# An Explainable Comparative Analysis of Machine Learning with Rule Based Models in Sentiment Tasks: Implication for Polarization Studies.

**Anonymous ACL submission**

## Abstract

This research systematically examines whether post-hoc Explainable AI (XAI) techniques can render transformer-based sentiment models sufficiently interpretable for measuring societal polarization from unstructured text data. Further, we compare explainable transformers with rule-based sentiment lexicons to illustrate which model is better suited to societal polarization data. Our results suggest that, XAI techniques help to explain and interpret the model decision and we also found that, transformer perform better in comparison to rule-based lexicons.

## 1 Introduction

Societal polarization is a key problem (Arbatli and Rosenberg, 2021). Researchers have examined such polarization using social media as a proxy for society with methods including surveys (Bail et al., 2018) and experiments (Levy, 2021). However, analyzing the unstructured social media text data has also emerged as an important method (Pereira et al., 2025). Utilizing the automated text analysis, researchers measure polarization as constructs using both machine learning (ML) and rule-based lexicon models (Garzón-Velandia and Pennebaker, 2025; Susanto et al., 2025). Such a research stream is effective from a social and policy standpoint as social media is not just a primary source of information but also allows citizens to express their opinion and thoughts that can be an antecedent of wider citizens' behavior (Levy, 2021). Thus, it helps capture nuanced insights into the nature of polarization, which can assist policymakers in devising appropriate policy interventions to promote an inclusive society (Németh, 2023).

The methodological choice between rule-based lexicons and transformer models involves a fundamental trade-off between interpretability and performance that has been recognized as one of the central tensions in applied machine learning (Rudin, 2019). While the machine learning community has increasingly emphasized that interpretability should not be treated as a monolithic concept, but rather as context-dependent and multi-dimensional (Lipton, 2018), polarization researchers have yet to systematically examine whether XAI techniques adequately address their specific needs for transparency and replicability. This gap is particularly consequential because policy-relevant research requires not merely accurate predictions, but explanations that can withstand public scrutiny and inform democratic deliberation.

In measuring polarization, sentiment ML models and lexicons are widely employed instruments. The resulting classification of positive and negative is also grounded in the theory of polarization measurement (Esteban and Ray, 1994; Lu and Lee, 2025). However, both of the measuring instruments are constrained by relevant limitations. Rule and lexicon-based models tend to measure polarization statically, miss contextual elements inherent in big unstructured text data (Humphreys and Wang, 2018). On the other hand, transformer-based sentiment models capture sentiment classes more effectively, leading to precise and nuanced polarization measurement (Alaparthi and Mishra, 2021). But, its black box nature inhibits replication necessary for both epistemological and policy perspectives. The post-hoc explainable artificial intelligence (XAI) method can help to "Open Blackbox" of these sentiment models (Shrivastava et al., 2024). Thus, systematic comparison of transformer sentiment models with XAI techniques against lexicons is necessary to determine whether post-hoc XAI makes transformers as explainable as inherently interpretable lexicons.

Therefore, it raises three research questions that need to be answered: *RQ1*: How can XAI techniques be used to determine model explainability? *RQ2*: Which of the local or global XAI explanations can be reliant? *RQ3*: Do transformer-based

sentiment models combined XAI techniques outperform lexicon-based models in sentiment prediction and polarization measurement?

## 2 Review of Related Work

### 2.1 Sentiment Analysis use in Polarization

Sentiment analysis has become a powerful and crucial tool for conducting studies on polarization (Kenyon-Dean et al., 2018; Xiang and Zhou, 2014). The theoretical foundation for using sentiment as a polarization metric stems from classic work in polarization measurement, which conceptualizes polarization as the clustering of opinions at opposing extremes (Esteban and Ray, 1994). In recent years, the possibilities of measuring polarization using sentiment analysis have emerged. In the following section we describe key measurement instrument.

