Enjoy the Salience: Towards Better Transformer-based Faithful Explanations with Word Salience

George Chrysostomou **Nikolaos Aletras**
Department of Computer Science, University of Sheffield
United Kingdom
{gchrysostomou1, n.aletras}@sheffield.ac.uk

Abstract

Pretrained transformer-based models such as BERT have demonstrated state-of-the-art predictive performance when adapted into a range of natural language processing tasks. An open problem is how to improve the faithfulness of explanations (rationales) for the predictions of these models. In this paper, we hypothesize that salient information extracted a priori from the training data can complement the task-specific information learned by the model during fine-tuning on a downstream task. In this way, we aim to help BERT not to forget assigning importance to informative input tokens when making predictions by proposing SALOSS; an auxiliary loss function for guiding the multi-head attention mechanism during training to be close to salient information extracted *a priori* using TextRank. Experiments for explanation faithfulness across five datasets, show that models trained with SALOSS consistently provide more faithful explanations across four different feature attribution methods compared to vanilla BERT. Using the rationales extracted from vanilla BERT and SALOSS models to train inherently faithful classifiers, we further show that the latter result in higher predictive performance in downstream tasks.[1]

1 Introduction

Pretrained transformer-based (Vaswani et al., 2017) language models (LMs) such as BERT (Devlin et al., 2019), have achieved state-of-the-art results in various language understanding tasks (Wang et al., 2018, 2019). Despite their success, their highly complex nature consisting of millions of parameters, makes them difficult to interpret (Jain et al., 2020). This has motivated new research on understanding and explaining their predictions.

Previous work has explored whether LMs encode syntactic knowledge by studying their multi-head attention distributions (Clark et al., 2019; Htut et al., 2019; Voita et al., 2019). Recent studies have evaluated the faithfulness of explanations[2] for predictions made by these models (Vashishth et al., 2019; Atanasova et al., 2020; Jain et al., 2020). In general, LMs can provide faithful explanations, particularly using attention (Jain et al., 2020), but still fall behind other simpler architectures (Atanasova et al., 2020) possibly due to increased information mixing and higher contextualization in the model (Brunner et al., 2019; Pascual et al., 2021; Tutek and Snajder, 2020). Recent studies have attempted to improve the explainability of non transformer-based models, by guiding them through an auxiliary objective towards informative input importance distributions (e.g. human or adversarial priors) (Ross et al., 2017a; Liu and Avci, 2019; Moradi et al., 2021).

In a similar direction, we propose **Sa**lient **Loss** (SALOSS), an auxiliary objective that allows the multi-head attention of the model to learn from salient information (i.e. token importance) during training to reduce the effects of information mixing (Pascual et al., 2021). We compute a priori token importance scores (Xu et al., 2020) using TEXTRANK (Mihalcea and Tarau, 2004) (i.e. an unsupervised graph-based method) and penalize the model when the attention distribution deviates from the salience distribution. Our contributions are as follows:

- We demonstrate that models trained with SALOSS generate more faithful explanations in an input erasure evaluation.
- We finally show that rationales extracted from SALOSS models result in higher predictive performance in downstream tasks when used as the only input for training inherently faithful classifiers.

[1] Code: https://github.com/GChrysostomou/saloss.

[2] A faithful explanation represents the true reasons behind a model's prediction (Jacovi and Goldberg, 2020).

2 Related Work

Model Explainability Explanations can be obtained by computing importance scores for input tokens to identify which parts of the input contributed the most towards a model's prediction (i.e. feature attribution). A common approach to attributing input importance is by measuring differences in a model's prediction between keeping and omitting an input token (Robnik-Šikonja and Kononenko, 2008; Li et al., 2016b; Nguyen, 2018a). Input importance can also be obtained by calculating the gradients of a prediction with respect to the input (Kindermans et al., 2016; Li et al., 2016a; Sundararajan et al., 2017; Bastings and Filippova, 2020). We can also use sparse linear meta-models that are easier to interpret (Ribeiro et al., 2016; Lundberg and Lee, 2017). Finally, recent studies propose using feature attribution to extract a fraction of the input as a rationale and then use it to train a classifier (Jain et al., 2020; Treviso and Martins, 2020).

Faithfulness of Pretrained LM Explanations Brunner et al. (2019) criticize the ability of attention in providing faithful explanations for the inner workings of a LM, by showing that constructed adversary attention maps do not impact significantly the predictive performance. Pruthi et al. (2020) show similar outcomes by manipulating attention to attend to uninformative tokens. Pascual et al. (2021) and Brunner et al. (2019) argue that this might be due to significant information mixing in higher layers of the model, with recent studies showing improvements in the faithfulness of attention-based explanations by addressing this (Chrysostomou and Aletras, 2021; Tutek and Snajder, 2020).

Atanasova et al. (2020) evaluate faithfulness of explanations (Jacovi and Goldberg, 2020) by removing important tokens and observing differences in prediction, showing that generally gradient-based approaches for transformers produce more faithful explanations compared to sparse meta-models (Ribeiro et al., 2016). However, transformer-based explanations are less faithful compared to simpler models due to their highly parameterized architecture. Atanasova et al. (2020) also show that explanation faithfulness does not correlate with how plausible it is (understandable by humans) corroborating arguments made by Jacovi and Goldberg (2020). Jain et al. (2020) show that attention-based feature attributions, in general, outperform gradient-based ones.

A different branch of studies introduced adversarial auxiliary objectives to influence attention-based explanations during training (Kennedy et al., 2020; Wiegreffe and Pinter, 2019; Ross et al., 2017b; Liu and Avci, 2019). These objectives have typically been used as a tool for evaluating explanation faithfulness generated by attention (Kennedy et al., 2020; Wiegreffe and Pinter, 2019; Pruthi et al., 2020; Ghorbani et al., 2019) while others used auxiliary objectives to improve the faithfulness of explanations generated by non-transformer based models (Ross et al., 2017b; Liu and Avci, 2019; Moradi et al., 2021; Mohankumar et al., 2020; Tutek and Snajder, 2020). The auxiliary objectives guide the model using human annotated importance scores (Liu and Avci, 2019), or allow for selective input gradient penalization (Ross et al., 2017b). Such studies illustrate the effectiveness of auxiliary objectives for improving the faithfulness of model explanations suggesting that we can also improve explanation faithfulness in transformers using appropriate prior information.

