FEVEROUS: Fact Extraction and VERification Over Unstructured and Structured information

Rami Aly, Zhijiang Guo, Michael Schlichtkrull, James Thorne, Andreas Vlachos
Department of Computer Science and Technology
University of Cambridge
{rmya2,zg283,mss84,jt719,av308}@cam.ac.uk

Christos Christodoulopoulos
Amazon Research Cambridge
chrchrs@amazon.co.uk

Oana Cocarascu
Department of Informatics
King's College London
oana.cocarascu@kcl.ac.uk

Arpit Mittal
Facebook
arpitmittal@fb.com

A Supplementary Material

A.1 Access to the dataset

The FEVEROUS dataset can be accessed from the official website of the FEVER Workshop `https://fever.ai/dataset/feverous.html` and is hosted in an the same AWS S3 Bucket as the FEVER dataset, which has been publicly available since 2018. As the authors of this paper manage the workshop's website they can ensure proper maintenance and access to the dataset. The hosting of the dataset includes a retrieval corpus, as well as each split of the dataset. At the time of this paper's submission, the shared task is still ongoing with the unlabeled test set being kept hidden until the last week of the shared task. Elementary code to process the data from both annotations and the provided Wikipedia DB (e.g. extracting context for a given element, getting a table from a cell ID etc..) is publicly available on `https://github.com/Raldir/FEVEROUS`. The repository also contains the code of the annotation platform as well as the baseline's code. The DOI for the FEVEROUS dataset is `10.5281/zenodo.4911508` and structured metadata has been added to the webpage.

The training and development data is hosted in Jsonlines format. Jsonlines contains a single JSON per line, encoded in UTF-8. This format allows to process one record at a time, and works well with unix/shell pipelines. Each entry consists of five fields: The training and development data contains 5 fields:

- id: The ID of the sample
- label: The annotated label for the claim. Can be one of SUPPORTS|REFUTES|NOT ENOUGH INFO.
- claim: The text of the claim.
- evidence: A list (at maximum three) of evidence sets. Each set consists of dictionaries with two fields (content, context).
  - content: A list of element ids serving as the evidence for the claim. Each element id is in the format "[PAGE ID]_[EVIDENCE TYPE]_[NUMBER ID]". [EVIDENCE TYPE] can be sentence, cell, header_cell, table_caption, item.

35th Conference on Neural Information Processing Systems (NeurIPS 2021) Track on Datasets and Benchmarks.

  - context: A dictionary that maps each element id in content to a set of Wikipedia elements that are automatically associated with that element id and serve as context. This includes an article's title, relevant sections (the section and sub-section(s) the element is located in), and for cells the closest row and column header (multiple row/column headers if they follow each other).

- annotator_operations: A list of operations an annotator used to find the evidence and reach a verdict, given the claim. Each element in the list is a dictionary with the fields (operation, value, time).
  - operation: Any of the following
    * start, finish: Annotation started/finished. The value is the name of the operation.
    * search: Annotator used the Wikipedia search function. The value is the entered search term or the term selected from the automatic suggestions. If the annotator did not select any of the suggestions but instead went into advanced search, the term is prefixed with "contains..."
    * hyperlink: Annotator clicked on a hyperlink in the page. The value is the anchor text of the hyperlink.
    * Now on: The page the annotator has landed after a search or a hyperlink click. The value is the PAGE ID.
    * Page search: Annotator search on a page. The value is the search term.
    * page-search-reset: Annotator cleared the search box. The value is the name of the operation. Highlighting, Highlighting deleted: Annotator selected/unselected an element on the page. The value is ELEMENT ID.
    * back-button-clicked: Annotator pressed the back button. The value is the name of the operation.
    * value: The value associated with the operation.
    * time: The time in seconds from the start of the annotation.
- expected_challenge: The challenge the claim generator selected will be faced when verifying the claim, one out of the following: Numerical Reasoning, Multi-hop Reasoning, Entity Disambiguation, Combining Tables and Text, Search terms not in claim, and Other.
- challenge: The main challenge to verify the claim, one out of the following: Numerical Reasoning, Multi-hop Reasoning, Entity Disambiguation, Combining Tables and Text, Search terms not in claim, and Other.

The retrieval corpus is provided to annotators in either Jsonlines format, or as an SQLite3 database. The latter allows faster retrieval for articles by their name, which is helpful for instance when mapping annotation ids to their contents. Each Wikipedia article contains 2 base fields:

- title: The title of the Wikipedia article
- order: A list of elements on the Wikipedia article in order of their appearance. Elements can be: section, table, list, sentence.

Each element specified in order is a field. A sentence field contains the text of the sentence.

A section element is a dictionary with following fields:

- value: Section text
- level: The level/depth of the section.

A table element is a dictionary with following fields:

- type: Whether the table is an infobox or a normal table
- table: The content of the table. The table is specified as a list of lists. Each element in a list is a cell with the fields (id, value, is_header, row_span, column_span).
- caption: Only specified if the table contains a caption.

A list element consists of following fields:

- type: Whether the list is an ordered or unordered list
- list: A list of dictionaries, with fields being (id, value, level, type). level is the depth of the list item. The level increments with each nested list. type specifies type of a nested list, which is starting after the item specifying the type. Field is only specified if the next item is in a nested list. Hyperlinks in text are indicated with double square brackets. If an anchor text is provided, it is the text on the right hand side of a vertical bar in the square backets

Example dataset and retrieval corpus entries can be found on `https://fever.ai/dataset/feverous.html`.

A.2 Ethics statement

The FEVEROUS dataset was collected with approval and following the practices outlined by the Ethics Committee of the Computer Lab of the University of Cambridge (reference number 1842). Furthermore, the external contractor has a well-outlined policy regarding their code of ethics to ensure the well-being of all annotators in our experiment. Their Code of Ethics consists of: Fair Pay, Inclusion, Crowd Voice (i.e. Feedback mechanisms), Privacy and Confidentiality, Communication, and Well-Being.

We anticipate that FEVEROUS will be used for the development of fact checking systems that might be applied in real world contexts to assign truth/false labels, similar to those on fact checking websites run by journalists. We use the labels supported/refuted (by evidence) instead of true/false to be clear that we do not make any judgements about the truth of a statement in the real-world, but only consider Wikipedia as the source of evidence to be used. And while Wikipedia is a great collaborative resource, it has mistakes and noise of its own similar to any encyclopedia or knowledge source. Thus we discourage users of FEVEROUS to make absolute statements about the claims being verified, i.e. avoid using it to develop truth-tellers. Finally, we require systems to predict when the evidence is not sufficient to make a judgement, in which case it would be useful to look beyond Wikipedia for evidence.

We did not collect personal data of the participants in any way. A participant is only identified using an identification number to access our online tool. Generated claims must only include information on Wikipedia or considered to be general world knowledge, while all evidence is taken from Wikipedia directly, thus not including any personally identifiable information or offensive content.

A.3 Data Statements

We follow the data statements structure of Bender and Friedman [2018] to give additional insights into the dataset and its construction.

Curation Rationale. In order to study fact extraction and verification on both unstructured and structured information, we use the entire English Wikipedia as the knowledge base. Wikipedia is a large-scale collaboratively created encyclopedia, covering a large extent of knowledge/topics and is as such considered to be a suitable testbed for our purpose. Only articles that have been flagged by Wikipedia to have issues, miss references and/or citations have been excluded. The rationale behind this decision is to compile a retrieval corpus with information that is consistent across pages. The entire content of an article is considered, with exception to sections that were flagged as aforementioned as well sections that are named 'References', 'Citations', 'Sources', 'Further reading', 'External links', 'Works', 'Gallery', 'Citations and references', 'Bibliography', or 'External links & References' as we consider these sections to be out-of-scope for our task. Sentence and Table highlights given to annotators were sampled randomly from the entire collection of English Wikipedia articles.

