The Role of Syntactic Span Preferences in Post-Hoc Explanation Disagreement

Jonathan Kamp[1], **Lisa Beinborn**[1,2], **Antske Fokkens**[1,3]
[1]Computational Linguistics and Text Mining Lab, Vrije Universiteit Amsterdam
[2]Institute of Computer Science and Campus Institute Data Science, University of Göttingen
[3]Dept. of Mathematics and Computer Science, Eindhoven University of Technology
{j.b.kamp, antske.fokkens}@vu.nl, lisa.beinborn@uni-goettingen.de

Abstract

Post-hoc explanation methods are an important tool for increasing model transparency for users. Unfortunately, the currently used methods for attributing token importance often yield diverging patterns. In this work, we study potential sources of disagreement across methods from a linguistic perspective. We find that different methods systematically select different classes of words and that methods that agree most with other methods and with humans display similar linguistic preferences. Token-level differences between methods are smoothed out if we compare them on the syntactic span level. We also find higher agreement across methods by estimating the most important spans dynamically instead of relying on a fixed subset of size k. We systematically investigate the interaction between k and spans and propose an improved configuration for selecting important tokens.

Keywords: interpretability, spans, agreement

1. Introduction

Transformer-based models learn to map features in the input to some output. When training an NLP system, the model learns to identify the most important features (in our case tokens) for the final prediction. Post-hoc explanation methods such as LIME (Ribeiro et al., 2016) and Integrated Gradient (Sundararajan et al., 2017) aim to attribute an importance score to the individual features to interpret the model's decisions. Generally, these methods tend to disagree with each other when ranking token importance on a set of top-k tokens based on attribution scores (Neely et al., 2022). Given their disagreement, and assuming that explanations that are faithful to the transformer's inner mechanisms should be agreeable (Jain and Wallace, 2019), the faithfulness of these methods comes under question. However, methods might agree more than initially appears. For example, Figure 1 shows that none of the methods selects the same top-4 tokens and that 12 of the 13 tokens appear in at least one top-4 selection, indicating a high variance across methods. Intuitively though, methods seem to target the verb phrases *are standing* and *are unloading* to a high degree as the vast majority highlights at least one of the tokens in each of these phrases. Similarly, some methods tend to agree on the noun phrases *shipyard workers* (first occurrence) and *the ships*, and even more so on different tokenised subwords of the same word, namely *un* and *##loading*. This leads us to hypothesise that agreement between methods is systematically higher when we look at the linguistic spans they are targeting: the constituents to which tokens syntactically belong.

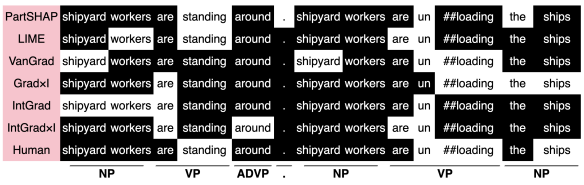


Figure 1: Top-k highlights (light background) per attribution method and human preference for $k = 4$. The syntactic spans are given underneath.

This example shows that a single method may have a specific *preference* for one word class over another, e.g. noun over adjective, auxiliary over inflected verb form or modifier over head. For example, Ramnath et al. (2020) report part-of-speech (POS) preference statistics for the different layers of BERT (Devlin et al., 2019) for the Integrated Gradient method. However, the extent to which preferences differ across methods remains unclear, as well as its impact on method–method agreement.

A methodological aspect that directly affects agreement is the selection of the top-k most important tokens for each method to compare. k is a relatively under-explored parameter and is defined as the number of features that are assigned highest scores by the attribution method, relative to all the features in the input example. A common way of picking k is by selecting a fixed number, generally in the range $[1, 10]$. Intuitively, a k that is fixed across instances (e.g. 4) is suboptimal, and the selection process of k is often overlooked (Jesus et al., 2021; Camburu et al., 2019) or obtained by an approximation (Krishna et al., 2022). As an alternative, k can be estimated dynamically across instances (Pruthi

LREC-COLING 2024, pages 16066–16078

et al., 2022; Kamp et al., 2023), but different conceptual settings for this approach and their effect on agreement have not been investigated yet. Instead of ranking tokens by attribution score and manually setting a k, Kamp et al. (2023) propose to automatically detect tokens that are signal peaks in the input. Hypothesising that spans are better suited for agreement than tokens conceptually overlaps with this dynamic k approach. Precisely, the latter suggests that solely focusing on token-level attribution scores, semi-arbitrary importance cut-offs and the consequent agreement measurements between k tokens may be undesirable for interpreting model behaviour.

In this paper, we aim to disentangle the interdependencies between word class preference, span-level agreement, and the determination of k. We show that methods systematically select different word classes and that methods that agree most with other methods and with humans exhibit similar word class preferences. We also find that dynamic k and spans work well in combination, and that an adapted threshold for dynamically selecting the k most important tokens passes our baseline tests for both token- and span-level k estimation. Our main contributions are: i) a linguistic analysis of disagreement on the token-level and on the span-level and ii) an improvement to the dynamic-k estimation algorithm.[1]

2. Related Work

In this section, we place our work in the context of prior work on interpretability (§2.1), the patterns of linguistic information that attribution methods reveal (§2.2) and top-k estimation (§2.3).

2.1. Model Interpretation

Tracing the decision processes in neural models poses difficulties due to various factors, including their non-linear nature and the absence of explicit human-defined rules to link patterns in the input features with output labels. Different research lines exist to interpret different aspects of the model (Choudhary et al., 2022; Räuker et al., 2023), such as the linguistic information that might implicitly be learned by the model, or the importance that single input features might have had towards the model's decision (Madsen et al., 2022).

To address the latter, post-hoc attribution methods in NLP have been developed to assign a score to each token in the input, creating an *attribution profile* over the tokens. While these methods are often being used in error analyses (Bongard et al., 2022, i.a.), their reliability is questionable. In fact, attribution profiles obtained from different methods can differ strongly even on the same input. This leads to an overall low inter-method agreement (Neely et al., 2022), which has also been found for domains outside of NLP (Krishna et al., 2022). Diverging experimental results of such methods on different models, datasets and tasks provide additional evidence on their inconsistency. For example, when trying to identify the attribution methods that best align with human preferences–the most *plausible* (Jacovi and Goldberg, 2020) methods–, Atanasova et al. (2020) and Attanasio et al. (2022) come to fundamentally opposing conclusions. Roy et al. (2022) characterise disagreement between methods in a software defect prediction task as being highest in terms of top-k feature importance, followed by rank, then sign. Similarly to Pirie et al. (2023), they propose aggregation schemes for different explanation methods that aim to tackle disagreement in real-world use cases.