3 Improving Explanation Faithfulness with Word Salience

Even though attention scores are more faithful than other feature attribution approaches (Jain et al., 2020), they usually pertain to their corresponding input tokens in *context* and not individually due to information mixing (Tutek and Snajder, 2020; Pascual et al., 2021). As such, we hypothesize that we can improve the ability of a pretrained LM in providing *faithful* explanations, by showing to the model alternative distributions of input importance (i.e. word salience). We assume that by introducing the salience distribution via an auxiliary objective (Ross et al., 2017b), we can reduce information mixing by "shifting" the model's attention to other informative tokens. In a similar direction to ours, Xu et al. (2020) showed that by computing attention together with salience information from keyword extractors improves text summarization.

Computing Word Salience We compute word salience $\boldsymbol{\sigma}$ using TEXTRANK (Mihalcea and Tarau, 2004), an unsupervised graph-based model for keyword extraction. TEXTRANK calculates indegree centrality of graph nodes iteratively based on a Markov chain, where each node is a wordpiece and each edge links wordpiece pairs within a context

window (Xu et al., 2020). For each input document X, we construct an undirected graph and apply TEXTRANK to compute the local salience scores (σ_i) of its words by:

$$\sigma_i = (1-d) + d \sum_{j \in In(V_i)} \frac{\sigma_j}{|Out(V_j)|} \quad (1)$$

where d is the damping coefficient, $In(V_i)$ and $Out(V_j)$ are the incoming and outgoing nodes. Our intuition is that by using the task-agnostic TEXTRANK, we can extract words that are important in the context of the sequence and as such offer an alternative view of token importance.[3]

Salience Loss We propose Salient Loss (SALOSS), an auxiliary objective which allows the model to learn attending to more informative input tokens jointly with the task. SALOSS penalizes the model when the attention distribution ($\boldsymbol{\alpha}$) deviates from the word salience distribution ($\boldsymbol{\sigma}$).[4] For $\boldsymbol{\alpha}$ we compute the average attention scores of the CLS token from the last layer (Jain et al., 2020). The joint objective for adapting a LM to a downstream classification task with SALOSS is:

$$\mathcal{L} = \mathcal{L}_c + \lambda \mathcal{L}_{sal} \quad (2)$$

where $\mathcal{L}_c$ is the Cross-Entropy Loss for a downstream text classification task and λ a regularization coefficient for the proposed SALOSS ($\mathcal{L}_{sal}$) which can be tuned in a development set. $\mathcal{L}_{sal}$ is defined as the KL divergence between $\boldsymbol{\alpha}$ and $\boldsymbol{\sigma}$:

$$\mathcal{L}_{sal} = KL(\boldsymbol{\alpha}, \boldsymbol{\sigma}) = \sum \boldsymbol{\alpha}(\log \boldsymbol{\alpha} - \log \boldsymbol{\sigma}) \quad (3)$$

We assume a standard text classification setting where a set of labeled documents is used for fine-tuning a pretrained LM by adding an extra output classification layer. We normalize the salience scores for compatibility with the KL divergence.

4 Experimental Setup

Datasets We consider five natural language understanding tasks (see dataset statistics in Appx. A): SST (Socher et al., 2013); AGNews (AG) (Del Corso et al., 2005); Evidence Inference (EV.INF.) (Lehman et al., 2019); MultiRC (M.RC) (Khashabi et al., 2018) and Semeval 2017 Task 4 Subtask A (SEMEVAL) (Rosenthal et al., 2017).

[3] We also considered the use of TFIDF and χ^2 scores observing comparable but lower performance in early experimentation. We hypothesize that TextRank performs well due to its effectiveness in improving performance in text summarization (Xu et al., 2020). See also Appx. H for TFIDF and χ^2 results on input erasure experiments.

[4] $\boldsymbol{\alpha} \in \mathrm{R}^t$; $\boldsymbol{\sigma} \in \mathrm{R}^t$, where t is the sequence length.

DATASET	BASELINE	λ	SALOSS
SST	.91 (.00)	1E-3	.91 (.00)
AG	.93 (.00)	1E-4	.93 (.00)
EV.INF	.82 (.01)	1E-4	.80 (.02)
M.RC	.76 (.01)	1E-3	.76 (.00)
SEMEVAL	.58 (.01)	1E-3	.57 (.03)

Table 1: F1 macro averaged across 3 seeds for vanilla LMs (BASELINE) and SALOSS models. λ represents the regularization coefficient of our proposed objective.

Models Similar to Jain et al. (2020) we use: BERT (Devlin et al., 2019) for (SST, AG, SEMEVAL); SCIBERT (Beltagy et al., 2019) for EV.INF.; ROBERTA (Liu et al., 2019) for M.RC.

Evaluating Explanation Faithfulness We evaluate the faithfulness[5] of model explanations using two standard approaches:

- **Input Erasure:** We first compute the average fraction of tokens required to be removed (in decreasing importance) to cause a change in prediction (decision flip) (Serrano and Smith, 2019; Nguyen, 2018b).
- **FRESH:** We also compute the predictive performance of a classifier trained on rationales extracted with feature attribution metrics (see §4) using FRESH (Jain et al., 2020). We extract rationales by; (1) selecting the top-k most important tokens (TOPK) and (2) selecting the span of length k that has the highest overall importance (CONTIGUOUS).