Language Variety. The extracted evidence aligns with English Wikipedia's characteristic on language variety. A section on this, describing the lack of standardization can be found here. For claim generation, half of the annotators were native US-English speakers, while the other half were English speakers from the Philippines. For claim verification, all annotators were native US-English speakers. The internal screening by the external contractor ensured that the variety of English used is very similar across annotators, being en-us.

Speech Situation. The retrieval corpus was compiled based on a December 2020 version of English Wikipedia. Wikipedia is a collaborative encyclopedia, and as such regularly edited. Wikipedia describes in detail the requirements and recommendation of texts in articles, which can be found for instance here, here, and here. Claims were generated between March and May 2021, with very detailed guidelines regarding content and structure. A claim is described in the guidelines as *a single well-formed sentence. It should end with a period; it should follow correct capitalization of entity names (e.g. 'India', not 'india'); numbers can be formatted in any appropriate English format (including as words for smaller quantities).* They further *must* ***not be subjective*** *and be* ***verifiable*** *using publicly available information/knowledge*. Claims further should as unambiguous as possible, and must not contain any idioms, figures of speech, similes, or verbose language (see Section A.9).

Text characteristics. Since highlights were sampled randomly from Wikipedia articles, the distribution of topics of generated claims roughly corresponds to the underlying English Wikipedia distribution of articles (i.e. people, geography, history, and sports being the main topics). We restrict the topic in some instances, such as: *Claims should not be about contemporary political topics (e.g. contemporary Wars (from the second world war and onwards), or disputed topics).*

Annotator Demographic Annotator candidates were screened specifically for our task, with multiple screening and calibration-stages as described in the paper. This ensures that annotators are aware of the constraints and guidelines when generating claims and verifying them. All annotators were paid above their local minimum wage.

- Age: Claim generation: 11 people between 18-24 years, 20 people between 25-34 years, 8 people between 35-44 years, 6 people between 45-54, and 12 people unspecified. Claim verification: 4 people between 18-24 years, 17 people between 25-34 years, 9 people between 35-44 years, 12 people between 45-54, 5 people between 55-64, and 12 people unspecified.
- Gender: Claim generation: 11 male, 42 female, and 4 unspecified. Claim verification: 15 male, 36 Female, and 3 unspecified.
- Race/ethnicity: -
- Native language: Claim generation: 33 people are native en-us speakers, 24 annotators are native en-ph (English (Philippines)) speaker. Claim verification: All annotators are native en-us speakers.
- Socioeconomic status: -
- Training in linguistics/other relevant discipline: Claim generation: English speakers from the Philippines are language-aware (an upper education degree in a language-related subject). Claim verification: all annotators are language-aware.

A.4 Licensing

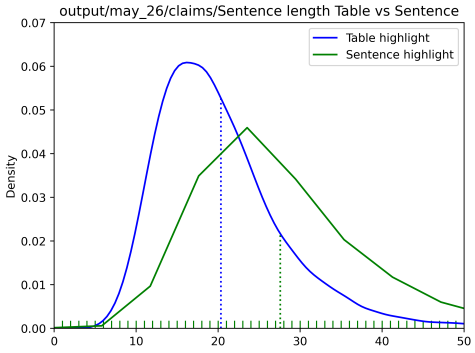
These data annotations incorporate material from Wikipedia, which is licensed pursuant to the Wikipedia Copyright Policy. These annotations are made available under the license terms described on the applicable Wikipedia article pages, or, where Wikipedia license terms are unavailable, under the Creative Commons Attribution-ShareAlike License (version 3.0), available at `http://creativecommons.org/licenses/by-sa/3.0/` (collectively, the "License Terms"). You may not use these files except in compliance with the applicable License Terms. Credits to the contents of a page go to the authors of the corresponding Wikipedia article. Since article names in the dataset are unchanged, the authors can be found on the respective article on Wikipedia (`https://www.wikipedia.org/wiki` + TITLE_ID). The associated code to FEVEROUS (i.e. annotation platform, baseline code) are licensed under Apache 2.0.

A.5 Detailed dataset and annotation statistics

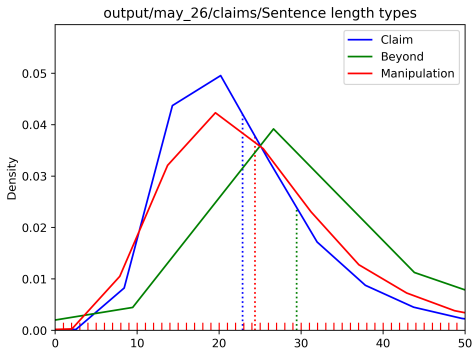
A.5.1 Claim generation

An average annotation (i.e. generating three claims) took an annotator 373 seconds. A total of 61058, and 32700 claims were created using table and sentence highlights, respectively. 47300 annotations were prompted to use information from the same page, and 46428 from different pages. The average

length of a claim is 23, 29, and 24 for Type I, Type II and Type III, respectively. Annotators used on average 0.71 hyperlinks and 0.15 search queries. For Type II claims that require multiple pages, annotator used on average 1.2 hyperlinks and 0.2 searches. Sentence length by claim type is shown in figure 1a. Average sentence length for both table and sentence highlights is seen in figure 1b.



(a) Sentence length of claims given sentence versus table highlight.

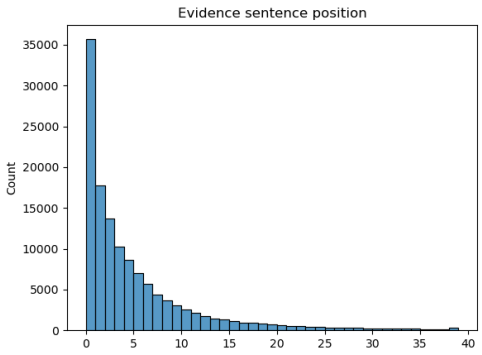


(b) Sentence length for different claim types.

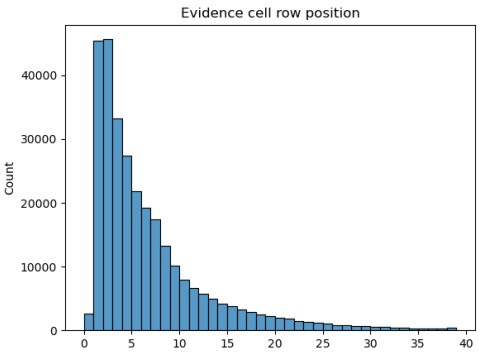
A.5.2 Claim verification

A single claim verification took on average 165 seconds. Claims are selected uniformly from the pool of different claim types, resulting in an claim verification set of about equal claims for each claim type. On average an annotation has 1.1 evidence sets, with a total of 7468 annotations having more than one evidence set. Annotators needed on average 1.34 search queries and 0.72 hyperlinks. On average 0.1 advanced searches were used (i.e. searches for which no of the given page suggestions matches, so that annotators had to go to the advanced search page that uses 'in page' matches with Elasticsearch). In about 84% of claims do the pages from which evidence was retrieved directly match a word or phrase in the claim itself. 69% of all pieces of evidence are table cells, 29% are sentences, 1% are list items, and 1% are table captions.

Plot 2a and 2b show the evidences' sentence positions and row positions of cells in tables, respectively. Plot 3a shows the distribution of evidence numbers in the dataset. Plot 3b shows the section number where evidence is located, with -1 being the introduction section.