One question that, to our knowledge, remains under-explored, is *why* attribution methods in NLP disagree. A key to answering this would be comparing methods on their *faithfulness*, i.e. the degree to which methods are reflecting the model's decision making process, as recent work (Atanasova et al., 2023, i.a.) aims to assess. However, directly measuring faithfulness might only find glimpses of the model's inner workings rather than providing a conclusive answer (Jacovi and Goldberg, 2020). Therefore, we think that the first step should be explaining disagreement by the observable output of the methods, i.e. the attribution profiles. We aim to provide a linguistic comparison by quantifying the kind of features that are targeted, expecting different methods to consistently target different classes of words.

2.2. Linguistic Patterns in Attributions

Identifying the linguistic preferences of models is important in order to pinpoint the cues upon which models depend during inference time. Only a handful of studies have explored POS preference. Especially in a feature attribution setting, there is little evidence that shows certain preferences by different attribution methods and how these preferences differ. Lai et al. (2019) find that different *models* (i.e. LSTM, XGBoost and SVM) have different POS preferences on the same data and task, but they do not explore preferences for different *attribution methods*. Ramnath et al. (2020) examine the top-5 most important tokens in each layer and find that BERT (Devlin et al., 2019) primarily focuses on nouns in all 12 layers, followed by verbs and adjectives. Interestingly, both punctuation tokens and stop words each correspond to 10% in the top-5 selections. However, only Integrated Gradient (Sundararajan et al., 2017) was used in this experiment,

[1] All analyses are available at: https://github.com/jbkamp/repo-Span-Pref

limiting the generalisability of their findings. Our analyses differ from theirs in that we compare different methods and investigate the overlap between agreement and linguistic preference.

Language (and model behavior) can often not be explained by merely highlighting individual tokens. Rather, we would ideally observe how features act in combination with each other and, for example, if they do so hierarchically. As an alternative way of analysing the attributions of tokens in isolation, we find a growing line of research on feature *interactions*. Jumelet and Zuidema (2023) find evidence of attribution methods faithfully reflecting linguistic structure in language models. Sikdar et al. (2021) combine token-wise attribution scores into scores assigned to syntactic parent constituents. Similarly, Babiker et al. (2023) train a model on intermediate representations in a hierarchical fashion. Song et al. (2023) aim to capture the causal effect of word group combinations on the prediction but limit their scope to the Integrated Gradient method. Pruthi et al. (2022) anticipate that certain spans of tokens should be highlighted by attribution methods in a sentiment analysis task. While their intuition is on point, the relatively broad expectations found in the latter underscore the relevance of a clear definition of token spans and their role in demonstrating how neighboring features are grouped.

As far as we know, there is no prior work that covers a linguistic analysis of the token selections targeted by different attribution methods. To the best of our knowledge, we are also the first to investigate the relation between disagreement on the linguistic level to overall disagreement among methods. We provide a linguistic analysis in terms of individual tokens, and also in terms of spans that have a clear syntactic definition. In particular, we link disagreement to linguistic preference on the token level and within spans.

2.3. Top-k Estimation

We analyse the factors of disagreement through an additional scope, namely top-k estimation. k represents the number of most important tokens in the attribution profile. Studies reporting on consistent disagreement between methods do not take the impact of the k number of selected tokens into account (Pruthi et al., 2022; Krishna et al., 2022; Neely et al., 2022). A common way of selecting k is approximating it to a *low* value, e.g. 1 or 2 (Bastings et al., 2022), 5 (Ramnath et al., 2020), 5 or 10 (Camburu et al., 2019), 25% of the average input length (Krishna et al., 2022). However, a k that is fixed does not account for variability among instances. A k that is too low can exclude important tokens from the comparison, whereas a k that is too high will include non-important tokens while artificially boosting agreement between methods. Keeping k relatively low also helps users to more easily digest the explanations in a real-world scenario.

The value of k has also been estimated dynamically. Pruthi et al. (2022) set k to 10% of the input length, assuming that longer inputs have a higher number of important features than shorter inputs. Kamp et al. (2023) propose a k that varies dynamically based on properties of the attribution profile of each instance, aiming to include features that display above average importance and that focus more on the targeted region of the input instead of the specific token. While their method estimates a value for k that is close to human preference, we find that their algorithm necessitates further experiments and refinement. Different importance thresholds are possible and need baseline benchmarking. Also, as of now, prior methods for determining dynamic k do not explicitly account for negative attribution scores.

We adopt and improve the **dynamic** k estimation by Kamp et al. (2023) throughout §4, when measuring agreement at the span level compared to the token level. Formally, this dynamic approach defines a strong signal in the attribution profile as a score that is higher than its neighboring scores according to two principles: *local importance* and *global importance*. *Local importance* requires that a score must be higher than its strict neighbors (± 1 window) to reduce redundancy of tokens belonging to the same signal. In other words, a set of adjacent tokens with relatively high scores is converted to a single important signal and the highest attribution in the set is kept as the peak of the signal. Similarly, the *global importance* principle requires important signals to be minimally above average signal strength, i.e. $> \mu_{ap}$, where ap is the attribution profile. By only adopting the *global importance* threshold, the inclusion of groups of (redundant) neighboring tokens with high attribution scores is expected to increase k, unnecessarily boosting the agreement scores. Therefore, the addition of a *local importance* setting, which we keep unaltered for our remaining experiments, is necessary to estimate signal *peaks*. As for *global importance*, we keep the threshold constant in §4.2 to compare span-level agreement to token-level agreement in previous work, and explore different settings in §4.3.

3. Linguistic Analysis

We hypothesize that one of the reasons attribution methods disagree is that different methods have different preferences for the classes of words they target. Following from this, we expect that differences in word class preferences are put under a different light when we look at the syntactic spans they are assigned to.

3.1. Setup

To analyse the disagreement problem, we consider six different attribution methods on a natural language inference task. For the sake of testing our hypothesis against the agreement results from prior work, we follow Kamp et al. (2023) for the experimental setup. For the backbone model, we use the default training split (549,361 instances) of the e-SNLI dataset (Camburu et al., 2018) to finetune DistilBERT (Sanh et al., 2019) 10 times on 10 different random seeds. We then use the model (0.89 F1) with the least variation in attribution profiles on the default test split (9,842 instances) for analysis. One instance in the dataset corresponds to the concatenation of a premise followed by a hypothesis. The possible output labels are contradiction, entailment and neutral, making it a multi-class problem. Classes are balanced and indicate the relation between premise and hypothesis.

