A Neural Model to Jointly Predict and Explain Truthfulness of Statements

ERIK BRAND, The University of Queensland, Australia
KEVIN ROITERO, University of Udine, Italy
MICHAEL SOPRANO, University of Udine, Italy
AFSHIN RAHIMI, The University of Queensland, Australia
GIANLUCA DEMARTINI, The University of Queensland, Australia

Automated fact-checking (AFC) systems exist to combat disinformation, however their complexity usually makes them opaque to the end user, making it difficult to foster trust in the system. In this paper, we introduce the E-BART model with the hope of making progress on this front. E-BART is able to provide a veracity prediction for a claim, and jointly generate a human-readable explanation for this decision. We show that E-BART is competitive with the state-of-the-art on the e-FEVER and e-SNLI tasks. In addition, we validate the joint-prediction architecture by showing 1) that generating explanations does not significantly impede the model from performing well in its main task of veracity prediction, and 2) that predicted veracity and explanations are more internally coherent when generated jointly than separately. We also calibrate the E-BART model, allowing the output of the final model be correctly interpreted as the confidence of correctness. Finally, we also conduct and extensive human evaluation on the impact of generated explanations and observe that: explanations increase human ability to spot misinformation and make people more skeptical about claims, and explanations generated by E-BART are competitive with ground truth explanations.

ACM Reference Format:
Erik Brand, Kevin Roitero, Michael Soprano, Afshin Rahimi, and Gianluca Demartini. 2018. A Neural Model to Jointly Predict and Explain Truthfulness of Statements. *ACM J. Data Inform. Quality* 1, 1, Article 1 (January 2018), 19 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

Automated fact-checking (AFC) makes use of natural language processing (NLP) techniques to determine the veracity of a claim. The problem is defined in the following way: given a statement (claim) and some evidence, determine whether the statement is true with respect to the evidence [32]. This is a challenging task for a human, let alone an autonomous system [9]. However, AFC systems are able to approximate this process of evidence retrieval and synthesis with some degree of success [32, 37]. The benefits and applications of an AFC system are numerous. The problem of disinformation is not new, however the rate of which it propagates has continued to increase, largely aided by the increasing popularity of social media platforms [21]. AFC systems are starting to become a critical tool in combating the sheer quantity of claims that need to be verified.

Authors' addresses: Erik Brand, e.brand@uq.net.au, The University of Queensland, Brisbane, Australia; Kevin Roitero, roitero.kevin@spes.uniud.it, University of Udine, Udine, Italy; Michael Soprano, michael.soprano@uniud.it, University of Udine, Udine, Italy; Afshin Rahimi, a.rahimi@uq.edu.au, The University of Queensland, Brisbane, Australia; Gianluca Demartini, demartini@acm.org, The University of Queensland, Brisbane, Australia.

1936-1955/2018/1-ART1 $15.00
https://doi.org/10.1145/1122445.1122456

While accurate [24, 32], AFC systems have been unable to supplement traditional fact-checkers due to a limitation in their design. A user may not accept to believe in a statement without first understanding the concepts and facts underpinning that statement. Such justifications are expected when reading journalistic fact-checking outcomes such as on Politifact; the fact-check outcome is accompanied by an explanation informing the reader of how the decision was reached. Without providing users with an explanation, the decision provided by an automated system is far less likely to be trusted [35], especially as it is not generated by humans.

Automated systems have recently been developed to this effect, and have demonstrated promising initial results [9]. While these initial results are unquestionably impressive, critical evaluation of the work reveals that many of these systems use separate models for veracity prediction and explanation generation. We argue that systems such as these are not actually describing their own actions and decision processes, and that the veracity prediction model is not made any more transparent.

In this paper, we propose and experimentally evaluate a system that jointly makes a veracity prediction and provides an explanation within the same model. This is novel as compared to classic post-hoc explainability methods that are built on top of existing machine learning models. As such, the generated explanations more closely reflect the decisions made by the veracity prediction model. In addition to this, we show that large transformer models are flexible enough to multitask, and are thus able to explain their actions without detriment to the original task. This allows human end users to better interface with transformer models, fostering a more trustworthy relationship between humans and deep learning models.

We specifically address the following research questions:

- RQ1: How can we design a deep learning model to classify information truthfulness and, at the same time, generate a natural language explanation supporting its classification decision?
- RQ2: Can such model result in both accurate classification decisions and high quality natural language explanations?
- RQ3: Are machine-generated explanations useful for humans to better assess information truthfulness?
- RQ4: Can we calibrate the deep learning model in a way that outputs reliable confidence scores for the truthfulness predictions?

By creating an automated system that is capable of both evaluating the truthfulness of a statement and simultaneously generating a human-interpretable explanation for this decision, it is hoped that automated fact-checking systems will become more widely adopted.

2 RELATED WORK

2.1 Existing Explainable-AFC Models

A number of techniques for generating explanations to accompany AFC decisions have been proposed. Saliency-based methods, such as those proposed by Shu et al. [30] and Wu et al. [38], use attention mechanisms to highlight the input that is most useful in determining the veracity prediction and present this information to the end user as a form of explanation. Logic-based approaches make use of graphs [7], rule mining, and probabilistic answer set programming [1] to output a series of logical rules that result in a veracity prediction. This set of rules constitutes an explanation. While these methods are highly transparent and logical, the resulting explanation is not always human-readable [1].

Summarisation techniques provide an explanation by summarising the retrieved evidence. The system proposed by Atanasova et al. [2] utilises DistilBERT [28] to pass contextual representations of the claim and evidence to two task-specific feed-forward networks which produce a classification and an extractive summary. Kotonya and Toni [13] take a similar approach but tailor their model to the public health domain. The pipeline utilises Sentence-BERT [26] to filter the evidence, a BERT-based veracity predictor, and a separate BERT-based summarisation model. The work by Kotonya and Toni [13] differs from Atanasova et al. [2] as it produces *abstractive* explanations, which are generally more coherent and similar to the way a human would generate a summary, rather than *extractive* explanations which take sentences verbatim from the evidence.

The framework proposed by Stammbach and Ash [32] also produces abstractive explanations, but places higher emphasis on the evidence retrieval process. The framework consists of two components: 1) an evidence retrieval and veracity prediction module, and 2) an explanation generation module. The first component is an enhanced version of the DOMLIN system [33], which uses separate BERT-based models for evidence retrieval and veracity prediction. For explanation generation, GPT-3 [5], a large pertained multi-purpose NLP model based on the Transformer, is used in 'few-shots' mode to generate a summary of the evidence with respect to the claim.

The system we present in this paper differs to the existing literature as rather than using two separate models for the veracity prediction and explanation generation, a single model is used to output both a veracity prediction and an abstractive summarisation.

2.2 BART Transformer Architecture

BART [16] is a transformer [36] model that aims to generalise the capabilities of both BERT [8] and GPT-style models. It consists of a bi-directional encoder, similar to BERT, as well as an auto-regressive decoder, similar to GPT. BART is pre-trained on a de-noising task whereby input text is corrupted and the model aims to reconstruct the original document, minimising the reconstruction loss. In contrast to existing de-noising models, BART is more flexible in that it is not trained to rectify a specific type of input corruption, but rather any arbitrarily corrupted document.

