Explaining Predictive Uncertainty by Looking Back at Model Explanations

Hanjie Chen, Wanyu Du, Yangfeng Ji

Department of Computer Science
University of Virginia
{hc9mx, wd5jq, yangfeng}@virginia.edu

Abstract

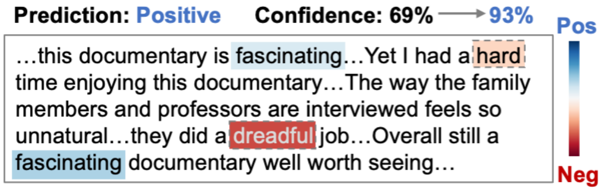
Predictive uncertainty estimation of pre-trained language models is an important measure of how likely people can trust their predictions. However, little is known about what makes a model prediction uncertain. Explaining predictive uncertainty is an important complement to explaining prediction labels in helping users understand model decision making and gaining their trust on model predictions, while has been largely ignored in prior works. In this work, we propose to explain the predictive uncertainty of pre-trained language models by extracting uncertain words from existing model explanations. We find the uncertain words are those identified as making negative contributions to prediction labels, while actually explaining the predictive uncertainty. Experiments show that uncertainty explanations are indispensable to explaining models and helping humans understand model prediction behavior.

1 Introduction

Pre-trained language models (e.g., BERT; Devlin et al. 2019) have been indispensable to natural language processing (NLP) due to their remarkable performance (Liu et al. 2019; Yang et al. 2019; Gururangan et al. 2020; Brown et al. 2020). Predictive uncertainty estimation of pre-trained language models is an important measure of how likely people can trust their predictions (Desai and Durrett 2020; Xu, Desai, and Durrett 2020).

A typical way of measuring predictive uncertainty is to calibrate model outputs with the true correctness likelihood (Guo et al. 2017; Kong et al. 2020; Zhao et al. 2021), so that the output probabilities well represent the confidence of model predictions. In this case, higher prediction confidence indicates lower uncertainty (Xu, Desai, and Durrett 2020; Jiang et al. 2021).

However, little is known about what makes a model prediction uncertain. Explaining predictive uncertainty is important to understanding model prediction behavior and complementary to explaining prediction labels for gaining users' trust, while has been largely ignored (Antorán et al. 2020). Most works on model explanations focus on explaining a model from the post-hoc manner by identifying important features in inputs that contribute to model predicted labels (Ribeiro, Singh, and Guestrin 2016; Lundberg and Lee 2017; Sundararajan, Taly, and Yan 2017; Chen, Zheng, and Ji 2020; Chen et al. 2021). Figure 1 shows an example of model explanation for sentiment classification, where the model makes the correct prediction (POSITIVE) with a relatively low confidence 69%. The top two salient words highlighted in blue color explain the predicted label. However, users may still wonder what compromises the prediction confidence?



Prediction: Positive Confidence: 69% → 93%

…this documentary is fascinating…Yet I had a hard time enjoying this documentary…The way the family members and professors are interviewed feels so unnatural…they did a dreadful job…Overall still a fascinating documentary well worth seeing…

Figure 1: An illustration of model explanation for sentiment classification, where the model makes the correct prediction (POSITIVE) with a relatively low confidence 69%. The top and bottom salient words with respect to the predicted label are highlighted in blue and red colors respectively, indicating different sentiment polarities. Darker color implies larger attribution. Removing the two bottom salient words in dashed boxes can improve the model prediction confidence to 93%.

This work is the first to explain model predictive uncertainty in NLP. Specifically, this work is based on a simple observation that bottom salient words in model explanations (e.g., `dreadful` and `hard` in Fig. 1) identified as making negative contributions to predicted labels actually explain model predictive uncertainty. The two bottom salient words in Fig. 1 indicate the opposite sentiment (NEGATIVE) to the model predicted label. Removing them can improve the model prediction confidence from 69% to 93%. We argue that both top and bottom salient words are indispensable to explaining model predictions. We name top salient words as *important words*, explaining model predicted labels; and bottom salient words as *uncertain words*, explaining model predictive uncertainty. In other words, a comprehensive prediction explanation should consist of *label explanation* with important words and *uncertainty explanation* with uncertain words.

The goal of this work is to demonstrate **the benefits of**

comprehensive explanations and the necessity of including uncertainty explanations. In the empirical study, we adopt two explanation methods, Leave-one-out (Li, Monroe, and Jurafsky 2016) and Sampling Shapley (Strumbelj and Kononenko 2010), to explain two pre-trained language models, BERT (Devlin et al. 2019) and RoBERTa (Liu et al. 2019) on three tasks. Experiments show the effectiveness of the two methods in identifying uncertain words for explaining model predictive uncertainty. Besides, human evaluations illustrate the indispensability of uncertainty explanations in helping humans understand model prediction behavior.

2 Related Work

The problem of predictive uncertainty estimation has been well studied (Kuleshov and Liang 2015; Gal and Ghahramani 2016; Pereyra et al. 2017; Kumar, Sarawagi, and Jain 2018; Liu et al. 2020; Kong et al. 2020; Jiang et al. 2021). However, little is known about what causes predictive uncertainty. Extensive literatures on model explanations focus on explaining model predicted labels, while ignoring predictive uncertainty (Ribeiro, Singh, and Guestrin 2016; Lundberg and Lee 2017; Sundararajan, Taly, and Yan 2017; Chen, Zheng, and Ji 2020; Chen et al. 2021). However, explaining predictive uncertainty is an important complement to explaining predicted labels for improving model trustworthiness (Antorán et al. 2020; Perez et al. 2022).