The words in every instance are also annotated as being important or not important towards the output label (3 annotators per instance, 4±3 important words on average), producing so-called human *rationales* (Carton et al., 2020). From these human rationales, we derive word-level aggregation scores comprised in the interval $[0, 1]$ indicating the proportion of annotators that found the word important. These scores are used to compare attribution scores to human preference when considering a top-k selection (see *Human* in Figures 1, 2 and 3). As for the attribution methods, we use both *gradient-based* approaches by including Vanilla Gradient (Simonyan et al., 2014), Integrated Gradient (Sundararajan et al., 2017), and both versions multiplied with the input (Shrikumar et al., 2017), as well as *perturbation-based* approaches, by including Partition SHAP (Lundberg and Lee, 2017) and LIME (Ribeiro et al., 2016).[2]

3.2. Preference for a Word Class

The first step in our analysis compares word class preference of different attribution methods on top-k tokens. We set k to 4 which corresponds to the average number of tokens that were highlighted by humans in e-SNLI. This value is reflected by a comparable value of averaged dynamic k and comparable method–method agreement levels as found by Kamp et al. (2023). Figure 2 illustrates the occurrence of different word classes among the tokens with the highest attribution values (i.e. *important* tokens) for each method and for human aggregated annotations. We compare the ratio of important stop words (Figure 2a), punctuation tokens (2b), and the distribution of the five most preferred POS tags by humans: NOUN, VERB, ADJ, ADP, DET (Figure 2c). Interestingly, with regards to Integrated Gradient, Gradient × Input and Integrated Gradient × Input, roughly 10% in each top-4 selection on average consists of punctuation. Despite question answering and natural language inference being different tasks, we replicate the findings on punctuation preference for Integrated Gradient by Ramnath et al. (2020). Notably, these findings do not generalise to the other methods.

Intuitively, this preference seems to be inherent to the method and not to the underlying model, as each instance normally is a concatenation of two sentences tailed by a full stop each; hence, it is very unlikely that the model is using punctuation as shortcut signals to the output labels. This might suggest that some methods pick up information about the approximate location of a signal in the sentence (*locality* information), rather than the precise token (*lexical* information). While punctuation may be a simple symptom of locality, it is important to further examine this phenomenon in the broader context of spans. We do so through a linguistic analysis of spans of locally adjacent tokens, the use of dynamic k, and their intersection in §4.

Stop words on the other hand do not display a similar preference as found by Ramnath et al. (2020) (40% versus 10% for Integrated Gradient), indicating that this difference might be task-related. For the other POS tag preferences, we do not observe a clear overlap with prior research for Integrated Gradient (NOUN: no overlap; VERB: overlap; ADJ: no overlap; ADP: cannot compare; DET: cannot compare). What we do observe from Figure 2, is the systematic different preference for stop words, punctuation and most frequent POS tags by Integrated Gradient, Gradient × Input and Integrated Gradient × Input (Group 1), compared to the other methods and to humans (Group 2). Hence, this intuitively leaves us with two groups displaying different word class preferences.

Assuming that methods (including human rationales) are independent, we apply Chi-Square tests to method–method (and human–method) pairs' preference distributions.[3] For each pair, we measure whether there is a significant difference between stop word distributions, between punctuation distributions and between POS tag distributions. The tests confirm our initial observations that most distributions from one group are significantly different from the other group (25/36 pairs,[4] with $p < .05$) and that no significant differences are found within groups. Most of the exceptions arise for pairs involving Integrated Gradient × Input, with 3 out of 3 non-significant differences found in com-

[2] Ferret package v0.4.1 (Attanasio et al., 2023).

[3] The full Chi-Square tests are given in Appendix A.

[4] A total of twelve Group 1 – Group 2 comparisons are possible for each of the three word classes (stop words, punctuation and POS), resulting in 36 pairs.

bination with Partition SHAP,[5] 2 out of 3 with LIME and 1 out of 3 with human rationales. Hence, Integrated Gradient × Input explains half (6/11) of the non-significant differences found and can roughly be placed in between the two groups. Additionally, punctuation preferences account for half (6/11) of the non-significant differences between groups. This might be due to the small numbers of the punctuation frequencies, which may have affected the Chi-Square statistics.

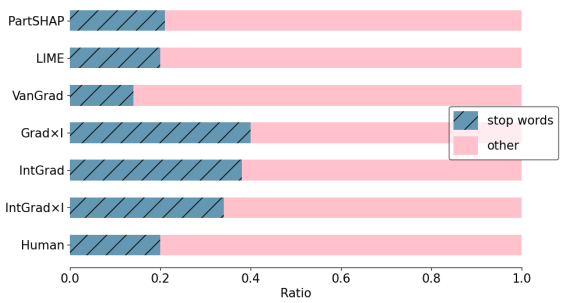
Primarily Integrated Gradient and Gradient × Input, followed by Integrated Gradient × Input, are indeed the methods for which Kamp et al. (2023) find that method–method and human–method agreement are lowest. This shows that the high similarity in terms of word class preference for the methods in Group 1 results in consistently lower agreement. Simultaneously, the similar preference for methods in Group 2, which happens to be close to human preference, correlates with higher agreement. From the opposite perspective: methods that are similar in terms of agreement scores exhibit similar word class preferences.

3.3. Span Definition

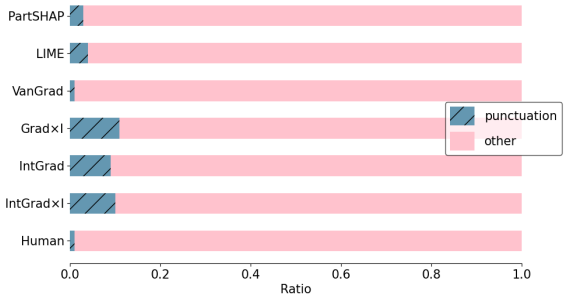
We obtain syntactic spans by shallow parsing the data with Flair chunker (Akbik et al., 2018), similarly to Zhou et al. (2020) who use parsed constituents as pre-processed spans for a parsing experiment. Chunking is commonly adopted in Named Entity Recognition where usually noun phrases or verb phrases are the focus of interest (Taufiq et al., 2023). For our task, the advantage of this method over full constituency parsing (Kitaev et al., 2019, e.g.) or dependency parsing (Chen and Manning, 2014, e.g.) is that the chunker output of discrete non-overlapping units facilitates direct alignment with attribution values. Punctuation tokens are ignored by the parser; we treat them as separate spans.

Sikdar et al. (2021) use constituency parsing (Mrini et al., 2020) as a basis for hierarchically attributing feature importance scores from tokens to phrases (including any subphrases). However, different methods can have different word class preferences (e.g. a noun modifier may systematically be attributed more importance over its head) and it is therefore questionable whether score aggregation of any kind is a sensible approach. Having clearly defined, non-overlapping *phrases* is instead crucial to our initial hypothesis.