The pre-trained BART model can be fine-tuned to a number of downstream tasks. The authors noted that the model performs comparably to other models, such as RoBERTa [18], on natural language inference tasks. They also note that BART outperforms current state-of-the-art models on natural language generation tasks, such as summarisation [16, 29]. Its ability to perform well on these two contrasting tasks made it an attractive choice as the base model for a system that can jointly predict the veracity of a claim, an inference task, and provide an explanation, a generative task.

3 A MODEL FOR JOINTLY PREDICTING AND EXPLAINING TRUTHFULNESS

Many of the systems in the reviewed literature use separate Transformer models for veracity prediction and explanation generation. Outlined here is our proposed architecture, E-BART, that jointly outputs a veracity prediction, as well as a human-readable, abstractive explanation addressing *RQ1*.

To adapt the BART-large encoder-decoder model to this downstream task, a 'joint prediction' head was developed, shown in Figure 1. This model is novel, and it has been developed specifically for the purpose of dual modeling detailed in this work, making it possible to perform both a discriminative and generative task at the same time. This head sits atop the BART model, and manipulates the transformer hidden states into the form of the desired output. Both the BART base model and the joint prediction head can be fine-tuned as a single unit to customise pre-trained BART weights to the joint prediction task. The joint prediction head is also shown using the green color in Figure 3. The head takes as input the final decoder hidden state embeddings. It then passes all embeddings to a single feed-forward layer to produce a series of logits which form the basis of the predicted explanation. To facilitate classification, the hidden state embeddings corresponding to the final sequence separator token (</s> in BART) are extracted and passed to

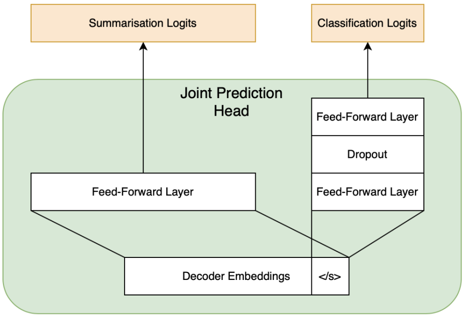


Fig. 1. Joint prediction head.

a small feed-forward network to shape the output to the desired number of classes. The logits obtained from this are then passed to a final soft-max layer to produce probabilities for each class. Unlike in BERT which uses embeddings corresponding to the `[cls]` token which is pre-pended to the input to perform classification, in BART the final sequence separator token is used instead as the decoder can only attend to the left of the current token. This conditions the classification on the entire input sequence. The summarisation component of the joint prediction head consists of a single feed-forward layer with input dimension equal to the decoder embedding dimension of 768, while the output dimension is equal to the vocabulary size of the model. The argmax of the raw logits are used by the head during the greedy generation process. It is instructive to consider the training and inference processes separately, as they differ slightly due to the auto-regressive nature of the BART decoder.

Figure 2 shows the inference process. Running inference on the model begins by running the encoder with the tokenised input to generate the encoder hidden states, as before. In contrast to the training process, the decoder is presented with the start sequence token (<s> in BART), and generates logits auto-regressively, guided by a beam search. The final phase of inference runs the decoder with the entire generated sequence presented at its input. At this point, the joint prediction head extracts the embeddings corresponding to the token immediately before the final sequence separator token from the generated sequence. This is done to mirror the training process (discussed in the following). These embeddings are passed to the classification component of the joint prediction head, and then to a soft-max layer to produce the final classification.

The training process is as follows. During the training phase, the encoder generates hidden states from the tokenised input that are then injected into the decoder. The tokenised gold summary is presented to both the input and summarization output of the decoder, with the input shifted right by one token. This conditions the decoder to predict the next token given the current token. Concurrently, the classification labels are presented to the classification output of the joint prediction head. Note that the model weights for each of the two main 'pathways' in the joint prediction head (i.e., generation and classification path) are not shared but up until the head (like in the base BART model), the weights are shared for both classification and generation. The loss is calculated as the weighted sum (with parameters α and $(1 - \alpha)$) of the Cross Entropy Loss computed between the summarisation logits and the gold summary, and the Cross Entropy Loss between the classification logits and the ground truth classification. Thus, the first objective function is to corrupt the text of the summary with a noising function and training the model to learn to reconstruct the original summary text, while the second objective function is to adjust the weights of the network such that the logits from the last layer are as close as possible to the class representation of the model

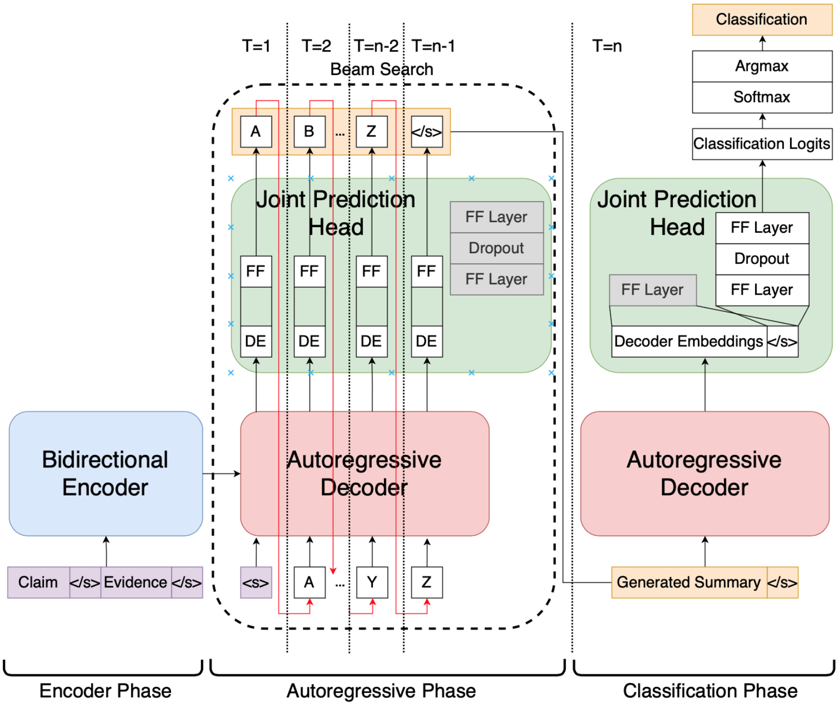


Fig. 2. E-BART Inference process.

(i.e., a classical classification task). This way, the the loss is optimized for both the generation task and the classification task jointly.

We rely on the code, training method, and regularization parameters used in the original paper (i.e., [16]) and repository[1] detailing the BART model, for further details see [16, Section 5]. More in detail, we start with the BART pre-trained weights, and then perform further training on the BART and the developed joint prediction head jointly.

While the detailed architecture is heavily based on BART, it must be noted that it has some important differences. Differently from BART, our model does not simply apply a classification layer n the top of the original BART model, but rather uses the "Joint Prediction Head" (see Figures 2 and 3) that makes it possible to perform a joint modeling and thus be able to both classify the truthfulness of a statement and the generate an explanation for that at the same time.

4 EXPERIMENTAL EVALUATION

4.1 Datasets

To evaluate the proposed models we make use of different datasets. The FEVER dataset consists of 185,445 claims, associated evidence, and veracity labels. The claims were generated by manipulating sentences taken from Wikipedia, and are labelled with either "Supports", "Refutes", or "Not_enough_info" based on whether the evidence entails the claim [34].