### 2.2 Rule-based and transformer for measuring polarization

Rule-based lexicon are pre-defined dictionaries of words which can be associated with emotional tone and subsequent sentiments (Taboada et al., 2011; Catelli et al., 2022). Popular lexicons employed in polarization research include AFINN and TextBlob (Loria et al., 2019; Aljedaani et al., 2022; Nielsen, 2011), as well as Valence Aware Dictionary and sEntiment Reasoner (VADER), specifically designed for social media text with attention to emoticons, slang, and informal expressions (Hutto and Gilbert, 2014). In contrast, transformer methods overcome some of those limitations by being trained on large corpora, enlarging the contextual nuances expressed in the language. For instance, BERT uses an encoder-only transformer with self-attention mechanisms, pre-trained on massive unlabeled text via masked language modeling (predicting masked words) and next sentence prediction (Pangtey et al., 2025; Kang et al., 2024; Devlin et al., 2019). It was found that transformer models achieve better results than lexicon-based approaches (Zhang et al., 2021). However, transformer methods lack explainability and transparency of results. They also need high computational resources for training (Leon, 2025). Explainable AI (XAI) has emerged as a response to these interpretability challenges, offering techniques that aim to make the decision-making processes of complex models more transparent and understandable to human users (Hwang and Lee, 2021).

### 2.3 Overview of XAI techniques

Most XAI techniques are referred to as post-hoc, meaning they are applied after the model has been trained. One such techniques is Shapley Additive exPlanations (Shap) which is based on game theory (Fryer et al., 2021) where each token is treated as a player in a cooperative coalitional game and payoff is the model's prediction (Parisineni and Pal, 2024). Local Interpretable Model-agnostic Explanations (LIME) explains models locally by generating perturbed input versions, obtaining predictions, and fitting a simple interpretable model to approximate local behavior (Ribeiro et al., 2016; Wang et al., 2025). The Layer-wise Relevance Propagation (LRP) is one another XAI techniques that tend to explain the model by propagating prediction backward through the network layers, decomposing the prediction into relevance scores for each input feature (Luo et al., 2024). Finally, Attention Visualization tend to explain the model, especially the transformer by illustrating parts of the neural network that receive focus during prediction (Vig, 2019). It is a useful tool in examining which words or phrases the model attends to when classifying sentiment (Seo et al., 2025). Commonly used techniques include BertViz in attention visualization (Vig, 2019).

The use of attention weights as explanations remains contested. Jain and Wallace (2019) demonstrated that attention weights frequently do not correlate with gradient-based measures of feature importance, raising questions about whether attention provides faithful explanations of model behavior. Wiegreffe and Pinter (2019) offered a partial rebuttal, arguing that attention can provide meaningful explanations under certain conditions, but acknowledged that attention should not be naively interpreted as providing definitive feature importance. This debate suggests researchers should exercise caution with attention visualization and triangulate findings with other XAI methods. More broadly, the question of explanation faithfulness - whether explanations accurately represent the model's actual reasoning process - has emerged as a central concern in XAI research (Jacovi and Goldberg, 2020). Explanations that appear plausible to human observers may not reflect the true computational mechanisms driving predictions, a problem that is particularly acute for post-hoc explanation methods applied to complex neural architectures. This concern motivates the present study's multi-method

approach to explainability assessment, combining global and local techniques with stability and consistency analyses

## 3 Method and material

### 3.1 Selection of Rule Based and ML Models

The open source three-rule-based lexicons and two transformer-based sentiment models, as shown in Table 1 are selected. Three open-source rule-based lexicons—TextBlob, VADER, and Afinn—calculate sentence polarity based on predetermined emotional tones of constituent words. Negative scores indicate negative sentiment, while positive scores indicate positive sentiment.

| Model | Type | Reference |
|---|---|---|
| TextBlob | Lexicon | (Loria et al., 2019) |
| VADER | Lexicon | (Hutto and Gilbert, 2014) |
| Afinn | Lexicon | (Nielsen, 2011) |
| Twitter-RoBERTa | Transformer | (Cardiff NLP, 2022; Loria et al., 2019) |
| BERTweet | Transformer | (Pérez, 2021) |

Table 1: Sentiment models used in this study.

### 3.2 XAI techniques

We employ four XAI techniques to evaluate model explainability: SHAP, LRP, LIME, and attention visualization.

### 3.3 Experimental Data

We use multiple datasets for different purposes. For general sentiment prediction (rule-based and transformer models), we utilize Telegram data obtained through its API. For XAI experiments (SHAP and LRP), we use a subset due to computational constraints. For comparing lexicons with transformers, we employ two human-labeled datasets and one benchmark sentiment dataset. Appendix 1 provides detailed data information.

### 3.4 Analytical techniques

Our XAI analysis using each of four XAI techniques consist two stages of analysis.

