Are Emotion Models Right for the Right Reasons?

Anonymous ACL submission

Abstract

Emotion classification models can predict the correct label while relying on spurious cues rather than the true emotion cause. For instance, a model that correctly predicts *disgust* for "a child was sent to my childrens school with covid" may attend to "a child" rather than "with covid," getting the right answer for the wrong reason, since "a child" is not what indicates *disgust*. We investigate how often a fine-tuned BERT emotion classifier produces correct predictions for the wrong reasons. The study requires two ingredients: phrase-level SHAP explanations to expose what the model attends to, and human judgment to decide whether those features match the true emotion cue. We carry this investigation out on 2,673 correctly predicted samples from a 15-class emotion corpus. Four annotators each rate a disjoint subset of approximately 585 SHAP attributions, with an additional 100 shared samples used to measure inter-annotator agreement. Annotators judge 12.7% of these correct predictions as made for the wrong reason at the phrase level. We show that automatically classifying whether the emotion predictor outputs the correct class for the right reasons is possible with F1 of 0.7. Such automation will enable large-scale analysis and model improvements in the future.

1 Introduction

An emotion classifier that correctly predicts *disgust* for "a child was sent to my childrens school with covid" may do so because it attends to "with covid" or because it picks up on "a child." The prediction is correct either way, but only in the first case does the model rely on the actual emotion cue. In the first case, the model is *right for the right reasons* (Ross et al., 2017; McCoy et al., 2019). In the second, it is *right for the wrong reasons*. This may happen because the training data contains spurious correlations between phrases like "a child" and the label *disgust*, which the model picks up as a shortcut. When models produce correct labels from the wrong evidence, they appear reliable on existing data but can fail unpredictably when those spurious cues are absent.

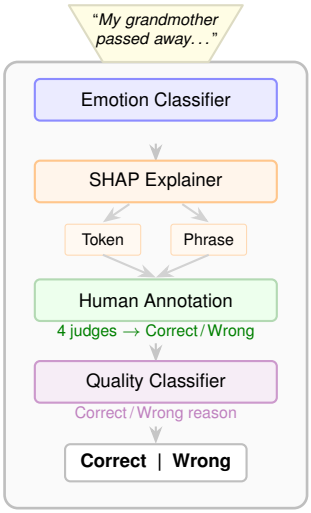

Figure 1: Experimental setting. A sentence enters the system, is classified by the emotion model, explained by SHAP at token and phrase level, reviewed by human annotators as right for the right or wrong reasons, and this judgment is automated by a quality classifier.

This concern is well documented in other areas of natural language processing (NLP). Geirhos et al. (2020) show that deep networks frequently learn shortcut strategies, and McCoy et al. (2019) find that natural language inference (NLI) models rely on lexical overlap heuristics rather than genuine reasoning. Ross et al. (2017) formalize the principle that models should be "right for the right reasons." But despite this growing awareness, there is little empirical work that measures how often a model is actually right for the right or wrong reasons on a specific task. A few studies have done this for specific tasks: Mathew et al. (2021) for hate speech detection. DeYoung et al. (2020) introduce a benchmark for evaluating whether model explanations match human-annotated rationale spans across seven NLP tasks; this is the closest large-

scale precedent, although they do not frame it as right-for-the-right-reasons. For emotion classification, however, we are not aware of any study that systematically evaluates whether the model's reasoning aligns with human judgment about the emotion cause.

We use SHapley Additive exPlanations (SHAP) (Lundberg and Lee, 2017) and human annotation to answer three questions:

1. **RQ1:** How often are emotion models right for the right reasons?

2. **RQ2:** Can the distinction between correct and wrong reasons be automated?

3. **RQ3:** Does reasoning quality vary across emotions, and what causes the variation?

Our approach has three components (Figure 1). We generate SHAP explanations for correctly predicted samples from a 15-class emotion corpus. Four human annotators then judge whether the highlighted words reflect the genuine emotion cause or spurious cues. Finally, we train a BERT-based (Devlin et al., 2019) classifier to automate this judgment.

Our contributions are:

1. A large-scale annotation study covering 2,338 correct predictions, with four annotators evaluating whether the SHAP explanations behind them reflect the genuine emotion cause or rely on spurious cues, finding that 12.7% of correct predictions are made for the wrong reason.

2. Evidence that phrase-level SHAP explanations are more reliable to evaluate than token level ones for emotion: only 3.1% of phrase-level explanations are marked unclear by annotators (73 of 2,338), compared to 4.8% at the token level (113 of 2,338).

3. An automatic classifier that distinguishes correct-reason from wrong-reason predictions with macro F1 of 0.70 and Cohen's $\kappa = 0.39$, within the range of human pairwise agreement (0.30 to 0.42).

2 Related Work

This section reviews four lines of work that frame our study. We first survey post-hoc explanation methods (Section 2.1), since SHAP is the tool we use to expose what the model attends to. We then turn to shortcut learning and right-for-the-right-reasons evaluation (Section 2.2), the broader phenomenon our work measures for emotion classification. Section 2.3 reviews how explanation quality has been evaluated in NLP, and Section 2.4 narrows to explainability work in emotion classification specifically, where we identify the gap our paper addresses.

2.1 Post-Hoc Explanation Methods

Post-hoc explanation methods assign importance scores to input features after a model has been trained. The two most widely used perturbation-based methods are Local Interpretable Model-agnostic Explanations (LIME, Ribeiro et al., 2016) and SHapley Additive exPlanations (SHAP, Lundberg and Lee, 2017). LIME fits a local linear surrogate by perturbing the input and observing output changes. SHAP computes Shapley values from cooperative game theory, satisfying efficiency, symmetry, dummy, and additivity (we give the formal definition in Section 3.1). Kokalj et al. (2021) adapt SHAP to transformer-based classifiers, and Mosca et al. (2022) survey SHAP variants for NLP and discuss their limitations.

Gradient-based methods offer a different family of explanations. Integrated Gradients (IG, Sundararajan et al., 2017) accumulates gradients along a path from a baseline input to the actual input. While computationally efficient, gradient methods operate at the token level and do not naturally capture phrase-level semantics.

Attention weights from transformers (Vaswani et al., 2017) are sometimes treated as explanations, but whether they constitute faithful explanations remains debated (Jain and Wallace, 2019; Wiegreffe and Pinter, 2019). Bastings and Filippova (2020) argue that saliency methods should be preferred when the goal is to identify which tokens drive a prediction.

