MPhil/other), Doctorate degree (PhD/other)

- Subject: Biochemistry (Molecular and Cellular), Biological Sciences, Biology, Biomedical Sciences, Chemistry, Dentistry, Health and Medicine, Medicine, Nursing, Pharmacology, Science

A.3 Corpus statistics

A.3.1 Number of claims per tweet

Figure 3 visualizes the number of claims per tweet.

A.3.2 Verdict co-occurrence

Figure 4 shows the pairwise co-occurrence of fact-checking verdicts in BEAR-FACT tweets with more than one claim.

A.4 Aggregation

When computing the majority vote for the fact-checking verdicts, we collapse the PARTIALLY SUPPORTED and PARTIALLY REFUTED verdicts into one group, respectively.

Criterion	Description
Incorrectly extracted from the tweet's context	Instances for which the entity-based claim extraction lead to the original statement being mis-represented, e.g., if relevant context is omitted when extracting the claim triplet.
Repetitions	Removing claim duplicates in the same tweet. The repetions are artifacts of the annotation aggregation strategy employed in BEAR (Wührl and Klinger, 2022b).
Contain relation "somehow related to"	Removing claims with this relation, because they are highly unspecific and therefore not check-able.
Off-topic claims	Claims that discuss medical conditions in animals.

Table 8: Filtering criteria for removing irrelevant claims before the fact-checking annotation.

Filtering step	Method	Result for example instance
Contains a claim	Claim classifier (Wührl and Klinger, 2021)	True
Contains <1 med. relation	Annotations in BEAR (Wührl and Klinger, 2022b)	True
Claim extraction	Extracting entity-based claims (Wührl and Klinger, 2022a)	'Females' negative influence on 'leukopenia'
Manual filtering	Manual inspection	passed
Correcting grammar	Manual	Being female has a negative influence on leukopenia.

Table 9: Data filtering process exemplified with an instances from the dataset. The input tweet reads: *Females tend to have greater relapses, leukopenia, more arthritis, and Raynaud phenomenon.* The claim we obtain as a result of the final filtering step is what the annotators verify during the a study. Note that we provide them with the claim as well as with the full tweet for context.

A.5 Evidence refinement

Table 10 shows example claims along with their confidence ratings w.r.t. if they think evidence exists and could be discovered given unlimited time and resources for research.

id	Claim	Ev. Exists Confidence
1	anti-everythings are the cause of being moodier	very sure *no* ev. exists
2	Strontium in chem trails is/are the cause of Cancer	pretty sure *no* ev. exists
3	trapped in fires is/are the cause of ptsd	pretty confident ev. *exists*
4	Magnesium glycinate 100-200mg has a pos. influence on immune system	very confident ev. *exists*

Table 10: Example claims and annotators' confidence ratings w.r.t. if evidence exists.