#### 3.4.1 First Stage of Analysis

The first stage consists of basic descriptive statistics and visualizations provided by each XAI technique (both local and global), including SHAP, LIME, LRP, and attention visualization.

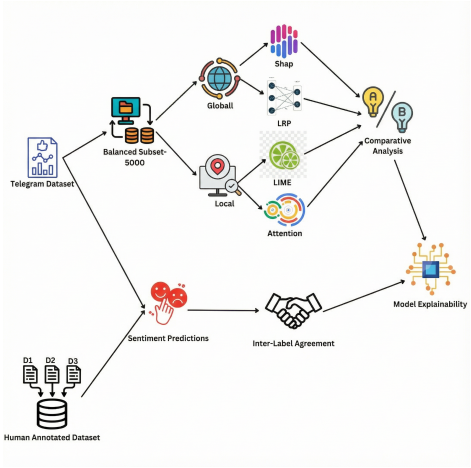



Figure 1: Experimental Setup

#### 3.4.2 Second layer of Analysis

The second stage of analysis is applied to evaluate the model by examining changes in performance resulting from the alteration (modification and removal) of dominant features (tokens), along with statistical analysis using complementary evaluation measures.

### 3.5 Comparative Analysis

Finally present research has undertaken a comparative analysis of transformer model with rule based lexicon models. First, an inter-lexicon and inter-transformer model agreement is developed based on experimental telegram dataset using different techniques. Secondly, their prediction is compared with benchmark (Maas et al., 2011), aspect based human annotated sentiment dataset (Guo et al., 2023) and human annotated amazon review dataset (Alghamdi and Alhasawi, 2024). Such comparative analysis revealed performance on models using criteria of accuracy and agreement that offers significant epistemological and policy implications.

## 4 Experimental Setup

Figure 1 depicts our experimental setup, which consists of three levels. First, we predicted sentiment using rule-based and ML models on 2 million Telegram posts and analyzed inter-model agreement. Second, we created a balanced 5,000-sample subset from transformer predictions and applied global explainability methods (SHAP and LRP) and local methods (LIME and attention visualization), comparing results to assess model explainability. Finally, we compared rule-based and transformer models against three human-annotated datasets.

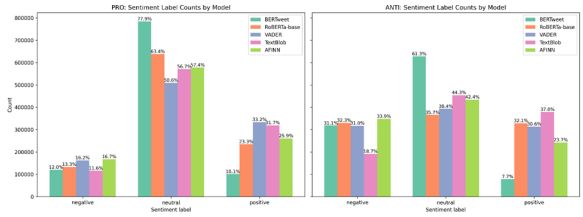



Figure 2: Sentiment Predictions

## 5 Results and Analysis

### 5.1 Prediction of Sentiment

The figure 2 shows the results of the sentiment prediction by each of the rule and transformer based models. The results convey two consistent trends. First, in both datasets neutral sentiment is being predicted highly compared to polarizing positive and negative sentiment. Secondly, the Bertweet-base model predicts neutral sentiment more frequently than all other models.

#### 5.1.1 Inter label Agreement

The inter model agreement of Cohen-Kappa calculated using Sci-kit learn python framework of both rule and transformer based have shown the mixed results. The result suggests that, both Vader and Afinn model tend to predict sentiment labels more closely with higher agreement level of 0.761 in pro and 0.78 in anti-dataset. On the other hand, two transformer model tend to have lesser comparison than 0.50 in both pro (0.457) and anti (0.428). The overall Fleiss Kappa also suggests that, lexicons (>0.50) tend to have agreement as compared to transformer models (<0.50).

| Model 1 | Model 2 | Cohen-Pro | Cohen-Anti |
|---|---|---|---|
| Roberta-Base | Bertweet-base | 0.457 | 0.428 |
| Vader | TextBlob | 0.420 | 0.490 |
| Vader | Afinn | 0.761 | 0.780 |
| TextBlob | Afinn | 0.432 | 0.478 |
| **Fleiss-Kappa (Models)** | | 0.446 | 0.399 |
| **Fleiss-Kappa (Lexicons)** | | 0.539 | 0.580 |

Table 2: Inter-model agreement measured using Cohen's Kappa and Fleiss' Kappa.

### 5.2 Global XAI Analysis

The present research employed two key techniques to assess global explainability of our models.

