Explainable Automated Fact-Checking for Public Health Claims

Neema Kotonya and **Francesca Toni**
Department of Computing
Imperial College London, United Kingdom
{nk2418,ft}@ic.ac.uk

Abstract

Fact-checking is the task of verifying the veracity of claims by assessing their assertions against credible evidence. The vast majority of fact-checking studies focus exclusively on political claims. Very little research explores fact-checking for other topics, specifically subject matters for which expertise is required. We present the first study of explainable fact-checking for claims which require specific expertise. For our case study we choose the setting of public health. To support this case study we construct a new dataset PUBHEALTH of 11.8K claims accompanied by journalist crafted, gold standard explanations (i.e., judgments) to support the fact-check labels for claims[1]. We explore two tasks: veracity prediction and explanation generation. We also define and evaluate, with humans and computationally, three *coherence* properties of explanation quality. Our results indicate that, by training on in-domain data, gains can be made in explainable, automated fact-checking for claims which require specific expertise.

1 Introduction

A great amount of progress has been made in the area of automated fact-checking. This includes more accurate machine learning models for veracity prediction and datasets of both naturally occurring (Wang, 2017; Augenstein et al., 2019; Hanselowski et al., 2019) and human-crafted (Thorne et al., 2018) fact-checking claims, against which the models can be evaluated. However, a few blind spots exist in the state-of-the-art. In this work we address specifically two shortcomings: the narrow focus on political claims, and the paucity of explainable systems.

One subject area which we believe could benefit from expertise-based fact-checking is public health – including the study of epidemiology, disease prevention in a population, and the formulation of public policies (Turnock, 2012). Recent events, including the COVID-19 pandemic, demonstrate the significant potential harm of misinformation in the public health setting, and the importance in accurately fact-checking claims. Unlike political and general misinformation, specific expertise is required in order to fact check claims in this domain. Oftentimes this expertise may be limited, and thus claims which surround public health may be inaccessible (e.g., because of the use of jargon and biomedical terminology) in a way political claims are not. Nonetheless, like political misinformation, the public health variety is also potentially very dangerous, because it can put people in imminent danger and risk lives.

Typically, statements which are candidates for fact-checking originate in the political domain (Vlachos and Riedel, 2014; Ferreira and Vlachos, 2016; Wang, 2017), and tend to surround more general topics or be non-subject specific (Thorne et al., 2018). This follows the trend of the rising interest in political fact-checking in the last decade (Graves, 2018). There are on-going efforts with respect to fact-checking scientific claims (Grabitz et al., 2017). Fact-checking in domains where specific subject expertise is required presents an interesting challenge because general purpose fact-checking systems will not necessarily adapt well to these domains.

The second shortcoming we look to address is the paucity of explainable models for fact-checking (of any kind). Explanations have a particularly important role to play in the task of automated fact-checking. The efficacy of journalistic fact-checking hinges on the credibility and reliability of the fact-check, and explanations (e.g., provided by model agnostic tools such as LIME (Ribeiro et al., 2016)) can strengthen this by communicating fidelity in

[1]Data and code are available here: https://github.com/neemakot/Health-Fact-Checking

Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pages 7740–7754,
November 16–20, 2020.

predictive models. Explainable models can also aid the end users' understanding as they further elucidate claims and their context.

In this study we explore the novel case of explainable automated fact-checking for claims for which specialised expertise or in-domain knowledge is essential. For our case study we examine the the public health (biomedical) context.

The system for veracity prediction we aim to produce must fulfil two requirements: (1) it should provide a human-understandable explanation (i.e., judgment) for the fact-checking prediction, and (2) that judgement should be understandable for people who do not have expertise in the subject domain. We list the following as our **three main contributions** in this paper:

1. We present a novel dataset for explainable fact-checking with gold standard fact-checking explanations by journalists. To the best of our knowledge, this is the first dataset specifically for fact-checking in the public health setting.

2. We introduce a framework for generating explanations and veracity prediction specific to public health fact-checking. We show that gains can be made through the use of in-domain data.

3. In order to evaluate the quality of our fact-checking explanations, we define three coherence properties. These can be evaluated by humans as well as computationally, as approximations for human evaluations of fact-checking explanations.

The explanation model trained on in-domain data outperforms the general purpose model on summarization evaluation and also when evaluated for explanation quality.

2 Related Work

A number of recent works in automated fact-checking look at various formulations of fact-checking and its analogous tasks (Ferreira and Vlachos, 2016; Hassan et al., 2017; Zlatkova et al., 2019). In this paper, we choose to focus on the two specific aspects of concern to us, which have not been thoroughly explored in the literature. These are domain-specific and expertise-based claim verification and explainability for automated fact-checking predictions.

2.1 Language Representations for Health

Fewer language resources exist for medical and scientific applications of NLP compared with other NLP application settings, e.g., social media analysis, NLP for law, and computational journalism and fact-checking. We consider the former below.

There are a number of open source pre-trained language models for NLP applications in the scientific and biomedical domains. The most recent of these pre-trained models are based on the BERT language model (Devlin et al., 2019). One example is BIOBERT, which is fine-tuned for the biomedical setting (Lee et al., 2020). BIOBERT is trained on abstracts from PubMed and full article texts from PubMed Central. BIOBERT demonstrates higher accuracies when compared to BERT for named entity recognition, relation extraction and question answering in the biomedical domain.

SCIBERT is another BERT-based pre-trained model (Beltagy et al., 2019). SCIBERT is trained on 1.14M Semantic Scholar articles relating to computer science and biomedical sciences. Similar to BIOBERT, SCIBERT also shows improvements on original BERT for in-domain tasks. SCIBERT outperforms BERT in five NLP tasks including named entity recognition and text classification.

Given that models for applications of NLP tasks in the biomedical domain, e.g., question answering, show marked improvement when domain-specific, we hypothesize that public health fact-checking could also benefit from the language representations suited for that specific domain. We will make use of both SCIBERT and BIOBERT in our framework.

2.2 Explainable Fact-Checking.

A number of in-roads have been made in developing models to extract explanations from automated fact-checking systems. To our knowledge, the current state of the art in explainable fact-checking mostly looks to produce extractive explanations, i.e., explanations for veracity predictions in relation to inputs to the system. Instead, our focus in this paper is on abstractive explanations. We choose this approach, which aims to distill the explanation into the most salient components which form it, as more amenable to users with limited domain expertise, as we discuss below.