Feature Attribution Approaches We opt using the following popular metrics to allocate importance to input tokens: (1) Normalized attention scores ($\boldsymbol{\alpha}$); (2) Attention scores scaled by their gradient ($\boldsymbol{\alpha}\nabla\boldsymbol{\alpha}$) (Serrano and Smith, 2019); (3) Gradients of the input scaled by the input ($\mathbf{x}\nabla\mathbf{x}$) (Kindermans et al., 2016; Atanasova et al., 2020); and (4) Integrated Gradients which compute the accumulated gradients along a path from a baseline to the input (**I.G.**) (Sundararajan et al., 2017).[6]

5 Experimental Results

Predictive Performance Table 1 shows F1 macro scores averaged over three runs with standard deviation across tasks, for vanilla pretrained

[5] We do not conduct human experiments, as faithfulness and plausibility (human understandability of explanations) do not correlate (Atanasova et al., 2020; Jacovi and Goldberg, 2020; Wiegreffe and Pinter, 2019).

[6] where $\nabla\alpha_i = \frac{\partial \hat{y}}{\partial \alpha_i}$ and $\nabla x_i = \frac{\partial \hat{y}}{\partial x_i}$

	METRIC	SST	AG	EV.INF.	M.RC	SEMEVAL
	RAND.	.66	.67	.51	.44	.54
BASELINE	α	.55	.43	.25	.40	.43
	$\mathbf{x}\nabla\mathbf{x}$	.65	.64	.42	.40	.55
	$\alpha\nabla\alpha$	.57	.52	.25	.38	.48
	I.G.	.63	.63	.42	.42	.50
SALOSS	α	**.42**†	.53	.14†	**.19**†	**.39**†
	$\mathbf{x}\nabla\mathbf{x}$	.61†	.59†	.38†	.30†	.51†
	$\alpha\nabla\alpha$	.48†	.50†	**.12**†	.24†	.41†
	I.G.	.61†	.57†	.33†	.33†	.45†

Table 2: Average fraction of tokens required to cause a decision flip across datasets and feature attribution metrics (lower is better). **Bold** denotes the best method in each dataset. † denotes a significant difference compared to BASELINE using the same attribution metric (Wilcoxon Rank Sum, $p < .05$).

LMs (BASELINE) and models with our proposed objective SALOSS. Results demonstrate that models trained with our proposed salience objective[7] achieve similar performance to the BASELINE models across datasets.

Input Erasure Table 2 shows results for the average fraction of input tokens required to be removed to cause a decision flip for BASELINE and SALOSS models in the test set. Results suggest that models trained with our proposed objective require a significantly lower fraction of tokens removed to cause a decision flip in 19 out of 20 cases (Wilcoxon Rank Sum, $p < .05$), with the exception of AG and α. This demonstrates that SALOSS obtains more faithful explanations in the majority of cases (Jacovi and Goldberg, 2020). For example in EV.INF., the BASELINE approach with α requires .25 fractions of tokens on average to observe a decision flip compared to .14 with SALOSS (approximately 40 tokens less). We also observe that in M.RC. where $\boldsymbol{\alpha}$ is not the most effective feature attribution method with BASELINE, with SALOSS it becomes the most effective. In fact, $\boldsymbol{\alpha}$ is the best performing feature attribution approach across most tasks and metrics using SALOSS, indicating the effectiveness of infusing salient information.

We also performed an analysis on the differences in Part-of-Speech (PoS) tags of the rationales selected by SALOSS and BASELINE,to obtain insights towards why rationales with SALOSS are shown to be more faithful to those from models trained without our proposed objective . In SST, we observe that SALOSS allocates more importance on adverbs and adjectives, which are considered important in sentiment analysis (Dragut and Fellbaum, 2014; Sharma et al., 2015). In EV.INF., we observe that SALOSS allocates importance to subordinating conjunction words such as *than*, which are indeed important for the task, which consists of inferring relationships (i.e. *higher than*). We thus hypothesize that SALOSS guides the model to other informative tokens, complementing the task specific information learned by the model.[8]

DATASET	BASELINE	SALOSS TEXTRANK	SALOSS UNIFORM
SST (20%)	.83 (.00)	**.87** (.00) †	.82 (.00)
AG (20%)	.92 (.00)	.92 (.00)	.92 (.00)
EV.INF. (10%)	**.82** (.00)	.81 (.00)	.78 (.00)
M.RC (20%)	.75 (.00)	.75 (.00)	.75 (.00)
SEMEVAL (20%)	.48 (.03)	**.53** (.01)†	.43 (.00)

Table 3: F1 macro on models trained with extracted rationales (TOPK and α) using FRESH for BASELINE and SALOSS models. **Bold** denotes best performance in each dataset. † indicates that SALOSS rationales perform significantly better (t-test, $p < .05$).

Rationale Extraction We finally compare our SALOSS models with vanilla LMs (BASELINE) on rationale extraction using FRESH (Jain et al., 2020), by measuring the predictive performance of the classifier trained on the extracted rationales. For completeness we also include an uninformative baseline for SALOSS, which comprise of a normalized uniform distribution over the input (i.e. all inputs are assigned the same salience score). For brevity, Table 3 presents results using the best performing metric from the erasure experiments α with TOPK.[9] Our approach significantly outperforms BASELINE in 2 out of 5 datasets (t-test, $p < 0.05$), whilst achieving comparable predictive performance on the rest. For example in SST we observe a 3% increase in F1 using the same ratio of rationales. It is notable that in M.RC, AG and EV.INF., performance of classifiers trained on rationales from both BASE. and SALOSS is comparable to that with full text (1-2% lower). We assume that this is due to the nature of the tasks, which likely do not require a large part of the input to reach high performance. This highlights the effectiveness of our approach, as a simple yet effective solution for improving explanation faithfulness.

[7] We treat λ as a hyper-parameter tuned on the development set, where $\lambda \in \{1e\text{-}2, 1e\text{-}3, 1e\text{-}4\}$.

[8] We include an extensive analysis in Appx. G.