#### 5.2.1 Shap Analysis

The first key result identifies the most dominant features (tokens) contributing to the prediction of both polarizing labels (positive and negative). The top 20 features are reported in Appendices 2.1 and 2.2. These results are re-validated (Table 3) by examining whether the identified features correspond to positive or negative sentiment using SentiWordNet (Baccianella et al., 2010). SentiWordNet is an interpretable and conservative lexical resource that assigns sentiment scores based on the emotional tone of individual words. Its use as a validation tool follows established practice in grounding feature-importance attributions with lexical sentiment resources (Baccianella et al., 2010). Although SentiWordNet does not capture contextual sentiment, its word-level scores provide a transparent benchmark for assessing whether SHAP-identified features align with human intuitions about sentiment-bearing vocabulary.

| Token Group | Roberta-Base | Bertweet-Base |
|---|---|---|
| Pro-Positive SHAP | 0.351 | 0.265 |
| Anti-Positive SHAP | 0.287 | 0.191 |
| Pro-Negative SHAP | -0.202 | -0.282 |
| Anti-Negative SHAP | -0.219 | -0.176 |

Table 3: Mean Sentiment score for Shap Features.

Finally, appendix 3.1 (for Roberta-base) and 3.2 (for Bert-base) shows beeswarm plots, a special visualization that helps to explain the model by interpreting and visualizing the key features that impact the model to yield a decision of a particular class of positive and negative sentiment. Two key interpretations here are to be noted. First, most features that impact model decisions of particular class (positive or negative) predictions are consistent with emotional tone. This suggests that plots from both models show internally consistent feature roles in prediction, with minor features influencing both positive and negative outcomes. Moreover, the overlap of most features across both models further reinforces the reliability of the results.

#### 5.2.2 Summary

The three different results of top 20 features by mean Shap values corresponding to positive and negative class, their mean explainable polarity score as using SentiWordNet and beeswarm plots present an explainable layer that clearly shows that model prediction of polarizing class is internally consistent. Therefore, it can be concluded that both models chosen are explainable at global level. However, further investigation is required to strength such findings.

### 5.2.3 Comparative Analysis

The comparative analysis has been undertaken to examine the relative interpretability of the models using average tokens required for each (positive and negative) class to be predicted by each model (Roberta-base and Bertweet-base) and confidence score. The average token required is calculated (see equation 1) as the average number of tokens in each sentiment class exceeding minimums score threshold (-0.01+0.01)

$$\text{avg\_tokens}^c = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{T} 1\Big(\Big|v_{i,j}^{(c)}\Big| > t\Big) \quad (1)$$

The results in table 4 clearly demonstrate that, Bertweet base model performs better relatively than Roberta-base model as in both datasets and sentiment class, it required fewer tokens to predict positive (pro=7.25, anti-=9.39) and negative (pro=8.09 , anti=14.04) and higher overall confidence score (pro=0.822, anti=0.828)

| Model / Dataset | Avg. Pos | Avg. Neg | Confidence | Uncertainty |
|---|---|---|---|---|
| Roberta-(Pro) | 10.56 | 9.59 | 0.731 | 0.269 |
| Roberta-(Anti) | 13.11 | 15.29 | 0.746 | 0.254 |
| Bertweet-(Pro) | 7.25 | 8.09 | 0.822 | 0.178 |
| Bertweet-(Anti) | 9.31 | 14.04 | 0.828 | 0.172 |

Table 4: Average number of tokens required.

### 5.2.4 Stability and Consistency

We conducted a stability analysis by removing the top 20 global features, then the top positive and negative sentiment features, and finally TF-IDF–extracted features, calculating the confidence score at each stage. The results as shown in table 5 suggest no significant impact of removing features on confidence score. The results suggest and consistent with established notion about transformer ability to understand and classify language contextually so removing these feature does not impact on overall model's confidence to understand and classify the text.

| Model / Dataset | Global-Shap Top 20 | Top 20-Shap by Sentiment | Top 20-Shap by TFIDF |
|---|---|---|---|
| RoBERTa-(Pro) | 0.7296 | 0.7288 | 0.73116 |
| RoBERTa-(Anti) | 0.7448 | 0.7445 | 0.7464 |
| BERTweet-(Pro) | 0.8297 | 0.8213 | 0.8220 |
| BERTweet-(Anti) | 0.8213 | 0.8260 | 0.8283 |

Table 5: Model performance after removing top 20 features.