There is limited work on studying the source of predictive uncertainty in NLP. For example, previous works on explaining uncertainty estimates mainly focus on tabular and image data (Antorán et al. 2020; Ley, Bhatt, and Weller 2021; Perez et al. 2022). Feng et al. (2018) observed that prediction confidence increases with input reduction, while focusing on model pathologies as reduced inputs lack predictive information. Differently, we focus on identifying uncertain words in inputs for explaining model predictive uncertainty. To the best of our knowledge, this is the first work on explaining predictive uncertainty of pre-trained language models in NLP.

3 Explaining Predictive Uncertainty

In this work, we consider models that are calibrated, such that their prediction confidence is aligned with their prediction probability. Let $f(\cdot)$ denote a model. Given an input $\boldsymbol{x} = [x_1, \dots, x_N]$ consisting of N words, the model prediction probabilities on $\boldsymbol{x}$ over classes are $[f_1(\boldsymbol{x}), \dots, f_C(\boldsymbol{x})]$, where $f_c(\boldsymbol{x}) = P(y = c \mid \boldsymbol{x})$ and C is the total number of classes. As model f is calibrated, the probability on the predicted class $\hat{y}$, i.e. $f_{\hat{y}}(\boldsymbol{x})$, represents the model prediction confidence on this label. As prediction confidence and predictive uncertainty are negative correlated (higher confidence implies lower uncertainty), we explain model predictive uncertainty by answering the question: *What drags model prediction confidence down?* We answer the question based on a simple observation on model prediction explanations (Ribeiro, Singh, and Guestrin 2016; Li, Monroe, and Jurafsky 2016; Lundberg and Lee 2017).

When a feature is identified with negative contribution, removing it can improve model prediction confidence, as shown in Fig. 1. Similar to the definition of prediction explanation, we consider this feature explains predictive uncertainty. Furthermore, given a ranking of input word contributions produced by an explanation method, we name top-ranked words as *important words*, explaining model predicted labels; and bottom words (with negative contributions) as *uncertain words*, explaining model predictive uncertainty. In other words, a comprehensive prediction explanation should consist of *label explanation* with important words and *uncertainty explanation* with uncertain words. As mentioned before, the goal of this study is to demonstrate the benefits of comprehensive explanations and the necessity of including uncertainty explanations. In this work, we focus on extracting uncertain words from *existing* explanation methods, with the expectation of stimulating further research on explaining predictive uncertainty in NLP.

3.1 Explanation Methods

With the previous discussion, we adopt two perturbation-based explanation methods, Leave-one-out (Li, Monroe, and Jurafsky 2016) and Sampling Shapley (Strumbelj and Kononenko 2010), for uncertainty explanations. Other explanation methods can be easily adapted to explaining predictive uncertainty.

Leave-one-out (LOO). This method evaluates the effect of each word on model prediction by leaving it out and observing the output probability change on the predicted class. We define a contribution score for each word as

$$S_i = f_{\hat{y}}(\boldsymbol{x}) - f_{\hat{y}}(\boldsymbol{x}_{\setminus i}), \tag{1}$$

where $\boldsymbol{x}_{\setminus i}$ denotes the input with the word $\boldsymbol{x}_i$ removed. The contribution score S_i quantifies how much the model prediction confidence decreases when $\boldsymbol{x}_i$ is left out.

Sampling Shapley (SS). This method computes feature contributions in a more sophisticated way by considering coalitions between words. Specifically, for a word $\boldsymbol{x}_i$, its contribution score is computed as

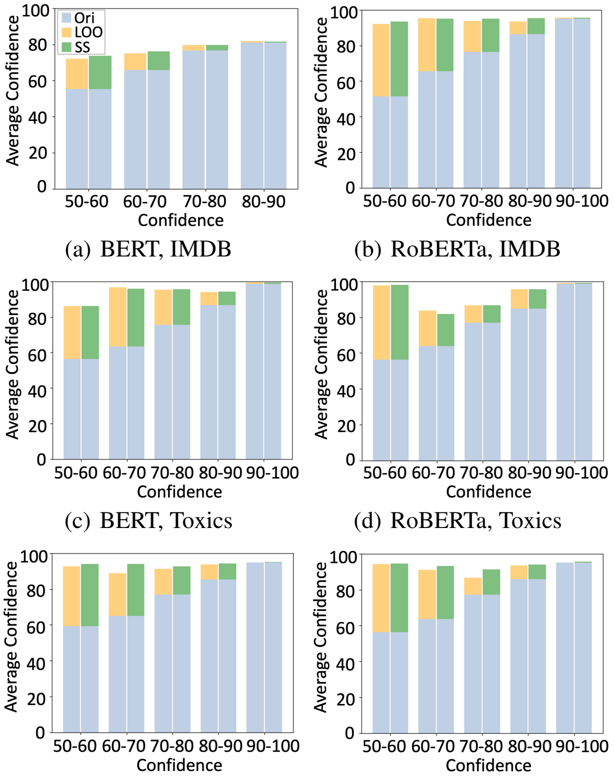
$$S_i = \frac{1}{M} \sum_{m=1}^{M} f_{\hat{y}}(\boldsymbol{x}_{\setminus i}^{(m)} \cup \{x_i\}) - f_{\hat{y}}(\boldsymbol{x}_{\setminus i}^{(m)}), \tag{2}$$

where M is the number of samples, and $\boldsymbol{x}_{\setminus i}^{(m)} \subseteq \boldsymbol{x}_{\setminus i}$ contains a subset of words in $\boldsymbol{x}_{\setminus i}$. The contribution score quantifies the overall contribution of the word $\boldsymbol{x}_i$ to the predicted label over M ensembles. In experiments, we set $M = 200$.