The e-FEVER dataset by Stammbach and Ash [32] augments the original FEVER dataset [34] with explanations generated by their framework. It consists of 50,000 examples from the FEVER

[1] https://github.com/pytorch/fairseq/blob/main/examples/bart/README.md.

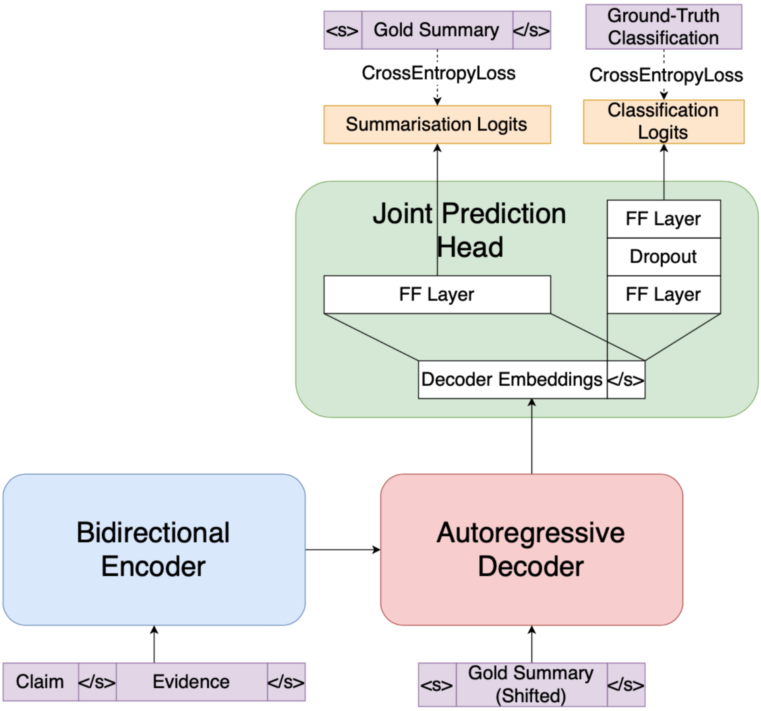


Fig. 3. E-BART Training configuration.

train set, and 17,687 from the development set. This provides a resource with claims, retrieved evidence, veracity labels, and explanations.

The e-SNLI dataset [6] extends the SNLI dataset [4] with human-generated explanations for each of the 570k examples. The SNLI task is to take two sentences and predict whether one entails, contradicts, or is neutral with respect to the other. e-SNLI adds complexity by also requiring a generated explanation for the label.

Summarizing, the statistics for the three datasets are as follows: FEVER dataset: 185,445 total examples, 165,447 for the training and 19,998 for the development set, eFEVER dataset: 67,687 total examples, 50,000 for the training and 17,687 for the development set, and eSNLI dataset: 570,152 total examples, 550,152 for the training and for the 20,000 development set.

4.2 Training Methodology

To investigate *RQ2* and evaluate the performance of the proposed model on the FEVER and extended e-FEVER tasks, two different versions of the model were trained. Due to retrieval evidence policies, the e-FEVER dataset contains some 'null' explanations [32]. Our first model, **E-BARTSmall**, was trained on the subset of the e-FEVER training set that did not include null explanations. This resulted in 40,702 examples. To process the data, the "+" character used to separate page titles from evidence was removed. The model inputs were tokenised and formatted as: "<s> claim </s> evidence </s>". The veracity labels were made numerical and explanations were tokenised in a similar manner. The processed dataset was used to fine-tune the BART-large model with joint prediction head for 3 epochs. Our second model, **E-BARTFull**, was trained in exactly the same

Table 1. Ground truth label is “Not_enough_info” and predicted label is “Refutes”.

Claim	**Evidence**	**Generated Explanation**
Marnie was directed by someone who was “The Master of Nothing”.	Alfred Hitchcock Sir Alfred Joseph Hitchcock (13 August 1899-29 April 1980) was an English film director and producer, at times referred to as “The Master of Suspense”. Marnie (film) Marnie is a 1964 American psychological thriller film directed by Alfred Hitchcock.	Marnie was directed by Alfred Hitchcock, who was “The Master of Suspense”.

way as the first, however it was trained using the entire e-FEVER training set, including examples with null explanations.

The models were trained on two platforms: Google Colab with a NVIDIA T4 GPU with 16GB of memory and Microsoft Azure with a 12GB NVIDIA Tesla K80 GPU. The BART-based models have roughly 375m parameters and took approximately 5 hours to train (i.e., fine-tune BART and the Joint Prediction Head for 3 epochs) on the NVIDIA T4 GPU using the eFEVER dataset.

4.3 Evaluation Methodology

The development split of the e-FEVER dataset was prepared identically to the training split, producing e-FEVER_Full and e-FEVER_Small which do, and do not, include examples with null explanations, respectively.

When evaluating the veracity prediction accuracy of the models, it was noted that including the “Not_enough_info” class could under-represent the actual classification performance. Take the example in Table 1, which has a ground truth label of “Not_enough_info”. Manual inspection shows that the explanation and evidence indicate that the claim is indeed refuted, which was correctly predicted by our model. Hence we report two sets of results, one with, and one without examples that have a e-FEVER label of “Not_enough_info”.

In the following sections we perform a set of experiments to evaluate the proposed approach. First, we test the effectiveness of the E-BART model on FEVER (Section 4.4), e-FEVER (Section 4.5), and e-SNLI (Section 4.6) datasets. Then, we validate the usage of the joint models (Sections 4.7 and 4.8). Finally, we test the impact of generated explanations (Section 4.9), and we calibrate the network to derive meaningful confidence scores (Section 4.10).

4.4 Evaluation Results on Original FEVER

To compare with existing models, we report the classification performance of E-BART on the original FEVER development set. The DOMLIN system [33] was used for evidence retrieval (discarding its veracity predictions) to provide evidence for 17k out of the 20k examples in the development set. We use our E-BART models to generate veracity predictions for the 17k examples, and then label the remaining with ‘Not_enough_info,’ as specified in the DOMLIN paper. Results are reported for the development set rather than the test set, as ground-truth labels were not published for the latter.

On the FEVER dataset, E-BARTSmall and E-BARTFull achieved label accuracies of **75.0** and **75.1**, respectively, outperforming state-of-the-art methods. For comparison, other published model accuracies on this dataset include: BERT-BASED 74.6 [31], DOMLIN 72.1 [33], UCL MR 69.7 [39],

Table 2. Effectiveness of the models on the e-FEVER dataset.

Model	Dataset	Accuracy no N.E.I	Accuracy full	ROUGE 1	ROUGE 2	ROUGE L	ROUGE Sum
E-BARTSmall	eFEVER_Small	87.2	78.2	73.581	64.365	71.434	71.585
E-BARTSmall	eFEVER_Full	85.4	77.1	59.447	50.177	57.697	57.782
E-BARTFull	eFEVER_Small	87.1	78.1	64.530	55.283	62.691	62.820
E-BARTFull	eFEVER_Full	85.2	77.2	65.511	57.598	64.071	64.144

UNC 69.6 [20], and UKP-Athene 68.5 [11]. E-BART compares favourably to the existing literature despite the e-FEVER training set having 95k less examples compared to FEVER, which the other models were trained on. It is hypothesised that the performance improvements are derived from using BART as a base model, and from requiring the model to further attend to the most relevant evidence in forming an explanation. The most noteworthy comparison is between E-BART and DOMLIN, which use identical evidence retrieval mechanisms, thus isolating the contribution of E-BART over standard veracity predictors.