This finding resonates with concerns raised by Hooker et al. (2019), who demonstrated that removing features identified as important by attribution methods often fails to degrade model performance as expected - a phenomenon they term the "masking problem." Our stability analysis suggests that for transformer-based sentiment classification, individual token-level importance may be less meaningful than the holistic contextual representation, which has implications for how researchers should interpret and communicate XAI outputs in polarization studies

A consistency analysis was conducted to examine whether semantically similar sentence exhibit comparable SHAP values, and whether embedding similarity can predict SHAP value similarity. So, we first calculated embeddings using All-MiniLM-L6-v2 (Reimers et al., 2024). Subsequently, we extracted 100 sentence pairs per sentiment class exhibiting cosine value of 0.60 to 0.90 and re-calculated their Shap values. The mean cosine similarity in the Roberta-base was 0.7347 and Bertweet-base 0.7383. The Jensen-Shannon Divergence (JSD) score (Menéndez et al., 1997) and Kolmogorov-Smirnov (KS) (Massey Jr., 1951) test were employed. The mean JSD scores (see Table 5) are approximately equivalent across both models, with lower values observed overall and within individual sentiment classes (positive and negative), indicating a higher degree of similarity in SHAP value distributions between sentence pairs. Notably, sentence pairs from the negative sentiment class exhibit higher similarity in Shap values compared to the overall and the positive sentiment. These findings are corroborated by the KS test results, where the mean KS statistics are also approximately equivalent across both models.

Finally, we also examined whether embedding similarity could be used to predict Shap value similarity. So, we calculated the cosine similarity of the Shap Values as well and then computed the Spearman rank correlation between embedding similarity and SHAP value similarity. As shown in Table 6, both models exhibit statistically significant Spearman rank correlations, demonstrating that embedding similarity is a reliable predictor of SHAP value similarity. This indicates consistency in model explanations: sentence pairs with similar semantic representations (embeddings) receive similar feature importance attributions (SHAP values), reinforcing that the sentiment-related patterns captured in embeddings are consistently reflected in the mod-

| Category | Roberta-Base | | | Bertweet-Base | | |
|---|---|---|---|---|---|---|
| | Mean JSD | Std. Dev | Mean KS (p>0.05) | Mean JSD | Std. Dev | Mean KS (p>0.05) |
| Overall | 0.339 | 0.143 | 0.287 | 0.341 | 0.150 | 0.282 |
| Negative | 0.271 | 0.083 | 0.295 | 0.270 | 0.084 | 0.300 |
| Positive | 0.406 | 0.192 | 0.225 | 0.425 | 0.186 | 0.224 |

Table 6: Comparison of Roberta-Base and Bertweet-Base models across categories

els' decision-making processes.

| Model | P-value | Correlation Coefficient |
|---|---|---|
| RoBERTa-Base | 1.16e-05 | 0.251 |
| BERTweet-Base | 0.002225 | 0.176 |

Table 7: Embedding and SHAP similarity estimation.

### 5.2.5 Quality of Explanations

Finally to assess the quality and robustness of Shap explanation stability, we undertaken perturbation of dataset using techniques of random deletion, token shuffling, and synonym replacement. In random deletion, 10 percnet of tokens were removed at random (at least one token per sentence). In token shuffling, all tokens were retained but their order was randomly permuted. In synonym replacement, 10 percent of tokens were replaced with WordNet-derived synonyms matching the part-of-speech (POS) of the original token. First we calculated the stability score of three perturbed dataset i.e. deletion, shuffling and synonym. The results as shown in table 7 suggest that, less change in confidence score suggests that, both models are relatively strong in making correct predictions

| Model | Original | Deletion | Shuffle | Synonym |
|---|---|---|---|---|
| Roberta-(Pro) | 0.731 | 0.731 | 0.726 | 0.727 |
| Roberta-(Anti) | 0.746 | 0.743 | 0.730 | 0.741 |
| Bertweet-(Pro) | 0.822 | 0.825 | 0.827 | 0.821 |
| Bertweet-(Anti) | 0.828 | 0.829 | 0.825 | 0.825 |

Table 8: Model performance under perturbations.