For each prediction, both methods produce an explanation with input word contributions, from which we extract important and uncertain words as label and uncertainty explanations respectively.

4 Setup

Models and datasets. We evaluate two pre-trained language models, BERT (Devlin et al. 2019) and RoBERTa (Liu et al. 2019), on three tasks, including sentiment analysis, toxic comments detection and political bias classification. We utilize the IMDB (Maas et al. 2011) dataset for sentiment analysis, Wikipedia Toxicity Corpus (Toxics) (Wulczyn, Thain, and Dixon 2017) for toxic comments detection,



(a) BERT, IMDB

(b) RoBERTa, IMDB

(c) BERT, Toxics

(d) RoBERTa, Toxics

(e) BERT, Politics

(f) RoBERTa, Politics

Figure 2: Average confidence (%) changes with uncertain words removed. X-axis shows different bins of original confidence. Ori: original confidence; LOO: Leave-one-out; SS: Sampling Shapley.

and Senator Tweets (Politics) [1] for political bias classification. More details about the models and datasets are in Appendix A.1.

Posterior calibration. We follow Desai and Durrett (2020) and calibrate the models on each dataset via temperature scaling (Guo et al. 2017), so that their output probabilities on predicted labels well represent prediction confidence. More details of model calibration are in Appendix A.2.

5 Experiments

In our experiments, we focus on the three research questions: (1) How effectively existing model explanation methods can identify uncertain words? (2) What insights we can obtain from uncertainty explanations in addition to label explanations? (3) Whether users appreciate uncertainty explanations in understanding model prediction behavior?

5.1 Quantitative Evaluation

For each dataset, we randomly select 1000 test examples and generate explanations for model predictions on them (see visualizations in Table 8). The following two results answer the research question (1) and (2) respectively.

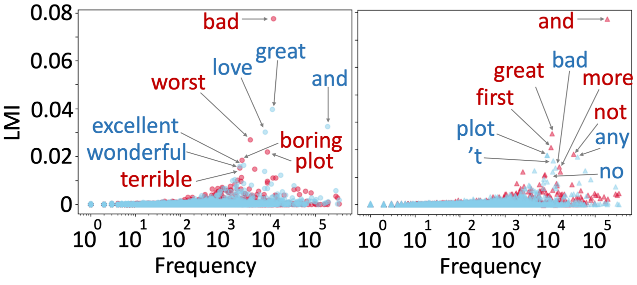


Figure 3: LMI distributions based on important words (a) and uncertain words (b). The x-axis represents word frequency in the vocabulary built on the IMDB dataset. We use blue and red colors to distinguish features associated with the POSITIVE and NEGATIVE labels respectively. Top 5 tokens in each distribution are pointed out.

Existing model explanation methods effectively identify uncertain words that limit model prediction confidence. We extract top k uncertain words identified by model explanations and remove them from inputs and then compute the average prediction confidence change in each bin of original confidence. We empirically set $k = 10$ for IMDB and $k = 5$ for Toxics and Politics based on their average sentence lengths in Table 2. Figure 2 shows that both LOO and SS capture uncertain words that limit prediction confidence. Overall, SS performs better than LOO in identifying uncertain words.

Important words and negations can result in uncertain predictions. We analyze feature statistics of model explanations via local mutual information (LMI) (Schuster et al. 2019; Du et al. 2021). LMI quantifies the association between a feature (an important/uncertain word) and a prediction label in model explanations (Chen et al. 2022). The details of computing LMI are in Appendix A.3. We analyze explanations generated by SS for RoBERTa on the IMDB dataset. Figure 3 shows LMI distributions based on important and uncertain words in explanations respectively. Some important words for model predictions on a specific label (e.g., great for POSITIVE, bad for NEGATIVE in (a)) become uncertain words for the other label in (b). This indicates models may get confused by important words corresponding to different labels in inputs. Besides, negation words (e.g., not, no) pointed out in (b) are not shown in (a), which means they may not be used by models for making predictions but can highly cause model predictive uncertainty. We observe similar results on other datasets in Table 7.

5.2 Human Evaluation

To answer the research question (3), we conduct human evaluation on both important and uncertain words in model explanations through the Amazon Mechanical Turk (AMT). The details of human evaluation are in Appendix A.4. The following two observations illustrate the effectiveness and indispensability of uncertainty explanations.

