Submission Number: 29

Submission Title: Enhancing textual counterfactual explanation intelligibility through Counterfactual Feature Importance

Dear reviewers,

The authors have received the evaluation of our submission and wish to express their sincere appreciation for the pertinence of the review and the relevance of the points raised. Therefore, we are very pleased to be able to submit a revised version with the changes made.

Please find hereafter the point-by-point responses to the suggested revisions grouped by comment category. We also indicate the adequate changes made in the submission document.

*How does the method compare with IG, SHAP and LIME? Are the important words same in both approaches?* *(review 1)*

We would like to thank reviewer 1 for this interesting question. We did not perform such a comparative analysis. We actually already use IG to perform CFI by fixing the instance to be explained as the reference point. We believe that CFI could also be computed based on other underlying local feature importance methods such as LIME, SHAP or DeepLift. However, it would probably lead to different explanation since CFI would not be computed in the same way. For example, computing CFI with SHAP would require to compute Shapley values related to x0 on one hand and xcf on the other. Then, CFI would be the difference between these two breakdowns on a token-by-token basis. CFI could also be computed based on DeepLift by fixing the instance to be explained as the reference point in the same way than IG. We are currently thinking of experimentally benchmarking these different approaches to assess which gives the best results.

*It would be interesting to results on fairness studies? Such as when predicting occupation is the counterfactual substitution of gendered token important? (review 1)*

We thank reviewer 1 for this interesting comment. We believe indeed that counterfactual generation can have applications in fairness. The analysis suggested by reviewer 1 could be done by applying CFI to counterfactual examples generated to explain an occupation classifier. Gendered tokens with high CFI could therefore highlight some classifier biases.

*Overall, the idea of model agnostic counterfactuals is very similar to adversarial text generation [1,2]. These methods also generate counterfactual-like examples. Some may even provide scores for the likelihood of a word changing the prediction. Hence, more discussion about how these are related to the current work would improve the paper. (review 2)*

We thank reviewer 2 for this pertinent comment. There is indeed similarities between counterfactual examples and adversarial attacks. We believe that the main difference is that adversarial attacks seek to fool a model without any explanatory purpose, leading to inaccurate adversarial predictions. On the contrary, counterfactual explanations seek to find instances that can be used as an explanation. Therefore, we have focused only on methods for generating counterfactual examples. However, Counterfactual Feature Importance can be applied to adversarial examples as well to better understand which token changes explain the most the label flipping. We have added a paragraph dealing with this matter in the Related Work and in the Discussion.

*This citation is incomplete: 2023. TIGTEC : Token Importance Guided TExt Counterfactuals. (review 2)*

We thank reviewer 2 for highlighting this mistake. We corrected it in the updated paper.

In conclusion, the authors would like to thank the reviewers for their time and consideration through their valuable comments and feedback on our paper. We have carefully considered each suggestion and have amended our work accordingly.

Kind regards,

The authors.