A separate line of work targets phrase-level explanations. Singh et al. (2019) propose contextual decomposition, attributing predictions to phrases by decomposing hidden states. Chen et al. (2020) detect feature interactions to build hierarchical explanations. Most recently, Amara et al. (2024) introduce SyntaxSHAP, which constrains Shapley value coalitions using dependency trees for text generation. Our phrase-level approach is related but differs in two ways: we define groups through a syntactic parsing algorithm rather than constraining coalitions, and we target classification rather

than generation. We motivate why these existing phrase-level approaches do not fit our setting in Section 3.1.

2.2 Shortcut Learning and Right for the Right Reasons

Deep networks frequently learn shortcut strategies that exploit superficial patterns rather than the intended decision rule (Geirhos et al., 2020). In NLP, McCoy et al. (2019) show that natural language inference models rely on heuristics like lexical overlap, achieving high accuracy when these heuristics happen to work but failing on adversarial examples. Ross et al. (2017) propose gradient-based penalties to train models that are "right for the right reasons."

The right-for-the-right-reasons question has received more sustained attention outside NLP. Schramowski et al. (2020) train plant disease classifiers that rely on biologically meaningful regions of the leaves rather than artifact backgrounds, providing a textbook example of using explanations to enforce correct reasoning during training. NLP has been comparatively slow to take up this evaluation question.

These studies diagnose the problem or propose training fixes. Our work takes a different angle: we directly measure how often an existing model's explanations align with human judgment, and we build a classifier to automate this measurement. To our knowledge, no prior work has applied this kind of systematic right-for-the-right-reasons evaluation to emotion classification.

2.3 Evaluating Explanation Quality

Evaluating whether an explanation is faithful to the model's reasoning or merely plausible to humans remains an open challenge (Lyu et al., 2024). DeYoung et al. (2020) introduce the ERASER benchmark with human-annotated rationales for several NLP tasks. Atanasova et al. (2020) compare LIME, SHAP, Integrated Gradients, and attention on text classification and find that no single method dominates across all evaluation metrics.

Human evaluation of explanations has been explored for specific tasks. Camburu et al. (2018) collect natural language explanations for NLI, focusing on free-text justifications rather than feature attributions. Mathew et al. (2021) annotate hate speech data with rationale spans and find that models with high classification accuracy do not necessarily produce faithful explanations. Our work follows this line but focuses on emotion classification: rather than collecting rationales, we ask annotators whether SHAP-highlighted features correspond to the genuine emotion cause.

2.4 Explainability for Emotion Classification

Emotion classification has advanced substantially with transformer-based models (Devlin et al., 2019), and several large-scale corpora now exist: GoEmotions (Demszky et al., 2020) with 58K Reddit comments, MELD (Poria et al., 2019) with multimodal dialogue data, and ISEAR (Scherer and Wallbott, 1994) with cross-cultural self-reports. Plaza-del Arco et al. (2024) survey 154 emotion analysis publications and identify several gaps, including limited work on model interpretability. The crowd-enVent corpus (Troiano et al., 2023) was built on appraisal theories of emotion: crowd workers described real events that trigger specific emotions, and a separate group of readers reconstructed the emotion from the text alone. Because participants describe the situation without naming the emotion, the corpus is well suited for studying whether models attend to genuine emotion triggers rather than emotion keywords. Our study uses a unified corpus that combines the Super Emotion Dataset (de Fortuny, 2025) (an aggregation of the corpora above) with crowd-enVent.

Explainability methods have been applied to many NLP tasks, but rarely to emotion classification specifically. Kumar and Raman (2022) propose a dual-channel BERT system for emotion recognition that uses attention visualization to explain predictions, but they do not evaluate whether the explanations reflect the actual emotion cause. Large language models can produce free-text rationales for their predictions, but whether those rationales reflect the actual reasoning process is itself an open faithfulness question, and one that has received limited attention for emotion classification.

3 Method

To answer whether emotion classifiers are right for the right reasons, we need three ingredients: explanations that expose what the model attends to, human judgments that decide whether those features match the genuine emotion cause, and a way to scale this assessment beyond a hand-annotated set. This section describes our approach to each. Section 3.1 introduces phrase-level SHAP, which extends standard token-level Shapley value attribu-

tions to syntactic phrase groups so that multi-word emotion cues are scored as units rather than fragmented across subwords. Section 3.2 describes the annotation study in which four annotators judge token-level and phrase-level explanations side by side, producing the labels we use to characterize how often the model is right for the right reasons (RQ1). Section 3.3 introduces a family of BERT-based explanation quality classifiers that learn to automate this judgment, addressing whether the distinction between correct and wrong reasons can be predicted from the SHAP attributions themselves (RQ2).

3.1 Phrase-Level SHAP

SHAP computes each feature's contribution by observing how the model's prediction changes when that feature is absent, so the attribution scores reflect the model's own behavior rather than an external judgment. Standard token-level SHAP treats each word as an independent feature, which fragments multi-word expressions. The idiom "get under my skin" (meaning *to irritate*) would receive separate attributions for "get," "under," "my," and "skin." Individually, none of these words explains *anger*; only together do they form a meaningful expression. This fragmentation can lead annotators to judge an explanation as wrong reason even though the model may actually attend to the right phrase. Token-level evaluation can thus be unfair to a model that reasons correctly at the phrase level but has no way to express this when each word is scored in isolation.

We address this problem by grouping syntactically related tokens into phrase-level features before computing Shapley values. Given a sentence with M tokens, we define a set of phrase groups $\mathcal{G} = \{g_1, \ldots, g_G\}$ where $G \leq M$ is the number of groups. The coalition vector $z' \in \{0,1\}^G$ controls groups rather than individual tokens: all tokens within a group are masked or unmasked together. The Shapley value for group g_j is:

$$\phi_j^{\text{rel}} = \sum_{S \subseteq \mathcal{G} \setminus \{g_j\}} \frac{|S|!\,(G - |S| - 1)!}{G!} \cdot [f(S \cup \{g_j\}) - f(S)] \quad (1)$$

Because the number of features is reduced from M to G, the KernelSHAP approximation also becomes more accurate for the same number of perturbations.

As a concrete example, for the sentence "people get under my skin like on almost a daily occurrence" (predicted: *anger*), token-level SHAP produces scattered values: "get" = 0.026, "under" = 0.030, "my" = 0.028, "skin" = 0.038. Relational SHAP groups the idiom as a single feature and yields "get under my skin" = 0.102, giving a coherent attribution for *anger*.