4.5 Evaluation Results on e-FEVER

Table 2 shows the results obtained on the development e-FEVER dataset. To the best of our knowledge, there have been no other results reported on this recent dataset, hence we present a comprehensive snapshot of E-BART's performance.

Perhaps unsurprisingly, both our models performed better on e-FEVER_Small, which contained less inconclusive examples. More surprising is the consistency of E-BART's performance regardless of whether it was trained on e-FEVER_Small or e-FEVER_Full. This indicates that E-BART is robust to situations where evidence is sparse. Table 3, qualitatively shows that the model can even express the fact that it was not able to find relevant evidence. In the generated explanation, the phrase 'however it does not say whether...' indicates in a human-readable way that the evidence retrieval mechanism was not able to provide sufficient evidence for the model to produce a confident prediction.

The ROUGE metrics evaluate the consistency between the generated and e-FEVER dataset explanations, but are not necessarily representative of explanation quality. For instance, the explanation generated by GPT-3 may include some additional information compared to E-BART. Whether this additional information results in a better explanation compared to something more succinct is largely subjective and dependent on the system's use case. In Tables 1, 3 and 4, we present examples from the development set. Looking at the example shown in Table 4, it is clear that the model has identified a contradiction between the claim that Janet Leigh was incapable of writing, and the evidence which suggested that she had written four books. The model has generated a coherent explanation that details this finding, extracting relevant facts from the evidence to support the justification. Finally, it has generated the correct label for the claim, which, importantly, is also consistent with the generated explanation.

4.6 Evaluation Results on e-SNLI

The e-SNLI task presents a similar challenge to e-FEVER, whereby the entailment between two sentences is predicted (similar to predicting veracity of a claim with respect to evidence), and an explanation is generated.

A different version of the E-BART model was trained specifically on this dataset. The data was prepared by enumerating the labels, removing noisy data, and tokenising the summaries. The

Table 3. Ground truth label is “Supports” and predicted label is “Supports”.

Claim	Evidence	Generated Explanation
CBS is the network that aired The Millers	The Millers The multi-camera series aired from October 3, 2013 to July 18, 2015 and ran 34 episodes over two seasons on CBS. CBS announced the cancellation of The Millers on November 14, 2014, four episodes into the show’s second season.	The Millers aired on CBS, however it does not say whether it was the network that aired it.

Table 4. Ground truth label is “Refutes” and predicted label is “Refutes” (E-BART_Small Model).

Claim	Evidence	Gold Explanation	Generated Explanation
Janet Leigh was incapable of writing.	Janet Leigh Janet Leigh (born Jeanette Helen Morrison; July 6, 1927-October 3, 2004) was an American actress, singer, dancer and author. Janet Leigh She also wrote four books between 1984 and 2002, including two novels.	Janet Leigh was capable of writing.	Janet Leigh wrote four books between 1984 and 2002, including two novels.

first and second sentences were concatenated and tokenised in the same way as the claim and explanation for the e-FEVER evaluation.

On the test e-SNLI dataset, E-BART achieved a label accuracy of **90.1** and a BLEU score of **32.70**. The model proposed in conjunction with the e-SNLI dataset, e-INFERSENT, achieved an accuracy of 84.0 and BLEU score of 22.4 [6]. In calculating the BLEU metric for the explainable models, the first two gold explanations were used as references as per [6]. As a further comparison, the following are the best performing models published in the literature[2] which, however, do not provide explanations: CA-MTL 92.1 [22], SemBERT 91.9 [43], MT-DNN 91.6 [17], SJRC 91.3 [42], D-CRCo-AN 90.1 [12], and LMTransformer 89.9 [25].

The performance of E-BART compares favourably with the state-of-the-art for this different task, corroborating the result from the FEVER task, and further confirming that E-BART is flexible enough to provide explanations without compromising its performance on the main task. To compliment the BLEU metric, we provide some examples in Tables 5 and 6 for manual verification of explanation quality. The generated explanation in Table 5 provides insight into E-BART’s capability to focus only on the important details in the input. Here, Sentence 1 also contains superfluous information about an additional person in a wool hat, whereas Sentence 2 makes no mention of this. The generated explanation reflects this aspect, as it only discusses the differences between the two sentences in terms of the tan girl, rather than the person in the wool hat. Similarly, Table 6 provides an example of where the model has to extract a small, difficult detail in the input sentences that suggests the two sentences may not perfectly entail each other. The fact that Sentence 2 makes a stronger claim about the family going to the beach means that the text in Sentence 1 is not wholly

[2] https://nlp.stanford.edu/projects/snli/

Table 5. Ground truth label is “Entailment” and predicted label is “Entailment”.

Sentence 1	Sentence 2	Generated Explanation
One tan girl with a wool hat is running and leaning over an object, while another person in a wool hat is sitting on the ground.	A tan girl runs leans over an object	One tan girl with a wool hat is running and leaning over an object is a rephrasing of a tan girl runs leans over an object.

Table 6. Ground truth label is “Neutral” and predicted label is “Neutral”.

Sentence 1	Sentence 2	Generated Explanation
A young family enjoys feeling ocean waves lap at their feet.	A young man and woman take their child to the beach for the first time.	Just because a young family enjoys feeling ocean waves lap at their feet does not mean that they take their child to the beach for the first time.

representative of the situation. The E-BART model was able to identify this detail and thus label the sentences as ’neutral’. The generated explanation reflects this, providing insight into the fact that the model has identified that Sentence 2 contains important details omitted in Sentence 1.

4.7 Validating the Use of Joint Models: Experiment 1

To further investigate *RQ2* and test the ability of our joint models, we compare the performance of E-BART to a pipeline that produces a veracity prediction and generates an explanation using two independent models. To ensure that the results between the two methods are comparable, the architecture and training methodology was kept as consistent as possible. The separate pipeline, which we term Separate-BART, uses a BART-based sequence classifier, and a BART-based model for language generation. Both E-BART and Separate-BART were initialized with the same pre-trained weights, and were trained and evaluated on e-FEVER_Small. The BART classifier is trained with the claim and evidence as input to both the encoder and decoder, in contrast to E-BART which uses the much shorter gold summary as input to the decoder. This meant that due to memory constraints, the inputs were truncated to a maximum length of 256 tokens (which only truncated 4.56% of examples). In addition to this, a virtual batch size of 32 was used (batch size four, with eight gradient accumulation steps) to overcome convergence issues. When training the sequence generator model, a batch size of two with two gradient accumulation steps was used, also due to memory restrictions on available hardware for this experiment. In comparison, the joint model was trained with a batch size of four and no additional gradient accumulation. Before detailing the results, it is worth remarking the trade-off between effectiveness and resources needed to train a single model instead of two separate models. Training two separate models would require either the double amount of resources if they are trained in parallel (since each model needs to be fitted into a GPU device independently), or the double amount of time if they are trained sequentially. On the contrary, a single model has to cope with more parameters, but it can be trained on two tasks at the same time.

Table 7. Effectiveness of the joint and separate models on the eFEVER_Small dataset.

Model	**Accuracy no N.E.I**	**Accuracy full**	**Rouge 1**	**Rouge 2**	**Rouge L**	**Rouge Sum**
E-BART	87.2	78.2	73.581	64.365	71.434	71.585
Separate-BART	88.1	78.9	73.070	63.634	71.005	71.136

Table 8. Internal consistency of the joint and separate models on the eFEVER_Small dataset.