Various methods have been applied to the explainable fact-checking task. These methods span the gamut form logic-based approaches such as

probabilistic answer set programming (Ahmadi et al., 2019) and reasoning with Horn rules (Ahmadi et al., 2019; Gad-Elrab et al., 2019) to deep learning and attention-based approaches, e.g., leveraging co-attention networks and human annotations in the form of news article comments (Shu et al., 2019a). The outputs of these systems also take a number of forms including Horn rules (Ahmadi et al., 2019), saliency maps (Shu et al., 2019a; Popat et al., 2018), and natural language generation (Atanasova et al., 2020).

All approaches produce explanations which are a distillation of the most relevant portion of the system input. In this paper we expand on the work by Atanasova et al. as we formulate explanation generation as a summarization exercise. However, our work differs from the existing literature as we construct a framework for joint extractive and *abstractive* explanation generation, as opposed to a purely extractive model. We choose an abstractive approach as we hypothesize that particularly in the case of public health claims, where specific expertise is required to understand the context, abstractive explanations can make the explanation more accessible, particularly for those with little knowledge of the subject matter. In this way we take into account the nature of the claims, something other explainable fact-checking systems do not consider.

2.3 Evaluation of Explanation Quality

Only a few explainable fact-checking systems employ thorough evaluation in order to assess the quality of explanations produced. In the cases where evaluations are provided, these primarily take the form of human evaluation, e.g., enlisting annotators to score the quality of explanations with respect to some properties (Atanasova et al., 2020; Gad-Elrab et al., 2019) or through the use of an established evaluation metric in the case where explanation generation is modelled as another task (Atanasova et al., 2020).

There is also work on the evaluation of explanation quality more broadly, independently of the task for which explanations are sought. Notably, Sokol and Flach (2019) present explainability fact-sheets for evaluating (machine learning) explanations along five axes, including usability. One of the usability criteria discussed by Sokol and Flach is *coherence*, which we use to develop our three explanation quality properties (see Section 5.3). Whereas Sokol and Flach discuss coherence in general, we provide concrete definitions and use them for evaluating our methods for explaining veracity predictions for public health claims.

3 The PUBHEALTH dataset

We constructed a dataset of 11,832 claims for fact-checking, which are related a range of health topics including biomedical subjects (e.g., infectious diseases, stem cell research), government healthcare policy (e.g., abortion, mental health, women's health), and other public health-related stories (see *unproven*, *false* and *mixture* examples in Table 1), along with explanations offered by journalists to support veracity labelling of these claims. The claims were collected from two sources: fact-checking websites and news/news review websites. An example dataset entry is shown in Table 1.

To the best of our knowledge, this is the first fact-checking dataset to explicitly include gold standard texts provided by journalists specifically as explanation of the fact-checking judgment. We describe below how the data was collected and processed to obtain the final PUBHEALTH dataset, and provide an analysis of the dataset.

3.1 Data collection

Initially, we scraped 39,301 claims, amounting to: 27,578 fact-checked claims from five fact-checking websites (Snopes[2], Politifact[3], TruthorFiction[4], FactCheck[5], and FullFact[6]); 9,023 news headline claims from the health section and health tags of Associated Press[7] and Reuters News[8] websites; and 2,700 claims from the news review site Health News Review (HNR)[9].

We scraped data for two text fields which are essential for fact-checking: 1) the full text of the fact-checking or news article discussing the veracity of the claim, and 2) the fact-checking justification or news summary as explanation for the veracity label of the claim. We also collected the URLs of sources cited by the journalists in the fact-checking and news articles. For each URL, in the case where the referenced sources could be accessed and read, we also scraped the source texts.

[2] https://www.snopes.com/
[3] https://www.politifact.com/
[4] https://www.truthorfiction.com/
[5] https://www.factcheck.org/
[6] https://fullfact.org/
[7] https://apnews.com/
[8] https://uk.reuters.com/news/health
[9] https://www.healthnewsreview.org/

Claim	Label	Explanation
Blue Buffalo pet food contains unsafe and higher-than-average levels of lead.	UNPROVEN	Aside from a single claimant's lawsuit against Blue Buffalo and an unrelated recall on one variety of Blue Buffalo product in March 2017, we found no credible information suggesting that Blue Buffalo dog food was tested and found to have abnormally high levels of lead.
Children who watch at least 30 minutes of "Peppa Pig" per day have a 56 percent higher probability of developing **autism**.	FALSE	Talk of a Harvard study linking the popular British children's show "Peppa Pig" to **autism** went viral, but neither the study nor the scientist who allegedly published it exists.
Expired boxes of cake and pancake mix are dangerously toxic.	MIXTURE	What's true: Pancake and cake mixes that contain mold can cause **life-threatening allergic reactions**. What's false: Pancake and cake mixes that have passed their expiration dates are not inherently dangerous to ordinarily **healthy** people, and the yeast in packaged baking products does not "over time develops spores."
Families tell U.S. lawmakers of **heparin** deaths.	TRUE	A man who said he lost his wife and a son to reactions from tainted **heparin** made with ingredients from China urged U.S. lawmakers on Tuesday to protect **patients** from other unsafe **drugs**.

Table 1: Example of claims and explanations for PUBHEALTH dataset entries. Vocabulary from the public health glossary which are contained in the claims and explanations are highlighted in **bold**.

All claims make reference to articles published between October 19 1995 and May 14 2020. In addition to the claim, article texts, explanation texts, and the date on which the fact-check or news article was published, we scraped meta-data related to each claim. These meta-data include the tags (single or multiple tokens) which may, for example, categorize the topics of the claim or indicate the source of the claim (see Appendix A.1), and the names of the fact-checkers and news reporters who contributed to the article.

3.2 Data processing and analysis

The data processing involved three tasks: standardizing the veracity labels, filtering out non-biomedical claims from the dataset, and finally removing claims with incomplete and brief explanations.

Labels for news headline claims did not require standardization, as we assumed all news headline claims (coming from reputable sources as they were) to be verified and thus labelled these *true*, but filtered out from the dataset news entries with the headline prefixes "AP EXCLUSIVE", "Correction", "AP Interview", and "AP FACT CHECK". Indeed, it would be difficult to label the veracity of the claim in this type of entries. On the other hand, fact-check and news claims, which were associated with 141 different veracity labels, did require compression. We standardized the original labels for 4-way classification (see Appendix A.1). The chosen 4 labels are *true*, *false*, *mixture*, and *unproven*. We discounted claims with labels that cannot be reduced to one of these 4 labels. The distribution of labels in the final PUBHEALTH is shown in Table 2. The dataset consists of a majority false claims. Unproven claims are the least common in the dataset.