Syntactic Phrase Grouping. Phrase groups are obtained from the dependency parse: manually annotated on the first 500 samples and produced by an automated dependency-based parser on the remaining 2,361. Appendix K gives the five-phase algorithm, the running example, and a comparison with four alternative parsers.

Existing phrase-related approaches do not fit our setting: contextual decomposition (Singh et al., 2019) is built for recurrent and convolutional networks, hierarchical interaction grouping (Chen et al., 2020) produces groupings that need not align with linguistic structure, and SyntaxSHAP (Amara et al., 2024) keeps features at the token level while constraining coalitions for text generation. We instead redefine the feature space so that syntactic phrases are the units being scored, which is what makes phrase-level annotation feasible.

Throughout the remainder of the paper, we treat the phrase-level SHAP annotations as the gold-standard target for evaluating explanation quality, since the phrase-level representation aligns more reliably with how annotators identify the emotion cause (Section 5.1).

3.2 Human Annotation

We recruit four annotators (one Business Informatics student and three Digital Humanities students) to judge whether SHAP explanations reflect the correct reason for the model's emotion prediction. They are employed on a German student mini-job basis at eight hours per week (federal minimum wage), are aged 19–24, and have high English proficiency but no strong background in emotion analysis. Before beginning, they were trained through multiple live sessions with one of the authors and through written guidelines.

A prediction is labeled right for the right reasons when the SHAP-highlighted features correspond to the genuine emotion trigger. For example, the model predicts *fear* for "I was having a panic attack for the first time and thought I would always feel that way," with phrase-level SHAP assigning the

Metric	Token	Phrase
Fleiss' κ	0.45	0.37
Avg. pairwise Cohen's κ	0.45	0.37
Unanimous agreement	56%	70%
Majority agreement	81%	87%

Table 1: Inter-annotator agreement on 100 shared samples (4 annotators). Both SHAP types show moderate agreement (Landis and Koch, 1977).

highest score to "was having a panic attack" (0.61). If the same model had instead highlighted "for the first time," it would produce the correct label from the wrong evidence.

Each annotator is assigned a dedicated subset of approximately 585 samples, and an additional 100 samples are shared among all four annotators to measure inter-annotator agreement. The annotators reach a Fleiss' κ (Fleiss, 1971) of 0.447 for token-level SHAP and 0.366 for phrase-level SHAP on these shared samples, both corresponding to moderate agreement on the Landis and Koch (1977) scale (Table 1). Unanimous agreement (all four annotators give the same label) is 56% for token-level and 70% for phrase-level SHAP. Majority agreement (at least three of the four annotators give the same label) is 81% for token-level and 87% for phrase-level SHAP. The lower κ for phrase-level SHAP despite higher unanimity reflects the skewed distribution: since 84.2% of phrase-level explanations are judged as correct reason, chance agreement is higher, which compresses the κ statistic.

In total, 2,338 of 2,861 samples are annotated (81.7%). After removing samples labeled as Unclear, 2,145 usable annotations remain. Further details about the annotation tool and interface are in Appendix L. As noted in Section 3.1, the phrase-level annotations serve as the gold standard for the subsequent evaluations.

3.3 Explanation Quality Classifier

We train BERT-based binary classifiers that predict whether the emotion model's SHAP explanation reflects the correct or wrong reason. We compare four input variants for the classifier. The **Baseline** variant conditions on only the sentence text and the predicted emotion label, without any SHAP values; it tests whether the sentence alone is sufficient to predict reasoning quality. The **SHAP-Token** variant additionally provides the token-level SHAP attributions to the classifier. The input concatenates the original sentence with the predicted emotion label, the model's confidence, and the list of input tokens paired with their SHAP scores, sorted by absolute attribution and filtered at $|\phi| > 0.001$. The **SHAP-Phrase** variant uses the same input format but with phrase-level SHAP values. The **SHAP-Joint** variant concatenates both the token-level and phrase-level SHAP representations. All four variants are trained against the same target, the phrase-level annotation label, so differences across variants isolate the contribution of the input representation.

Train/Test Splits. All four variants share the same 2,145 samples (1,716 train, 429 test) after excluding samples labeled Unclear, using one stratified 80/20 split applied identically across variants. We use 5-fold stratified cross-validation on the training portion for model selection and report results on the held-out test set.

4 Experimental Setup

4.1 Corpus and Emotion Model

We construct a unified emotion corpus by combining the Super Emotion Dataset (de Fortuny, 2025) (which aggregates six corpora: GoEmotions (Demszky et al., 2020), MELD (Poria et al., 2019), ISEAR (Scherer and Wallbott, 1994), CrowdFlower (CrowdFlower, 2016), SemEval Affect (Mohammad et al., 2018), and Twitter Emotion (Mohammad, 2012)) with the crowd-enVent corpus (Troiano et al., 2023). Labels are harmonized to a 15-class taxonomy. After filtering for a minimum length of 15 words and removing emoji-heavy texts, 34,732 samples remain.[1]

We fine-tune `bert-base-uncased` (Devlin et al., 2019) (110M parameters) on these samples for 15-class emotion classification (learning rate 2×10^{-5}, batch size 16, early stopping at epoch 3, stratified split: 26,049 train, 3,474 val, 5,209 test). Table 2 shows test performance: 81.6% accuracy and 0.81 weighted F1. Performance varies across classes, with frequent emotions like *sadness* (F1 = 0.91) and *joy* (F1 = 0.89) predicted reliably and rare categories like *guilt* and *shame* poorly recognized.

For the SHAP analysis, we restrict to crowd-enVent samples where the model's prediction matches the ground-truth label, yielding 2,673 correctly predicted samples. Because crowd-enVent participants describe situations rather than stating

[1]The annotation data, trained model weights, and analysis code will be released upon paper acceptance.

Emotion	Support	Prec.	Rec.	F1
sadness	1,195	.94	.88	.91
joy	1,544	.89	.88	.89
anger	756	.81	.85	.83
fear	558	.76	.82	.79
love	506	.74	.77	.76
surprise	243	.65	.75	.70
boredom	35	.56	.71	.63
trust	41	.50	.63	.56
neutral	157	.47	.45	.46
pride	33	.48	.39	.43
relief	37	.36	.38	.37
disgust	37	.28	.41	.33
Macro avg	5,209	.50	.53	.51
Weighted avg	5,209	.81	.82	.81

Table 2: Emotion classifier test performance, restricted to classes with at least 30 test samples. Test accuracy = 81.6%. Macro and weighted averages are computed across all 15 classes. Full per-class results in Appendix B.

the emotion explicitly, the model must rely on contextual cues. The per-emotion sample counts are listed in Appendix A (Table 8). SHAP explanations are computed using KernelSHAP with 500 perturbations per sample, with absent tokens replaced by the `[MASK]` embedding.