Finally, we evaluated the stability of Shap values by computing them for each perturbed dataset and performing paired t-tests against the original SHAP values (Appendices 4.1 and 4.2). Statistical significance ($p < 0.05$) was observed for nearly all pairwise comparisons, indicating that perturbations induced significant changes in Shap values across both models and datasets. The sole exception was the deletion perturbation in Bertweet-base on the pro dataset. Despite statistical significance, the associated t-statistics and Cohen's D effect sizes indicate that the practical magnitude of these changes is moderate. These results demonstrate that Shap values are dynamic and sensitive to sentence structure modifications; however, the limited magnitude of change reflects transformers' ability to maintain contextual understanding despite surface-level perturbations

### 5.2.6 LRP Analysis

We employed the simplified method of LRP focusing on classification head of model. which explain the model by illustrating dominant features in model through relevance score which is calculated by (a) calculating predicted logit that (b) propagated backward to model's linear classification head using the Z-LRP rule (c) which in turn assign relevance value proportional to its contribution to the final prediction. The appendix 5.1 and 5.2 shows results of dominant features yielding the sentiment class prediction by their relevance score.

Table 9 and Appendix 5.3 present mean absolute relevance (average token contribution strength), sparsity (fraction of highly relevant tokens), and entropy (token concentration). RoBERTa-Base shows stronger, more focused contributions for Pro texts, while BERTweet exhibits higher, balanced relevance across both classes. Both models display very low sparsity, indicating distributed evidence rather than reliance on few tokens. Sentiment-level analysis reveals that RoBERTa-Base shows higher relevance and lower entropy for positive Pro texts, indicating focused contributions, while negative Pro texts show weaker, diffuse evidence. Anti texts display balanced but higher entropy across sentiments. BERTweet shows comparable scores across datasets, with negative sentiment receiving higher relevance and lower entropy in both, suggesting focused model attention on this class. Overall, sentiment-level LRP analysis reveals fine-grained explainability differences..

| Metric | RoBERTa Base | | BERTweet Base | |
|---|---|---|---|---|
| | Pro | Anti | Pro | Anti |
| Mean Absolute Relevance | 2.01 | 1.77 | 2.38 | 2.33 |
| Sparsity Score | 0.00 | 0.00 | 0.001 | 0.000 |
| Entropy Score | 10.0250 | 11.50 | 11.34 | 11.78 |

Table 9: Global Explainability Metrics.

### 5.3 Local XAI Analysis

The present research have also employed local XAI techniques which assess individual prediction of model. The LIME, Shap and attention mechanism visualization is employed. All of these techniques are visualization based where LIME, Shap help us to identify the most influential feature deriving specific class prediction and attention mechanism visualization help us to assess neural network layers processing the input data. Appendices 6.1-6.8 present SHAP and LIME results for both models on pro- and anti-datasets. Both techniques accurately explain each model's sentiment predictions. For instance, appendices 6.1 (LIME) and 6.2 (SHAP) identify features like "beauty," "largest," and "giants" as key drivers of positive sentiment predictions for both RoBERTa-base and BERTweet-base. Attention mechanism visualization further illuminates how the models' neural network layers process input data. Both models contain 12 layers (0-11), each with multiple attention heads. Appendix 7.1 (RoBERTa-base) and 7.2 (BERTweet-base) demonstrate how the 12 heads in layer 1 process token sequences. Appendix 7.3 identifies the most effective layer-head combinations: Layer 2, Head 8 for BERTweet and Layer 9, Head 4 for RoBERTa. These insights help researchers better understand text classification model behavior

### 5.4 Concluison of XAI Analysis

The global and local explanation of the XAI techniques have helped us to understand the behavior of both of our model across different set of data. Although no such metric exist that can help us to determine the level of accuracy in the explanation of models using variety of XAI techniques but experimenting with these techniques have enabled to understand such behavior in a different and complimentary ways. For example, top global feature identified by Shap were reconfirmed their sentiment tone with more conservative SentiWordNet lexicon. Secondly, it was also observed no change in confidence when these dominant (top 20) were changed or removed. It reconfirms the basic notion of human language in which certain words explains the overall context but context itself is macro level which transformer are capable of understanding and generating. Consistently, it is also confirmed a relationship in between vector of embeddings and vectors Shap which validates the notion that, both are interdependent enhancing their overall interpretability