[9] For CONTIGUOUS see Appx. F

Example 1	**Data.:AG Id: test_239**
[BASELINE]: NEW YORK (Reuters) - Shares of **Google Inc. will make their Nasdaq stock market** debut on Thursday after the year 's most anticipated initial public offering priced far below initial estimates , raising $1.67 billion .	
[SALOSS (Ours)]: NEW YORK (Reuters) - Shares of Google Inc. will make their Nasdaq stock market debut on Thursday after the year 's most anticipated initial public offering **priced far below initial estimates , raising $1.67** billion .	
[Topic]: Business	
Example 2	**Data.:SST Id: test_78**
[BASELINE]: If nothing else this **movie introduces a** promising unusual kind of psychological horror.	
[SALOSS (Ours)]: If nothing else this movie introduces **a promising unusual** kind of psychological horror.	
[Sentiment]: Positive	
Example 3	**Data.:EV.INF. Id: 4118506_0**
[BASELINE]: ... analgesics . **ABSTRACT.AIM : : The aim of this study is to evaluate the efficacy of fentanyl along with LA field infiltration in controlling pain and discomfort associated with CVC insertion . ABSTRACT**.SETTINGS AND DESIGN : :...	
[SALOSS (Ours)]: ... ABSTRACT.RESULTS : : The median interquartile range pain score is worst for placebo group after LAI (5 [3 - 6]) and in the immediate postprocedure period (5 [4 - 5]) **which was significantly attenuated by addition of fentanyl (3.5 [2 - 5] and 3 [2 - 4]) (P = 0.009 and 0.001** respectively) ...	
[Intervention || **Comparator** || **Outcome]:** Fentanyl || Normal saline || Pain score	
[Relationship]: Significantly decreased	

Table 4: True examples of extracted rationales from models using our proposed approach (SALOSS) and from models that do not (BASELINE)

6 Qualitative Analysis

In Table 4 we present examples of extracted rationales from a model trained with our proposed objective (SALOSS) and without (BASELINE) using $\alpha\nabla\alpha$, to gain further insights to complement the PoS analysis. For clarity we present rationales of CONTIGUOUS type.

In AG we observed similar performance between models trained with SALOSS and without. Example 1 illustrates such a case, where both models predicted correctly but attended to different parts of the input. Despite in different locations, both segments are closely associated with the label of "Business". Example 2 is an instance from the SST dataset, were the SALOSS rationale points to a phrase that is more associated with the task ("*a promising unusual*") compared to the BASELINE. This also aligns with previous observations from the PoS analysis, that models trained with our proposed objective attend to more adjectives compared to BASELINE. Example 3 considers an instance from the Ev.Inf. dataset, which shows that the model trained with SALOSS and BASELINE attended to two different sections. In fact what we observed in agreement with the PoS analysis, is that models with SALOSS attend mostly to segments including words related to relationships, such as "*significantly attenuated*" in this particular example.

7 Conclusion

We introduced **Sa**lient **Loss** (SALOSS), an auxiliary objective to incorporate salient information to attention for improving the faithfulness of transformer-based prediction explanations. We demonstrate that our approach provides more faithful explanations compared to vanilla LMs on input erasure and rationale extraction. In the future, we plan to explore additional objectives to better optimize for contiguity of rationales.

Acknowledgments

NA is supported by EPSRC grant EP/V055712/1, part of the European Commission CHIST-ERA programme, call 2019 XAI: Explainable Machine Learning-based Artificial Intelligence.

References

Leila Arras, Franziska Horn, Grégoire Montavon, Klaus-Robert Müller, and Wojciech Samek. 2016. Explaining predictions of non-linear classifiers in NLP. In *Proceedings of the 1st Workshop on Representation Learning for NLP*, pages 1–7.

Leila Arras, Grégoire Montavon, Klaus-Robert Müller, and Wojciech Samek. 2017. Explaining recurrent neural network predictions in sentiment analysis. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 159–168.

Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. 2020. A diagnostic study of explainability techniques for text classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3256–3274, Online. Association for Computational Linguistics.

Jasmijn Bastings and Katja Filippova. 2020. The elephant in the interpretability room: Why use attention as explanation when we have saliency methods? In *Proceedings of the Third BlackboxNLP Workshop*

on Analyzing and Interpreting Neural Networks for NLP, pages 149–155, Online. Association for Computational Linguistics.

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

Gino Brunner, Yang Liu, Damian Pascual, Oliver Richter, Massimiliano Ciaramita, and Roger Wattenhofer. 2019. On identifiability in transformers. In *International Conference on Learning Representations*.

George Chrysostomou and Nikolaos Aletras. 2021. Improving the faithfulness of attention-based explanations with task-specific information for text classification. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 477–488, Online. Association for Computational Linguistics.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What does BERT look at? an analysis of BERT's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286, Florence, Italy. Association for Computational Linguistics.

Gianna M Del Corso, Antonio Gulli, and Francesco Romani. 2005. Ranking a stream of news. In *Proceedings of the 14th international conference on World Wide Web*, pages 97–106.

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

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. 2020. ERASER: A benchmark to evaluate rationalized NLP models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4443–4458, Online. Association for Computational Linguistics.

Eduard Dragut and Christiane Fellbaum. 2014. The role of adverbs in sentiment analysis. In *Proceedings of Frame Semantics in NLP: A Workshop in Honor of Chuck Fillmore (1929-2014)*, pages 38–41, Baltimore, MD, USA. Association for Computational Linguistics.

Amirata Ghorbani, Abubakar Abid, and James Zou. 2019. Interpretation of neural networks is fragile. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):3681–3688.

Christopher Grimsley, Elijah Mayfield, and Julia R.S. Bursten. 2020. Why attention is not explanation: Surgical intervention and causal reasoning about neural models. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1780–1790, Marseille, France. European Language Resources Association.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Phu Mon Htut, Jason Phang, Shikha Bordia, and Samuel R Bowman. 2019. Do attention heads in bert track syntactic dependencies? *arXiv preprint arXiv:1911.12246*.