Website	tru.	fal.	mix.	unp.	total
AP News	2,132	0	0	0	2,132
FactCheck	0	50	29	8	87
FullFact	65	39	16	48	168
HNR	819	839	745	0	2,433
Politifact	671	1,339	423	0	2,433
Reuters	1,971	0	0	0	1,971
Snopes	386	1,131	405	220	2,142
TruthOrFict.	132	172	120	72	496
Total	6,176	3,570	1,526	299	11,832

Table 2: Summary of the distribution of true (**tru.**), false (**fal.**), mixture (**mix.**) and unproven (**unp.**) veracity labels in PUBHEALTH, across the original sources from which data originated.

The second step in processing the data was to remove claims with no biomedical context. This step was especially crucial for the claims which originated from fact-checking websites where the bulk of fact-checks concern political and economic claims. Health claims are easier to acquire from news websites, such as Reuters, as they can be quickly identified by the section of the website in which they were located during the data collection process. Although we mentioned that a sizeable number of claims from fact-checking sources are re-

lated to political events, some are connected to both political and health events or other mixed health context, and we collected claims whose subject matter intersects other topics in order to obtain a subject-rich dataset (see Appendix A.1).

Claims in the larger dataset were filtered according to a lexicon of 7,000 unique public health and health policy terms scraped from five health information websites (See Appendix A.1).

Furthermore, we manually added 65 more public health terms that were not retrieved during the initial scraping, but which we determined would positively contribute to the lexicon because of their relevance to the COVID-19 pandemic (see Appendix A.1). These claims were identified through exploratory data analysis of bigram and trigram collocations in PUBHEALTH.

In order to filter out the entries which are not health-related, we kept only claims with main article texts that mentioned more than three unique terms in our lexicon. Specifically, let L be our lexicon, and A_c and T_c, respectively, be the article text and claim text accompanying a candidate dataset entry c. Then, we included in PUBHEALTH only the following set C of claim entries, with accompanying information:

$$C_A = \{c \mid \{l_1, ..., l_n\} = A_c \cap L, n > 3\}$$

$$C_T = \{c \mid \{l_1, ..., l_n\} = T_c \cap L, n > 3\}$$

$$C = C_A \cup C_T \quad (1)$$

As we already knew that all Reuters health news claims qualify for our dataset, we used the lower bound frequency of words from our lexicon present in these article texts to determine our lower bound of three unique terms. We acknowledge that there might be disparities in the amount of medical information present in entries. However, analysis of the dataset shows, quite promisingly, that on average claims' accompanying article texts have 8.92 ± 5.54 unique health lexicon terms and claim texts carry 4.45 ± 0.88 unique terms from the health lexicon.

Claims and explanations in the entries in the dataset were also cleaned. Specifically, we also ensured all claims are between 25 and 400 characters in length. We removed explanations less than 25 characters long as we determined that very few claims shorter than this length contained fully formed claims; we removed claims longer than 400 characters to avoid the complexities of dealing with texts containing multiple claims. We also omitted claims and explanations ending in a question mark to ensure that all claims are statements, i.e., clearly defined.

Note that one aspect of the explanations' quality which we chose not to control, was the intended purpose of the text we labelled as the explanation: as shown in Table 7 in Appendix A.1, there was a wide variation across the websites we crawled.

Table 3 shows the Flesch-Kincaid (Kincaid et al., 1975) and Dale-Chall (Chall and Dale, 1995) readability evaluations of claims from our fact-checking dataset when compared to four other fact-checking datasets. The results show that PUBHEALTH claims are, on average, the most challenging to read. Claims from our dataset have a mean Flesch-Kincaid reading ease score of 59.1, which corresponds to a 10th-12th grade reading level and fairly difficult to read. The other fact-checking datasets have reading levels which fit into the 6th, 7th and 8th grade categories. Similarly for the Dale-Chall readability metric, on average our claims are more difficult to understand. Our claims have a mean score of 9.5 which is equivalent to the reading age of college student, whereas all other datasets' claims have an average score which indicates that they are readable by 10th to 12th grade students. Both these results support our earlier assertion about the complexity of public health claims relative to political and more general claims.

Dataset	**Flesch-Kincaid**		**Dale-Chall**	
	μ	σ	μ	σ
Wang (2017)	61.9	20.2	8.4	2.2
Shu et al. (2019b)	67.1	24.3	8.9	3.0
Thorne et al. (2018)	71.7	24.9	8.2	3.3
Augenstein et al. (2019)	60.8	22.1	8.9	2.5
Our dataset	59.1	23.3	9.5	2.6

Table 3: Comparison of readability of claims presented in large fact-checking datasets (i.e., those with $>$ 10K claims). We compute the mean and standard deviation for Flesch-Kincaid and Dale-Chall scores of claims for LIAR (Wang, 2017), FEVER (Thorne et al., 2018), MultiFC (Augenstein et al., 2019), FAKENEWSNET (Shu et al., 2019b), and also our own fact-checking dataset. The sample sizes used for evaluation for each dataset are as follows, LIAR: 12,791, MultiFC: 34,842, FAKENEWSNET: 23,196, FEVER: 145,449, and 11,832 for our dataset.

4 Methods

In this section we describe in detail the methods we employed for devising automated fact-checking models. We trained two fact-checking models: a classifier for veracity prediction, and a second summarization model for generating fact-checking explanations. The former returns the probability of an input claim text belonging to one of four classes: true, false, unproven, mixture. The latter uses a form of joint extractive and abstractive summarization to generate explanations for the veracity of claims from article text about the claims. Full details of hyperparameters chosen and computer infrastructure which was employed can be found in Appendix A.2.

4.1 Veracity Prediction

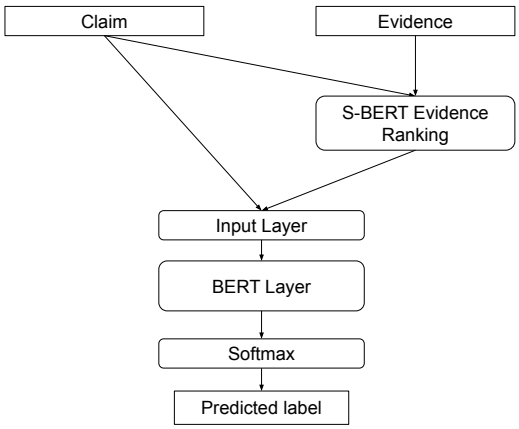


Figure 1: Architecture of veracity prediction.