4.2 Classifier Training

All classifier variants use `bert-base-uncased` as the backbone, trained with learning rate 2×10^{-5}, batch size 16, maximum sequence length 256, weighted cross-entropy loss, and early stopping with patience 3. We use 5-fold stratified cross-validation for model selection and report results on a held-out test set. The optimizer is AdamW (Loshchilov and Hutter, 2019) with 10% linear warmup. All training runs on a single NVIDIA RTX 4090; each variant completes in approximately 5–8 minutes including 5-fold cross-validation.

5 Results

5.1 How Often Are Emotion Models Right for the Right Reasons?

To answer this question, four annotators independently judged 2,338 correctly predicted samples. For each sample, they saw the original sentence alongside both token-level and phrase-level SHAP explanations and decided whether the highlighted features reflect the genuine emotion cause (right for the right reasons) or whether the model relies on incidental cues (right for the wrong reasons).

	Token-level		Phrase-level	
Label	**N**	**%**	**N**	**%**
Correct reason	1,605	68.6	1,969	84.2
Wrong reason	620	26.5	296	12.7
Unclear	113	4.8	73	3.1
Total	**2,338**		**2,338**	

Table 3: Annotation distributions for token-level and phrase-level SHAP explanations. Four annotators judged whether highlighted features correspond to the genuine emotion cause (correct reason) or incidental cues (wrong reason).

Table 3 shows the results for token-level SHAP. Annotators judged 68.6% of explanations as right for the right reasons, 26.5% as right for the wrong reasons, and 4.8% as Unclear. When tokens are grouped into phrases, the Unclear rate is 3.1% (73 samples), indicating that phrase-level SHAP is a clearer method for annotators to judge whether the model decided for the right or the wrong reasons. Because phrase-level judgments are more decisive and produce fewer Unclear cases, we adopt the phrase-level annotation as the gold-standard label against which the explanation quality classifiers in Section 5.2 are trained and evaluated.

Why Do Wrong-Reason Predictions Occur? Some wrong-reason predictions occur because the model relies on spurious cues rather than on the genuine cause of the emotion. For example, the *disgust* prediction on "a child was sent to my childrens school with covid" assigns 0.083 to the phrase "a child" and ranks the emotion-bearing phrase "with covid" only fourth.

5.2 Can the Distinction Between Correct and Wrong Reasons Be Automated?

A substantial share of the SHAP explanations our annotators reviewed in Section 5.1 fall into the wrong-reason category. We ask whether a classifier can learn to make this distinction automatically. If it can, the same judgment can be applied at scale to predictions we have not annotated by hand, surfacing where and on which cues the model decides for the wrong reason. Those patterns are the input to any later intervention, whether through data, training objective, or model engineering. We trained four BERT-based classifier variants that receive increasingly rich input representations (Section 3.3). All four variants are trained and evaluated against the same target: the phrase-level annotation label

Model	Test F1
Baseline	0.63
SHAP-Token	0.61
SHAP-Phrase	0.68
SHAP-Joint	0.70

Table 4: Classifier comparison on the held-out test set (macro F1). Full per-class metrics in Appendix C.

that Section 5.1 identifies as the more reliable judgment signal. Differences across variants therefore isolate the contribution of the input representation, not of the target. The question is not just whether automation is possible, but how closely the classifier agrees with the human annotators on this inherently subjective task. Table 4 compares the four variants.

Phrase-level SHAP input is the load-bearing signal. The two variants that include phrase SHAP, SHAP-Phrase and SHAP-Joint, both reach 0.68–0.70 test macro F1, while Baseline and SHAP-Token reach only 0.63 and 0.61. SHAP-Joint is the strongest on the held-out test set with macro F1 = 0.70 and Cohen's κ = 0.39, while SHAP-Phrase is the cross-validation winner (CV F1 = 0.70 vs. 0.68 for SHAP-Joint). The 0.07 macro F1 gap between SHAP-Joint and the Baseline is modest in absolute terms but consistent with the finding of Section 5.1 that phrase-level grouping yields the more reliable judgment signal.

SHAP-Token on its own is not enough. With token-level SHAP as the only attribution input and phrase-level annotations as the target, SHAP-Token reaches 0.61 test macro F1, slightly below the Baseline at 0.63. Token attributions fragment multi-word causes into isolated pieces (Section 5.1), and the classifier cannot reconstruct the phrase the annotators based their judgment on. The same token attributions become useful only when paired with phrase-level SHAP in SHAP-Joint, where wrong-reason recall jumps from 0.36 to 0.59 and Cohen's κ rises from 0.22 to 0.39.

Comparison to Human Agreement. To put the classifier's performance in context, we treat each classifier as an additional annotator and compute Cohen's κ against the gold-standard labels. SHAP-Joint reaches $\kappa = 0.394$ and SHAP-Phrase $\kappa = 0.361$, while SHAP-Token reaches $\kappa = 0.216$ and the Baseline $\kappa = 0.259$. Among our four human annotators on the same gold standard, the six pairwise κ values range from 0.297 to 0.421 with a mean of 0.348 (Table 1). The two phrase-input classifiers fall within this range of human disagreement, with SHAP-Joint near the upper bound and SHAP-Phrase essentially at the human mean.

5.3 Does Reasoning Quality Vary Across Emotions?

The overall numbers hide substantial differences between emotion categories. Table 6 shows the SHAP-Joint classifier's per-emotion performance; the annotation-side correct-reason rate ranks the emotions in the same order (full breakdown in Appendix E). To explain that ranking we examined what the model attends to when annotators flag a wrong reason, both at the token and at the phrase level, and identified four recurring patterns of wrong-reason failure. Each pattern is concentrated in a small set of emotions. Table 5 shows one representative example per pattern; the full set is in Appendix F.

Sadness, *anger*, and *joy* sit at the top because the cause is typically carried by a distinctive word or short phrase. Words like "died," "funeral," "birthday," and "married" anchor the attribution; verb-object phrases of violation ("took my," "broke my") anchor *anger*. When a single lexical cue carries the emotion, both annotators and the classifier converge.

Trust, *pride*, and *relief* share a different failure mode. The wrong cue is most often a function word: the first-person pronoun "I" for self-referential *pride* and *trust* statements, the copulas "was" or "were" that begin most narratives, and intensifiers like "very". These emotions lack distinctive vocabulary, so the model attends to what is plentiful. The classifier's lower macro F1 on these three emotions in Table 6 reflects this thinness of usable cues.