Alon Jacovi and Yoav Goldberg. 2020. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4198–4205, Online. Association for Computational Linguistics.

Sarthak Jain, Sarah Wiegreffe, Yuval Pinter, and Byron C. Wallace. 2020. Learning to faithfully rationalize by construction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4459–4473, Online. Association for Computational Linguistics.

Brendan Kennedy, Xisen Jin, Aida Mostafazadeh Davani, Morteza Dehghani, and Xiang Ren. 2020. Contextualizing hate speech classifiers with post-hoc explanation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5435–5442, Online. Association for Computational Linguistics.

Daniel Khashabi, Snigdha Chaturvedi, Michael Roth, Shyam Upadhyay, and Dan Roth. 2018. Looking beyond the surface: A challenge set for reading comprehension over multiple sentences. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 252–262, New Orleans, Louisiana. Association for Computational Linguistics.

Pieter-Jan Kindermans, Kristof Schütt, Klaus-Robert Müller, and Sven Dähne. 2016. Investigating the influence of noise and distractors on the interpretation of neural networks. *arXiv preprint arXiv:1611.07270*.

Eric Lehman, Jay DeYoung, Regina Barzilay, and Byron C. Wallace. 2019. Inferring which medical treatments work from reports of clinical trials. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational*

Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 3705–3717, Minneapolis, Minnesota. Association for Computational Linguistics.

Jiwei Li, Xinlei Chen, Eduard Hovy, and Dan Jurafsky. 2016a. Visualizing and understanding neural models in NLP. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 681–691.

Jiwei Li, Will Monroe, and Dan Jurafsky. 2016b. Understanding neural networks through representation erasure. *arXiv preprint arXiv:1612.08220*.

Frederick Liu and Besim Avci. 2019. Incorporating priors with feature attribution on text classification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6274–6283, 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. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, pages 4768–4777.

Rada Mihalcea and Paul Tarau. 2004. TextRank: Bringing order into text. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 404–411, Barcelona, Spain. Association for Computational Linguistics.

Akash Kumar Mohankumar, Preksha Nema, Sharan Narasimhan, Mitesh M. Khapra, Balaji Vasan Srinivasan, and Balaraman Ravindran. 2020. Towards transparent and explainable attention models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4206–4216, Online. Association for Computational Linguistics.

Pooya Moradi, Nishant Kambhatla, and Anoop Sarkar. 2021. Measuring and improving faithfulness of attention in neural machine translation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2791–2802, Online. Association for Computational Linguistics.

Dong Nguyen. 2018a. Comparing automatic and human evaluation of local explanations for text classification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1069–1078.

Dong Nguyen. 2018b. Comparing automatic and human evaluation of local explanations for text classification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1069–1078, New Orleans, Louisiana. Association for Computational Linguistics.

Damian Pascual, Gino Brunner, and Roger Wattenhofer. 2021. Telling BERT's full story: from local attention to global aggregation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 105–124, Online. Association for Computational Linguistics.

Danish Pruthi, Mansi Gupta, Bhuwan Dhingra, Graham Neubig, and Zachary C. Lipton. 2020. Learning to deceive with attention-based explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4782–4793, Online. Association for Computational Linguistics.

Marco Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 97–101, San Diego, California. Association for Computational Linguistics.

Marko Robnik-Šikonja and Igor Kononenko. 2008. Explaining classifications for individual instances. *IEEE Transactions on Knowledge and Data Engineering*, 20(5):589–600.

Sara Rosenthal, Noura Farra, and Preslav Nakov. 2017. SemEval-2017 task 4: Sentiment analysis in Twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 502–518, Vancouver, Canada. Association for Computational Linguistics.

Andrew Slavin Ross, Michael C. Hughes, and Finale Doshi-Velez. 2017a. Right for the right reasons: Training differentiable models by constraining their explanations. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pages 2662–2670.

Andrew Slavin Ross, Michael C. Hughes, and Finale Doshi-Velez. 2017b. Right for the right reasons: Training differentiable models by constraining their explanations. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pages 2662–2670.

Sofia Serrano and Noah A. Smith. 2019. Is attention interpretable? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2931–2951.

Raksha Sharma, Mohit Gupta, Astha Agarwal, and Pushpak Bhattacharyya. 2015. Adjective intensity and sentiment analysis. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2520–2526, Lisbon, Portugal. Association for Computational Linguistics.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 3319–3328. JMLR.org.

Marcos Treviso and André F. T. Martins. 2020. The explanation game: Towards prediction explainability through sparse communication. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 107–118, Online. Association for Computational Linguistics.

Martin Tutek and Jan Snajder. 2020. Staying true to your word: (how) can attention become explanation? In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 131–142, Online. Association for Computational Linguistics.

Shikhar Vashishth, Shyam Upadhyay, Gaurav Singh Tomar, and Manaal Faruqui. 2019. Attention interpretability across NLP tasks. *arXiv preprint arXiv:1909.11218*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. 2019. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5797–5808, Florence, Italy. Association for Computational Linguistics.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems*, pages 3266–3280.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Sarah Wiegreffe and Yuval Pinter. 2019. Attention is not not explanation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 11–20.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Song Xu, Haoran Li, Peng Yuan, Youzheng Wu, Xiaodong He, and Bowen Zhou. 2020. Self-attention guided copy mechanism for abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1355–1362, Online. Association for Computational Linguistics.

A Datasets

For our experiments we use the following tasks (see dataset details in Table 5.):

SST (Socher et al., 2013): Binary sentiment classification with removed neutral sentences.

AG News (Del Corso et al., 2005): News articles categorized by the following topics; Science, Sports, Business, and World.

Ev.Inf (Evidence Inference) (Lehman et al., 2019): Abstract-only biomedical articles describing randomized controlled trials. The task is to infer the reported relationship between a given intervention and comparator with respect to an outcome.