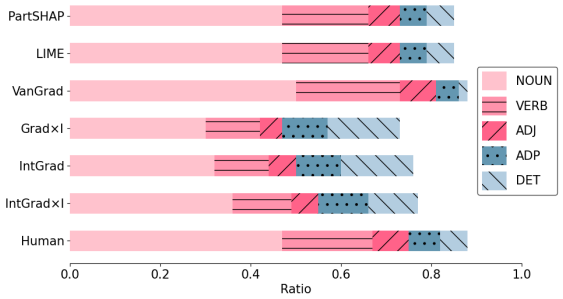
In our dataset, each sentence contains on average 24.4 tokens (6–73), which are grouped into 15.3 spans (3–45). The average ratio of spans over tokens is 0.63 (0.23–1.0). A targeted span is a span that contains at least one token included in the top-k selection by the attribution method. During agreement evaluation we treat spans as atomic units, meaning that a span is assigned 1 if targeted, otherwise 0 (similarly to tokens in top-k selection). For a fixed k set to 4, the average number of targeted spans in a sentence is slightly lower: Partition SHAP 3.5, LIME 3.6, Vanilla Grad 3.5, Grad × Input 3.6, Integrated Gradient 3.7, Integrated Gradient × Input 3.5, Human 3.3. The average over methods is 3.5.



(a) Relative frequency for important stop words, $k = 4$.



(b) Relative frequency for important punctuation, $k = 4$.



(c) Relative frequency for important POS, $k = 4$. We consider the 5 most preferred POS tags by humans.

Figure 2: Preference for different word classes per attribution method.

3.4. Head vs. Modifier Preference

We have seen that Gradient × Input and Vanilla Gradient exhibit complementary linguistic preferences

[5] For each pair of methods, there are three word classes for which significance can be tested. We therefore compute the ratio 'i out of 3'.

for noun tokens (the lowest versus highest ratio of noun tokens in the top-4). We zoom in on this phenomenon and investigate the attribution patterns in noun phrases (NPs), focussing on methods that select the head over its modifier and vice versa.

We examine a subset of noun phrase spans that are grouped according to $k = 4$ by Gradient × Input and Vanilla Gradient. The NPs must span a minimum of two tokens to make the preference analysis for different word classes possible. To add some consensus stability to this subset, the spans under question should also be targeted by highly agreeing methods Partition SHAP and LIME. We compare the attribution profiles of Gradient × Input and Vanilla Gradient on the token and the span level for the specific [DET, NOUN] construction, the most prevalent among length-2 noun phrases (73%, 1,963). Interestingly, of the cases where Vanilla Gradient targeted NOUN (99%, 1,951), Gradient × Input targeted DET half of the times (899). This example clearly illustrates how methods do not only target different word classes in absolute terms, but also how that translates to systematic, alternating differences within syntactic spans.

Furthermore, the ratio of targeted tokens in the [DET, NOUN] NPs is comparable: 57% for Vanilla Gradient versus 60% for Gradient × Input. This detail strengthens the claim of systematic preference in that the DET–NOUN alternation, i.a., is usually *exclusive*. In other words, it is uncommon for the two described methods to target both tokens from the NPs. This increases the prominence of the preference phenomenon in cases where one selects the DET and the other the NOUN.

4. Agreement at the Span Level

We showed that different methods have different word class preferences and that the preference can be strong in the case of syntactic noun phrases. A consistently strong preference by two methods leads to a strong disagreement at the token level. The expectation that methods should agree on the token level might therefore be too strict. Given these insights, we measure method–method and human–method agreement at the span level, expecting a relative improvement compared to token-level agreement.

4.1. Setup

The dataset, model configurations and pool of attribution methods that we use are identical to those described in the linguistic analysis (§3). In addition, we adopt the definition for spans given in §3.3. Our data therefore has a version where the instances are divided into tokens and one where instances are split into spans. The details of dynamic k correspond to those described in §2.3.

4.2. The Effect of Dynamic k on Spans

We compare the effect of dynamic k on the span level versus dynamic k on the token level. We measure the *effect* as the increase in agreement i) versus a baseline to assess overall difficulty of the task and ii) versus fixed $k = 4$ to assess the ability of the dynamic approach to detect important spans. We expect dynamic k to be better suited than fixed k to identify linguistic spans that the model considers important in the instance. Specifically, the *local importance* setting (in combination with *global importance*) appears to work as a pooling operator, highlighting the distinct important parts of the instance rather than few concentrated parts. We assume here that for the specific NLI task, > 1 parts of the input should be considered important.

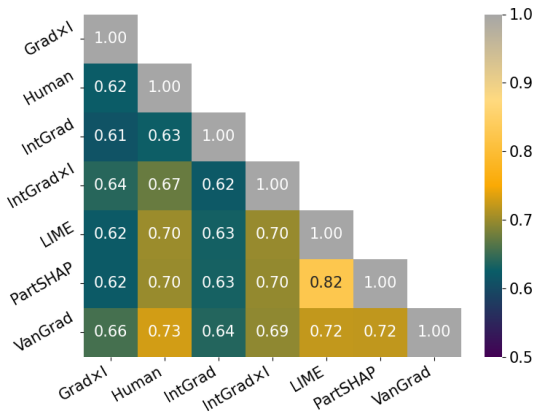
Agreement is measured as follows. We denote an attribution method as $\mathbf{A}$. $\mathbf{A}$ assigns an attribution profile $\mathbf{a} = \{a_1, a_2, ..., a_n\}$ to the input sequence of tokens $\mathbf{s} = \{w_1, w_2, ..., w_n\}$ so that each a_i indicates the importance of token w_i towards the inferred class. The subset of k tokens with the highest attribution values are formalized as $topk_A = \{t_1, t_2, ..., t_k\}$. We compare m attribution methods $A_1, ..., A_m$ in pairs by calculating sentence-level agreement@k. Agreement@k is based on the relevance of a each token. Relevance for a token w_i is equal to the ratio of methods that include the token in their respective $topk$ subsets. Agreement@k ignores perfect agreement on non-important tokens (where relevance $= 0$) in order not to inflate the score. For our experiments, we report mean agreement@k, the averaged agreement over instances in the dataset $\mathbf{D} = \{s_1, s_2, \ldots, s_d\}$.

$$\text{Relevance } r(w_i) = \frac{\sum_{A_j=1}^{m} [w_i \in topk_{A_j}]}{m} \quad (1)$$