Anger and *disgust* fail in a different mode: the model anchors on a real content word that names the situation but not its emotional cause. On *anger* this is most often a narrative noun (*people*, *coworker*, *government*) or a verb of action (*refusing*, *found out*); on *disgust* it is most often a context noun (*food*, *school*) that surrounds the trigger. The classifier reaches strong macro F1 on both emotions because the wrong cue is identifiable enough to learn from.

A third pattern is the absence of any clear anchor. On *joy*, *sadness*, and *fear* a notable share of the Wrong-Reason samples received no feature-level flag from annotators: the SHAP attribution was spread across the sentence with no peak that read

Pattern	Emotion	Sentence	Wrong phrase	SHAP
Function-word anchor	pride	"I won a scholarship."	*i*	0.13
Content phrase, not the cause	fear	"My friend collapsed in a night club."	*in a night club*	0.22
Diffuse (no anchor)	joy	"When my niece was born because she was the first baby born that was close to me."	—	—
Ellipsis artefact	surprise	"I felt . . . when my husband . . . me and flew to Portugal to meet me on holiday and propose."	"..."	0.22

Table 5: Right-for-wrong-reason (RWR) examples for each pattern at the phrase level. The Diffuse row has no specific wrong phrase because the SHAP attribution is spread across the sentence without a peak. Full set of examples in Appendix F.

Emotion	N	W	Acc.	F1	RWR
sadness	34	1	.97	.83	.67
anger	58	4	.97	.82	.67
joy	26	3	.92	.81	.67
disgust	37	6	.87	.77	.62
pride	55	10	.84	.74	.57
boredom	27	6	.78	.71	.57
fear	70	10	.81	.60	.32
relief	40	6	.68	.60	.44
trust	36	7	.64	.60	.48
surprise	45	4	.87	.59	.25

Table 6: SHAP-Joint classifier performance per emotion. N = test samples, W = wrong-reason count, RWR = wrong-reason class F1. Sorted by macro F1.

as the wrong cue. The model is wrong by omission here rather than by anchoring on the wrong word. This shows up as high accuracy but uneven wrong-reason F1 on these emotions in Table 6. Two corpus artefacts, the template word *felt* and the "..." ellipsis that follows it, recur across emotions and contribute additional wrong cues that are not specific to any single category.

6 Conclusion and Future Work

This paper investigated whether emotion classification models make correct predictions for the correct reasons. Four annotators judged 2,338 correctly predicted samples against token-level and phrase-level SHAP explanations; phrase-level explanations were clearer to judge and we adopted them as the gold standard. The annotation study and the explanation-quality classifier converge on the same picture: most correct emotion predictions are made for the genuine emotion cue, a minority are not, and the wrong-reason cases are patterned rather than random. The shortcut takes a form specific to emotion classification: function-word anchors and writing-prompt artefacts rather than the lexical overlap heuristics observed in NLI.

Feature granularity matters more than feature richness: SHAP-Joint and SHAP-Phrase outperform SHAP-Token because human judgment operates at the phrase rather than at the token. When the explanation fragments a multi-word cue across its parts, the classifier loses the unit the gold label refers to. The clearest evidence is that SHAP-Token falls below the Baseline that has no SHAP input at all: token attribution scored against a phrase target misleads the classifier rather than helping it. Adding token information on top of phrase information in SHAP-Joint yields only a marginal further gain, confirming that phrase structure rather than token detail is the load-bearing signal. The SHAP-Joint classifier reaches macro F1 = 0.70 and Cohen's κ = 0.39, within the range of pairwise agreement between the four annotators (0.30 to 0.42). Our approach turns *is this model right for the right reasons?* from a manual annotation task into a scalable prediction problem.

Several directions follow from this work. The framework should be extended beyond a single BERT classifier to a range of architectures, and to large language models used through prompting, so that the right-for-the-right-reasons measurement becomes a property of the task rather than of one model. Scaling the explanation-quality classifier to label predictions at corpus scale would turn the wrong-reason patterns identified here into a recurring signal: which cues a model anchors on when it is wrong, which emotions concentrate the failure, and which corpora contribute template artefacts. Those patterns are the input to interventions. Once known at scale, training data can be rebalanced and learning objectives can target the genuine emotion cue, raising right-for-the-right-reasons performance alongside accuracy. The case for that intervention is strongest in sensitive domains such as clinical decision support and mental health triage, where being correct for the right reason matters as much as being correct.

Limitations

Our study has several limitations. First, inter-annotator agreement is moderate (Fleiss' κ = 0.447 for token-level SHAP, 0.366 for phrase-level SHAP), measured on 100 shared samples. While majority agreement is high (81–87%), judging whether highlighted features constitute the genuine emotion cause is inherently subjective. Different annotators may reasonably disagree about what counts as the emotion trigger, especially for emotions expressed through ambiguous or context-dependent language. Second, the annotated dataset (2,145 usable samples after removing Unclear labels) is relatively small for training a binary classifier, which likely limits the achievable accuracy. Larger annotation efforts could improve classifier performance, especially for emotions with few wrong-reason examples such as *sadness* and *joy*. Third, our study covers a single fine-tuned BERT classifier. We do not extend the analysis to other architectures or to prompting-based LLMs, because each additional classifier produces its own SHAP attributions and would have required a fresh round of human annotation. Replicating the annotation effort for every model was beyond the scope of this work. Finally, KernelSHAP with 500 perturbations is an approximation; different random seeds may yield slightly different Shapley values, introducing noise into the explanations.

Ethics Statement

Data. We use seven existing emotion corpora. The crowd-enVent corpus (Troiano et al., 2023) was collected from consenting crowd workers who described events that triggered particular emotions. The six corpora aggregated in the Super Emotion Dataset (de Fortuny, 2025) are public research releases drawn from Reddit, TV dialogue, self-report, and Twitter sources. None contain personally identifiable information. The four annotators who labeled SHAP explanations were paid at the German federal minimum wage on a student mini-job contract and were informed about the research purpose.

Implications. The explanation-quality classifier is a diagnostic tool. It surfaces predictions whose attributions look spurious to human judges, not a guarantee that the model reasons correctly when the classifier flags it as right. We caution against deploying emotion classification models in high-stakes contexts such as mental health monitoring or hiring decisions without thorough validation of both prediction accuracy and explanation quality.