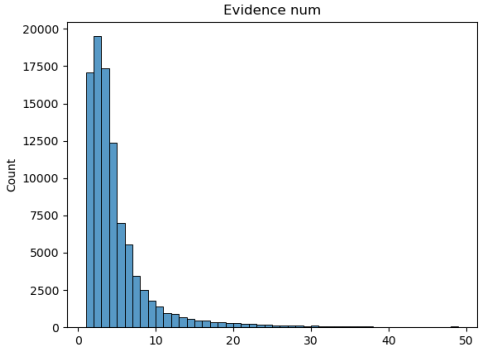
(a) Sentence position distribution in evidence.



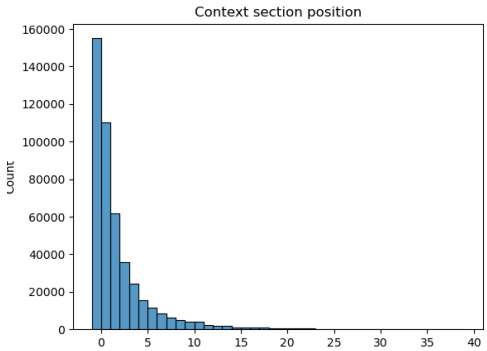
(b) Row position of cells in tables in evidence.

A.5.3 Claim Verification Challenges

Table 1 shows the distribution of verification challenges in the FEVEROUS dataset, both the expected challenges as selected by the claim generators as well as the verification challenges by the verification annotators. The latter constitutes the actual distribution of challenges in FEVEROUS. As seen, the



(a) Distribution of number of evidence pieces in an evidence set.



(b) Section position of evidence pieces.

distribution is relatively similar across the splits, with about 10% of all claims having numerical reasoning, 16% multi-hop reasoning, 14% Combining Tables and text, 2% Entity Disambiguation, and 1.3% Search terms not in claim as their main verification challenge. Figure 4 shows the confusion matrix between expected and actual challenges, normalized along the x-axis. It is apparent that the claim generators overpredicted *Other* as the main challenge, indicating that the generators were frequently not aware of the challenge their claim poses when generating them. This particularly applies to *Entity Disambiguation* and *Search terms not in claim*, which have almost never been predicted correctly by the generators, most likely due to the generators not searching for the pages themselves, but are given a highlight on a page to generate their claim. Interestingly, there is also some discrepancy between numerical claims as well as claims that need both tables and text. This might be explained by information redundancy, having generated claims using both tables and text, not knowing that there is a sentence that contains the information of both. Further analyzing the challenges might lead to highly interesting insights on interactions and discrepancies between the expected difficulties from someone generating a claim and an annotator actually verifying it.

Following we show an example claim for each challenge category, taken from the dataset:

- **Numerical Reasoning** As of the 2011 Indian census, Nimbapur –located in the Indian state of Maharashtra, which is the second-most populous Indian state – has a population of 1903, with nearly half of the residents being non-workers. (*Calculation of the ratio between total population and residents who are non-workers*)
- **Multi-hop Reasoning** Belgium's Léon Schots, a Belgian former long-distance runner who competed in track and cross country running competitions, was the fastest athlete in the senior men's race (12.3km) at the 1977 IAAF World Cross Country Championships. (*Evidence to verify the claim are from two different articles*)
- **Entity Disambiguation** VUKOVI is a rock band from Scotland that plays pop rock, noise pop music and is formerly called Wolves. *Disambiguation of the term Wolves*
- **Search terms not in claim** In 2011, Evans signed with the Cincinnati Bengals after going undrafted in the NFL draft; but in November 2011, Evans was suspended for four games. (*To retrieve evidence, annotator first searched for any page containing "Evans signed with the Cincinnati Bengals", until finding the page for the entity's full name "DeQuin Evans".*)
- **Combining Tables and Text** Braeden Lemasters, an American actor, musician, and voice actor, appeared in six films since 2008 and also appeared in TV shows such as Six Feet Under where he starred as Frankie. (*Needing evidence from both tables and text*)
- **Other** Aquarion Logos is an anime series produced by Satelight which is a Japanese animation studio which serves as a division of pachinko operator Symphogear Group. (*Neither of the above five challenges apply*)

Challenge Category	Train	Dev	Test	Total
Expected Challenges				
Numerical Reasoning	7798	1024	842	9664
Multi-hop Reasoning	17248	1871	2011	21130
Entity Disambiguation	826	143	77	1046
Combining Tables and Text	7775	975	769	9519
Search terms not in claim	405	57	90	552
Other	37239	3820	4056	45115
Verification Challenges				
Numerical Reasoning	7214	873	740	8827
Multi-hop Reasoning	11624	1281	1195	14100
Entity Disambiguation	1353	201	200	1754
Combining Tables and Text	10083	1035	940	12,058
Search terms not in claim	824	131	193	1148
Other	40193	4369	4577	49139

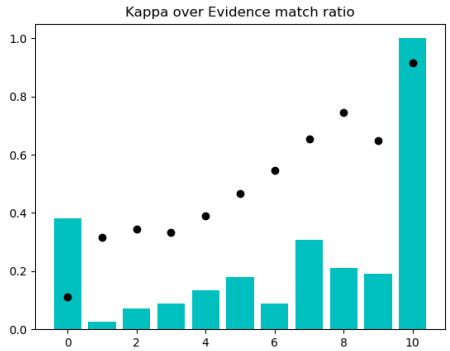
Table 1: Distribution of verification challenges in the FEVEROUS dataset. Top: Expected verification challenges, selected during claim generation. Bottom: Verification challenges, selected by annotator after a claim was verified.

Figure 4: Confusion matrix for expected challenges versus actual challenges. Numbers are normalized across the x-axis.

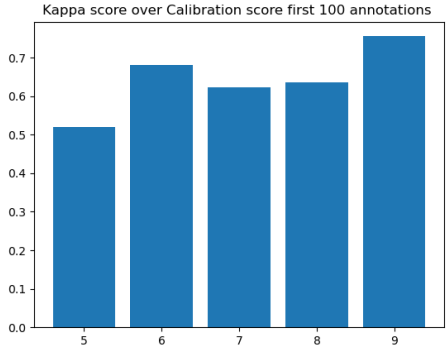
A.5.4 Details on QA statistics

In addition to the overall agreement, we measured the annotator agreement over evidence match ratios. As seen in Figure 5a, in the case of exact evidence match, the kappa agreement κ is 0.92, linearly decreasing with an agreement of 0.11 in the case of completely distinct evidence.

We further measured the annotation agreement sorted by annotator calibration score (more specifically the verdict accuracy) for the first 100 full-scale annotations of an annotator. As seen in Figure 5b, annotators with a calibration score of over 0.9 have an overall higher kappa score, of around 0.8 while annotators with a score of below 0.6 only achieved a kappa score of about 0.5. This indicates that the calibration score is indicative of the performance of annotators in the beginning. Looking at the score over all annotations, we however noticed, that annotators continue to align their annotations with annotators having very similar agreement irregardless of calibration score (except annotators with very calibration score above 0.9 still having higher agreement).



(a) Kappa score κ measured for QA-claims with different evidence match ratio. Blue bars indicate the proportion of total annotations with that evidence match. The match is calculated as as $\frac{2*|E_1 \cap E_2|}{|E_1|+|E_2|}$. Shown x-values (similarity) are multiplied by ten.



(b) Kappa agreement κ over calibration score for the first 100 full-scale annotations. Shown x-values (calibration scores) are multiplied by ten.