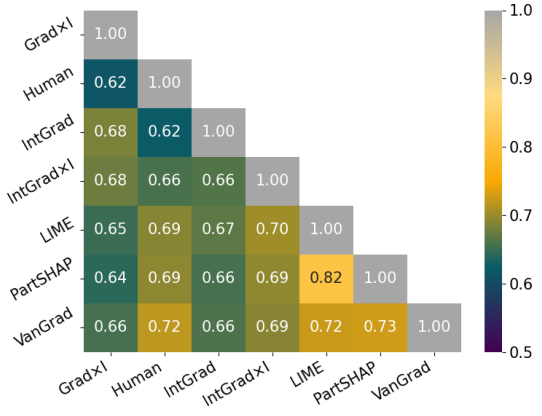
$$\text{Agreement@}k(s_i) = \frac{\sum_{w_i=1}^{n} r(w_i)}{\sum_{w_i=1}^{n} [r(w_i) > 0]} \quad (2)$$

$$\text{Agreement@}k(D) = \frac{\sum_{s_i=1}^{d} \text{agreement@}k(s_i)}{d} \quad (3)$$

The average pair-wise agreement (all method–method combinations) for dynamic k is 0.61 on the token level and 0.69 on the span level. 0.5 indicates perfect disagreement and 1.0 perfect agreement. While agreement seems relatively low, it might still suggest that consistently the same, few types of signals are identified by a pair of methods.



(a) Mean span agreement@$k = 4$.



(b) Mean span agreement@$k =$ dynamic.

Figure 3: Span agreement for fixed and dynamic k.

We compute a baseline to measure how likely methods are to agree on the token and span level with a pseudo-random attribution method. In other words, we measure task difficulty of making two vectors with a subset of important tokens and spans to agree given a low value of k. For fixed $k = 4$, 16% of the tokens in a sentence would be highlighted on average; consequently, 23% of the spans would be highlighted on average. For the token-level baseline, we then randomly shuffle two binary vectors of 100 elements, 16 of which 1s, and compute pairwise agreement. We repeat the process 10^3 times. For the span-level baseline we adopt the same procedure, with the exception that the 1s in the vector are 23. The resulting baselines are 0.54 and 0.57, respectively, indicating that agreeing on tokens and spans is similarly difficult at low values of k. We thus observe the token-wise baseline for fixed k being outperformed by 0.07 (0.54→0.61), whereas the span-wise baseline is outperformed by a relatively larger increase of 0.12 (0.57→0.69).

The comparison results between span agreement on fixed k versus dynamic k are given in Figure 3. While dynamic k provided marginal boosts (+0.00, +0.01, +0.02 compared to fixed $k = 4$) on the token level for Gradient × Input and Integrated Gradient (compare Kamp et al. (2023)), it proves to have a larger positive effect on the span level. Specifically, the span agreement for method–method pairs that include Gradient × Input and/or Integrated Gradient remains constant or increases (changes from +0.00 to +0.07 compared to fixed k). At the same time, other method–method and human–method agreement scores remain constant or marginally decrease (+0.01, +0.00, -0.01).

With regards to the largest difference observed between dynamic and fixed k, namely Integrated Gradient versus Gradient × Input, this can also be explained through the concentration levels of targeted tokens within spans. In fact, dynamic k scatters the important tokens so that more spans are targeted compared to selecting an average fixed k. While $k = 4$ yields 3.7 and 3.6 spans on average for the two methods, dynamic k yields 6.9 and 6.5. Since it becomes easier for methods to agree when more tokens (and therefore more spans) are targeted, we investigate the settings of dynamic k further (§4.3).

4.3. Adjusting Dynamic k

How can we validate or improve the dynamic k algorithm? A solid global importance threshold should meet two conditions: i) resulting values of k should be *low*, preferably close to human preference average of 4±3; ii) they should outperform a baseline.

We explore multiple thresholds: different combinations of $\mu[+,-][0,1,2]\sigma$, typical distances from the mean in a distribution; the *median*, which is more robust to outliers than μ. The thresholds are calculated for (a) all scores and (b) positive scores. Thresholds for positive scores should ignore attributions with negative importance towards the inferred class. These are common in methods such as Integrated Gradient × Input. The influence of negative values and peaks in the attribution profiles is not accounted for by the current threshold set at μ.

The resulting values of k for different thresholds are given in Table 1. We find that for different thresholds, resulting ks are comparable across methods, which might indicate that the attribution profiles have overall similar distributions. The three thresholds that yield closest ks to human preference are μ, $\mu > 0$ and *median* > 0. Closeness corresponds to the averaged Euclidean distance between the mean±stdev pairs and human preference of 4±3, for each threshold column.[6] Among these three, μ had already proven to keep k low and close to ground truth average (Kamp et al., 2023).

Even if the estimated ks by the three candidate thresholds are relatively low, it could be, for ex-

[6] See Appendix A for an overview of the distances.

		Thresholds					
	Method	μ	$\mu+\sigma$	$\mu+2\sigma$	$\mu-\sigma$	$\mu-2\sigma$	*median*
all	PartSHAP	4.54±1.73	2.16±0.95	1.25±0.65	7.36± 1.89	7.37±1.89	6.19±1.62
	LIME	5.34±2.35	2.24±1.05	1.23±0.65	8.31±2.86	8.32±2.87	7.13±2.48
	VanGrad	4.58±1.68	2.41±1.02	1.39±0.61	7.63±2.68	7.64±2.69	6.20±2.08
	Grad×I	6.83±2.59	2.39±1.12	0.68±0.65	8.21±2.82	8.28±2.83	7.08±2.51
	IntGrad	7.30±2.63	2.66±1.23	0.64±0.63	8.41±2.88	8.46±2.9	7.41±2.58
	IntGrad×I	5.68±2.37	2.27±1.08	1.02±0.62	8.04±2.80	8.07±2.82	6.83±2.39
>0	PartSHAP	3.34±1.33	1.86±0.82	1.07±0.55	7.00±2.01	7.28±1.93	5.01±1.61
	LIME	3.56±1.56	1.87±0.87	1.06±0.53	7.95±2.91	8.25±2.89	5.59±2.04
	VanGrad	4.58±1.68	2.41±1.02	1.39±0.61	7.63±2.68	7.64±2.69	6.20±2.08
	Grad×I	3.51±1.56	1.75±0.83	0.73±0.56	6.81±2.82	7.86±2.92	4.69±1.88
	IntGrad	3.47±1.60	1.67±0.81	0.62±0.55	6.60±2.81	7.81±3.05	4.54±1.86
	IntGrad×I	3.83±1.69	1.91±0.90	0.98±0.53	7.57±2.74	7.99±2.81	5.38±1.96

Table 1: Values of *k* for different global importance thresholds. The three methods that yield values of k closest to human preference are visually indicated with a dark background.

ample, that a method-specific k is too high, positively biasing the agreement score. A high k would even give high agreement for a pseudo-random attribution profile, which should not be possible if the threshold is properly set. Hence, we compare each method's agreement scores with other methods to the method's agreement with a baseline. This gives us an indication of how well a specific threshold works with different attribution profiles. We do this both on the token level and on the span level. The baseline method operates pseudo-randomly by assigning attribution scores to the tokens without knowledge about token importance. For each method, we randomly shuffle the scores in each attribution profile. Each method has its own baseline so that the different distributional properties of the attribution profiles are preserved. We then compute agreement@dynamic-k between original and shuffled attribution profiles, which are consequently averaged over the dataset. If the threshold for k-estimation is strong, the agreement with the baseline for each method should be lower than the agreement with other methods.