[1] https://huggingface.co/datasets/m-newhauser/senator-tweets

Model	Dataset	LOO		SS	
		Label	Unc	Label	Unc
BERT	IMDB	63.33	**86.67**	67.50	**87.50**
	Toxics	60.00	**86.67**	86.67	**90.00**
	Politics	66.67	**83.33**	82.50	75.00
RoBERTa	IMDB	66.67	**83.33**	80.00	**86.67**
	Toxics	80.00	**83.33**	80.00	**86.67**
	Politics	63.33	60.00	75.00	65.00

Table 1: Human prediction performance (%) on label explanations (Label) and uncertainty explanations (Unc).

Humans perform better on understanding uncertainty explanations than label explanations. First, we provide inputs with important words highlighted and ask evaluators to guess model prediction labels. Then we show model predictions with confidence and ask evaluators whether removing uncertain words can improve prediction confidence or not. Table 1 shows the results of human performance on predicting model prediction labels and confidence change. Overall, humans have better performance on understanding model predictive uncertainty based on uncertain words. This indicates the effectiveness of uncertainty explanations in helping users understand model predictions. Besides, SS produces more understandable explanations to humans than LOO. This is also reflected by the evaluation results where evaluators score (from 1-5) the quality of explanations with the average values 3.7 and 4.0 for LOO and SS respectively.

Humans prefer to see uncertainty explanations in addition to label explanations. We ask evaluators to vote whether they want to include uncertainty explanations in addition to label explanations for understanding model decision making. Most (71%) evaluators prefer to see uncertainty explanations. Besides, evaluators mark 72.6% of uncertainty explanations identify the words that largely limit model prediction confidence. This implies that uncertainty explanations are indispensable to explaining model prediction behavior.

6 Conclusion

In this paper, we propose to explain model prediction uncertainty by extracting uncertain words from existing model explanations. We adopt two explanation methods to explain BERT and RoBERTa on three tasks. Experiments show the effectiveness of uncertainty explanations in explaining models and helping humans understand model predictions.

Limitations

One limitation is that we adopted two perturbation-based explanation methods, Leave-one-out and Sampling Shapley, identifying word-level features. Utilizing high-level explanation methods (e.g., hierarchical explanations (Chen, Zheng, and Ji 2020)) may capture more semantic information that explains model predictive uncertainty. Another limitation is that we identified uncertain words by removing them and observing whether model prediction confidence increases. An alternative way is replacing those words with their synonyms, hence maintaining the original semantic meaning. But this may lead to adversarial examples, which we leave to future work.

Ethics Statement

Regarding ethical concerns, this work utilized a sensitive dataset (Wikipedia Toxicity Corpus (Wulczyn, Thain, and Dixon 2017)) which contains toxic comments. Before conducting human evaluation on this sensitive dataset, we had reported potential participant risks to Institutional Review Board (IRB) and gotten approval of continuing this research. We will provide the link to IRB approval with the publication of this paper. The other two datasets, IMDB and Senator Tweets, do not have higher risks than those encountered in daily life and daily online activities. For all human evaluation experiments, we did not collect any personal information (e.g. demographic and identity characteristics) of participants.

References

Antorán, J.; Bhatt, U.; Adel, T.; Weller, A.; and Hernández-Lobato, J. M. 2020. Getting a clue: A method for explaining uncertainty estimates. *arXiv preprint arXiv:2006.06848.*

Brown, T. B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165.*

Chen, H.; Feng, S.; Ganhotra, J.; Wan, H.; Gunasekara, C.; Joshi, S.; and Ji, Y. 2021. Explaining Neural Network Predictions on Sentence Pairs via Learning Word-Group Masks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 3917–3930. Online: Association for Computational Linguistics.

Chen, H.; Zheng, G.; Awadallah, A. H.; and Ji, Y. 2022. Pathologies of Pre-trained Language Models in Few-shot Fine-tuning. *arXiv preprint arXiv:2204.08039.*

Chen, H.; Zheng, G.; and Ji, Y. 2020. Generating Hierarchical Explanations on Text Classification via Feature Interaction Detection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5578–5593. Online: Association for Computational Linguistics.

Desai, S.; and Durrett, G. 2020. Calibration of Pre-trained Transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 295–302. Online: Association for Computational Linguistics.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186. Minneapolis, Minnesota: Association for Computational Linguistics.

Du, M.; Manjunatha, V.; Jain, R.; Deshpande, R.; Dernoncourt, F.; Gu, J.; Sun, T.; and Hu, X. 2021. Towards Interpreting and Mitigating Shortcut Learning Behavior of NLU models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 915–929. Online: Association for Computational Linguistics.

Feng, S.; Wallace, E.; Grissom II, A.; Iyyer, M.; Rodriguez, P.; and Boyd-Graber, J. 2018. Pathologies of Neural Models Make Interpretations Difficult. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 3719–3728. Brussels, Belgium: Association for Computational Linguistics.

Gal, Y.; and Ghahramani, Z. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, 1050–1059. PMLR.

Guo, C.; Pleiss, G.; Sun, Y.; and Weinberger, K. Q. 2017. On calibration of modern neural networks. In *International Conference on Machine Learning*, 1321–1330. PMLR.

Gururangan, S.; Marasović, A.; Swayamdipta, S.; Lo, K.; Beltagy, I.; Downey, D.; and Smith, N. A. 2020. Don't Stop Pretraining: Adapt Language Models to Domains and Tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 8342–8360. Online: Association for Computational Linguistics.