A.6 Dataset Processing & Implementation Details

A.7 Dataset Processing

Wikipedia articles were split into sentences using the NLTK unsupervised sentence tokenizer[1]. We trained the unsupervised tokenizer on Wikipedia text to extract a large list of abbreviation words used on Wikipedia. These can be simply abbreviations of names (e.g. John F. Kennedy) or glossing abbreviations (for instance 'e.g.'). Due to the extensive use of Wikipedia templates for tables and the difficulty in resolving/parsing them, we opted in extracting articles from Wikipedia directly. We used Scrapy for this [2] and maintained a date stamp for each site. We limited the extracted tables to the classes 'wikitable' and 'infobox'. This restriction was set as there are contents of Wikipedia categorized as HTML tables while being highly specifically formatted, such as climate tables or tournament brackets. FEVEROUS maintains the diversity of Wikipedia tables/lists, only filtering ones out with formatting errors or that are empty (e.g. due to only containing images).

The FEVEROUS Wikipedia retrieval corpus was processed by keeping only hyperlinks with an associated article in the corpus. We replaced hyperlinks that are references to redirect pages with the respective page that the redirect page references to. URLS are replaced with a special token and text has been cleaned using the clean-text library[3].

[1] `https://www.nltk.org/api/nltk.tokenize.html`
[2] https://scrapy.org/
[3] `https://pypi.org/project/clean-text/`

For the annotation platform, we populated a MediaWiki 1.31 database with the extracted articles as well as Wikipedia redirects. We installed the CirrusSearch extension [4] to enable the search engine to use Elasticssearch as the back-end search. The annotations where stored in an SQL database using MariaDB.

A.7.1 Implementation & Evaluation Details

Retriever. We use Spacy[5], specifically the `en_core_web_sm` model) to extract entities from claims. We match extracted entities against all titles of our Wikipedida database and extract pages with an exact match. The TF-IDF part of our retriever is largely based on DrQA [Chen et al., 2017], computing the cosine similarity between the binned unigram and bigram TF-IDF vectors of claim and the introductory section of a Wikipedia article. The same TF-IDF approach is used to extract sentences and tables, however, restricted to the top k extracted pages. We excluded lists from the retrieval for our baseline to minimize computation time, considering that only 1% of annotated evidence is located in lists.

The cell retrieval model uses pre-trained RoBERTa$_{base}$ from Huggingface[6]. Parts of the table that were longer than the maximum input length of RoBERTa were simply cut-off. To prevent this from happening during training we use row-sampling. We concatenate rows that contain relevant cell evidence first, before considering irrelevant rows.

The cell retrieval RoBERTa classifier was fine-tuned using binary cross-entropy. The batch-size was set to 16, with weight decay of 0.1, a learning rate of $5e^{-5}$, and a total of 1 training epochs. These hyperparameters are largely taken from recommendations [Devlin et al., 2019] and have not been further fine-tuned as the baseline's purpose is not to achieve the highest possible scores, but rather to provide a working, intuitive model that motivates further exploration of the dataset. As such the models are not the main part of the paper.

Verdict predictor The verdict predictor uses RoBERTa$_{large}$, particularly the model pre-trained on multiple NLI datsets by Nie et al. [2020], which can be found here. Each piece of evidence is separated using `</s>`. We linearize cell evidence the following : `[CONTEXT-HEADER] is [CELL]`, similar to [Schlichtkrull et al., 2020]. Our model is fine-tuned using a batch-size of 16, a weight decay of 0.01, a learning rate of $1e^{-5}$ for 1 epoch. Similar to the cell retrieval model, these values are largely taken from reference and have not been fine-tuned. Same rationale here as stated above.

Experiments using RoBERTa have been repeated twice and the average was reported, with very low variance (around $2e^{-5}$). All experiments were done in Python3.7. We fine-tuned all models on a single Quadro RTX 8000. Fine-tuning the cell extractor took around 1.5h, while fine-tuning the verdict predictor took around 4h. The TF-IDF retrieval needed around 10h on a Xeon Gold 5218 8 cores.

A.8 Annotation details

The annotation process to create FEVEROUS is visualized in Figure 6.

A.8.1 Annotation interfaces

Navigation To find relevant pages annotators can make use of the MediaWiki search functionality, a custom page search functionality, as well as hyperlinks. We aimed to create an ecosystem as realistic as possible, so annotators were motivated to approach this problem naturally: *How would you search for relevant information to check the truthfulness of a statement/claim given to you?* The search bar shows relevant articles to annotator's search as soon as they start typing. They are further allowed to use the given recommendations and entering the main search page (i.e. clicking on 'Containing ...'). Tere are three kinds of hyperlinks: i) Hyperlinks embedded in a sentence, table or list, ii) Hyperlinks in the content box of each Wikipedia article, iii) Hyperlinks below section headers that refer to the main article or to a more specialized article. Moreover, annotators had the option modify previous annotations.

[4] `https://www.mediawiki.org/wiki/Extension:CirrusSearch`

[5] `https://spacy.io/`

[6] `https://huggingface.co/`

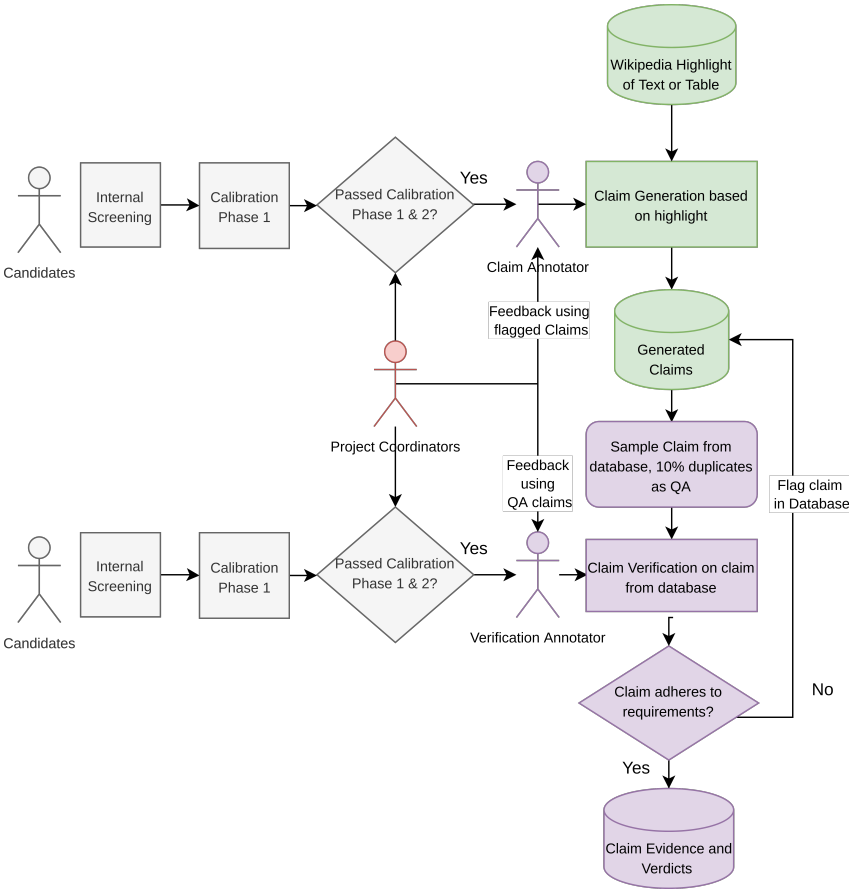


Figure 6: A schematic view on the annotation process of FEVEROUS. The claim generation phase is highlighted in green, the claim verification phase is noted in purple, and the screening of annotators is highlighted in gray.

Operating the search engine The search engine allows annotators to simply type in words or phrases that they are looking for. If they type in a query into the engine it will show them suggestions if a **title matches the query**. If it cannot find a matching title, they can start a "full text search" i.e. **searching through the actual content of a page by clicking on "*containing...*" on the very bottom of the suggestions**. Doing so redirects them to a search page, with suggestions and highlights in articles where the query could be (partially) matched. While queries can simply be words or phrases, annotators could further modify their search queries with some operators (see Annotation Guidelines, however, these have been used only very rarely.