Veracity prediction is composed of two parts: evidence selection and label prediction (see Figure 1).

For evidence selection, within fact-checking and news articles, we employ Sentence-BERT (S-BERT) (Reimers and Gurevych, 2019). SBERT is a model for sentence-pair regression tasks which is based on the BERT language model (Devlin et al., 2019), to encode contextualized representations for each of the evidence sentences and then rank these sentences according to their cosine similarity with respect to the contextualized representation of the claim sentence. We then select the top k sentences for veracity prediction. As with sentence selection approaches from the fact-checking literature (Nie et al., 2019; Zhong et al., 2019), we choose $k = 5$.

The claim and selected evidence sentences form the inputs for the label prediction part of our model (see Figure 1). We fine-tuned, on the PUBHEALTH dataset, pre-trained models for the downstream task of fact-checking label prediction. We employed four pre-trained models: original BERT uncased, SCIBERT, BIOBERT v1.0, and also BIOBERT v1.1. The two versions of BIOBERT differ slightly in that the earlier version is trained for 470K steps on PubMed abstracts and PubMed Central (PMC) full article texts, whereas BIOBERT v1.1 is trained for 1M steps on PubMed abstracts.

4.2 Explanation Generation as Abstractive Summarization

We make use of extractive-abstractive summarization (Liu and Lapata, 2019) in developing the explanation model. We choose this architecture because explanations for claims which concern a specific topic area having a highly complex lexicon can benefit from the ability to articulate judgment in simpler terms. In order to deploy the model proposed by (Liu and Lapata, 2019) we also implemented an explanation generation model.

Just as is the case for the predictor model, the explanation model is fine-tuned for the task on evidence sentences ranked by S-BERT. However, for the explanation model we use all article sentences as well as the claim sentence to fine-tune a BERT-based summarization model pre-trained on the Dailymail/CNN news article and summaries dataset (Hermann et al., 2015). One of our models, EXPLAINERFC, is fine-tuned using non-public health data, which we extract from the portion of the 39.3K originally crawled fact-checks, news reviews, and news articles not included in PUBHEALTH. For fairness, we ensure these data have the same proportion of claims from each website and the number of examples is the same as PUBHEALTH. The second model, EXPLAINERFC-EXPERT, is fine-tuned on PUBHEALTH. Also, we evaluate both models on PUBHEALTH test data. Table 2 shows an example of the explanations generated by the two methods.

5 Results

We conducted experiments to evaluate the performance of both predictor(s) and explainer(s). The performance of the (various incarnations of the) prediction model is evaluated using an automatic approach, whereas the performance of the (two incarnations of the) explainer is assessed using both automatic and human evaluation.

Claim **Label: FALSE**

Under Obamacare, patients 76 and older must be admitted to the hospital by their primary care physicians in order to be covered by Medicare.

Gold explanation

Obamacare does not require that patients 76 and older must be admitted to the hospital by their primary care physicians in order to be covered by Medicare.

EXPLAINERFC explanation

What's true: nothing in the Affordable Care Act requires that a primary care physician admit patients 76 or older to a hospital in order for their hospital care to be treated under Medicare. What's false: none of the provisions or rules put an upper age limit on medicare coverage.

EXPLAINERFC-EXPERT explanation

The Affordable Care Act does not require Medicare to admit patients to a hospital after paying the Part B deductible. It's not the same age limit on medicare coverage. But the evidence doesn't specifically set an upper age limit.

Figure 2: Example of model-generated explanations as compared to the gold standard from our fact-checking dataset.

5.1 Prediction

We split the PUBHEALTH dataset as follows: 9,466 training examples, 1,183 examples for validation and 1,183 examples for testing.

We evaluated veracity prediction using macro-F1, precision, recall and accuracy metrics as shown in Table 4. We employ two baselines: a randomized sentence selection approach with BERT (bert-base-uncased) classifier, and lastly a BERT model, also using pre-trained uncased BERT, which does not make use of sentence selection and instead makes use of the entire article text to fine-tune for the fact-checking task.

Out of the four BERT-derived models, SCIBERT achieves the highest macro F1, precision and accuracy scores on the test set. BIOBERT v1.1 achieves the second highest scores for F1, precision and accuracy. As expected, BIOBERT v1.1 outperforms BIOBERT v1.0 on all four metrics. The standard BERT model achieves the highest precision score of the four models, however it also achieves the lowest recall and F1 scores. This supports the argument we presented in Section 1 that subject-specific fact-checking can benefit from training on in-domain models.

Model	Pr.	Rc.	F1	Acc.
BERT (rand. sents.)	38.97	39.38	39.16	20.99
BERT (all sents.)	56.50	56.50	56.50	55.40
BERT (top k sents.)	**77.39**	54.77	63.93	66.02
SCIBERT	75.69	**66.20**	**70.52**	**69.73**
BIOBERT 1.0	73.93	57.57	64.57	65.18
BIOBERT 1.1	75.04	61.68	67.48	68.89

Table 4: Veracity prediction results for the two baselines and four BERT-based models on the test set. Model performance is assessed against precision (Pr.), recall (Rc.), macro F1, and accuracy (Acc.) metrics.

5.2 Explanations

We use two methods for evaluating the quality of explanations generated by our methods: automated evaluation and qualitative evaluation, in turn amounting to human and computational evaluation of explanation properties.

5.2.1 Automated Evaluation

We make use of ROUGE summarization evaluation metrics (Lin, 2004). Specifically we use the F1 values for ROUGE-1, ROUGE-2, and ROUGE-L, to evaluate the explanations generated by the EXPLAINERFC and EXPLAINERFC-EXPERT models.

As in the setup employed by Liu and Lapata (2019), we compare our explanation models to two other methods: a LEAD-3 baseline, which constructs a summary out of the first three sentences of an article, and an extractive summarization-based ORACLE upper bound. The results of this evaluation are shown in Table 5. The EXPLAINERFC-EXPERT explanation model outperforms EXPLAINERFC. EXPLAINERFC-EXPERT achieves higher scores than EXPLAINERFC for R1, R2, and RL metrics.