Jiang, Z.; Araki, J.; Ding, H.; and Neubig, G. 2021. How can we know when language models know? on the calibration of language models for question answering. *Transactions of the Association for Computational Linguistics*, 9: 962–977.

Kong, L.; Jiang, H.; Zhuang, Y.; Lyu, J.; Zhao, T.; and Zhang, C. 2020. Calibrated Language Model Fine-Tuning for In- and Out-of-Distribution Data. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1326–1340. Online: Association for Computational Linguistics.

Kuleshov, V.; and Liang, P. S. 2015. Calibrated structured prediction. *Advances in Neural Information Processing Systems*, 28: 3474–3482.

Kumar, A.; Sarawagi, S.; and Jain, U. 2018. Trainable calibration measures for neural networks from kernel mean embeddings. In *International Conference on Machine Learning*, 2805–2814. PMLR.

Ley, D.; Bhatt, U.; and Weller, A. 2021. Diverse, Global and Amortised Counterfactual Explanations for Uncertainty Estimates. *arXiv preprint arXiv:2112.02646*.

Li, J.; Monroe, W.; and Jurafsky, D. 2016. Understanding neural networks through representation erasure. *arXiv preprint arXiv:1612.08220*.

Liu, J. Z.; Lin, Z.; Padhy, S.; Tran, D.; Bedrax-Weiss, T.; and Lakshminarayanan, B. 2020. Simple and principled uncertainty estimation with deterministic deep learning via distance awareness. *arXiv preprint arXiv:2006.10108*.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Lundberg, S. M.; and Lee, S.-I. 2017. A unified approach to interpreting model predictions. In *Proceedings of the 31st international conference on neural information processing systems*, 4768–4777.

Maas, A.; Daly, R. E.; Pham, P. T.; Huang, D.; Ng, A. Y.; and Potts, C. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, 142–150.

Pereyra, G.; Tucker, G.; Chorowski, J.; Kaiser, Ł.; and Hinton, G. 2017. Regularizing neural networks by penalizing confident output distributions. *arXiv preprint arXiv:1701.06548*.

Perez, I.; Skalski, P.; Barns-Graham, A.; Wong, J.; and Sutton, D. 2022. Attribution of Predictive Uncertainties in Classification Models. In *The 38th Conference on Uncertainty in Artificial Intelligence*.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. " Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 1135–1144.

Schuster, T.; Shah, D.; Yeo, Y. J. S.; Roberto Filizzola Ortiz, D.; Santus, E.; and Barzilay, R. 2019. Towards Debiasing Fact Verification Models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3419–3425. Hong Kong, China: Association for Computational Linguistics.

Strumbelj, E.; and Kononenko, I. 2010. An efficient explanation of individual classifications using game theory. *The Journal of Machine Learning Research*, 11: 1–18.

Sundararajan, M.; Taly, A.; and Yan, Q. 2017. Axiomatic attribution for deep networks. In *International Conference on Machine Learning*, 3319–3328. PMLR.

Wulczyn, E.; Thain, N.; and Dixon, L. 2017. Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th international conference on world wide web*, 1391–1399.

Xu, J.; Desai, S.; and Durrett, G. 2020. Understanding Neural Abstractive Summarization Models via Uncertainty. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 6275–6281. Online: Association for Computational Linguistics.

Yang, Z.; Dai, Z.; Yang, Y.; Carbonell, J.; Salakhutdinov, R. R.; and Le, Q. V. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.

Zhao, T. Z.; Wallace, E.; Feng, S.; Klein, D.; and Singh, S. 2021. Calibrate before use: Improving few-shot performance of language models. *arXiv preprint arXiv:2102.09690*.

A Supplement of Experiments

A.1 Models and Datasets

We adopt the pretrained BERT-base and RoBERTa-base models from Hugging Face[2]. For sentiment analysis, we utilize the IMDB (Maas et al. 2011) dataset which contains positive and negative movie reviews. For toxic comments detection, we test on the Wikipedia Toxicity Corpus (Toxics) (Wulczyn, Thain, and Dixon 2017). The task is to detect whether a comment is toxic or nontoxic. For political bias classification, we adopt the Senator Tweets dataset (Politics) [3], which collects all tweets made by US senators during 2021-2022. The task is to recognize the political bias of each tweet as Democratic or Republican. All datasets are in English. Table 2 shows the statistics of the datasets. We fine-tune the models on the three datasets and report their prediction performance in Table 3.

We implement the models in PyTorch 3.7. The numbers of parameters in the BERT and RoBERTa models are 109484547 and 124647170 respectively. We manually set hyperparameters as: learning rate is $1e-5$, maximum sequence length is 256, maximum gradient norm is 1, and batch size is 8. All experiments were performed on a single NVidia GTX 1080 GPU. The corresponding validation accuracy for each reported test accuracy is in Table 4. The time for training each model on each dataset is in Table 5. All training and evaluation are based on one run.

A.2 Posterior Calibration

A common way of measuring predictive uncertainty is by calibrating model outputs with the true correctness likelihood, so that the predictive probabilities well represent the confidence of model predictions being correct (Guo et al. 2017; Kong et al. 2020; Desai and Durrett 2020; Zhao et al. 2021). Lower prediction confidence indicates higher uncertainty (Xu, Desai, and Durrett 2020; Jiang et al. 2021). We follow the post-calibration methods and adopt the temperature scaling (Guo et al. 2017; Zhao et al. 2021) to calibrate the pre-trained language models (BERT and RoBERTa) in our experiments.