Model	**Accuracy no N.E.I**	**Accuracy full**
E-BART	91.8	86.8
Separate-BART	90.4	85.8

The results in Table 7 indicate that the prediction performance of both types of model is almost identical, with Separate-BART being slightly more effective. Manual inspection of the generated explanations revealed that both were of a similar quality in terms of expressiveness and cohesiveness. This experimental result reinforces what was seen in the practical evaluations on e-FEVER and e-SNLI: that E-BART is able to jointly provide an explanation without diminishing the performance on its main task.

4.8 Validating the Use of Joint Models: Experiment 2

This experiment aims to investigate whether the internal consistency between the predicted veracity and predicted explanation differs between the joint and separate models. We use the same E-BART and Separate-BART models from Experiment 1, but train an additional 'judge' model to predict the veracity of a claim, given an *explanation*. The ground truth veracity labels and dataset explanations from e-FEVER_Small were used to train the BART-based sequence classifier. As such, its weights are not conditioned on those of E-BART or Separate-BART, meaning that it is independent from both models.

We run the experiment by taking the claims from the development set and the predicted explanations from E-BART. The claims and explanations are then passed to the 'judge' model to produce a veracity prediction. This 'judge' veracity prediction is then compared against the veracity prediction from E-BART, and the accuracy is computed. The process was repeated for Separate-BART, and the results are presented in Table 8.

The results show a higher accuracy for E-BART as determined by the 'judge' model. This provides indication that the veracity prediction and explanation generated by E-BART are more consistent with each other than those generated by Separate-BART. Ultimately this means that joint models are one step closer to being truly interpretable compared to models that generate explanations separately in a post-hoc manner. While this is not conclusive proof, it does provide some evidence that there are consistency gains to be made when using joint prediction and explanation models.

4.9 Testing the Impact of Explanations

To address *RQ3*, we experimentally validated the benefit of explanations generated by our model with human annotators, performing a set of crowdsourcing studies as detailed in the following. The data used to perform this study can be found at https://github.com/KevinRoitero/predict-explain-truthfulness.

We collected the data using the Amazon MTurk crowdsourcing platform. To test the impact of machine-generated explanations of truthfulness, we deployed four versions of the same human annotation task. In each version of the task we provided participants with a claim from the FEVER dataset and we asked them to provide both a truthfulness assessment using the "True", "False" binary scale along with a sentence justifying their assessment, as this has been shown to improve assessment quality [15]. Each worker has been asked to assess the truthfulness of four claims, two labelled in the ground truth as "Supports", and two labelled as "Refutes". Each claim has been assessed by ten distinct human participants. To avoid bias, we performed a randomisation process while generating the claim-participant assignments (i.e., in the MTurk HITs). For consistency, we kept the same assignments (i.e., same HITs) for all the distinct versions of the task (apart for one case, as explained later). Participants were only allowed to complete one version of the task. To ensure high quality of the collected data and to avoid adversarial behaviour, we required participants to spend at least 2 seconds on each task page.

With the settings detailed above, we implemented and deployed the following four tasks. In the first version (Task 1), we provided participants with the claim from the FEVER dataset and we asked them to provide a truthfulness assessment and the justification. In the second version (Task 2), we provided participants with both the claim and the explanation generated by our E-BART system and ask for the truthfulness assessment and the justification. In the third version (Task 3), we provided participants with both the claim and the ground truth explanation, asking for the assessment and justification. Finally, the fourth and final version of the task (Task 4), we provided participants with the claim and both the ground truth and E-BART explanation, asking for an assessment and justification; we also asked them to indicate whether they preferred (i.e., found more informative) the ground truth explanation or the E-BART one. Inspecting the dataset, we found that for some HITs the ground truth and E-BART explanations where the same; thus, for this task we re-sampled the statements by requiring the two explanations to be different by at least 1 character.

The experimental setup detailed above allows us to make multiple comparisons, both implicit and explicit. By comparing Task 1 (i.e., no explanation) with Task 2 (i.e., E-BART explanation) and with Task 3 (i.e., ground truth explanation) we can both test the effect of showing the explanation to the worker, as well as to make an implicit comparison between the two explanations, the ground truth and the E-BART one. In addition, Task 4 (i.e., preferred explanation) allows us to to explicitly assess which explanation is preferred by the workers.

Figures 4 and 5 shows the external agreement between the ground truth and the crowd when considering both the individual participant judgments and the judgments aggregated over the ten participants assessing the same claim, using majority vote as aggregation function. By inspecting the plots, we can compute the accuracy scores for the different versions of the task, as follows. Task 1: 0.70 for raw and 0.83 for aggregated judgments; Task 2: 0.73 for raw and 0.90 for aggregated judgments; Task 3: 0.64 for raw and 0.65 for aggregated judgments; Task 4: 0.64 for raw and 0.71 for aggregated judgments. To account for statistical significance, we run a non-parametric ANOVA with post-hoc test[3]. Given that normality assumptions are violated, we use a non-parametric Kruskal-Wallis H test (one-way non-parametric ANOVA) to test the probability that samples came from the same distribution. Given that we obtained a p-value < 0.05 we may reject the null hypothesis that the population medians of all of the considered tasks are equal. To identify what tasks differ in their medians we need to run post-hoc tests; we considered the Conover's test. Results shows that the task pairs for which we can reject the null hypothesis (i.e., those which are statistically

[3] see https://scikit-posthocs.readthedocs.io/en/latest/.