M.RC (Multi RC) (Khashabi et al., 2018): A reading comprehension dataset composed of questions with multiple correct answers that depend on information from multiple sentences. Similar to DeYoung et al. (2020) and Jain et al. (2020) we convert this to a binary classification task where each rationale/question/answer triplet forms an instance and each candidate answer has a label of True or False.

SEMEVAL Rosenthal et al. (2017): The Semeval 2017 dataset for Task 4 Subtask A which consists of tweets and the task is to classify whether the message is of positive, negative, or neutral sentiment.

B TextRank Training

We run for 10 steps, or until convergence, with a window of 4 words, a damping coefficient of 0.85 and normalize the salience scores to make them more compatible to attention distributions.

C Model Hyper-Parameters

Table 6 presents the hyper-parameters used to train the models across different datasets, along with F1 macro performance on the development set. Models where finetuned across 3 runs for 10 epochs, with the exception of the SEMEVAL dataset which was finetuned for 20. We implement our models using the Huggingface library (Wolf et al., 2020) and use default parameters of the ADAMW optimizer apart from the learning rates and a linear scheduler. Each experiment is run on a single Nvidia Tesla V100 GPU.

We found that the learning rate of our proposed objective, does not impact significantly F1 macro performance. As such, since our objective is improving faithfulness, our λ selection includes training then evaluating on the development set the average fraction of tokens required to cause a decision flip. We use the model with the lowest fraction of tokens scores and report on the test set.

DATA	AV. $\|W\|$	C	SPLITS TRAIN/DEV/TEST
SST	18	2	6,920 / 872 / 1,821
AG	36	4	102,000 / 18,000 / 7,600
EV.INF.	363	3	5,789 / 684 / 720
M.RC	305	2	24,029 / 3,214 / 4,848
SEMEVAL	20	3	6,000 / 2,000 / 20,630

Table 5: Dataset statistics including average words per input, number of classes and splits (see also Appx. A).

DATASET	MODEL	lr^m	lr^c	F1
SST	BERT-BASE	1E-5	1E-4	.91 ± .00
AG	BERT-BASE	1E-5	1E-4	.93 ± .00
EV.INF.	SCIBERT	5E-6	2E-4	.84 ± .01
M.RC	ROBERTA-BASE	2E-6	2E-4	.75 ± .01
SEMEVAL	BERT-BASE	1E-5	1E-4	.59 ± .02

Table 6: Model and their hyper-parameters for each dataset, including learning rate for the model (lr^m) and the classifier layer (lr^c) and F1 macro scores on the development set across three runs.

D Further Details on Evaluating Faithfulness

Erasure (Serrano and Smith, 2019; Nguyen, 2018b): Jacovi and Goldberg (2018b) propose that an appropriate measure of *faithfulness* of an explanation can be obtained through *input erasure* (the most relevant parts of the input–according to the explanation–are removed). We therefore record the average fraction of tokens required to be removed across instances to cause a decision flip. Removal is conducted in descending token importance order at every 5% of the length in the sequence, as searching at every token is computationally expensive (Atanasova et al., 2020). Note that we conduct all experiments at the input level (i.e. by removing the token from the input sequence instead of only removing its corresponding attention weight) as we consider the scores from importance metrics to pertain to the corresponding input token following related work (Arras et al., 2016, 2017; Nguyen, 2018a; Vashishth et al., 2019; Grimsley et al., 2020).

FRESH (Jain et al., 2020): A pipeline composed of a *support model-extractor-classifier*, whereby the *support model* is the model trained on the full text and allocates importance to tokens, *extractor* the approach used and extract the rationales according to the importance from the *support model* and *classifier* the model trained on the rationales. The higher the *classifier*'s predictive performance the more faithful the rationales by the *support model*.

Similar to Jain et al. (2020), for FRESH we extract rationales of a fixed ratio compared to the sequence length by two thresholder approaches (THRESH.):

- TOPK: The *top-k* tokens as indicated by the corresponding importance metric, treating each word independently.
- CONTIGUOUS: The span of length *k* that results in the highest overall score as indicated by the importance metric.

E Further Details on Feature Attribution Approaches

- $\boldsymbol{\alpha}$: Importance rank corresponding to normalized attention scores.
- $\boldsymbol{\alpha}\nabla\boldsymbol{\alpha}$: Scales the attention scores α_i with their corresponding gradients $\nabla\alpha_i = \frac{\partial\hat{y}}{\partial\alpha_i}$. Serrano and Smith (2019)[10]
- $\mathbf{x}\nabla\mathbf{x}$ (InputXGrad) (Kindermans et al., 2016; Atanasova et al., 2020): Ranking words by multiplying the gradient of the input by the input with respect to the predicted class, where $\nabla x_i = \frac{\partial\hat{y}}{\partial x_i}$
- **I.G.** (Integrated Gradients) (Sundararajan et al., 2017): Ranking words by computing the integral of the gradients taken along a straight path from a baseline input to the original input, where the baseline input is a sequence of zero embedding vectors.

F Further Results on FRESH

In Table 7 we present complementary results on the F1 macro scores of the classifier trained on extracted contiguous rationales. Rational ratios for datasets SST, AG, EV.INF. and M.RC are from Jain et al. (2020), whilst for SEMEVAL we choose a 20% ratio.

DATASET	BASELINE	SALOSS TEXTRANK	SALOSS UNIFORM
SST (20%)	.82 (.00)	**.83** (.00) †	.80 (.00)
AG (20%)	**.90** (.00)	.89 (.00)	.89 (.00)
EV.INF. (10%)	**.79** (.00)	.78 (.00)	.78 (.00)
M.RC (20%)	.70 (.00)	.67 (.00)	**.71** (.00)
SEMEVAL (20%)	.46 (.03)	**.47** (.01)†	.42 (.00)

Table 7: F1 macro on models trained with extracted rationales (CONTIGUOUS and α) using FRESH for BASELINE and SALOSS models. **Bold** denotes best performance in each dataset. † indicates that SALOSS rationales perform significantly better (t-test, $p < 0.05$).