A.9 Claim generation

A.9.1 Guidelines

Generating Claim using highlight (Type I) The first claim should **exclusively use information from the highlighted table/sentences**. Only the page title and/or section title the highlight is located might be used for the claim as well. The claim must either align with the contents in the highlight or contradict them, indicated on the tool (i.e. true and false claims). A claim should adhere to following requirements:

- A claim based on a table highlight should combine information of multiple cells if possible. This includes comparisons (e.g. *X scored higher/lower than Y*, or *While X was the son of Z, Y was the son of Q.*), superlatives (*X scored the highest/lowest*, or *X was the first Japanese supercomputer.*), filters (*X, Y and Z scored more than 10 points*, or *X, Y, Z are manufactured*

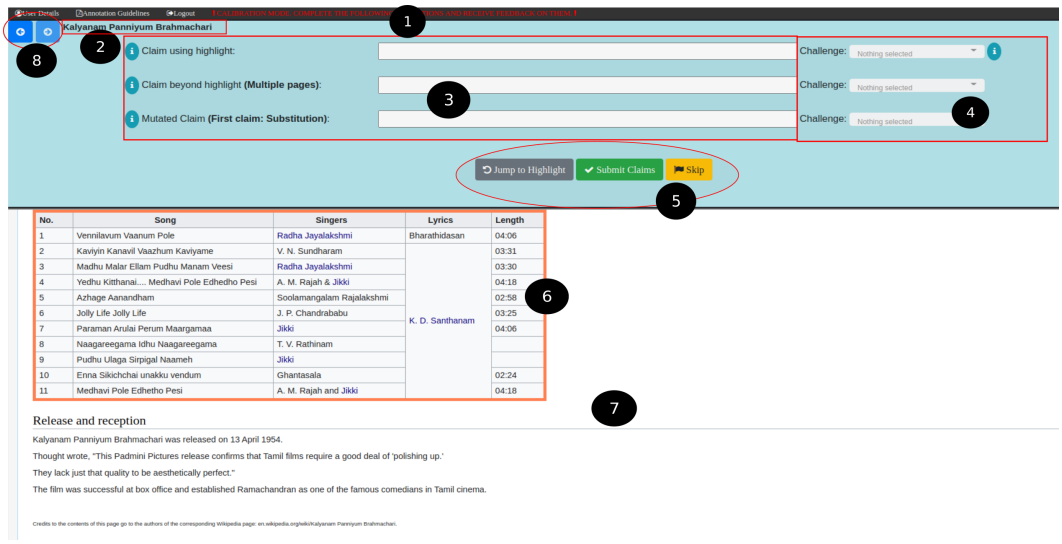


Figure 7: Claim generation interface. ① Menu bar with Access to User Information, Annotation Guidelines, and Logout. ② Title of current Wikipedia article. ③ Text field for writing claims. ④ Selection of expected challenges. ⑤ Buttons for i) Jumping back to the Wikipedia highlight, ii) submitting the written claims and selected challenges, iii) skip the current highlight. ⑥ Wikipedia article and navigation. ⑦ Article highlight to base the claims on. ⑧ Move between previously annotated claims.

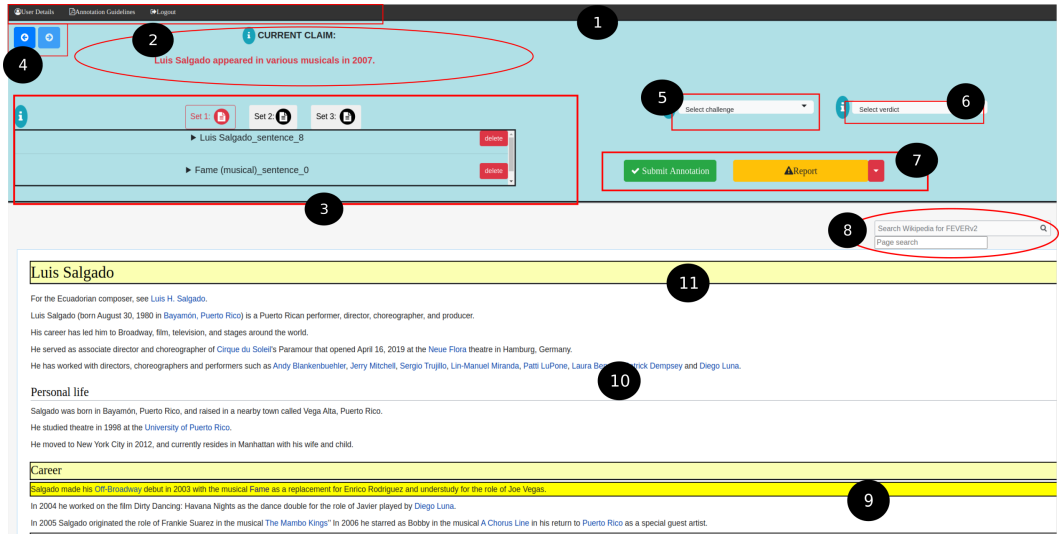


Figure 8: claim verification Interface. ① Menu bar with Access to User Information, Annotation Guidelines, and Logout. ② Current claim to verify and retrieve evidence for. ③ Move between previously annotated evidence. ④ Management of selected evidence. ⑤ Specifying annotation challenges ⑥ Selection of the claims veracity (Supported, Refuted, NotEnoughInformation) ⑦ Button for submitting annotating/reporting claim ⑧ Search bars for i) navigating through Wikipedia articles, ii) information filtering within a Wikipedia page. ⑨ WikiMedia interface ⑩ Selected evidence (yellow highlighting) ⑪ Corresponding evidence context (lighter yellow highlighting).

in Germany.), and arithmetic operations (*5 teams scored more than 10 points*, or *X was born 2 years and 8 months before Y.*).

- A claim based on highlighted sentences should not simply paraphrase a highlighted sentence or concatenate sentences. Instead, information of multiple sentences must be combined. Information from at least two sentences must be used for generating the claim.

- A claim should be a single well-formed sentence. It should end with a period; it should follow correct capitalization of entity names (e.g. 'India', not 'india'); numbers can be formatted in any appropriate English format (including as words for smaller quantities).
- Generated claims must **not be subjective** and be **verifiable** using publicly available information/knowledge.
  Don't: John Lennon was a more popular musician than Tommy Moore.
  Do: John Lennon's discography sold two times as many box sets as Tommy Moore in 1997.
  Don't: Sea Songs by Yadollah Royaee (born in 1932) is rich in symbolism and is deeply inspired by Persian mysticism. *Do*: Sea Songs by Yadollah Royaee (born in 1932) contains symbolism and is inspired by Persian mysticism.
- The claim should be **as unambiguous as possible** and avoid vague or speculative language (e.g. might be, may be, could be, rarely, many, barely or other indeterminate count words)
  Don't: The Olympic Games have rarely taken places in Europe
  Do: The Olympic Games were held three three times in Europe. *Don't*: Michael Ballack scored the most goals.
  Do: Michael Ballack scored the most goals in the Bundesliga 2004/2005 season.
- A claim must not contain any idioms, figures of speech, similes, or verbose language. *Don't*: The scientist Mary Lamb owned five sheep with fleece as black as coal, but they were not used in any of her experiments.
  Do: The scientist Mary Lamb owned five sheep with black fleece, but they were not used in any of her experiments.
- The claim must be understood by itself (i.e. no pronouns) – *[Note: in the case where the highlighted text does not contain a mention of the entity at question, you should use the title of the page or the header of the section for that information].*
  Don't: He played most of his football career for Chelsea.
  Do: Didier Drogba played most of his football career for Chelsea.
- Claims should not be about contemporary political topics (e.g. contemporary Wars (from the second world war and onwards), disputed topics) – skip pages where the highlighted area only discusses such topics.
  Don't: In 1974 Turkey had landed 30,000 troops on Cyprus and captured Kyrenia.
- In some cases highlighted Wikipedia information is not correct/consistent. These highlights are still valid for claim generation. For this workflow don't worry about the factual correctness of Wikipedia. If you think that the highlighted information is disputed, better skip it.
- Do not incorporate your own knowledge, believes or additional world knowledge into the claim. Focus only on the highlighted Wikipedia section given to you!