Specifically, we use the development set to learn a temperature T which corrects model output probabilities by dividing non-normalized logits before the softmax function. Then the learned T is applied to modify model outputs on the test set. In experiments, we linearly search for an optimal temperature T between $[0, 10]$ with a granularity of 0.01, which empirically performs well. We evaluate model calibration with Expected Calibration Error (ECE) (Guo et al. 2017). The ECE measures the difference between prediction confidence and accuracy, i.e.

$$ECE = \sum_{k=1}^{K} \frac{|B_k|}{n} |acc(B_k) - conf(B_k)|, \tag{3}$$

where the total n predictions are partitioned into K equally-spaced bins, B_k represents the predictions fall into the kth bin, $acc(\cdot)$ and $conf(\cdot)$ compute the average accuracy and confidence in each bin respectively. For a perfect calibration, $acc(B_k) = conf(B_k)$, $k \in \{1, \ldots, K\}$. In this work, we set $K = 10$. We report the learned temperature scalars and ECEs before and after calibration in Table 6. Temperature scaling performs effectively in decreasing model calibration errors. This enables us to further explain prediction uncertainty based on calibrated confidence. We apply temperature scaling to correct model outputs in experiments.

A.3 Local Mutual Information

To understand which features contribute to model predictions and which features cause prediction uncertainty, we follow (Schuster et al. 2019; Du et al. 2021; Chen et al. 2022) and analyze feature statistics of model explanations via local mutual information (LMI). LMI quantifies the association between a feature and a prediction label in model explanations. We compute LMI based on top 5 important and uncertain words in prediction and uncertainty explanations respectively. Specifically, for each group of features, we can get a set of unique features, $E = \{e\}$. The LMI between a feature e and a prediction label y is

$$\text{LMI}(e, y) = p(e, y) \cdot \log\left(\frac{p(y \mid e)}{p(y)}\right), \tag{4}$$

where $p(y \mid e) = \frac{count(e,y)}{count(e)}$, $p(y) = \frac{count(y)}{|E|}$, $p(e, y) = \frac{count(e,y)}{|E|}$, and $|E|$ is the number of occurrences of all features in E. Then we can get a distribution of LMI over all tokens in the vocabulary ($\{w\}$) built on the dataset, i.e.

$$P_{\text{LMI}}(w, y) = \begin{cases} \text{LMI}(w, y) & \text{if token } w \in E \\ 0 & \text{else} \end{cases} \tag{5}$$

We normalize the LMI distribution by dividing each value with the sum of all values. Table 7 records top 10 tokens in different LMI distributions of model explanations.

A.4 Human Evaluation

We conduct human evaluation on both important and uncertain words in model explanations through the Amazon Mechanical Turk (AMT). For each dataset, we randomly select 30 test examples to generate explanations for each pretrained language model. Each explanation (with 2-3 important and uncertain words extracted respectively) is assessed by 5 workers. We pay the workers \$0.3 for assessing each explanation. We have collected 900 annotations in total.

For each explanation, we ask the worker to answer the following 5 questions:

1. **Prediction on label explanations (multiple choices)**: Given the model input text, can you guess the model prediction label based on the highlighted tokens?
2. **Rating on label explanations (1-5 Liker scale)**: Given the model input text and model prediction label, how much do you think the highlighted tokens make sense to you?
3. **Prediction on uncertainty explanations (multiple choices)**: Given the model input text and model prediction probability, do you think removing the highlighted

[2] https://github.com/huggingface/transformers

[3] https://huggingface.co/datasets/m-newhauser/senator-tweets

Datasets	*L*	*#train*	*#dev*	*#test*	Label distribution
IMDB	231	20K	5K	25K	Positive: *train*(10036), *dev*(2414), *test*(12535) Negative: *train*(9956), *dev*(2583), *test*(12451)
Toxics	68	96K	32K	32K	Toxic: *train*(9245), *dev*(3069), *test*(3048) Nontoxic: *train*(86447), *dev*(29059), *test*(28818)
Politics	34	70K	7.8K	19K	Democratic: *train*(36222), *dev*(3982), *test*(10240) Republican: *train*(33796), *dev*(3789), *test*(9189)

Table 2: Summary statistics of the datasets, where L is average sentence length, and # counts the number of examples in the *train/dev/test* sets. For label distribution, the number of examples with a specific label in *train/dev/test* is noted in bracket.

Models	IMDB	Toxics	Politics
BERT	91.29	96.96	91.20
RoBERTa	93.36	96.75	91.32

Table 3: Prediction accuracy (%) of different models on the test sets.

Models	IMDB	Toxics	Politics
BERT	91.76	96.83	91.44
RoBERTa	93.30	96.69	91.53

Table 4: Validation accuracy (%) for each reported test accuracy.

tokens can further increase the model prediction probability or not?

4. **Rating on uncertainty explanations (1-3 Liker scale)**: How much do you think the current model prediction probability could be changed by removing the highlighted tokens?
5. **Comparison on label explanations and uncertainty explanations (multiple choices)**: Which type of model explanations can help you better understand the model prediction?

Figure 4 and Figure 5 show the interfaces of human evaluation on Q1 and Q3 respectively.