We can first observe that models trained on contiguous rationale extracted from models trained with SALOSS, obtain comparable performance to models without (BASE). Additionally, results show that classifier performance does not reach those with TOPK rationales. We can therefore assume that TOPK rationales result to inherently faithful classifiers with higher performance. It is encouraging to notice that in the datasets where performance is comparable with our approach (AG, EV.INF., M.RC), it is likely due to reaching close to FULLTEXT performance. For example, classifier performance trained on CONTIGUOUS rationales from BASE. in SST is at .82 compared to .83 with SALOSS rationales.

Results also suggest that our uninformative baseline (UNIF.), reduces the faithfulness of rationales in most cases resulting in lower classifier performance. We hypothesize that in cases where performance is comparable with BASE. and SALOSS, it is due to the task being relatively easy and as such the loss function not impacting the faithfulness of rationales. We consider this direction as an interesting area for future work.

G PoS Importance Allocation

We also conduct an analysis whereby we record the average importance scores under each Part of Speach (PoS) tag. We run a pretrained PoS tagger from `spaCy` (Honnibal et al., 2020) across the text and compute average importance calculated from a feature attribution approach for each PoS tag. We therefore aim to observe differences in allocation of importance in linguistic features between models trained with out our proposed approach (BASE.) and with (SALOSS). In Figure 1 we present distri-

[10] Serrano and Smith (2019) show that gradient-based attention ranking metrics ($\boldsymbol{\alpha}\nabla\boldsymbol{\alpha}$) are better in providing faithful explanations compared to just using attention ($\boldsymbol{\alpha}$).

bution of importance (calculated with $\alpha\nabla\alpha$) across PoS tags, on three datasets (SST, AG and EV.INF.).

Observing Figure 1a, we can see that $\alpha\nabla\alpha$ with SALOSS places greater importance on proper nouns (PROPN), auxiliary words (AUX), pronouns (PRON) and interjections (INTJ). In comparison the most prominent tags with BASE are INTJ, PROPN, coordinating conjunctions (CCONJ) and nouns (NOUN). In a sentiment analysis task, it is notable that both BASE. and SALOSS base high importance on average on interjections, which typically demonstrate feelings or emotions. Both appear to highlight particularly well adjectives, which we consider more important for sentiment analysis as they name attributes of other words. On the other end we also observe that SALOSS places lower importance on average to CCONJ and punctuation (PUNCT) compared to BASE. This suggests that for SST, SALOSS models possibly shift their importance to more informative for the task word groups.

Moving on to Figure 1b, we observe a very high peak on proper nouns (PROPN) and unidentified tokens (X) with SALOSS compared to BASE.. In a news classification task proper nouns such as the NATO and other organization or city names can indicate the topic of a sequence. We assume that for SALOSS to place such great importance on proper nouns, we manage with our approach to shift the model's attention to more informative for the task tokens. However we also observe unidentified symbols having large average importance scores with SALOSS. Whilst we do not study plausibility (human understandability of explanations), we consider this a limitation and we consider exploring and addressing this an interesting direction for future work.

Finally, examining Figure 1c, we observe that both SALOSS and BASE place very high importance on particle (PART) words such as *not*. We consider this encouraging, as large parts of the task is to infer if there was a significant difference or *not* based on an observation in the text. Additionally, we observe that SALOSS attends highly to subordinating conjunction (SCONJ) words such as than, which if placed in the context of "*significantly higher than*" directly relates to our task. Also with SALOSS we observe a reduction in attention to pronouns (PRON) compared to BASE, which we consider encouraging as PRON words are not directly related to the task of infering relationships.

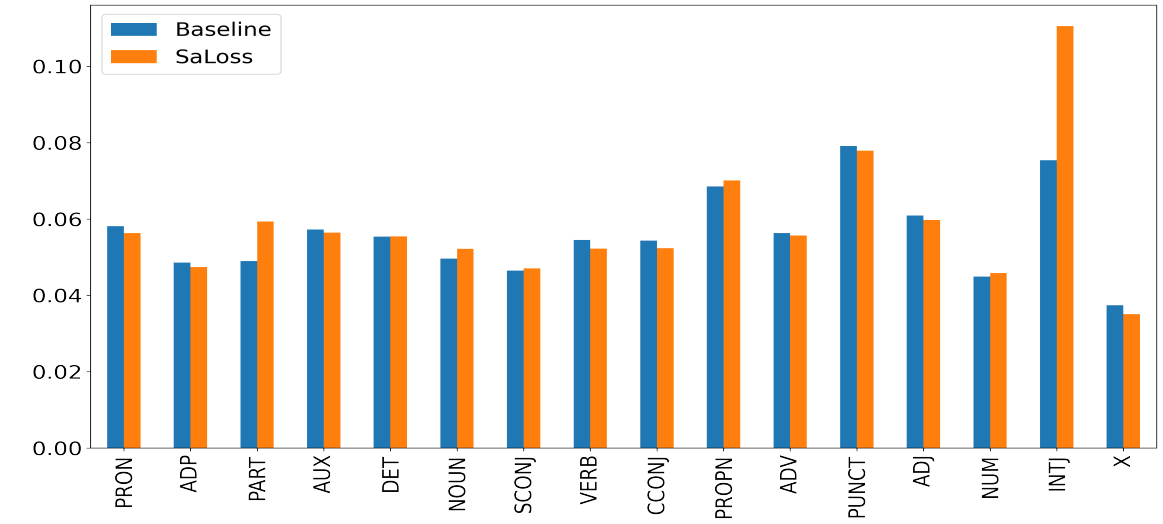
	METRIC	SST	AG	EV.INF.	M.RC	SEMEVAL
	RAND.	.66	.67	.51	.44	.54
TEXTRANK	α	.42	.53	.14	.19	.39
	$\mathbf{x}\nabla\mathbf{x}$	.61	.59	.38	.30	.51
	$\alpha\nabla\alpha$	.48	.50	.12	.24	.41
	I.G.	.61	.57	.33	.33	.45
CHISQUARED	α	.49	.67	.29	.38	.44
	$\mathbf{x}\nabla\mathbf{x}$	.60	.59	.47	.34	.54
	$\alpha\nabla\alpha$	.61	.71	.28	.33	.49
	I.G.	.58	.56	.48	.38	.47
TFIDF	α	.47	.43	.20	.33	.48
	$\mathbf{x}\nabla\mathbf{x}$	.62	.57	.41	.36	.57
	$\alpha\nabla\alpha$	.50	.47	.20	.37	.58
	I.G.	.58	.56	.40	.38	.53

Table 8: Average fraction of tokens required to cause a decision flip across datasets and feature attribution metrics (lower is better).