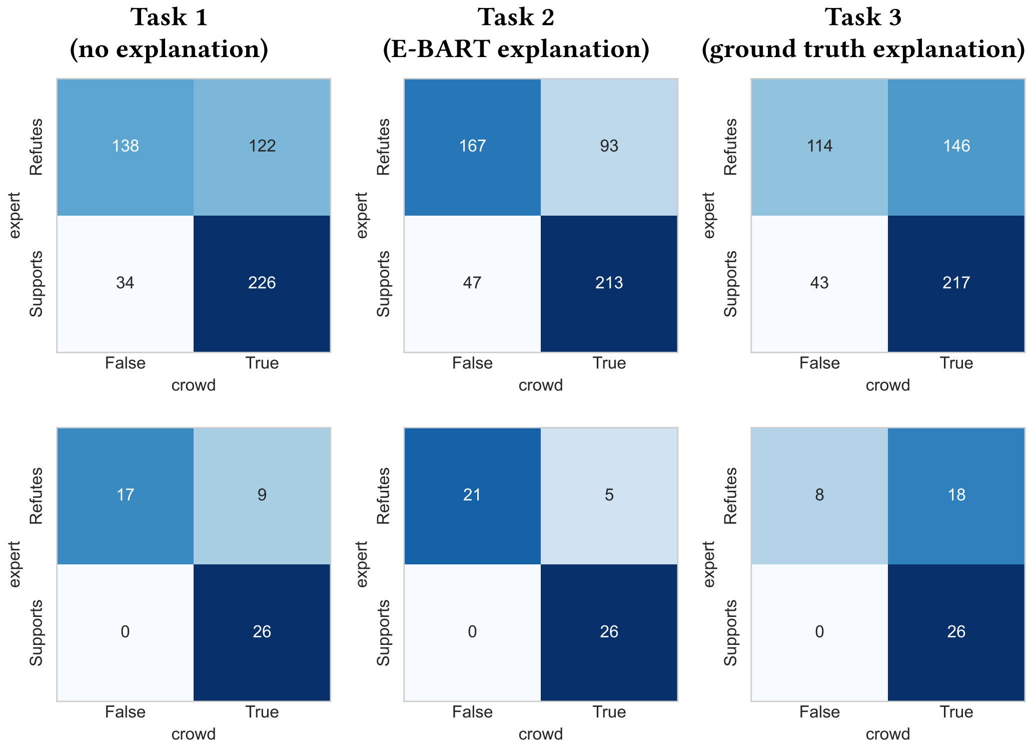


Fig. 4. External agreement between ground truth and crowd for raw (first row) and aggregated (second row) truthfulness assessments. Each cell represents either the count of judgments (first row), or claims (second row). Task 1 (first column) shows just the claim, Task 2 (second column) shows the claim and the natural language explanation generated by our E-BART model, and Task 3 (third column) shows the claim and the ground truth natural language explanation. Correctly classified statements lay on the main diagonal.

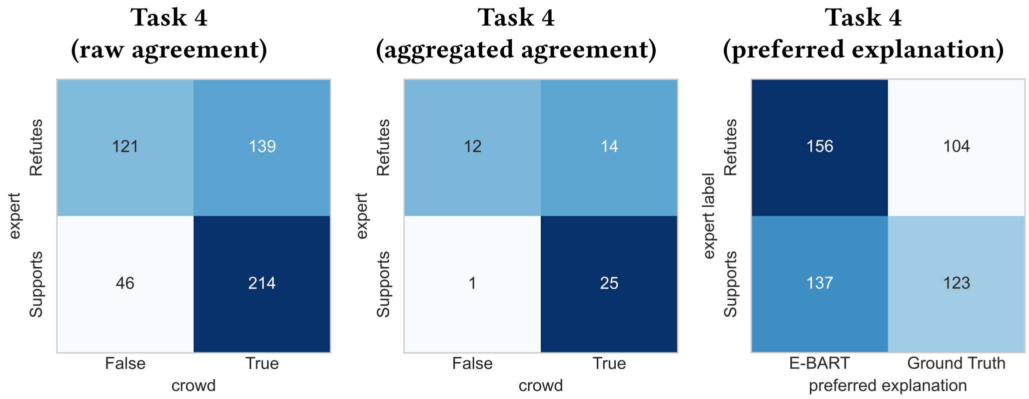


Fig. 5. External agreement between ground truth and crowd for raw (first column) and aggregated (second column) truthfulness assessments for Task 4, which shows the claim, ground truth, and E-BART explanations, asking workers for the preferred explanation. The worker preferences are shown in the third column. Each cell represents either the count of judgments (first and third columns), or claims (second column).

significantly different) are: Task 1 – Task 2 ($p < 0.05$), Task 2 – Task 3 ($p < 0.01$), and Task 2 – Task 4 ($p < 0.01$); those results hold for both raw and aggregated judgments.

Apart from the accuracy scores, it is worth inspecting and measuring the different levels of inter-annotator agreement. To this aim, we rely on Krippendorff's α, a well known and popular agreement measure used in crowdsourcing [14, 41]. Thus, we computed α on individual judgments and the agreement scores for the four tasks are respectively of 0.19 for Task 1, 0.24 for Task 2, 0.10 for Task 3, and 0.10 for Task 4. Those values indicate similar level of agreement between the tasks, which all have low agreement [3, 41].

From those results we can make the following remarks. It appears that showing the E-BART explanation (Task 2) is better than showing no explanation (Task 1), while this is not true when considering the ground truth explanation (comparison between Task 1 and Task 3). More than that, the implicit comparison between Task 2 and Task 3 shows that it appears that crowd workers are more accurate when considering the E-BART explanation in place of the ground truth explanation. Such result is also confirmed when making the explicit comparison (see last plot of Figure 5).

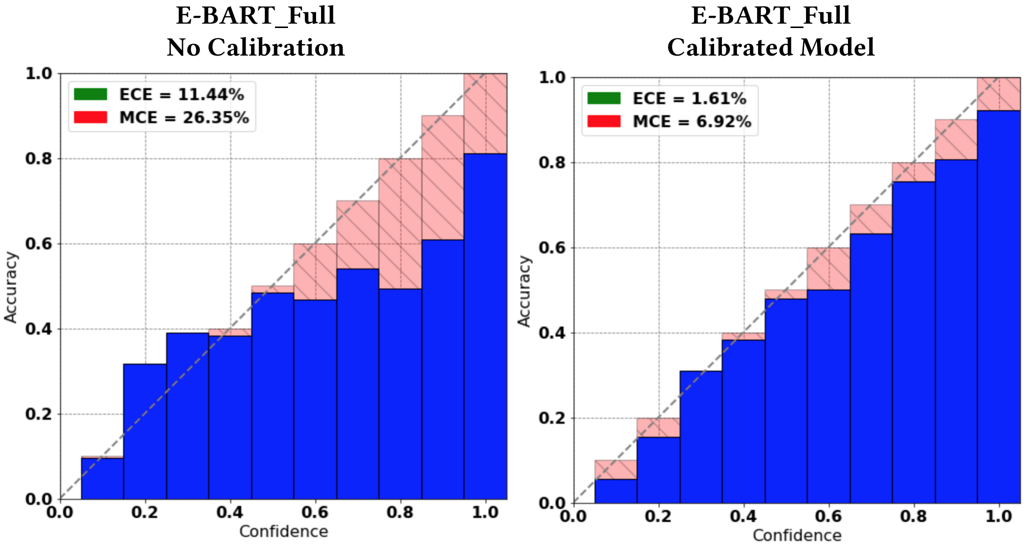
We can additionally observe that the display of E-BART explanation (i.e., Task 2) reduces the number of *false positives* (i.e., claims that are false but are erroneously perceived as being true by human subjects) from 122 to 93; such result is not true when considering the ground truth explanation. Thus, it appears that the explanations automatically generated by our E-BART model have the effect of making people more skeptical about claims (see also Table 3 for an example). This behavior does not hold for *false negatives* (i.e., claims that are true but are erroneously perceived as being false by human subjects). Note that in misinformation settings *false positives* are potentially more dangerous than *false negatives*, and it is better to be erroneously skeptical than to not recognize false statements.

By looking back at accuracy scores and their implications, we see that by performing a simple aggregation of crowd labels and under conditions of Task 2, we are able achieve 90% non-expert label accuracy, which is a promising step towards crowdsourced truthfulness annotations [27].