A.5 Visualizations

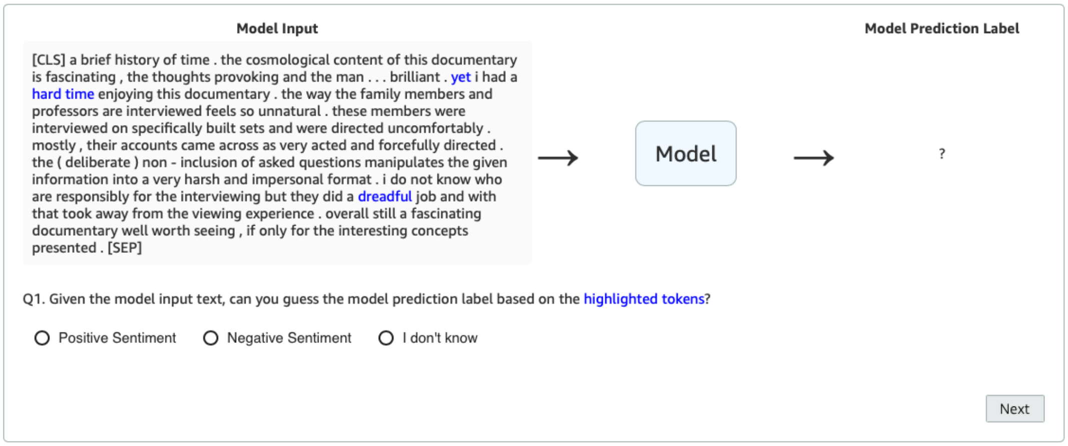
Table 8 shows visualizations of different model explanations with both important and uncertain words highlighted.

Models	IMDB	Toxics	Politics
BERT	856.43	3254.33	1483.65
RoBERTa	912.47	3467.39	1646.12

Table 5: The average runtime (s/epoch) of each model on each in-domain dataset.

Models	IMDB	Toxics	Politics
BERT:			
T	4.59	1.95	4.2
pre-ECE	8.45	2.36	8.56
post-ECE	2.85	0.89	3.83
RoBERTa:			
T	2.76	2.16	3.98
pre-ECE	6.36	2.90	8.45
post-ECE	2.50	1.13	4.29

Table 6: Posterior calibration results. T is the learned temperature. pre-ECE and post-ECE represent the ECEs on test sets before and after calibration respectively.

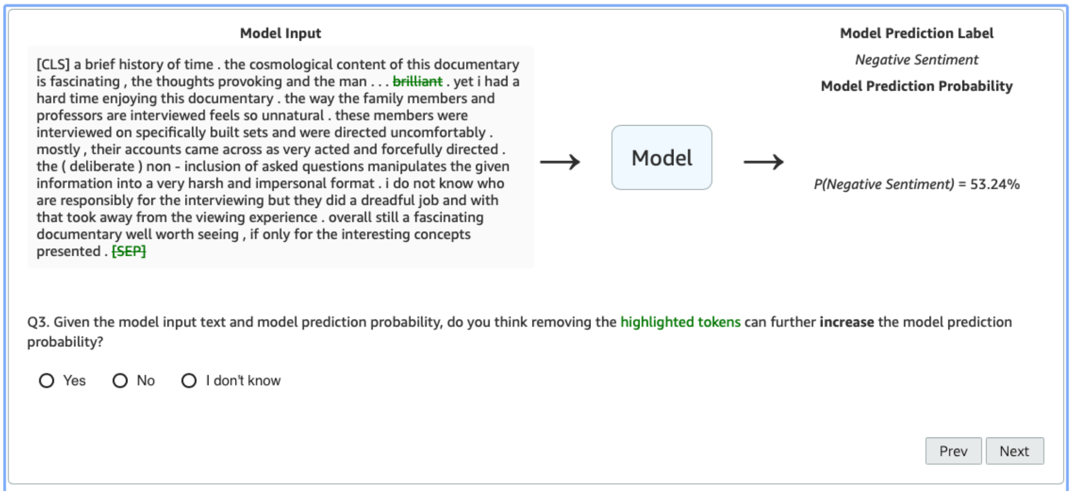


[CLS] a brief history of time . the cosmological content of this documentary is fascinating , the thoughts provoking and the man . . . brilliant . yet i had a hard time enjoying this documentary . the way the family members and professors are interviewed feels so unnatural . these members were interviewed on specifically built sets and were directed uncomfortably . mostly , their accounts came across as very acted and forcefully directed . the (deliberate) non - inclusion of asked questions manipulates the given information into a very harsh and impersonal format . i do not know who are responsibly for the interviewing but they did a dreadful job and with that took away from the viewing experience . overall still a fascinating documentary well worth seeing , if only for the interesting concepts presented . [SEP]

Q1. Given the model input text, can you guess the model prediction label based on the highlighted tokens?

○ Positive Sentiment ○ Negative Sentiment ○ I don't know

Figure 4: Interface of human evaluation on important words highlighted in blue color.



