E-BART: Jointly Predicting and Explaining Truthfulness

Erik Brand
The University of Queensland
e.brand@uq.net.au

Kevin Roitero
University of Udine
roitero.kevin@spes.uniud.it

Michael Soprano
University of Udine
michael.soprano@uniud.it

Gianluca Demartini
The University of Queensland
demartini@acm.org

Abstract

Automated fact-checking (AFC) systems exist to combat disinformation, however their complexity makes them opaque to the end user, making it difficult to foster trust. 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. Finally, we also conduct human evaluations on the impact of generated explanations and observe that explanations increase human ability to spot misinformation and make people more skeptical about claims.

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 (Stammbach and Ash, 2020). This is a challenging task for a human, let alone an autonomous system (Graves, 2018). However, AFC systems are able to approximate this process of evidence retrieval and synthesis with some degree of success (Stammbach and Ash, 2020; Vlachos and Riedel, 2014). 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 (Pennycook et al., 2021). AFC systems are starting to become a critical tool in combating the sheer quantity of claims that need to be verified.

While accurate (Stammbach and Ash, 2020; Portelli et al., 2020), 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 (Toreini et al., 2020), especially as it is not generated by humans.

Automated systems have recently been developed to this effect, and have demonstrated promising initial results (Graves, 2018). 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?

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. (2019) and Wu et al. (2020), 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 (Denaux and Gomez-Perez, 2020), rule mining, and probabilistic answer set programming (Ahmadi et al., 2019) 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 (Ahmadi et al., 2019).

Summarisation techniques provide an explanation by summarising the retrieved evidence. The system proposed by Atanasova et al. (2020) utilises DistilBERT (Sanh et al., 2019) 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 (2020) take a similar approach but tailor their model to the public health domain. The pipeline utilises Sentence-BERT (Reimers and Gurevych, 2019) to filter the evidence, a BERT-based veracity predictor, and a separate BERT-based summarisation model. The work by Kotonya and Toni (2020) differs from Atanasova et al. (2020) 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 (2020) 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 (Stammbach and Neumann, 2019), which uses separate BERT-based models for evidence retrieval and veracity prediction. For explanation generation, GPT-3 (Brown et al., 2020), 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 (Lewis et al., 2020) is a transformer (Vaswani et al., 2017) model that aims to generalise the capabilities of both BERT (Devlin et al., 2019) 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 (Liu et al., 2019b), 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 (Lewis et al., 2020; Shleifer and Rush, 2020). 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. 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 depicted in green in Figure 1. 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 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. 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.

During training, the encoder generates hidden

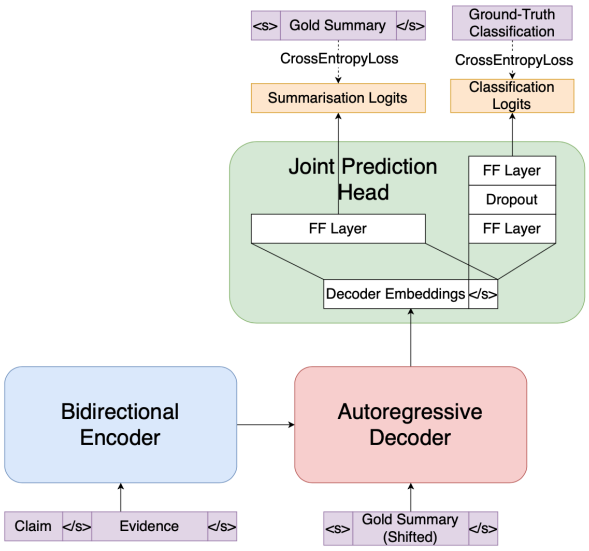

Figure 1: E-BART Training configuration.

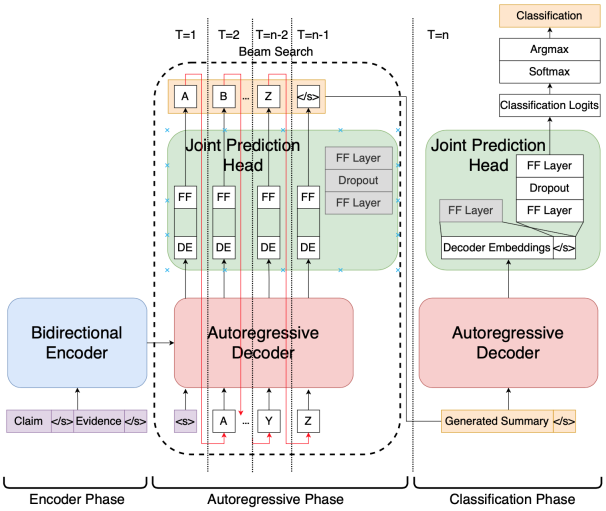

Figure 2: E-BART Inference process.

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. 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.

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. 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.

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 (Thorne et al., 2018).

The e-FEVER dataset by Stammbach and Ash (2020) augments the original FEVER dataset (Thorne et al., 2018) with explanations generated by their framework. It consists of 50,000 examples from the FEVER 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 (Camburu et al., 2018) extends the SNLI dataset (Bowman et al., 2015) 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.

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. In the e-FEVER dataset, if the GPT-3 component decided that the retrieved evidence was insufficient, it would produce a default 'null' explanation. 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 way as the first, however it was trained using the entire e-FEVER training set, including examples with null explanations.

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".

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 (Stammbach and Neumann, 2019) 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

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".

accuracies on this dataset include: BERT-BASED 74.6 (Soleimani et al., 2020), DOMLIN 72.1 (Stammbach and Neumann, 2019), UCL MR 69.7 (Yoneda et al., 2018), UNC 69.6 (Nie et al., 2019), and UKP-Athene 68.5 (Hanselowski et al., 2018). 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.