References

Kenza Amara, Rita Sevastjanova, and Mennatallah El-Assady. 2024. SyntaxShap: Syntax-aware explainability method for text generation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 4551–4566. 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*, pages 3256–3274. 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. Association for Computational Linguistics.

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-SNLI: Natural language inference with natural language explanations. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Hanjie Chen, Guangtao Zheng, and Yangfeng Ji. 2020. Generating hierarchical explanations on text classification via feature interaction detection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5578–5593. Association for Computational Linguistics.

CrowdFlower. 2016. Sentiment analysis: Emotion in text. Data for Everyone Library.

Enric Junqué de Fortuny. 2025. The super emotion dataset. *Preprint*, arXiv:2505.15348.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. GoEmotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054. Association for Computational Linguistics.

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. 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. Association for Computational Linguistics.

Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. 2020. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spACy: Industrial-strength natural language processing in Python.

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. Association for Computational Linguistics.

Enja Kokalj, Blaž Škrlj, Nada Lavrač, Senja Pollak, and Marko Robnik-Šikonja. 2021. BERT meets Shapley: Extending SHAP explanations to transformer-based classifiers. In *Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation*, pages 16–21. Association for Computational Linguistics.

Puneet Kumar and Balasubramanian Raman. 2022. A BERT based dual-channel explainable text emotion recognition system. *Neural Networks*, 150:392–407.

J. R. Landis and G. G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Qing Lyu, Marianna Apidianaki, and Chris Callison-Burch. 2024. Towards faithful model explanation in NLP: A survey. *Computational Linguistics*, 50(2):657–723.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. HateXplain: A benchmark dataset for explainable hate speech detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14867–14875.

R. Thomas McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448. Association for Computational Linguistics.

Saif Mohammad. 2012. #emotional tweets. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 246–255, Montréal, Canada. Association for Computational Linguistics.

Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. SemEval-2018 task 1: Affect in tweets. In *Proceedings of the 12th International Workshop on Semantic Evaluation*, pages 1–17.

Edoardo Mosca, Ferenc Szigeti, Stella Tragianni, Daniel Gallagher, and Georg Groh. 2022. SHAP-based explanation methods: A review for NLP interpretability. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4593–4603. International Committee on Computational Linguistics.

Flor Miriam Plaza-del Arco, Alba A. Cercas Curry, Amanda Cercas Curry, and Dirk Hovy. 2024. Emotion analysis in NLP: Trends, gaps and roadmap for future directions. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5696–5710. ELRA and ICCL.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 527–536. Association for Computational Linguistics.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144. Association for Computing Machinery.

Andrew Slavin Ross, Michael C. Hughes, and Finale Doshi-Velez. 2017. 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.

Klaus R. Scherer and Harald G. Wallbott. 1994. Evidence for universality and cultural variation of differential emotion response patterning. *Journal of Personality and Social Psychology*, 66(2):310–328.

Patrick Schramowski, Wolfgang Stammer, Stefano Teso, Anna Brugger, Franziska Herbert, Xiaoting Shao,

Hans-Georg Luigs, Anne-Katrin Mahlein, and Kristian Kersting. 2020. Making deep neural networks right for the right scientific reasons by interacting with their explanations. *Nature Machine Intelligence*, 2(8):476–486.

Chandan Singh, W. James Murdoch, and Bin Yu. 2019. Hierarchical interpretations for neural network predictions. In *International Conference on Learning Representations*.

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 of *Proceedings of Machine Learning Research*, pages 3319–3328. PMLR.

Enrica Troiano, Laura Oberländer, and Roman Klinger. 2023. Dimensional modeling of emotions in text with appraisal theories: Corpus creation, annotation reliability, and prediction. *Computational Linguistics*, 49(1):1–72.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

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. Association for Computational Linguistics.

A Corpus Details

Table 7 lists the six constituent sources of the Super Emotion Dataset.

Source	Domain	Size
GoEmotions	Reddit	58K
MELD	TV dialogues	13K
ISEAR	Self-reports	7K
CrowdFlower	Tweets	40K
SemEval Affect	Tweets	10K
Twitter Emotion	Hashtag-based	160K

Table 7: Source corpora in the Super Emotion Dataset (de Fortuny, 2025).

Table 8 shows the emotion distribution of the 2,673 correctly predicted crowd-enVent samples used for SHAP analysis and human annotation. Table 9 shows representative examples from the crowd-enVent corpus.

B Full Emotion Classifier Results

Table 10 shows the complete per-class results for the emotion classifier on the test set (5,209 samples).

Emotion	N	Emotion	N
fear	333	disgust	242
sadness	331	relief	191
anger	312	boredom	134
trust	302	no-emotion	7
surprise	281	guilt	1
joy	274		
pride	265		
Total			**2,673**

Table 8: Emotion distribution of correctly predicted crowd-enVent samples used for SHAP analysis and human annotation.

Label	Example
anger	"People get under my skin. For example, when an entitled customer demands to speak to my manager for a simple issue I could solve."
fear	"I was driving on the highway and someone cut me off and brake-checked me, almost causing a car accident."
sadness	"Someone misunderstood me and took it out of context."

Table 9: Examples from the crowd-enVent corpus (Troiano et al., 2023).

C Explanation Quality Classifier: Detailed and Cross-Target Metrics

Table 11 reports the full set of test-set metrics for the four classifier variants from Section 5.2, evaluated against the label corresponding to each variant's training target. Table 12 additionally reports a cross-target evaluation: the predictions of each SHAP variant scored against both the token-level and phrase-level annotation labels on the shared 429 test samples. No retraining is performed; only the gold label used for scoring changes.

D Inference on Unannotated and Unclear Samples

As an out-of-sample sanity check, we applied SHAP-Joint to the 528 samples that were excluded from the train/test pool because they were either unannotated or judged Unclear by all four annotators. The classifier predicts 84.8% correct reason and 15.2% wrong reason, closely matching the 13.6% wrong-reason rate in the annotated data. The match supports that the classifier's decisions on unseen samples are not systematically different from those on the training distribution. The classifier is more confident on correct-reason pre-

Emotion	Supp.	Prec.	Rec.	F1
sadness	1,195	.937	.878	.906
joy	1,544	.894	.883	.889
anger	756	.814	.849	.831
fear	558	.759	.824	.790
love	506	.744	.771	.757
surprise	243	.650	.749	.696
boredom	35	.556	.714	.625
trust	41	.500	.634	.559
neutral	157	.473	.446	.459
pride	33	.481	.394	.433
relief	37	.359	.378	.368
disgust	37	.283	.405	.333
guilt	19	.000	.000	.000
shame	18	.000	.000	.000
no-emotion	30	.000	.000	.000
Micro avg	5,209	.816	.816	.816
Macro avg	5,209	.497	.528	.510
Weighted avg	5,209	.812	.816	.813

Table 10: Full per-class emotion classifier test results. Classes with <20 support achieve near-zero performance due to insufficient training data. Micro F1 equals test accuracy.