Negative Sentiment

P(Negative Sentiment) = 53.24%

[CLS] a brief history of time . the cosmological content of this documentary is fascinating , the thoughts provoking and the man . . . ~~brilliant~~ . yet i had a hard time enjoying this documentary . the way the family members and professors are interviewed feels so unnatural . these members were interviewed on specifically built sets and were directed uncomfortably . mostly , their accounts came across as very acted and forcefully directed . the (deliberate) non - inclusion of asked questions manipulates the given information into a very harsh and impersonal format . i do not know who are responsibly for the interviewing but they did a dreadful job and with that took away from the viewing experience . overall still a fascinating documentary well worth seeing , if only for the interesting concepts presented . ~~[SEP]~~

Q3. Given the model input text and model prediction probability, do you think removing the highlighted tokens can further **increase** the model prediction probability?

○ Yes ○ No ○ I don't know

Figure 5: Interface of human evaluation on uncertain words highlighted in green color.

Model	Dataset	Label	Leave-one-out		Sampling Shapley	
			Important	Uncertain	Important	Uncertain
BERT	IMDB	Pos	this great best film a good excellent and it wonderful	i movie this was to just the would but not	great best and excellent love wonderful good this very enjoyed	movie just would bad but nothing not could off plot
		Neg	this worst movie bad not but no terrible just nothing	but is and not the it a this ’t great	bad worst this movie just boring terrible awful not nothing	and great very it is good in not his seen
	Toxics	Tox	you fuck hell fucking bullshit idiot dick suck stupid gay	the are so fuck of good have wow love for	you fuck hell gay fucking bullshit idiot dick suck stupid	the can in please if so certainly because know help
		NTox	please to i if not the of wikipedia can is thank	you your i the and a to please me is	please the can to if for in of thank use	you a i your me the my it van and
	Politics	Dec	and to climate must this child our so the in	the and to of we in american is americans i	this must climate that health now today more every to	the a is for american back ensure work and not
		Rep	democrats the is border bid great and communist inflation fox	to and the i this our my with in for	the a is bid and border for communist his fox	that this to more must you today every now your
RoBERTa	IMDB	Pos	this best and great not but good I film is	the not I is a for this and no was	great and love excellent wonderful best very amazing brilliant perfect	any plot ’t bad no nothing movie much never this
		Neg	this not bad worst boring just the even and no	bad ’t not and plot butwas a to me	bad worst plot boring terrible nothing stupid much no waste	and great first not more special very love life moments
	Toxics	Tox	you fuck stupid HELL suck Fuck YOU You fucking shit	to but if or Go an reported thanks ipedia should	you fuck You stupid HELL suck Fuck YOU fucking shit	to for reported OF but the about in help need
		NTox	to the Please article of please for Thank and in	you your is I a Your vandal not my me	to the for article use in please Please of If	you your a me is my You vandal I are
	Politics	Dec	and the to this in our a must for will	the americ an in to for is act this a	this and climate care child today that workers how families	americ is the will of back family they not would
		Rep	bid en is democr the border americ us great to	to the and our my i of in this a	americ border democr bid is spending great not inflation would	and this that to our it more my you families

Table 7: Top 10 tokens in different LMI distributions of model explanations. Important: statistics of top salient words in explanations; Uncertain: statistics of bottom salient words in explanations; Pos: postive; Neg: negative; Tox: toxic; NTox: nontoxic; Dec: democratic; Rep: republican. Warning: this table contains toxic tokens.

Model/Dataset	Method	Prediction	Explanation
BERT/IMDB	LOO	Negative (0.69 → 0.80)	i found it to be a complete disappointment . if i had of known this movie was going to be as stupid as it was , i would have stayed home and done something more entertaining ... the plot was a great idea , just could have been done in a much better way .
RoBERTa/IMDB	SS	Positive (0.51 → 0.86)	Not the best of the Lone Star series, but it moves along quickly with good performances. Introduced as ”Singin’ Sandy” in the main title, John Wayne as a ’singing cowboy’ isn’t successful...
BERT/Toxics	LOO	Nontoxic (0.83 → 0.97)	oh , and i have a question . why was the article on brad christian , a famous magician , deleted because of vandalism instead of simply restored ? i believe that many users on this site are biased towards magicians . i have come to the conclusion that wiki is a useless site that does nothing to help anyone . you are welcome to ban me longer , and i understand completely if you do , but this site is the worst piece of garbage i have ever found !
RoBERTa/Toxics	SS	Nontoxic (0.71 → 0.94)	You are so full of shit. First of all, you aren’t an admin, and for the sake of this site I hope you never will be. I know I will personally work against you if you ever decide to try for one. But I digress as you are not an administrator, and especially since you have no access to checkuser, you cannot determine who is or is not a sockpuppet nor do you have the authorization to place a tag on a user page .
BERT/Politics	SS	Democratic (0.64 → 0.95)	fantastic news . star plastics was founded in ravenswood and is continuing to invest in west virginia . this expansion will lead to economic development and growth in jackson county , and shows that wv is the perfect place for companies large and small .
RoBERTa/Politics	LOO	Republican (0.85 → 0.94)	happy national day , taiwan. your commitment to democracy and market economics is an effective model that can be relied upon to solve our collective problem .

Table 8: Visualizations of prediction explanations for different models on different datasets, where top two important and uncertain words are highlighted in blue and red colors respectively. The prediction confidence changes are shown in brackets when the highlighted uncertain words are removed. LOO: Leave-one-out; SS: Sampling Shapley. Warning: some examples may be offensive or upsetting.