Generating Claim beyond the highlight (Type II) The second claim should be based on the highlight, but **must include information beyond the highlighted table/sentences**. You are free in deciding to modify the previously created claim that uses only the highlight or to create an unrelated one (that still includes information from the highlight). Either way, the new claim must still adhere to the requirements mentioned above. The new claim can either be supported or refuted. So in general, you should not worry whether the new claim preserves the truth value of the first claim. However, please keep in mind that we aim for having a similar number of positive vs negative claims. Information to include must either be on the same page or from other Wikipedia pages, indicated on the tool:

1. **Same page:** Include information outside of the highlight but on the same page.
2. **Multiple pages:** Include information from other Wikipedia page(s). You can search freely through Wikipedia using the search function, available hyperlinks on the pages, and the *Return to highlight* button.

Moreover, for this claim it is allowed to use information/knowledge that might not be available in Wikipedia but you assume to be general knowledge, e.g. that 90s refers to the timespan from 1990 to 1999. Similarly to the previous claim, the claim can either align with the used information or

contradict it. We encourage you to create claims that are based on a combination of structured and unstructured information: tables, sentences, lists, captions, or section titles.

Example 1:
Claim using highlight: The Zuse Z3 was program-controlled by punched 35mm film stock.
Claim using more than highlight: Programs were executed on the Zuse Z3 by using punched 35mm film stock with manually entered initial values.

Example 2:
Claim using highlight: The player with the most number of total assists at Shrewsbury Town F.C in 2013 is Luke Summerfield.
Claim using more than highlight: The player with the most number of total assists at Shrewsbury Town F.C in 2013 also played for Liverpool.

Example 3:
Claim using highlight: Jeff Gordon had the most points at the 1998 Pepsi 400 stock car race.
Claim using more than highlight: Jeff Gordon's points at the end of the Winston cup in 1998 were higher than the points of all drivers at 1998 NAPA 500 combined.

Mutated Claim (Type III) We additionally ask the annotators to modify one of the two claims with one of the following *mutations types*: **More Specific, Generalization, Negation, Paraphrasing, Entity Substitution, Tense Shift**. Both the type of modification and which of the two claims to be modified are specified in the interface (see ③ in Figure 7). Similar to 'Claim beyond highlights', the modification can result in a claim that can either be supported or refuted. So in general, you should not worry whether the mutation will preserve the truth of the claim or not. Again, for this claim it is allowed to use information/knowledge that might not be available in Wikipedia but you assume to be general knowledge. Make sure that the new claim is still a single sentence! Here is an explanation for each mutation type:

1. **Generalization** Make the claim more general so that the new claim is a generalization of the original claim (by making the meaning less specific)
2. **More Specific** Make the claim more specific so that the new claim is a specialization (as opposed to a generalization) of the original claim (by making the meaning more specific).
3. **Negation** Negate the meaning of the claim. This is not to be confused with making claim false: negating the meaning of a claim could make a false claim true and vice versa!
4. **Paraphrasing** Rephrase the claim so that it has the same meaning
5. **Entity Substitution** Substitute an entity in the claim to alternative from either the same or a different set of things. If the object in the claim is an entity, replace this entity. Chose any entity in the claim otherwise.

Given the claim "*John E. Moss was a politician of the US Democratic party.*" Table 2 shows each modification for the example sentence, following table shows example modifications for each mutation type:

Type	**Modified Claim**
More specific	John E. Moss was a politician of the US Democratic party for California's 3rd congressional district.
Negation	John E. Moss has never ran for office.
Generalization	John E. Moss was a US American politician.
Paraphrase	John E. Moss was a US American politician of the Democratic party.
Entity Substitution	John E. Moss was a politician of the US Republican Party.

Table 2: Claim manipulation for the claim "*John E. Moss was a politician of the US Democratic party.*"

Expected Main Verification Challenge We want to know what you think is the main challenge for assessing the veracity and retrieving evidence for the claim you have created. You must select one of the given challenge categories you expect to be the main challenge: **Multi-hop Reasoning, Numerical Reasoning, Combining Text and Tables, Entity Disambiguation**, and **Search terms not in claim**. If the main challenge hasn't been any of these, select **Other**.

1. **Multi-hop Reasoning** Multi-hop reasoning expected to be the main challenge for verifying that claim, i.e. several pages/sections will be required for verification. e.g. "*The player who ranked 3rd at the US Open in 2010 played in the most populated city of Germany in 2014*".
2. **Numerical Reasoning** Numerical reasoning expected to be the main challenge to verify the claim, i.e. reasoning that involves numbers or arithmetic calculations. This also includes steps such as counting cells in tables. Example: Given a claim "A is older than B", and for both A and B only their birth dates are given, concluding the older person would require mathematical inference. Another example would be given the following scores in tennis '7-4', 2-6', and 6-1' to conclude that Player 1 won the match.
3. **Combining Tables and Text** Combining list(s)/table(s) with information from text (i.e. phrases, captions, sentences) outside tables is expected to be the main challenge, i.e. when the Text provides important context to Tables/List to be understood and vice versa (titles are excluded when talking about text in this challenge).
4. **Entity disambiguation** Disambiguating an entity is expected to be the main challenge for verifying a given claim. E.g. Adam Smith was a footballer for the Bristol Rovers (Wikipedia lists 4 Adam Smiths that played football).
5. **Search terms not in claim** The main challenge is expected to be finding relevant search terms to pages with required evidence to verify a given claim goes beyond searching for terms located in the claim itself, e.g. for the Claim "*Non college educated voters voted 67 percent for the democratic party in 1952*" the evidence is located on the page "New Deal Coalition" – challenging to deduce the page based on the claim. Evidence that can quickly be found by searching for an entity mentioned in the claim is most likely not a retrieval challenge (excluding entity mentions that could refer to many entities).
6. **Other** If none of the above challenges can be identified.

A.9.2 Examples

See Table 9 and 10 for examples.

A.10 Claim Verification

A.10.1 Guidelines

Evidence highlighting As soon as relevant information has been found in either *text (sentences or table captions), tables, or lists* you can add it as evidence to your annotation by clicking on it. For free text, the **entire sentence/phrase** will be selected as evidence. For tables, one **cell** is selected, and finally for lists, one **item** will be highlighted. Evidence from different Wikipedia can be freely combined – there are no restrictions. There is also no limitation in terms of evidence pieces required to validate a claim. However, an entire annotation for a single claim should not surpass **10 Minutes**. If it does, keep the already annotated evidence and submit it with the verdict *NotEnoughInformation*.