Model	F1	$F1_W$	$F1_C$	P_W	R_W	κ
Baseline	0.65	0.49	0.81	0.49	0.48	0.29
SHAP-Token	0.74	0.63	0.85	0.61	0.66	0.49
SHAP-Phrase	0.73	0.53	0.92	0.52	0.55	0.45
SHAP-Joint	0.73	0.59	0.86	0.65	0.54	0.45

Table 11: Full test-set metrics on the native target. F1 = macro F1, $F1_W$ / $F1_C$ = wrong-reason / correct-reason F1, P_W / R_W = wrong-reason precision / recall, κ = Cohen's κ against the gold labels. Baseline, SHAP-Token, and SHAP-Joint are scored against token-level labels; SHAP-Phrase is scored against phrase-level labels.

dictions (mean probability 0.881) than on wrong-reason predictions (mean 0.812), consistent with the asymmetry on the test set.

E Per-Emotion Annotation Rates

Table 13 reports the phrase-level correct-reason rate per emotion from the annotation study. The ranking matches the SHAP-Joint classifier's per-emotion macro F1 in Table 6.

F Wrong-Reason Pattern Examples

Table 14 gives representative examples of the four wrong-reason patterns and the two corpus artefacts identified in Section 5.3. *Wrong feature* is the SHAP feature the annotator flagged as a wrong cue, and *SHAP* is the model's attribution to that feature.

Model	Axis	F1	$F1_W$	R_W	κ
SHAP-Token	token (native)	0.74	0.63	0.66	0.49
SHAP-Token	phrase (transfer)	0.63	0.43	0.64	0.29
SHAP-Phrase	phrase (native)	0.73	0.53	0.55	0.45
SHAP-Phrase	token (transfer)	0.66	0.47	0.37	0.34
SHAP-Joint	token (native)	0.73	0.59	0.54	0.45
SHAP-Joint	phrase (transfer)	0.65	0.43	0.55	0.31

Table 12: Cross-target evaluation on the shared 429 test samples. Each model's predictions are scored against both axes without recalibrating its decision threshold. F1 = macro F1, $F1_W$ = wrong-reason F1, R_W = wrong-reason recall, κ = Cohen's κ against the gold labels for that axis. Each model performs best on its native axis; the transfer drop reflects both label disagreement between the token and phrase annotations and a threshold calibrated for the wrong-reason prior of the native axis (28% for token, 15% for phrase).

Emotion	N	Phrase
sadness	212	94.8
anger	315	91.7
pride	265	84.5
fear	339	84.1
disgust	255	82.7
surprise	283	82.7
joy	160	81.2
boredom	137	80.3
trust	170	78.8
relief	193	77.2

Table 13: Phrase-level correct-reason rate (%) from the annotation study per emotion. Sorted by phrase-level rate.

G KernelSHAP Formulation

KernelSHAP approximates Shapley values by formulating the problem as a weighted linear regression:

$$\min_g \sum_{z' \in Z} \pi(z') \left[f(h(z')) - g(z') \right]^2 \qquad (2)$$

where $z' \in \{0,1\}^M$ is a binary coalition vector, $h(z')$ maps the coalition to an input (replacing absent tokens with `[MASK]`), $g(z') = \phi_0 + \sum_{i=1}^{M} \phi_i z'_i$ is the linear explanation model, and the kernel weight is:

$$\pi(z') = \frac{M-1}{\binom{M}{|z'|} \cdot |z'| \cdot (M - |z'|)} \qquad (3)$$

This weighting assigns highest importance to coalitions near the extremes ($|z'| \approx 0$ or $|z'| \approx M$), ensuring convergence to the true Shapley values.

Pattern	Emotion	Sentence	Wrong feature	SHAP
Function-word anchor				
	pride	"I won a scholarship."	*i*	0.13
	pride	"I completed a project."	*i*	0.13
	trust	"When I was in an ambulance I . . . the professionals."	*i*	0.07
Content word, not the cause				
	anger	"people push my boundaries and attempt to control my way of living."	*people*	0.25
	anger	"my sister, after continuously ignoring my clear boundary of entering my private space . . . "	*refusing*	0.30
	disgust	"I saw a dead pigeon being eaten by a crow."	*saw*	0.08
	disgust	"I saw a chicken bone that had been left on the kitchen countertop by my housemate."	*saw*	0.07
Diffuse (no specific feature flagged)				
	joy	"Finding out that I passed my final masters exams with a first (distinction)."	—	—
	sadness	"Donald Trump won in 2016. I felt so . . . for the people of the USA . . . "	—	—
	fear	"I felt . . . when followed by an unknown vehicle in a foreign country."	—	—
Template artefact (token level)				
	fear	"I felt when my kid had her first asthma attack."	*felt*	0.20
	sadness	"my daughter was ill and was very sick through the night and i felt i couldn't help."	*felt*	0.17
Ellipsis artefact (phrase level)				
	surprise	"I felt . . . when my husband . . . me and flew to Portugal to meet me on holiday and propose."	"..."	0.22
	trust	"I felt . . . with my ex partner when I was giving birth to my babies . . . "	"..."	0.18

Table 14: Representative examples of the wrong-reason patterns from Section 5.3. Ellipsis marks within sentences originate from the enVent writing prompt ("I felt . . . when"). — marks samples flagged Wrong Reason by annotators without a specific feature being marked. SHAP values are rounded to two decimals.

For phrase-level SHAP, the same formulation applies with M replaced by the number of phrase groups G.

H SHAP Example: Anger

Consider sample #1: "people get under my skin. Like for example if an entitled customer shows up at my work and demands to speak to my manager..." (predicted: *anger*, confidence: 0.561).
Token-level SHAP (top features):

Token	ϕ_i	Token	ϕ_i
demands	0.178	under	0.030
shows	0.047	my	0.028
for	0.044	get	0.026
skin	0.038	that	0.030

Phrase-level SHAP groups the idiom as a single feature:

Phrase	ϕ_g
demands	0.114
get under my skin	0.102
on almost a daily occurrence	0.057
for example	0.056

The scattered token values ($0.026 + 0.030 + 0.028 + 0.038 = 0.122$) are consolidated into a single interpretable attribution ($\phi = 0.102$).