4.10 Network Calibration and Generation of Confidence Scores

To further enhance the transparency and interpretability of the E-BART predictions and address *RQ4*, it would be ideal to produce confidence scores along with the veracity classification. These confidence scores provide insight into how confident the model is in making its prediction. The confidence could be impacted by a number of factors, including the quality and quantity of the provided evidence, or the similarity of the claim to the training data. E-BART makes its predictions via a soft-max layer over a series of logits with dimensions equal to the number of target classes. While the output of the soft-max layer produces a 'probability' score for each class, it is unlikely that this probability is well-calibrated. That is, the output from the soft-max layer is not likely to be representative of the true probability of correctness [10]. Though this calibration error is present in most modern deep neural networks, it is not clear if such issue is present in all transformer-based models; thus, it is work investigating this issue for the E-BART model. A number of post-processing techniques exist to correct the calibration error, allowing the output of the final soft-max to be correctly interpreted as the confidence of correctness. Some techniques include histogram binning [40], Bayesian Binning into Quantiles [19], and Platt scaling [23]. Temperature scaling [10] is a technique that has demonstrated high efficacy on a range of neural networks including multi-class classifiers, and is relatively easy to implement. Temperature scaling introduces a single parameter, the temperature ($T > 0$), and uses it to produce a confidence prediction given a series of logits z_i:

$$\hat{q}_i = \max_k \sigma_{SM} \left(z_i/T\right)^k$$



E-BART_Full
No Calibration

E-BART_Full
Calibrated Model

Fig. 6. Confidence as a function of accuracy for the E-BART_Full model before and after calibration.

where σ_{SM} is the soft-max function. Applying temperature scaling to a model does not change its classification prediction, and therefore does not alter the accuracy of the model [10].

The temperature parameter must be tuned on the validation set rather than the original training set [10]. As there is no test split for the eFEVER dataset, the first 9999 items from the eFEVER development set were used for training the temperature parameter, with the remaining examples used for validation. The temperature parameter was inserted after the final fully-connected layer of the original model, and the original model parameters were frozen. The temperature parameter was then trained using the LBFGS optimiser for $10,000$ iterations. It is important to note that the E-BART model was run in "auto-regressive (inference) mode", to produce the logits for input to the temperature parameter. Finally, the model was tested using the held-out validation data. Testing proceeded by running E-BART in inference mode, and applying the temperature parameter prior to the final soft-max function.

To evaluate the calibration of the model, a reliability diagram was produced for the model before and after calibration. In addition, the Expected Calibration Error (ECE) and Maximum Calibration Error (MCE) were calculated using ten bins. Figure 6 shows the E-BART_Full model before and after calibration.

The dotted 45 degree line on the reliability diagrams specifies a perfect calibration. Deviation from this line indicates that the predicted confidence scores differ from the actual accuracy. Visual inspection of the reliability diagrams indicates that the original E-BART model was in fact not well calibrated, with ECE of 11.44% and MCE of 26.35%. However, following temperature scaling, the model performs more closely to the ideal calibration, with ECE and MCE reduced to 1.61% and 6.92%, respectively. This increase in calibration means that the output of the final soft-max layer can be more accurately interpreted as a confidence score. To demonstrate how this change impacts the model output, the examples in Tables 9 and 10 are provided: the first example shows a situation in which E-BART predicts the correct veracity, and indicates that it is confident in its prediction. The second example demonstrates a situation where the model is not confident in its prediction. Here, the ground truth veracity indicates that the claim is refuted, and even the ground truth explanation

Table 9. Ground truth label is “Supports” and predicted label is “Supports”, confidence score is 0.851, original confidence was 0.987 (E-BART_Full Model).

Claim	**Evidence**	**Gold Explanation**	**Generated Explanation**
Ekta Kapoor worked on an Indian soap opera that premiered in 2000.	Kyunki Saas Bhi Kabhi Bahu Thi (Because a mother-in-law was once a daughter-in-law , too) is an Indian soap opera that premiered on 3 July 2000 on Star Plus. Ekta Kapoor Some soap operas she had worked on include [...]. Pavitra Rishta (Sacred Ties) is a 2009 Indian soap opera produced by Ekta Kapoor of Bal-aji Telefilms , that aired on Zee TV. Kasautii Zindagii Kay (The criterion of life), often abbreviated as KZK, is an Indian soap opera created by Ekta Kapoor ’s Balaji Tele-films for the channel STAR Plus.	Ekta Kapoor worked on the Indian soap opera Kyunki Saas Bhi Kabhi Bahu Thi, which premiered in 2000.	Ekta Kapoor worked on the soap opera Kyunki Saas Bhi Kabhi Bahu Thi, which premiered in 2000.

Table 10. Ground truth label is “Refutes” and predicted label is “Supports”, confidence score is 0.432, original confidence was 0.550 (E-BART_Full Model).

Claim	**Evidence**	**Gold Explanation**	**Generated Explanation**
Henry III assumed the throne when he was 2 years old.	Henry III of France Henry III (19 September 1551 - 2 August 1589 ; born Alexandre Édouard de France , Henryk Walezy , Henrikas Valua) was a monarch of the House of Valois who was elected the monarch of the Polish-Lithuanian Commonwealth from 1573 to 1575 and ruled as King of France from 1574 until his death.	Henry III was elected the monarch of the Polish-Lithuanian Commonwealth when he was 2 years old.	Henry III was born on 19 September 1551 and died on 2 August 1589.

does a poor job of communicating the evidence. In this case, E-BART predicts that the claim is supported, however it indicates that it has low confidence in this prediction. Without providing the confidence, the only indication that perhaps the model is not accurate in this instance is the low quality of the explanation. Now, however, there is a quantitative representation of the confidence. This would demonstrate to the end user that in this instance, the model should not be trusted.

5 CONCLUSIONS

In this paper we explored the potential of AFC models jointly making a prediction and providing a human-readable explanation for that prediction. To this end, we proposed the E-BART architecture and evaluated its performance on the extended FEVER and SNLI tasks. Experimentation revealed that E-BART could achieve results comparable to the state-of-the-art and simultaneously generate coherent and relevant explanations. We argued that jointly predicting explanations makes AFC systems more transparent, and fosters greater trust in the system. Human evaluation of the impact of generated explanations revealed that the explanations provided by E-BART generally make people more accurate in detecting misinformation and more skeptical of a claim they encounter online. Finally, we calibrated the E-BART architecture in order to make it produce reliable confidence scores for the truthfulness predictions.

The model presented in this paper opens to a plenty of future work which is made possible because of the unique joint modeling technique used. In future work we plan to exploit saliency maps to investigate the effect of evidence on classification performance and vice-versa investigating the relationships among the two models used in our network, the discriminative and the generative one.

ACKNOWLEDGMENTS

This work is supported by a Facebook Research award, the Australian Research Council (Grant No. DP190102141), and by the ARC Training Centre for Information Resilience (Grant No. IC200100022).

REFERENCES

[1] Naser Ahmadi, Joohyung Lee, Paolo Papotti, and Mohammed Saeed. 2019. Explainable Fact Checking with Probabilistic Answer Set Programming. In *Proceedings of the 2019 Truth and Trust Online Conference*.

[2] 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*. Association for Computational Linguistics, 7352–7364.

[3] Dylan T Beckler, Zachary C Thumser, Jonathon S Schofield, and Paul D Marasco. 2018. Reliability in evaluator-based tests: using simulation-constructed models to determine contextually relevant agreement thresholds. *BMC medical research methodology* 18, 1 (2018), 1–12.

[4] 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*. Association for Computational Linguistics, Lisbon, Portugal, 632–642.

[5] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems*.

[6] Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-SNLI: Natural Language Inference with Natural Language Explanations. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems*, Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett (Eds.). Montréal, Canada, 9560–9572.

[7] Ronald Denaux and Jose Manuel Gomez-Perez. 2020. Linked Credibility Reviews for Explainable Misinformation Detection. In *The Semantic Web – ISWC 2020*. Springer International Publishing, 147–163.

[8] 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*. Association for Computational Linguistics, Minneapolis, USA, 4171–4186.