For every highlighted sentence/cell/list item some context is extracted automatically and shown to you in the interface. Article titles and sections (and subsections, subsubsections etc.) in which the evidence is located are always extracted. Additionally, if a cell has been highlighted the corresponding table headers are extracted as well. Due to the complexity and diversity in Wikipedia tables, it is possible that some additional table headers have not been highlighted automatically, but would still be needed to interpret the selected evidence correctly. **These headers need to be highlighted manually by you.**

You must apply common-sense reasoning to the evidence you read but avoid applying your own (world) knowledge. If possible, additional evidence should be highlighted which provides the missing

Claim using Highlight (true): **Warneford Place is a country house in England that dates as far back as the 17th century.**

Challenge: Other

Report Claim

Claim beyond Highlight: (Same page): **The Warneford Place was built in the 17th century in Wiltshire, England and neglected and rebuilt over the years and was owned for over 20 years by Ian Flemming who demolished the old house and rebuilt it.**

Challenge: Other

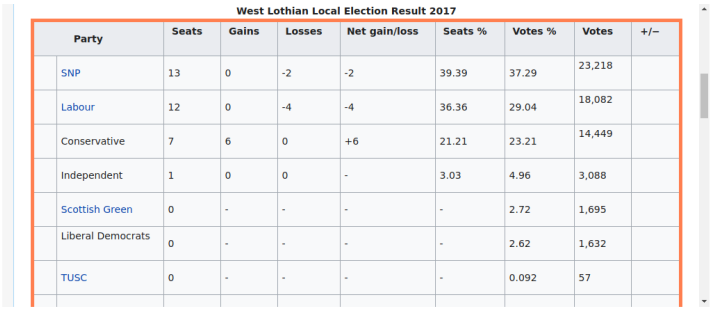
Report Claim

Manipulation: (Second claim: Substitution): **The Warneford Place was built in the 18th century in Wiltshire, England and neglected and rebuilt over the years and was owned for over 20 years by Ian Flemming who demolished the old house and rebuilt it.**

Challenge: Other

Report Claim

Figure 9: Example claim generation annotation, given a sentence highlight.



West Lothian Local Election Result 2017

Party	Seats	Gains	Losses	Net gain/loss	Seats %	Votes %	Votes	+/−
SNP	13	0	-2	-2	39.39	37.29	23,218	
Labour	12	0	-4	-4	36.36	29.04	18,082	
Conservative	7	6	0	+6	21.21	23.21	14,449	
Independent	1	0	0	-	3.03	4.96	3,088	
Scottish Green	0	-	-	-	-	2.72	1,695	
Liberal Democrats	0	-	-	-	-	2.62	1,632	
TUSC	0	-	-	-	-	0.092	57	

Claim using Highlight (true): **The Labour Party lost the 2017 West Lothian Council election having only 29.04% (18,082) of total votes.**

Challenge: Other

Report Claim

Claim beyond Highlight: (Multiple pages): **The SNP Party (Scottish National Party) won the 2017 West Lothian Council election with a total of 37.29% (23,218) of all votes.**

Challenge: Multi-hop Reasoning

Report Claim

Manipulation: (Second claim: Paraphrasing): **A total of 37.29% (23,218) of all votes was accounted forthe SNP (Scottish National Party), making them win the 2017 West Lothian Local Election.**

Challenge: Combining Tables/Lists and Text

Report Claim

Figure 10: Example claim generation annotation, given a table highlight.

information (e.g. that a *Democrat* is a politician of the Democratic Party, or that *60s* refers to the years 1960 – 1969). If this very general world knowledge cannot be found on Wikipedia you are nonetheless allowed to use it for the verdict or to find further evidence. Be careful that you do not use your knowledge to reach hasty conclusions, for instance given the evidence X is goalkeeper and the captain of Team Y, we do not have enough support that 'X is the starting goalkeeper for Y'. While it

is often the case that the first implies the second, it is not always true.

As a guide - you should ask yourself: ***If I was given only the selected sentences, table cells, and list items shown in the evidence overview ③, do I have strong enough reason to believe the claim is supported or strong enough reason to believe the claim is refuted. If I'm not certain, what additional information do I have to add to reach this conclusion and can I find it on Wikipedia?*** While claims that are *Supported* require evidence for each fact mentioned in that claim as far as possible, *Refuted* claims must only select evidence of the information that contradicts (parts of) the claim. If a refuted claim is partially supported, do not provide evidence for the partially supporting parts, unless it is necessary context for the refuting evidence, e.g. ensuring the correct entity is being referred to. If a claim is marked as *NotEnoughInformation* please still submit the evidence found in the process of reaching this verdict!

All your annotated evidence (excluding titles, and sections, but including table headers) is shown to you in the evidence overview ③, Figure 8. You initially see the ID of the annotated evidence, however, by clicking on the ID in the overview it will expand and show you the actual content of the element you selected. This way you can keep track of evidence from possibly multiple pages easily. If you change your mind and want to remove a piece of evidence simply click again on the now highlighted element.

Note!

- If the verification of a claim requires to include every entry in a table row/column (e.g. claims with universal quantification such as 'highest number of gold medals out of all countries'), you must highlight each cell of that row/column (c.f. Example 4, 7).
- Content on Wikipedia that contains qualifier or hedges (e.g. probably, likely, might) should not be used as evidence. For instance, a sentence such as *Michael Mueller was likely not involved in the 2012 scandal* should not be considered as evidence for the given claim.
- If you are not able to find any evidence for the given claim, you are still required to submit the annotation. As mentioned above, select the verdict *Not enough Information* in this case and challenges (as described below)
- Make sure that you find evidence to support each fact mentioned in a claim when selecting Supported, especially for longer claims. For instance given the claim "The scientist Mary Lamb owned five sheep with black fleece, but they were not used in any of her experiments.", the claim can be broken down into five pieces of information that all need to be verified in order to select supported:
  1. There exists a person named Mary Lamb
  2. Mary Lamb is a scientist
  3. Mary Lamb owned five sheep
  4. Those sheep had black fleece
  5. The sheep that Mary owned were not used in any of her experiments
- Do not take possible motives of Wikipedia editors into account when assessing the evidence – take the evidence as it is.
- When highlighting cells in very large tables there could be a delay until the cell and the automated context are highlighted. This is because the table has to be processed before the correct context is identified.
- There exists no interaction between different claim verification annotators in the interface – do not worry about this!
- Even if entire sentences are located in tables or lists, the finest granularity remains the cell or item, respectively. Therefore, the entire content of the cell will be added which is fine!

Ambigious & Misleading Claims In cases where you could find multiple ways of interpreting the claim which give rise to different verdicts, ask yourself the following question: ***Would you consider yourself misled by the claim given the evidence you found?*** For instance, take the claim "*Shakira is Canadian*". Even if the evidence only concludes that she is Colombian (not a direct contradiction to the claim), it is still okay to refute the claim as there is enough evidence to believe that the claim is

misleading, according to common perception. Similar case with a claim "Lamb owned five sheep", given the sentence "Lamb had a love of farming and owned many barnyard animals, including two hens and four sheep", we can conclude that the claim is misleading and thus refuted.

If you have doubt regarding your assessment go with NEI, e.g. given the claim 'Shakira was diagnosed with Diabetes Type II' and the evidence that Shakira was diagnosed with Diabetes when she was 10, it is clear to someone with specialized knowledge that the claim is false, however as it goes beyond common perception it is NEI. Do not include any knowledge about how the claims are generated when evaluating how misleading a claim is, e.g. that this claim is likely to be a corrupted version of the claim "*Shakira is Colombian*".

Reporting a claim It is possible to report and skip a given claim. It might be appropriate to flag a claim if i) the claim is personal, implausible, not verifiable, not understandable by itself, or too ambiguous ii) does not meet other aspects of the guidelines from the Claim generation task (i.e. not containing idioms, figures of speech, similes, verbose language, and not be about contemporary political topics)*, iii) ungrammatical claims or typographical errors, spelling mistakes iv) required evidence is not displayed correctly. When reporting a claim select the appropriate action from the menu or write an individual text. Do not skip a claim if it is phrased similarly to another one you have already annotated. We explicitly include paraphrased claims for annotation as we want to gather claim verifications for these too.

Main Verification Challenge We are interested in gaining more insights into the main challenge the annotator had for finding evidence for the given claim. You must select one of the given challenge categories: **Multi-hop Reasoning, Numerical Reasoning, Combining Tables and Text, Entity Disambiguation**, and **Search terms not in claim**. If the main challenge hasn't been any of these, select **Other**.