I SHAP Example: Sadness

Table 15 shows SHAP values for a *sadness* prediction, illustrating how phrase-level SHAP preserves the phrasal verb "did pass" (a euphemism for death).

Token-level		Phrase-level	
Token	ϕ	Phrase	ϕ
away	.260	away	.271
pass	.225	did pass	.237
felt	.098	felt	.098
very	.063	she	.055
she	.053	i	.048
funeral	.046		

Table 15: SHAP values for "I felt ... when I had to go to a funeral of a family member, she had been ill ... once she did pass away." Predicted: *sadness* (conf: 0.943).

J SHAP Statistics

Table 16 compares SHAP-derived statistics between correct-reason and wrong-reason samples, showing systematic differences that the classifier can exploit.

Statistic	Correct	Wrong	Diff
Top-1 agrees (T↔P)	0.696	0.387	+0.309
Overlap ratio	0.678	0.489	+0.189
T signif. features	0.451	0.317	+0.134
Model confidence	0.484	0.379	+0.105
Top-3 conc. (phrase)	0.744	0.655	+0.088
Max token SHAP	0.148	0.081	+0.068
Sentence length (words)	21.1	30.0	−8.9

Table 16: SHAP statistics for correct-reason vs. wrong-reason samples. T = token-level, P = phrase-level. Top-1 agreement between token-level and phrase-level SHAP is the strongest discriminator.

K Syntactic Phrase Grouping and Parser Comparison

The automated phrase grouping follows a five-phase algorithm applied to each sentence: (1) prepositional phrases, identified through prep+pobj dependency relations; (2) noun phrases, extracted from the dependency parser's noun chunk function; (3) copular constructions (auxiliary + adjectival predicate); (4) verb phrases and verb-object combinations, including auxiliaries, negation, and particles; (5) any remaining tokens become singleton groups. The output is a list of token lists. For example, "people get under my skin like on almost a daily occurrence" yields ["people"], ["get under my skin"], ["like on almost a daily occurrence"].

The implementation uses spaCy (Honnibal et al., 2020) with the `en_core_web_md` model, reaching 85.6% exact match accuracy and 99.8% token coverage against the 500-sample gold standard.

Table 17 compares five parsing approaches for syntactic phrase grouping against 500 gold-standard samples manually created by human annotators.

Parser	Exact Match	Token Cov.
spaCy dep tree	85.6%	99.8%
Benepar	72.1%	99.5%
Stanza	69.8%	99.3%
NLTK chunker	58.4%	97.2%
spaCy enhanced	54.2%	98.1%

Table 17: Parser comparison against 500 gold-standard phrase groupings. Exact match = percentage of samples where the parser produces the same groupings as the gold standard.

L Annotation Tool

The annotation tool is a custom Streamlit web application deployed on Streamlit Community Cloud. For each sample, the tool displays the original sentence with color-coded SHAP highlights (red for positive, blue for negative attribution), interactive bar charts showing the top-8 features by $|\phi|$ for both token-level and phrase-level SHAP, the phrase groupings from the syntactic phrase grouping step, and the predicted emotion with model confidence. The annotator provides (1) an overall judgment per SHAP type (correct reason, wrong reason, or unclear), and (2) per-feature annotations marking individual tokens or phrases as correct or wrong. All annotators receive written guidelines with examples before beginning annotation. Annotations persist to Google Sheets with a CSV backup. Figure 2 shows a screenshot of the interface.

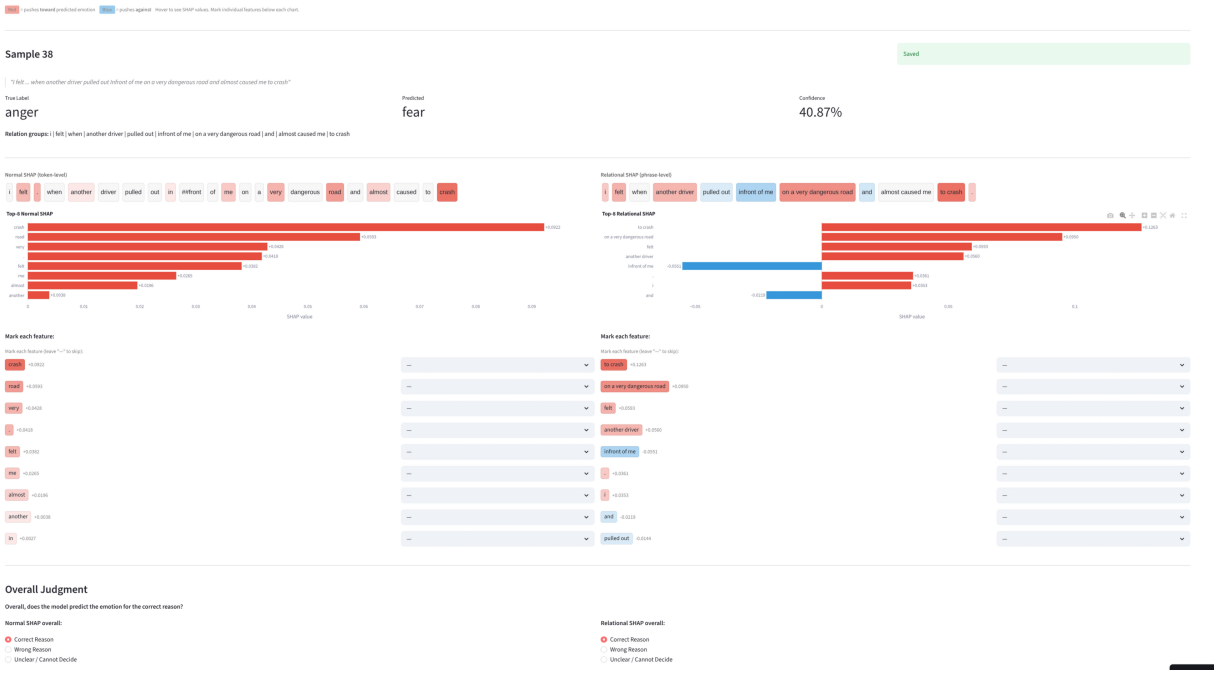

Figure 2: Screenshot of the annotation tool showing color-coded SHAP highlights, bar charts for token-level (left) and phrase-level (right) explanations, per-feature annotation dropdowns, and overall judgment radio buttons.