[9] Lucas Graves. 2018. Understanding the Promise and Limits of Automated Fact-Checking. *Reuters Institute for the Study of Journalism* (2018).

[10] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On Calibration of Modern Neural Networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70* (Sydney, NSW, Australia) *(ICML'17)*. JMLR.org, 1321–1330.

[11] Andreas Hanselowski, Hao Zhang, Zile Li, Daniil Sorokin, Benjamin Schiller, Claudia Schulz, and Iryna Gurevych. 2018. UKP-Athene: Multi-Sentence Textual Entailment for Claim Verification. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*.

[12] Seonhoon Kim, Inho Kang, and Nojun Kwak. 2019. Semantic Sentence Matching with Densely-Connected Recurrent and Co-Attentive Information. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence*. AAAI Press, Honolulu, Hawaii, USA, 6586–6593.

[13] Neema Kotonya and Francesca Toni. 2020. Explainable Automated Fact-Checking for Public Health Claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 7740–7754.

[14] Klaus Krippendorff. 2011. Computing Krippendorff's alpha-reliability. (2011).

[15] Mücahid Kutlu, Tyler McDonnell, Tamer Elsayed, and Matthew Lease. 2020. Annotator Rationales for Labeling Tasks in Crowdsourcing. *Journal of Artificial Intelligence Research* 69 (2020), 143–189.

[16] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 7871–7880.

[17] Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2019. Multi-Task Deep Neural Networks for Natural Language Understanding. In *Proceedings of the 57th Conference of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 4487–4496.

[18] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692* (2019).

[19] Mahdi Pakdaman Naeini, Gregory F. Cooper, and Milos Hauskrecht. 2015. Obtaining Well Calibrated Probabilities Using Bayesian Binning. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence* (Austin, Texas) *(AAAI'15)*. AAAI Press, 2901–2907.

[20] Yixin Nie, Haonan Chen, and Mohit Bansal. 2019. Combining Fact Extraction and Verification with Neural Semantic Matching Networks. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence*, Vol. 33. Honolulu, Hawaii, USA, 6859–6866.

[21] Gordon Pennycook, Ziv Epstein, Mohsen Mosleh, Antonio A Arechar, Dean Eckles, and David G Rand. 2021. Shifting attention to accuracy can reduce misinformation online. *Nature* 592, 7855 (2021), 590–595.

[22] Jonathan Pilault, Amine Elhattami, and Christopher J. Pal. 2021. Conditionally Adaptive Multi-Task Learning: Improving Transfer Learning in NLP Using Fewer Parameters & Less Data. In *Proceedings of the 9th International Conference on Learning Representations*.

[23] John Platt et al. 1999. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers* 10, 3 (1999), 61–74.

[24] Beatrice Portelli, Jason Zhao, Tal Schuster, Giuseppe Serra, and Enrico Santus. 2020. Distilling the Evidence to Augment Fact Verification Models. In *Proceedings of the Third Workshop on Fact Extraction and VERification (FEVER)*. Association for Computational Linguistics, 47–51.

[25] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving Language Understandingby Generative Pre-Training. (2018).

[26] 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*. Association for Computational Linguistics, Hong-Kong, China, 3980–3990.

[27] Kevin Roitero, Michael Soprano, Shaoyang Fan, Damiano Spina, Stefano Mizzaro, and Gianluca Demartini. 2020. Can The Crowd Identify Misinformation Objectively? The Effects of Judgment Scale and Assessor's Background. In *Proceedings of the 43rd International ACM SIGIR Conference*. Association for Computing Machinery, New York, NY, USA, 439–448.

[28] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108* (2019).

[29] Sam Shleifer and Alexander M Rush. 2020. Pre-trained Summarization Distillation. *arXiv preprint arXiv:2010.13002* (2020).

[30] Kai Shu, Limeng Cui, Suhang Wang, Dongwon Lee, and Huan Liu. 2019. dEFEND: Explainable Fake News Detection. In *Proceedings of the 25th ACM SIGKDD International Conference* (Anchorage, AK, USA). Association for Computing Machinery, New York, NY, USA, 395–405.

[31] Amir Soleimani, Christof Monz, and Marcel Worring. 2020. BERT for Evidence Retrieval and Claim Verification. In *Proceedings of the 42nd European Conference on IR Research*, Vol. 12036. Springer, Lisbon, Portugal, 359–366.

[32] Dominik Stammbach and Elliott Ash. 2020. e-FEVER: Explanations and Summaries for Automated Fact Checking. In *Proceedings of the 2020 Truth and Trust Online Conference (TTO 2020)*. Hacks Hackers, 32.

[33] Dominik Stammbach and Guenter Neumann. 2019. Team DOMLIN: Exploiting evidence enhancement for the FEVER shared task. In *Proceedings of the Second Workshop on Fact Extraction and VERification (FEVER)*. 105–109.

[34] James Thorne, Andreas Vlachos, Oana Cocarascu, Christos Christodoulopoulos, and Arpit Mittal. 2018. The Fact Extraction and VERification (FEVER) Shared Task. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)"*. Association for Computational Linguistics, Brussels, Belgium, 1–9.

[35] Ehsan Toreini, Mhairi Aitken, Kovila Coopamootoo, Karen Elliott, Carlos Gonzalez Zelaya, and Aad van Moorsel. 2020. The Relationship between Trust in AI and Trustworthy Machine Learning Technologies. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Barcelona, Spain). Association for Computing Machinery,

New York, NY, USA, 272–283.

[36] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Proceedings of the Annual Conference on Neural Information Processing Systems*. Long Beach, CA, USA, 5998–6008.

[37] 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*. Association for Computational Linguistics, Baltimore, MD, USA, 18–22.

[38] Lianwei Wu, Yuan Rao, Yongqiang Zhao, Hao Liang, and Ambreen Nazir. 2020. DTCA: Decision Tree-based Co-Attention Networks for Explainable Claim Verification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault (Eds.). Association for Computational Linguistics, 1024–1035.

[39] Takuma Yoneda, Jeff Mitchell, Johannes Welbl, Pontus Stenetorp, and Sebastian Riedel. 2018. UCL Machine Reading Group: Four Factor Framework For Fact Finding (HexaF). In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*. Association for Computational Linguistics, Brussels, Belgium, 97–102.

[40] Bianca Zadrozny and Charles Elkan. 2001. Obtaining Calibrated Probability Estimates from Decision Trees and Naive Bayesian Classifiers. In *Proceedings of the Eighteenth International Conference on Machine Learning (ICML '01)*. 609–616.

[41] Antonia Zapf, Stefanie Castell, Lars Morawietz, and André Karch. 2016. Measuring inter-rater reliability for nominal data–which coefficients and confidence intervals are appropriate? *BMC medical research methodology* 16, 1 (2016), 1–10.

[42] Zhuosheng Zhang, Yuwei Wu, Zuchao Li, and Hai Zhao. 2018. Explicit Contextual Semantics for Text Comprehension. *arXiv preprint arXiv:1809.02794* (2018).

[43] Zhuosheng Zhang, Yuwei Wu, Hai Zhao, Zuchao Li, Shuailiang Zhang, Xi Zhou, and Xiang Zhou. 2020. Semantics-Aware BERT for Language Understanding. In *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence*. AAAI Press, New Your, NY, USA, 9628–9635.