	Token	**Span**
Method	*BL:minAgr–maxAgr*	
PartSHAP	0.56:0.56–0.78	0.64:0.64–0.82
LIME	0.57:0.57–0.78	0.65:0.65–0.82
VanGrad	0.56:0.58–0.68	0.64:0.66–0.73
Grad×I	0.59:**0.56**–0.60	0.68:**0.64**–0.69
IntGrad	0.60:**0.58**–**0.59**	0.69:**0.66**–**0.68**
IntGrad×I	0.58:0.58–0.64	0.66:0.66–0.70

Table 2: Token and span agr. with other methods (range *minAgr* to *maxAgr*) versus baseline (*BL*), for threshold $= \mu$. Scores $<$ baseline in **bold**.

Results for μ are given in Table 2. We find that for μ, Integrated Gradient and Gradient × Input have higher baseline agreement than the other methods. This can be explained by the higher values of k for this threshold (i.e. 6.83 and 7.30 in Table 1). Importantly, both methods have method–method agreement scores that do not beat the baseline (which pseudo-randomly selects tokens), neither on the token level nor on the span level. With regards to *median* > 0,[7] multiple methods do not beat their baselines either. The threshold $\mu > 0$ instead does, for all methods and both on tokens and on spans. This is an indication of the fact that the latter might be a better threshold than μ for dynamic k estimation. An additional interpretation of why $\mu > 0$ works better than μ is that negative local maxima in the attribution profiles are hereby ignored, leading to less but more important k tokens (and spans) to be targeted. This baseline testing also shows that Gradient × Input and Integrated Gradient are unreliable methods: they have low agreement with other methods and often fail to beat a random baseline.

5. Discussion

Analysing disagreement from a linguistic perspective helps us to better understand the differences between attribution methods. We briefly discuss the implications of token- and span-level analyses on other tasks than NLI. With an eye on the ability and reliability of these methods to reflect the model's decision process, we also consider the implications for the *faithfulness* aspect in interpretability research.

[7] While reporting the baseline tests for threshold $= \mu$ in Table 2, we leave the overviews for thresholds $\mu > 0$ and *median* > 0 to Appendix A.

Generalisability of Spans Generally speaking, an NLI task is sufficiently challenging that it avoids sentences of different classes (e.g. contradiction, entailment) differing by exactly one word. It is therefore fair to expect methods to target the same span and not to penalise them for disagreeing on the token level. However, targeting a modifier instead of its syntactic head can make a big difference for other tasks. Additionally, the span-token ratio should determine the difficulty of assessing span-level agreement compared to tokens. The choice of considering spans rather than tokens should therefore be weighted against the type of task and data.

On a similar note, §3.2 describes the systematic differences in punctuation preferences. We may hypothesise that methods that consistently include full stops in their top-k are actually catching the signal's onset (*locality* information) rather than the full stop being itself a signal (*lexical* information). To this end, our choice of treating punctuation as separate spans might have influenced the span agreement of such methods. More research is necessary to disentangle locality from lexical information.

Agreement as a Proxy for Faithfulness Agreement is linked with both plausibility and faithfulness. We considered plausibility when estimating dynamic k thresholds, as we aimed for ks close to human preference. However, a more direct way of testing for plausibility in this context is by assessing human–method agreement, which we mostly left out of scope in this study. To that end, we did find that agreement results are constant on both tokens and spans, possibly suggesting that human–method agreement reaches a ceiling already at the token level (i.e. tokens are targeted that belong to different signals in the sentence). This interpretation might even hold for more faithful methods. In fact, models do often not rely on the same patterns as humans do, instead resorting to shortcut signals.

Measuring faithfulness, on the other hand, is less straightforward. Following Jain and Wallace (2019), who state that faithful attention-based explanations should be agreeable, we carefully extend their perspective in that agreement between method-*generic* explanations can be considered as a proxy for faithfulness. According to the principle of reproducibility in science (Popper, 2005), a finding that is confirmed through different means is, in principle, more likely to be correct. As such, if two attribution methods with distinct means yield similar results, they are likely similarly (un)faithful. If one method disagrees with the majority of the batch, either the one, the majority, or all are unfaithful. Because of the reproducibility principle, however, it is more likely that the majority is more faithful.

In this light, we could therefore speculate that Gradient × Input and Integrated Gradient were two of the less faithful methods in our study, an argument that is supported by their scarce agreement compared to a pseudo-random baseline. Given that some methods might highly correlate with other methods by design, one must be careful at drawing conclusions. Constructing a batch of methods that is representative of different ways of interpreting the model is, for this reason, not a simple task.

6. Conclusion and Future Directions

In this study, we approached post-hoc explanation disagreement from a syntactic perspective. We found that methods that agree most with other methods and with aggregated scores of human rationales have similar POS tag preferences for the targeted tokens. We then determined that attribution methods agree more at the span level than at the token level, which appear to be similarly difficult tasks at low values of k. One particular reason for disagreement is the consistent preference by one method to target the determiners instead of the noun head within the same noun phrase. We showed that dynamic k works well in combination with spans, as it seeks for non-neighboring important signals in the sentence. Finally, we empirically tested for different thresholds of the *global importance* setting of dynamic k, suggesting a value ($\mu > 0$) that accounts for both negative attribution scores and results in low ks.

One issue that dynamic k aims to tackle is the targeting of redundant tokens as signals in the same span. To complement this, a more in-depth analysis would provide a better understanding about the way that different methods concentrate their targeted tokens in the same spans. Intuitively, for a fixed k, some methods highlight tokens that are more sparse across the instance, whereas other more quickly concentrate targeted tokens within the same spans. To obtain such a concentration metric, one could measure how rapidly a set of tokens belonging to the most important ground truth span are being targeted, at increasing values of k.

Future directions of research include the exploration of different *local importance* criteria in the dynamic k algorithm, such as different windows (current ± 1 versus ± 2, ± 3). Another is to exploit (syntactic) span-based information to improve interpretability accuracy at the token level, or to improve explanation aggregation techniques. Finally, we advise future evaluation datasets based on multiple annotators' rationales to preserve specific instance–annotator mappings in the metadata. This would facilitate new directions in assessing the plausibility of attribution methods, specifically how variations in human subjectivity relate to agreement.