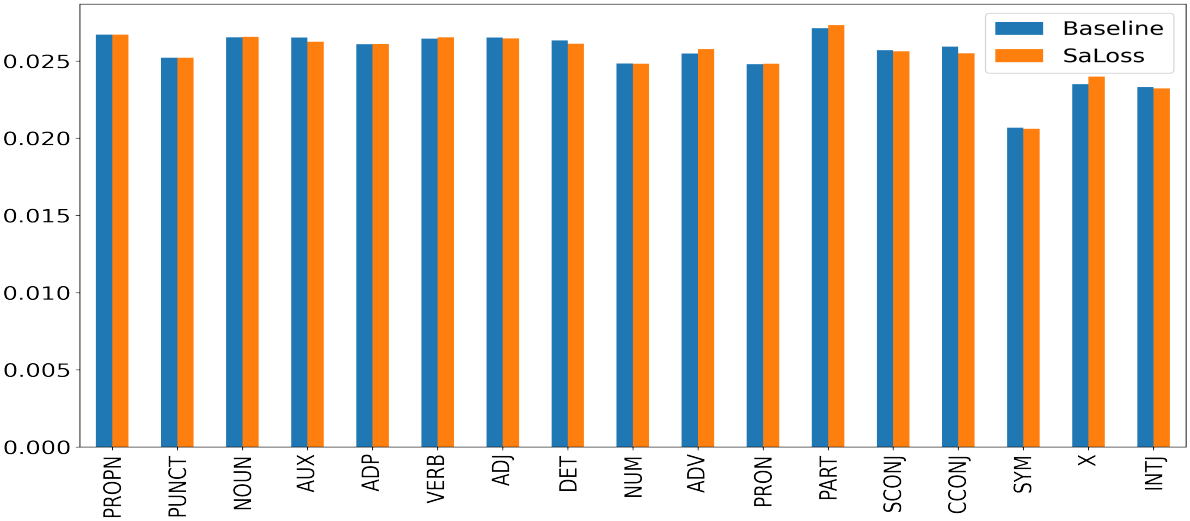
This indicates that our proposed objective manages to guide the model's attention away from uninformative tokens such as others and punctation, and towards more informative for the task token types (SCONJ, CCONJ).

H Input Erasure

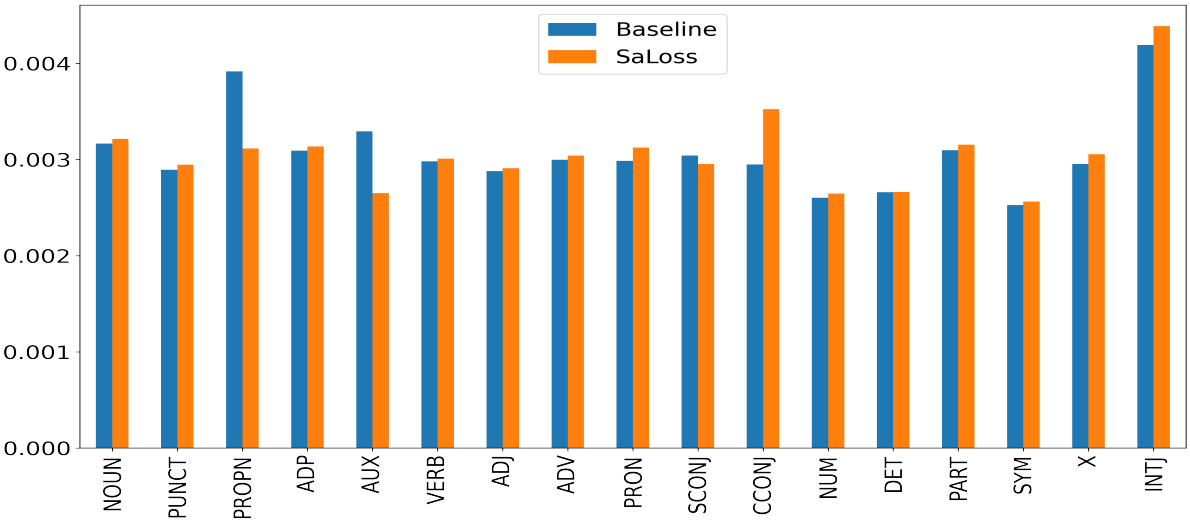
Table 8 presents the average fraction of tokens required to cause a prediction switch (decision flip), when training models with SALOSS and (1) TEXTRANK; (2) CHISQUARED; (3) TFIDF. We observe that when models are regularized with TEXTRANK scores, the feature attribution approaches result in a lower average fraction of tokens to cause a prediction switch compared to the other two salience functions. We also observe that TFIDF is comparable with TEXTRANK in most cases, outperforming CHISQUARED. We hypothesize that TFIDF performs poorer than TEXTRANK is due to the way these two approaches compute their "importance" scores. The first computes them globally, whilst the latter locally (at instance-level) which we assume is more beneficial for explanation faithfulness.



(a) (SST)



(b) (AG)



(c) (EV.INF.)

Figure 1: Average importance across Part of Speech (PoS) tags as indicated by $\alpha\nabla\alpha$ with BASELINE, with our proposed component SALOSS.