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.

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 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 (Camburu et al., 2018). In calculating the BLEU metric for the explainable models, the first two gold explanations were used as references as per (Camburu et al., 2018). As a further comparison, the following are the best performing models published in the literature[1] which, however, do not provide explanations: CA-MTL 92.1 (Pilault et al., 2021), SemBERT 91.9 (Zhang et al., 2020), MT-DNN 91.6 (Liu et al., 2019a), SJRC 91.3 (Zhang et al., 2018), D-CRCo-AN 90.1 (Kim et al., 2019), and LMTransformer 89.9 (Radford et al., 2018).

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

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

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

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.

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.

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 initialised with the same pre-trained weights, and were trained and evaluated on e-FEVER_Small. However 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. In comparison, the joint model was trained with a batch size of four and no additional gradient accumulation.

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

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.

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.

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 the crowdsourcing study detailed in the following. We collected the data using the Amazon MTurk crowdsourcing platform. To test the impact of machine-generated explanations of truthfulness, we deployed two versions of the same human annotation task. In the first version (Task 1), we provided participants with a claim from the FEVER dataset and we asked them to provide a truthfulness assessment on the true/false binary scale along with a sentence justifying their assessment, as this has been shown to improve assessment quality (Kutlu et al., 2020). 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 an assessment and justification. Thus, the only difference between Task 1 and 2 is the absence or presence of the E-BART explanation. 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 the two versions of the task. 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.

Figure 3 shows the external agreement between the ground truth and the crowd when considering both the individual participant judgments and the judgements aggregated over the ten participants assessing the same claim using majority vote. As we can see from the figure, the accuracy of the participants in Task 2 is higher than the ones of Task 1 (0.70 for Task 1 and 0.74 for Task 2 for the individual; 0.83 for Task 1 and 0.90 for Task 2 for the aggregated judgments). The differences are statistically significant according to the Mann-Whitney

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.

Table 7: Effectiveness of the joint and separate models.

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.

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

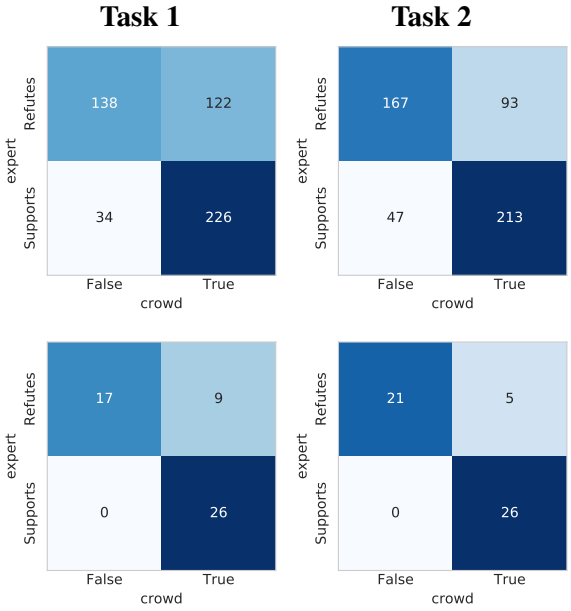

Figure 3: External agreement between ground truth and crowd for raw (first row) and aggregated (second row) truthfulness assessments. Task 1 shows just the claim while Task 2 shows the claim and the natural language explanation generated by our E-BART model.

U test at the $p < 0.05$ level for both the individual and the aggregated judgements. We can additionally observe that the display of explanations (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; 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). Performing simple aggregations and under condition of Task 2, we are able achieve 90% non-expert label accuracy, which is a promising step towards crowdsourced truthfulness annotations (Roitero et al., 2020).

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. Finally, 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.

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

References

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*.

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. 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.

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*.

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*, pages 9560–9572, Montréal, Canada.

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

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*, pages 4171–4186, Minneapolis, USA. Association for Computational Linguistics.

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

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)*.

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*, pages 6586–6593, Honolulu, Hawaii, USA. AAAI Press.

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*, pages 7740–7754. Association for Computational Linguistics.

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

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*, pages 7871–7880. Association for Computational Linguistics.

Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2019a. Multi-Task Deep Neural Networks for Natural Language Understanding. In *Proceedings of the 57th Conference of the Association for Computational Linguistics*, pages 4487–4496, Florence, Italy. 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. 2019b. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692*.

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*, volume 33, pages 6859–6866, Honolulu, Hawaii, USA.

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):590–595.

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*.

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)*, pages 47–51. Association for Computational Linguistics.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving Language Understandingby Generative Pre-Training.

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*, pages 3980–3990, Hong-Kong, China. Association for Computational Linguistics.

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*, page 439–448, New York, NY, USA. Association for Computing Machinery.

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*.

Sam Shleifer and Alexander M Rush. 2020. Pre-trained Summarization Distillation. *arXiv preprint arXiv:2010.13002*.

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*, page 395–405, New York, NY, USA. Association for Computing Machinery.

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*, volume 12036, pages 359–366, Lisbon, Portugal. Springer.

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)*, page 32. Hacks Hackers.

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)*, pages 105–109.

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)"*, pages 1–9, Brussels, Belgium. Association for Computational Linguistics.

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*, page 272–283, New York, NY, USA. Association for Computing Machinery.

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*, pages 5998–6008, Long Beach, CA, USA.

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.

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*, pages 1024–1035. Association for Computational Linguistics.

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)*, pages 97–102, Brussels, Belgium. Association for Computational Linguistics.

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

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*, pages 9628–9635, New Your, NY, USA. AAAI Press.