7. Ethical Considerations

We would like to reiterate that attribution scores cannot be blindly relied upon to precisely determine model functioning, as they can be influenced by experimental factors such as task and model performance. To avoid drawing generalised conclusions, it is advisable to employ multiple metrics when studying feature attribution.

8. Acknowledgements

Jonathan Kamp's research was funded by the Dutch National Science Organisation (NWO) through the project InDeep: Interpreting Deep Learning Models for Text and Sound (NWA.1292.19.399). Antske Fokkens was supported by the EU Horizon 2020 project InTaVia: In/Tangible European Heritage - Visual Analysis, Curation and Communication (http://intavia.eu) under grant agreement No. 101004825. Lisa Beinborn's work was funded by the Dutch National Science Organisation (NWO) through the VENI program (Vl.Veni.211C.039). We would like to thank the anonymous reviewers for their valuable contribution.

9. Bibliographical References

Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual string embeddings for sequence labeling. In *COLING 2018, 27th International Conference on Computational Linguistics*, pages 1638–1649.

Pepa Atanasova, Oana-Maria Camburu, Christina Lioma, Thomas Lukasiewicz, Jakob Grue Simonsen, and Isabelle Augenstein. 2023. Faithfulness tests for natural language explanations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 283–294, Toronto, Canada. Association for Computational Linguistics.

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.

Giuseppe Attanasio, Debora Nozza, Eliana Pastor, Dirk Hovy, et al. 2022. Benchmarking post-hoc interpretability approaches for transformer-based misogyny detection. In *Proceedings of NLP Power! The First Workshop on Efficient Benchmarking in NLP*. Association for Computational Linguistics.

Giuseppe Attanasio, Eliana Pastor, Chiara Di Bonaventura, and Debora Nozza. 2023. ferret: a framework for benchmarking explainers on transformers. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 256–266, Dubrovnik, Croatia. Association for Computational Linguistics.

Housam KB Babiker, Mi-Young Kim, and Randy Goebel. 2023. From intermediate representations to explanations: Exploring hierarchical structures in nlp. In *ECAI 2023*, pages 157–164. IOS Press.

Jasmijn Bastings, Sebastian Ebert, Polina Zablotskaia, Anders Sandholm, and Katja Filippova. 2022. "will you find these shortcuts?" a protocol for evaluating the faithfulness of input salience methods for text classification. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 976–991, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Leonard Bongard, Lena Held, and Ivan Habernal. 2022. The legal argument reasoning task in civil procedure. In *Proceedings of the Natural Legal Language Processing Workshop 2022*, pages 194–207.

Oana-Maria Camburu, Eleonora Giunchiglia, Jakob Foerster, Thomas Lukasiewicz, and Phil Blunsom. 2019. Can i trust the explainer? verifying post-hoc explanatory methods. *arXiv preprint arXiv:1910.02065*.

Samuel Carton, Anirudh Rathore, and Chenhao Tan. 2020. Evaluating and characterizing human rationales. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9294–9307, Online. Association for Computational Linguistics.

Danqi Chen and Christopher D Manning. 2014. A fast and accurate dependency parser using neural networks. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 740–750.

Shivani Choudhary, Niladri Chatterjee, and Subir Kumar Saha. 2022. Interpretation of black box nlp models: A survey. *arXiv preprint arXiv:2203.17081*.

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.

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.

Sarthak Jain and Byron C Wallace. 2019. Attention is not explanation. 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 3543–3556.

Sérgio Jesus, Catarina Belém, Vladimir Balayan, João Bento, Pedro Saleiro, Pedro Bizarro, and João Gama. 2021. How can i choose an explainer? an application-grounded evaluation of post-hoc explanations. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 805–815.

Jaap Jumelet and Willem Zuidema. 2023. Feature interactions reveal linguistic structure in language models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8697–8712, Toronto, Canada. Association for Computational Linguistics.

Jonathan Kamp, Lisa Beinborn, and Antske Fokkens. 2023. Dynamic top-k estimation consolidates disagreement between feature attribution methods. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6190–6197, Singapore. Association for Computational Linguistics.

Nikita Kitaev, Steven Cao, and Dan Klein. 2019. Multilingual constituency parsing with self-attention and pre-training. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3499–3505, Florence, Italy. Association for Computational Linguistics.

Satyapriya Krishna, Tessa Han, Alex Gu, Javin Pombra, Shahin Jabbari, Steven Wu, and Himabindu Lakkaraju. 2022. The disagreement problem in explainable machine learning: A practitioner's perspective. *arXiv preprint arXiv:2202.01602*.

Vivian Lai, Zheng Cai, and Chenhao Tan. 2019. Many faces of feature importance: Comparing built-in and post-hoc feature importance in text classification. 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 486–495.

Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.

Andreas Madsen, Siva Reddy, and Sarath Chandar. 2022. Post-hoc interpretability for neural nlp: A survey. *ACM Computing Surveys*, 55(8):1–42.

Khalil Mrini, Franck Dernoncourt, Quan Hung Tran, Trung Bui, Walter Chang, and Ndapandula Nakashole. 2020. Rethinking self-attention: Towards interpretability in neural parsing. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 731–742.

Michael Neely, Stefan F Schouten, Maurits Bleeker, and Ana Lucic. 2022. A song of (dis) agreement: Evaluating the evaluation of explainable artificial intelligence in natural language processing. In *HHAI2022: Augmenting Human Intellect*, pages 60–78. IOS Press.

Craig Pirie, Nirmalie Wiratunga, Anjana Wijekoon, and Carlos Francisco Moreno-Garcia. 2023. AGREE: a feature attribution aggregation framework to address explainer disagreements with alignment metrics. In *Proceedings of the Workshops at the 31st International Conference on Case-Based Reasoning (ICCBR-WS 2023)*, pages 184–199. CEUR.

Karl Popper. 2005. *The logic of scientific discovery*. Routledge.

Danish Pruthi, Rachit Bansal, Bhuwan Dhingra, Livio Baldini Soares, Michael Collins, Zachary C Lipton, Graham Neubig, and William Cohen. 2022. Evaluating explanations: How much do explanations from the teacher aid students? *Transactions of the Association for Computational Linguistics*, 10:359–375.

Sahana Ramnath, Preksha Nema, Deep Sahni, and Mitesh M Khapra. 2020. Towards interpreting bert for reading comprehension based qa. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3236–3242.

Tilman Räuker, Anson Ho, Stephen Casper, and Dylan Hadfield-Menell. 2023. Toward transparent ai: A survey on interpreting the inner structures of deep neural networks. In *2023 IEEE Conference*

on Secure and Trustworthy Machine Learning (SaTML), pages 464–483. IEEE.