Model	ROUGE-F		
	R1	R2	RL
ORACLE	39.24	14.89	32.78
LEAD-3	29.01	10.24	24.18
EXPLAINERFC	31.42	12.38	26.27
EXPLAINERFC-EXPERT	**32.30**	**13.46**	**26.99**

Table 5: ROUGE-1 (R1), ROUGE-2 (R2) and ROUGE-L (RL) F1 scores for explanations generated via our two explanation models.

5.3 Evaluation of Explanation Quality

As the explanations we generate are from heterogeneous sources (and therefore not directly comparable), evaluation using ROUGE does not present us with a complete picture of the usefulness or quality of these explanations. For this reason, we adapt to the task of explainable fact-checking three of the desirable usability properties for machine learning explanations offered by Sokol and Flach (2019). We define these properties formally and evaluate the quality of the generated explanations against them. These same properties are also used for our human evaluations and for a comparison between human and computational evaluation of the quality of our explanations. To the best of our knowledge, ours is the first systematic evaluation of the quality of explanations for fact-checking in terms of formal properties. We define the three explanation properties as (two forms of) global coherence and (a form of) local coherence, as follows.

Global Coherence refers to the suitability of fact-checking explanations with respect to both the claim and label to which it is associated. We consider two incarnations of global coherence:

- *Strong global coherence*. Let E be an explanation of the veracity label l for claim C, where $e_1, \ldots, e_N$ are all the individual sentences which make up E. Then, E satisfies strong global coherence iff $\forall e_i \in E$, $e_i \models C$. Put simply, for this property to hold for a generated fact-checking explanation, every sentence in the explanatory text must entail ($\models$) the claim.

- *Weak global coherence*. Let E be an explanation of the veracity label l for claim C, where $e_1, \ldots, e_N$ are all the individual sentences which make up E. Then, E satisfies weak global coherence iff $\forall e_i \in E$, $e_i \not\models \neg C$. For this property to hold for a generated fact-checking explanation, no sentence in the explanatory text should contradict the claim (by entailing its negation); from a natural language inference (NLI) perspective, for weak global coherence to hold all explanatory sentences should entail or have a neutral relation with respect to the claim.

When measuring coherence, we treat as *neutral* claims originally labelled as *false* if their claim is *contradicted* by its explanation. Note that if the *false* claim is *entailed* by its explanation we do not reassign the label, because doing so would impose too strong an assumption that the entailment is related to the veracity which we cannot verify.

Local Coherence. Let E be an explanation of the veracity label l for claim C, where $e_1, \ldots, e_N$ are all the individual sentences which make up E. Then, E satisfies local coherence iff $\forall e_i, e_j \in E$, $e_i \not\models \neg e_j$.

Local coherence is a measure of how cohesive sentences in an explanation are. For local coherence to hold any two sentences in an explanation must not contradict each other, i.e., there is no pairwise disagreement between sentences which make up the explanation.

Note that all three coherence properties relate to the usability property of coherence discussed by Sokol and Flach (2019). Local coherence draws specifically on the idea of avoiding internal inconsistencies in explanations. Figure 3 shows an example of evaluation of the three properties, for a specific claim-explanation pair. Schematic examples of explanations and evidence sentence relations which satisfy these coherence properties are shown in Appendix A.4.

5.3.1 Human & Computational Evaluations

We employ human evaluation in order to assess the quality of the gold and generated explanations with respect to these properties. Also, we conduct a computational evaluation of the three coherence properties using NLI.

For human evaluation, we randomly sampled 25 entries from the test set of PUBHEALTH, and enlisted 5 annotators to evaluate the quality of the gold explanations and explanations generated by EXPLAINERFC and EXPLAINERFC-EXPERT for these entries. We asked participants to annotate explanations according to the following criteria: 1) the agreement and disagreement between sentences in the explanation, and 2) relevance of the explanation to the claim. Further information, including an example from the questionnaire, can be found in Appendix A.3.

We conducted the computational evaluation on three pretrained NLI models: 1) a decomposable attention model (Parikh et al., 2016) using ELMo embeddings (Peters et al., 2018) trained on the Stanford Natural Language Inference (SNLI) corpus (Bowman et al., 2015), 2) RoBERTa (Liu et al., 2019) trained on SNLI, and 3) RoBERTa trained

on the Multi-Genre Natural Language Inference (MNLI) corpus (Williams et al., 2018). We implemented these evaluation methods using the AllenNLP platform (Gardner et al., 2018).

For the human evaluation we computed Randolph's free-marginal κ (Randolph, 2005) and overall agreement (O.A.) for all multiple choice questions. For the gold explanations, we computed κ (and O.A.) of 0.24 (62%), 0.48 (65.6%), and 0.39 (54.4%) for 2-, 3-, and 4-nary questions respectively. For EXPLAINERFC, 0.06 (53.2%), 0.17 (44.8%), and 0.12 (34%) for 2-, 3-, and 4-nary questions respectively. Lastly for EXPLAINERFC-EXPERT, we computed κ and O.A. of 0.36 (68%), 0.44 (62.73%), and 0.20 (40%) for 2-, 3-, and 4-nary questions. The computational evaluation was conducted on all examples from the test set. The results of both the human and computational evaluation of the three coherence measures are shown in Table 6. Our results suggest that the NLI approximation is a reliable approximation for weak global coherence and local coherence properties. However, entailment appears to be a poor approximation for strong global coherence. Further, a larger human evaluation study would be required in order to verify these results.

Evaluation Method	SGC	WGC	LC
	Gold explanations		
Human	76.80	98.40	65.60
DA+ELMo; SNLI	8.72	87.61	55.20
RoBERTa; SNLI	1.28	75.87	52.12
RoBERTa; MNLI	2.66	87.52	54.84
EXPLAINERFC generated explanations			
Human	53.60	88.80	58.10
DA+ELMo; SNLI	8.26	89.45	51.32
RoBERTa; SNLI	0.46	76.42	48.01
RoBERTa; MNLI	0.73	84.59	50.20
EXPLAINERFC-EXPERT generated explanations			
Human	60.4	76.80	59.30
DA+ELMo; SNLI	7.61	89.72	64.60
RoBERTa; SNLI	0.64	76.15	60.07
RoBERTa; MNLI	2.48	84.04	62.43

Table 6: % of explanations which satisfy strong global coherence (SGC), weak global coherence (WGC) and local coherence (LC) properties.