1. **Multi-hop Reasoning** Multi-hop reasoning was the main challenge challenge for verifying that claim, i.e. several pages/sections will be required for verification. e.g. "*The player who ranked 3rd at the US Open in 2010 played in the most populated city of Germany in 2014*".
2. **Numerical Reasoning** Numerical reasoning was the main challenge when verifying the claim, i.e. reasoning that involves numbers or arithmetic calculations. This also includes steps such as counting cells in tables. Example: Given a claim "A is older than B", and for both A and B only their birth dates are given, concluding the older person would require mathematical inference. Another example would be given the following scores in tennis '7-4', 2-6', and 6-1' to conclude that Player 1 won the match.
3. **Combining Tables and Text** Combining list(s)/table(s) with information from text (i.e. phrases, captions, sentences) outside tables was the the main challenge, i.e. when the Text provides important context to Tables/List to be understood and vice versa (titles and sections are excluded when talking about text in this challenge).
4. **Entity disambiguation** Disambiguating an entity was the main challenge for verifying a given claim. E.g. Adam Smith was a footballer for the Bristol Rovers (Wikipedia lists 4 Adam Smiths that played football).
5. **Search terms not in claim** The main challenge was finding relevant search terms to pages with required evidence to verify a given claim goes beyond searching for terms located in the claim itself, e.g. for the Claim "*Non college educated voters voted 67 percent for the democratic party in 1952*" the evidence is located on the page "New Deal Coalition" – challenging to deduce the page based on the claim. Evidence that can quickly be found by searching for an entity mentioned in the claim is most likely not a retrieval challenge (excluding entity mentions that could refer to many entities).
6. **Other** If none of the above challenges can be identified.

A.10.2 Examples

In addition to Figure 1 in the main paper, two further examples are shown in Figure 11.

Claim: Mike Ledwith (a professional baseball player) played one game in MLB and scored one run.

Evidence
P: `wiki/Mike Ledwith`
S0: `Introduction`
e_1: Michael Ledwith, was a professional baseball player who played catcher in one game for the 1874 Brooklyn Atlantics.
S0: `Introduction`

e_2:

Mike Ledwith	
MLB statistics	
Games played	1
Runs scored	1
Hits	1
Batting average	0.250

Verdict: Supported
Expected Challenge: Combining Tables and Text
Challenge: Combining Tables and Text

Claim: Braeden Lemasters, an American actor, musician, and voice actor, appeared in six films since 2008 and also appeared in TV shows such as Six Feet Under where he starred as Frankie.

Evidence
P: `wiki/Braeden_Lemasters`
S2: `Filmography`
e_1: Braeden Lemasters (born January 27, 1996) is an American actor, musician, and voice actor.
e_2:

Year	Film	Role
2008	Beautiful Loser	Jake
2009	The Stepfather	Sean Harding
2010	Easy A	8th Grade Todd
2012	A Christmas Story 2	Ralphie Parker
2017	Totem	Todd
2017	Flock of Four	Joey Grover

S1: `Life and career`
e_3: In 2005, Braeden started his career at age 9, as Frankie, on the TV show Six Feet Under.

Verdict: Supported
Expected Challenge: Combining Tables and Text
Challenge: Combining Tables and Text

Figure 11: Two examples from the FEVEROUS dataset that require both unstructured and structured information. The dataset contains both short, simple claims (left) and complex claims (right).

A.10.3 QA annotation interface

QA data was also used to recognise guidelines aspects that needed further clarification. Clarifications were communicated through updated guidelines as well as multiple FAQs. QA annotations were also used on an individual annotator level in combination with production reports, which measured statistics such as the number of claims an annotator generated that have been reported by verification annotators, to identify error patterns and giving annotators further individual feedback. An interface was provided to annotators to see the annotations that have been quality checked and to allow them to maintain an overview on their performance.

Figure 12 shows the QA interface for project managers. QA annotations with only partial agreement or complete disagreement are highlighted in the interface in red. The QA interface for annotators looks similar, with ID's being anonymized.

A.11 Author statement

The authors of this paper bear all responsibility in case of violation of copyrights associated with the FEVEROUS dataset.

References

Emily M. Bender and Batya Friedman. Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. *Transactions of the Association for Computational Linguistics*, 6:587–604, 12 2018. ISSN 2307-387X. doi: 10.1162/tacl_a_00041. URL `https://doi.org/10.1162/tacl_a_00041`.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. Reading Wikipedia to answer open-domain questions. In *Association for Computational Linguistics (ACL)*, 2017.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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*

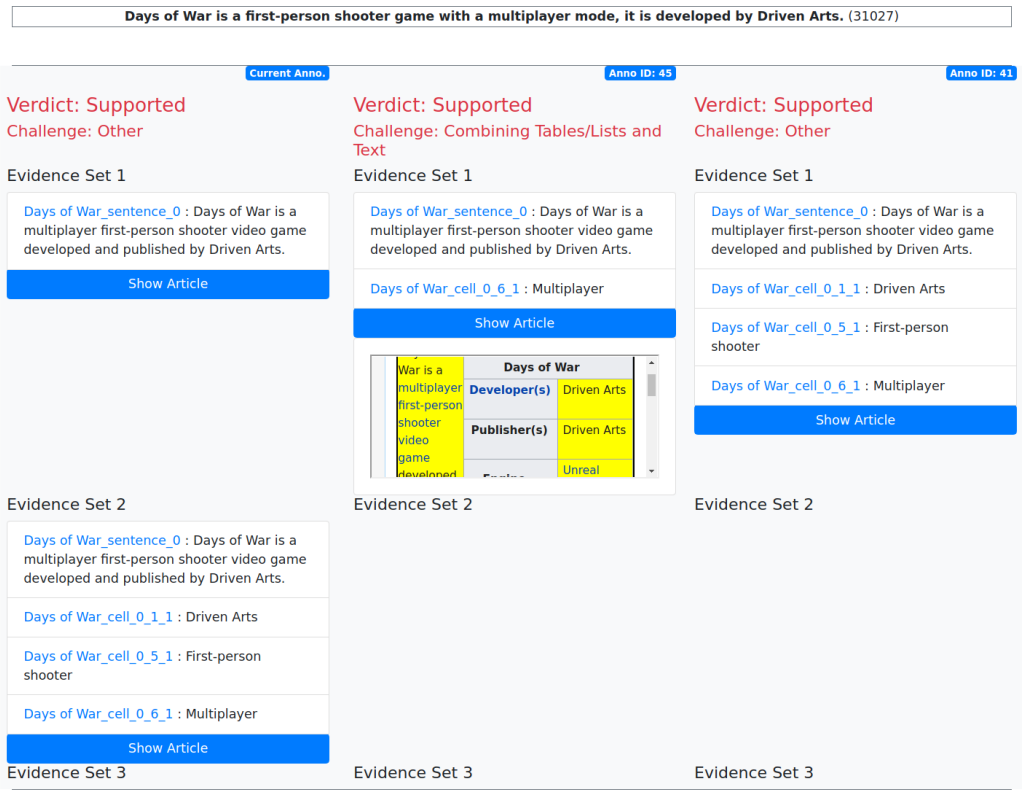


Figure 12: QA annotation interface for project managers. Interface for annotators looks similar, with ID's being anonymized

Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL `https://www.aclweb.org/anthology/N19-1423`.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2020.

Michael Schlichtkrull, Vladimir Karpukhin, Barlas Oğuz, Mike Lewis, Wen-tau Yih, and Sebastian Riedel. Joint verification and reranking for open fact checking over tables. *arXiv preprint arXiv:2012.15115*, 2020.