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.

Saumendu Roy, Gabriel Laberge, Banani Roy, Foutse Khomh, Amin Nikanjam, and Saikat Mondal. 2022. Why don't xai techniques agree? characterizing the disagreements between post-hoc explanations of defect predictions. In *2022 IEEE International Conference on Software Maintenance and Evolution (ICSME)*, pages 444–448. IEEE.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. 2017. Learning important features through propagating activation differences. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 3145–3153. JMLR.org.

Sandipan Sikdar, Parantapa Bhattacharya, and Kieran Heese. 2021. Integrated directional gradients: Feature interaction attribution for neural nlp models. 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 865–878.

K Simonyan, A Vedaldi, and A Zisserman. 2014. Deep inside convolutional networks: visualising image classification models and saliency maps. In *Proceedings of the International Conference on Learning Representations (ICLR)*. ICLR.

Rui Song, Fausto Giunchiglia, Yingji Li, and Hao Xu. 2023. Automatic counterfactual augmentation for robust text classification based on word-group search. *arXiv preprint arXiv:2307.01214*.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR.

Umar Taufiq, Reza Pulungan, and Yohanes Suyanto. 2023. Named entity recognition and dependency parsing for better concept extraction in summary obfuscation detection. *Expert Systems with Applications*, 217:119579.

Junru Zhou, Zuchao Li, and Hai Zhao. 2020. Parsing all: Syntax and semantics, dependencies and spans. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4438–4449, Online. Association for Computational Linguistics.

10. Language Resource References

Camburu, Oana-Maria and Rocktäschel, Tim and Lukasiewicz, Thomas and Blunsom, Phil. 2018. *e-snli: Natural language inference with natural language explanations*. GitHub Repository without PID/islrn: https://github.com/OanaMariaCamburu/e-SNLI.

A. Appendix

The baseline tests for importance thresholds $\mu > 0$ and *median* > 0 described in §4.3 are given in Table 3 and Table 4, respectively. The averaged Euclidean distances that led to selecting these thresholds (§4.3) are reported in Table 5. In Table 6, we find the results of the Chi-Square tests adopted within the linguistic analysis in §3.2.

	Token	**Span**
Method	*BL:minAgr–maxAgr*	
PartSHAP	0.55:0.55–0.81	0.60:0.61–0.83
LIME	0.55:0.55–0.81	0.60:0.61–0.83
VanGrad	0.56:0.58–0.68	0.64:0.64–0.72
Grad×I	0.55:0.55–0.58	0.60:0.60–0.65
IntGrad	0.55:0.55–0.58	0.59:0.61–0.64
IntGrad×I	0.55:0.57–0.65	0.61:0.61–0.69

Table 3: Token and span agr. with other methods (range *minAgr* to *maxAgr*) versus baseline (*BL*), for thresh. $= \mu > 0$. Scores $<$ baseline in **bold**.

	Token	**Span**
Method	*BL:minAgr–maxAgr*	
PartSHAP	0.57:**0.56**–0.76	0.65:**0.63**–0.80
LIME	0.57:**0.56**–0.76	0.65:**0.63**–0.80
VanGrad	0.59:**0.58**–0.68	0.69:**0.66**–0.74
Grad×I	0.56:0.56–0.59	0.63:**0.62**–0.67
IntGrad	0.56:0.56–0.58	0.62:0.62–0.66
IntGrad×I	0.57:0.57–0.64	0.65:0.66–0.74

Table 4: Token and span agr. with other methods (range *minAgr* to *maxAgr*) vs. baseline (*BL*), for thresh. $=$ *median* > 0. Scores $<$ baseline in **bold**.

	μ	$\mu+\sigma$	$\mu+2\sigma$	$\mu-\sigma$	$\mu-2\sigma$	*median*
all	12.286	15.200	22.782	24.165	24.342	17.596
>0	9.082	17.893	23.354	19.756	23.020	10.265

Table 5: The averaged Euclidean distances between the methods' mean±stdev values for each threshold, and human preference (4±3). We analyse further the three thresholds visually indicated with a dark background that have nearest distance to human preference.

	Stop words			*Punctuation*			*POS*		
Comparison	χ^2	p	df	χ^2	p	df	χ^2	p	df
PartSHAP vs LIME	0.0	1.0	1	0.0	1.0	1	0.0	1.0	4
PartSHAP vs VanGrad	1.247	0.264	1	0.255	0.614	1	2.580	0.630	4
PartSHAP vs Grad×I	7.642	0.006*	1	3.763	0.052	1	10.361	0.035*	4
PartSHAP vs IntGrad	6.155	0.013*	1	2.216	0.137	1	9.578	0.048*	4
PartSHAP vs IntGrad×I	3.611	0.057	1	2.962	0.085	1	5.219	0.266	4
PartSHAP vs Human	0.0	1.0	1	0.255	0.614	1	0.117	0.998	4
LIME vs VanGrad	0.886	0.347	1	0.820	0.365	1	2.580	0.630	4
LIME vs Grad×I	8.595	0.003*	1	2.595	0.107	1	10.361	0.035*	4
LIME vs IntGrad	7.018	0.008*	1	1.316	0.251	1	9.578	0.048*	4
LIME vs IntGrad×I	4.287	0.038*	1	1.920	0.166	1	5.219	0.266	4
LIME vs Human	0.0	1.0	1	0.820	0.365	1	0.117	0.998	4
VanGrad vs Grad×I	15.855	<0.001*	1	7.181	0.007*	1	20.485	<0.001*	4
VanGrad vs IntGrad	13.747	<0.001*	1	5.158	0.023*	1	19.476	<0.001*	4
VanGrad vs IntGrad×I	9.896	0.002*	1	6.157	0.013*	1	13.148	0.011*	4
VanGrad vs Human	0.886	0.347	1	0.0	1.0	1	2.635	0.621	4
Grad×I vs IntGrad	0.021	0.885	1	0.056	0.814	1	0.095	0.999	4
Grad×I vs IntGrad×I	0.536	0.464	1	0.0	1.0	1	1.544	0.819	4
Grad×I vs Human	8.595	0.003*	1	7.181	0.007*	1	10.212	0.037*	4
IntGrad vs IntGrad×I	0.195	0.659	1	0.0	1.0	1	1.242	0.871	4
IntGrad vs Human	7.018	0.008*	1	5.158	0.023*	1	9.381	0.052	4
IntGrad×I vs Human	4.287	0.038*	1	6.157	0.013*	1	4.876	0.300	4

Table 6: Chi-Square test results for comparing different methods on their preference for stop words, punctuation and POS. Asterisk (*) indicates statistical significance at the 0.05 level. A dark background visually highlights the hypothesised Group 1 – Group 2 comparisons.