6 Conclusion and Future work

In this paper, we explored fact-checking for claims for which specific expertise is required to produce a veracity prediction and explanations (i.e., judgments used for awarding the label/veracity prediction). To support this exploration we constructed PUBHEALTH, a sizeable dataset for public health fact-checking and the first fact-checking dataset to include explanations as annotations. Our results show that training veracity prediction and explanation generation models on in-domain data improves the accuracy of veracity prediction and the quality of generated explanations compared to training on generic language models without explanation.

We hope to explore the topics of explainable fact-checking and specialist fact-checking further. In order to do this, we hope to explore other subjects, in addition to public health, for which fact-checking requires a level of expertise in the subject area. Furthermore, we hope to explore the quality of fact-checking explanations with respect to properties other than coherence, e.g., actionability and impartiality. lastly, we plan to explore congruity between veracity prediction and explanation generation tasks, i.e., generating explanations which are compatible with the predicted label and vice versa.

Claim
A list of chemicals, written as if they were ingredients on a food label, accurately depicts the chemical composition of a banana.
Label: TRUE
Explanation
In sum, this graphic accurately depicts the chemicals that comprise a banana, using a variety of tactics to make that completely natural food appear to be full of "chemicals" — something originally created by a high school chemistry teacher as part of a lesson on chemophobia.

Figure 3: Example of explanation which satisfies all three coherence properties.

Acknowledgements

We would like to thank all those who participated in the explanation evaluation study for their valuable contributions. The first author is supported by a doctoral training grant from the UK Engineering and Physical Sciences Research Council (EPSRC).

References

Naser Ahmadi, Joohyung Lee, Paolo Papotti, and Mohammed Saeed. 2019. Explainable fact checking with probabilistic answer set programming. *arXiv preprint arXiv:1906.09198*, abs/1906.09198.

Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. 2020. Generating fact checking explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7352–7364, Online. Association for Computational Linguistics.

Isabelle Augenstein, Christina Lioma, Dongsheng Wang, Lucas Chaves Lima, Casper Hansen, Christian Hansen, and Jakob Grue Simonsen. 2019. MultiFC: A real-world multi-domain dataset for evidence-based fact checking of claims. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4685–4697, Hong Kong, China. Association for Computational Linguistics.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620, Hong Kong, China. Association for Computational Linguistics.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Jeanne Sternlicht Chall and Edgar Dale. 1995. *Readability revisited: The new Dale-Chall readability formula*. Brookline Books.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

William Ferreira and Andreas Vlachos. 2016. Emergent: a novel data-set for stance classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1163–1168, San Diego, California. Association for Computational Linguistics.

Mohamed H Gad-Elrab, Daria Stepanova, Jacopo Urbani, and Gerhard Weikum. 2019. Exfakt: a framework for explaining facts over knowledge graphs and text. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, pages 87–95. ACM.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke Zettlemoyer. 2018. AllenNLP: A deep semantic natural language processing platform. In *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*, pages 1–6, Melbourne, Australia. Association for Computational Linguistics.

Peter Grabitz, Yuri Lazebnik, Joshua Nicholson, and Sean Rife. 2017. Science with no fiction: measuring the veracity of scientific reports by citation analysis. *BioRxiv*, page 172940.

Lucas Graves. 2018. Boundaries not drawn: Mapping the institutional roots of the global fact-checking movement. *Journalism Studies*, 19(5):613–631.

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

Naeemul Hassan, Fatma Arslan, Chengkai Li, and Mark Tremayne. 2017. Toward automated fact-checking: Detecting check-worthy factual claims by claimbuster. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1803–1812.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Advances in neural information processing systems*, pages 1693–1701.

J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical report, Naval Technical Training Command Millington TN Research Branch.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3730–3740, Hong Kong, China. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Yixin Nie, Haonan Chen, and Mohit Bansal. 2019. Combining fact extraction and verification with neural semantic matching networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6859–6866.

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

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Kashyap Popat, Subhabrata Mukherjee, Andrew Yates, and Gerhard Weikum. 2018. DeClarE: Debunking fake news and false claims using evidence-aware deep learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 22–32, Brussels, Belgium. Association for Computational Linguistics.

Justus J Randolph. 2005. Free-marginal multirater kappa (multirater k [free]): An alternative to fleiss' fixed-marginal multirater kappa. *Online submission*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144.

Kai Shu, Limeng Cui, Suhang Wang, Dongwon Lee, and Huan Liu. 2019a. defend: Explainable fake news detection. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, pages 395–405, New York, NY, USA. ACM.

Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. 2019b. Fakenewsnet: A data repository with news content, social context and spatialtemporal information for studying fake news on social media.

Kacper Sokol and Peter Flach. 2019. Desiderata for interpretability: Explaining decision tree predictions with counterfactuals. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 10035–10036.

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

Bernard Turnock. 2012. *Public Health: What It Is and How It Works*. Jones & Bartlett Publishers, Gaithersburg, Md.

Andreas Vlachos and Sebastian Riedel. 2014. Fact checking: Task definition and dataset construction. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, pages 18–22, Baltimore, MD, USA. Association for Computational Linguistics.

William Yang Wang. 2017. "liar, liar pants on fire": A new benchmark dataset for fake news detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 422–426, Vancouver, Canada. Association for Computational Linguistics.

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

Wanjun Zhong, Jingjing Xu, Duyu Tang, Zenan Xu, Nan Duan, Ming Zhou, Jiahai Wang, and Jian Yin. 2019. Reasoning over semantic-level graph for fact checking.

Dimitrina Zlatkova, Preslav Nakov, and Ivan Koychev. 2019. Fact-checking meets fauxtography: Verifying claims about images. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2099–2108, Hong Kong, China. Association for Computational Linguistics.

A Supplementary Material

A.1 Dataset

Here we expand on the dataset analysis presented in Section 3. Figure 4 shows the most commonly occuring public health terms in the PUBHEALTH dataset entry texts. Figure 5 illustrates the distribution of claim and explanation lengths. Note that the nature and format of the explanations for each of the scraped websites differed slightly.

Table 7 shows the origin fact-checking explanations included in the PUBHEALTH dataset. In Table 8 we show examples of the subject-rich tags scraped alone with the claims. Table 9 shows the mapping between the standardized and original veracity labels.

Building the public health lexicon. In order to compile the lexicon we scraped health related terms from the following website sources. In total we scraped vocabulary from a number of pages across six websites. These websites are NHS Health A-Z,[10] Everyday Health, [11] Medline Plus,[12] Think Local, Act Personal,[13] National Careers Healthcare Job,[14] and the Mayo Clinic.[15]

Additional words added the health lexicon. The following are the extra words added to lexicon which we did not scraped. 'Centers for Disease Control and Prevention', 'abscess', 'adolescence', 'airborne', 'alimentation', 'alopecia', 'aneurysm', 'anorexia', 'anti-vaxxer', 'arrhythmia', 'bacteria', 'bacterium', 'biohazard', 'bioterrorism', 'bleeding', 'blood pressure', 'chickenpox', 'chloroquine', 'contagious', 'death', 'disease', 'embolism', 'endemic', 'environment', 'epidemiology', 'first aid', 'flatten the curve', 'flu', 'gallbladder', 'gangrene', 'heart attack', 'heparin', 'hospital', 'hydroxychloroquine', 'hygiene', 'hypertension', 'illness', 'immune', 'infant mortality rate', 'infect', 'influenza', 'lactose intolerance', 'liver', 'medicine', 'menstruation', 'mental health', 'nurse', 'organs', outbreak, pacemaker, 'pandemic', 'pathogen', 'patients', 'period poverty', 'public health', 'quarantine', 'sickness', 'smoking', 'stroke', 'surgical', 'tumour', 'vaccine', 'ventilator', 'virus', 'x-ray'.

[10] https://www.nhs.uk/conditions/

[11] https://www.everydayhealth.com/conditions/

[12] https://medlineplus.gov/encyclopedia.html

[13] https://www.thinklocalactpersonal.org.uk/Browse/Informationandadvice/CareandSupportJargonBuster/

[14] https://nationalcareers.service.gov.uk/job-categories/healthcare

[15] https://www.mayoclinic.org/diseases-conditions, https://www.mayoclinic.org/symptoms, https://www.mayoclinic.org/tests-procedures, https://www.mayoclinic.org/drugs-supplements

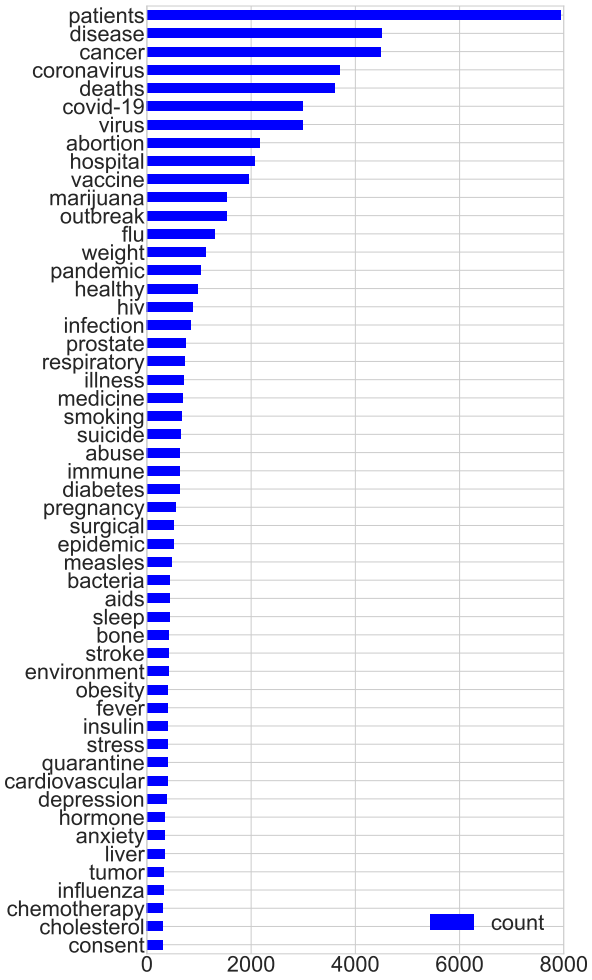


Figure 4: Vocabulary from the health lexicon which features $>$ 300 times in PUBHEALTH article texts.

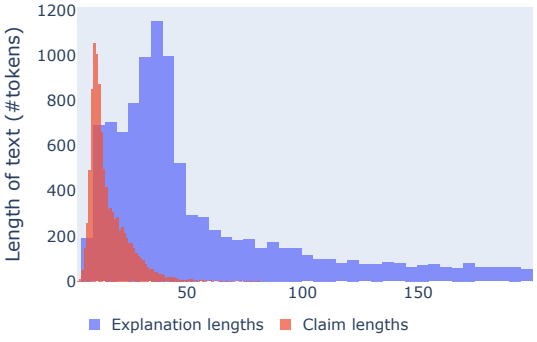


Figure 5: Histograms showing the distribution of lengths, measured by the number of tokens, for claims and explanations in the PUBHEALTH dataset.

Website	Explanations
AP News (*n*)	Leading paragraph.
FactCheck (*f*)	Summarizing paragraph.
FullFact (*f*)	Fact-check conclusions.
HNR (*r*)	Summary of reliability judment.
Politifact (*f*)	Fact-check ruling/rating comments.
Reuters (*n*)	Leading paragraph.
Snopes (*f*)	Fact-check what's true / false / undetermined or concluding paragraph.
TruthOrFict. (*f*)	Summarizing paragraph.

Table 7: Format of explanations scraped from fact-checking (*f*), news (*n*), news review (*r*) websites.

A.2 Reproducibility

Here we provide further information about the experiments described in Section 4.

Prediction models hyperparameters. We perform hyper-parameter grid search as part of validation for batch sizes from {8, 16, 32}, learning rates from {1e-5, 5e-6, 1e-6}, and epochs {2, 3, 4}. We optimize our veracity prediction model on cross entropy loss. The hyper-parameters we selected from this grid search are a batch size of 16, learning rate 1e-6 and 4 epochs for model training.

Computing Infrastructure. All experiments were run on a machine with a dual Intel(R) Core(TM) i9-9900X 3.50GHz CPU. The GPU used for experiments is the Nvidia GeForce RTX 2080 Ti model. Additional information about the software packages used in the development of the explanation generation and veracity prediction models can be found in the GitHub repository, the link to which is given in Footnote 1.

A.3 Human Evaluation Questionnaire

The following are example question and response pairs typical of those presented to participants in the human evaluation questionnaire (see Section 5.3). Question and response pairs are related to the claim and explanation presented below.

1. **Question**: Are there any sentences or phrases in the explanation which disagree with each other?

   Response options: {Yes, No}.

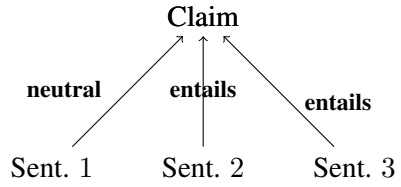
2. **Question**: Which veracity label would you give to the claim taking into account the entire explanation?

   Response options: {Mixture, false, true, unproven}.

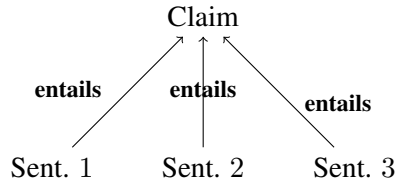
Claim
State reports new findings of mosquito-borne illnesses.
Explanation
Rhode Island health officials say a second mosquito case tested positive for eastern equine encephalitis has been confirmed in the state, marking the first human case of the equine encephalitis in Rhode Island in more than two years.

A.4 Coherence properties

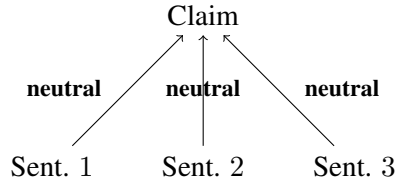
Figure 6 shows examples of the three coherence properties mentioned in Section 5.3, shown schematically in graphical form.



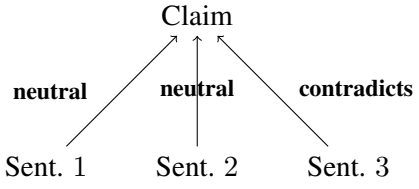
(a) Strong Global Coherence not satisfied



(b) Strong Global Coherence satisfied.



(c) Weak Global Coherence satisfied.

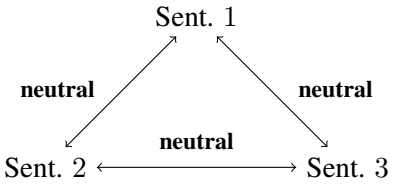


(d) Weak Global Coherence not satisfied.

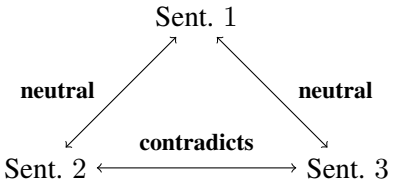
Figure 6: Schematic representations of strong and weak global coherence properties.

Claim: Judge dismisses lawsuit over release of vaccination data. **Label:** **TRUE** **Tags:** Immunizations, Health, General News, Public health, Connecticut, Hartford, Bristol, Lawsuits **Date published**: September 30, 2019
Claim: FDA allows marketing of cooling cap to reduce hair loss during chemotherapy. **Label:** **MIXTURE** **Tags:** Breast cancer, FDA, medical devices, Women's health **Date published**: December 15, 2015
Claim: Clinical study shows that retinal imaging may detect signs of Alzheimer's disease. **Label:** **MIXTURE** **Tags:** Alzheimer's disease, NeuroVision Imaging LLC, retinal imaging **Date published**: August 24, 2017
Claim: Salt lamps, because they emit negatively charged ions, impart myriad health benefits including reduced anxiety, improved sleep, increased energy, and protection from an "electric smog." **Label**: **FALSE** **Tags:** medical, salt lamps **Date published**: December 22, 2016

Table 8: Examples of tag metadata for entries in the PUBHEALTH dataset.



(a) Local coherence satisfied.



(b) Local coherence not satisfied.

Figure 7: Schematic representations of local coherence.

Standardized	Fact-checking and news review veracity labels
false	'0 Star', '1 Star', '2 Star', 'barely-true', 'digital manipulations!', 'disputed', 'disputed!', 'false', 'fiction', 'fiction!', 'fiction! & disputed!', 'fiction! satire!', 'full-flop', 'inaccurate attribution!', 'incorrect attribution!', 'incorrect authorship!', 'incorrectly attributed!', 'misattributed', 'mostly fiction!', 'mostly-false', 'not true', 'pants-fire', 'pants-on-fire!', 'reported as fiction!', 'reported fiction!
mixture	'3 Star', 'cherry picks', 'confirmed authorship! but inaccurate attribution!', 'decontextualized', 'depends on where you vote!', 'distorts the facts', 'exaggerates', 'half-flip', 'half-true', 'lacks context', 'misleading', 'misleading!', 'mixed', 'mixture', 'not the whole story', 'outdated', 'outdated!', 'previously truth! & now resolved!', 'previously truth! but now resolved!', 'reported as truth! & disputed!', 'spins the facts', 'truth & fiction!', 'truth! & disputed!', 'truth! & fiction!', 'truth! & fiction! & disputed!', 'truth! & fiction! & unproven!', 'truth! & misleading!', 'truth! & outdated!', 'truth! & unproven!', 'truth! and fiction!', 'truth! and unproven!', 'truth! but decision reversed!', 'truth! but inaccurate description!', 'truth! but misleading!', 'truth! but obama quote is fiction!', 'truth! but overturned!', 'truth! but resolved!', 'truth! but she denies it reflects her views!', 'truth! fiction! & disputed!', 'truth! fiction! & satire!', 'truth! fiction! & unproven!', 'truth!, fiction!, and unproven!', 'truth!, unproven!, & fiction!'
true	'4 Star', '5 Star', 'authorship confirmed!', 'commentary!', 'confirmed authorship', 'confirmed authorship!', 'correct attribution!', 'correct-attribution', 'correctly attributed!', 'mostly truth!', 'mostly-true', 'no-flip', 'official!', 'reported to be true!', 'reported to be truth!', 'true', 'truth but an opinion!', 'truth!', 'truth! but an opinion!', 'truth! but not intentionally!', 'truth! but not the one you think!', 'truth! but now resolved!'
unproven	'investigation pending!', 'no evidence', 'pending investigation!', 'unconfirmed attribution!', 'unknown', 'unofficial!', 'unproven', 'unproven!', 'unsupported'

Table 9: These are the four standardized labels we defined for veracity prediction (left) and lists (right) of the original fact-checking labels provided by the fact-checking and news review websites we scraped, mapped to our four standardized labels

.