## 6 Comparison

In the final stage, transformer models are compared against lexicon models for sentiment predictions using two human annotated dataset of aspect base (Guo et al., 2023) sentiment amazon product review dataset (Alghamdi and Alhasawi, 2024) and benchmark sentiment dataset (Maas et al., 2011). Two key measured of accuracy and Cohen Kappa employed. The accuracy measure number of instance each of rule based and transformer model predict accurately or matches with each of human annotated and benchmarks dataset. The equation 2 shows the mathematical expression of accuracy.

$$accuracy = \frac{1}{N}\sum_{i=1}^{N} I(\hat{y}_i = y_i) \qquad (2)$$

The results of accuracy are shown in table 10 suggest three insights. First, both transformer models outperform compared to rules based models on accuracy. Second, Bertweet models perform better on aspect based sentiment (Guo et al., 2023) and amazon review dataset (Alghamdi and Alhasawi, 2024) abd underperforms against Roberta-base in benchmark dataset (Maas et al., 2011). Finally, rule based lexicons given their limitations have not underperformed overall. All three lexicons outperform the Bertweet-based model on benchmark sentiment dataset, their performance at Amazon review dataset is also not low (above 50 on average).

| Dataset | Roberta | Bertweet | VADER | TextBlob | Afinn |
|---|---|---|---|---|---|
| Aspect-Based | 0.44 | 0.49 | 0.30 | 0.36 | 0.34 |
| Amazon Review | 0.69 | 0.71 | 0.54 | 0.49 | 0.53 |
| Benchmark | 0.71 | 0.65 | 0.70 | 0.69 | 0.69 |

Table 10: Model performance across different datasets.

Secondly, we have assessed the performance of sentiment models using Cohen's Kappa coefficient (McHugh, 2012) which assess the level of agreement between predicted label by models and labels in each human-annotated and benchmark datasets.

To account for the ordinal nature of sentiment classes, weighted Cohen's Kappa with quadratic weighting (Doewes et al., 2023) was employed (see equation 3).

$$K = \frac{P_o - P_e}{1 - P_e} \tag{3}$$

The results in table 11 show model performance and transformer consistently achieve higher agreement with human annotations which indicates their advantage in capturing sentiment nuances. Secondly, discussing the level of performance on each of dataset, it is high on Amazon Review dataset, moderate on the Benchmark Sentiment Dataset, and lowest on the Aspect-Based sentiment which suggest that aspect-level sentiment classification is a more challenging task. Finally, among the transformer models, Bertweet-base slightly outperforms Roberta-base on all three datasets. Lexicon-based models (VADER, TextBlob, and Afinn) show substantially lower agreement with human annotations than transformer models, with kappa values generally below 0.50. Among rule-based lexicons, VADER performs best in sentiment classification.

| **Dataset** | **Roberta** | **Bertweet** | **VADER** | **TextBlob** | **Afinn** |
|---|---|---|---|---|---|
| Aspect-Based | 0.255 | 0.291 | 0.102 | 0.118 | 0.097 |
| Amazon Review | 0.779 | 0.786 | 0.502 | 0.402 | 0.477 |
| Benchmark | 0.656 | 0.686 | 0.392 | 0.376 | 0.422 |

Table 11: Model performance comparison

## 7 Conclusion

The measurement of polarization from unstructured text data especially on social media presents an opportunity to undertake policy and socially relevant research. However, as the use of ML models are being preferred, questions regarding their Blackbox nature generate confusion, especially with regard to the replicability of research. Our research concludes that XAI techniques can effectively be used to understand the model's behavior that only can help us to infer the replicability and other issues.

For polarization researchers specifically, these findings suggest a three practical workflow. (1) report global feature importance using SHAP or LRP to demonstrate that the model attends to sentiment-relevant vocabulary, (2) conduct stability analyses to characterize model robustness, and (3) validate predictions against established benchmarks. However, our findings also underscore that XAI techniques do not eliminate the interpretability-performance trade-off but rather shift its terms. Researchers facing strict replicability requirements may still prefer lexicon-based approaches, despite their lower performance on certain benchmarks. This aligns with recent arguments that model selection should be driven not solely by predictive accuracy but by the broader sociotechnical context of deployment (Selbst et al., 2019).

Global and local XAI techniques help open the black box by identifying key features driving sentiment predictions. SHAP and LRP reveal features consistent with sentiment tone, providing insights into model behavior. Local techniques (LIME and attention) strengthen these inferences, though more research is needed to generalize local explanations to overall model behavior. Therefore, SHAP and LRP are more reliable for polarization studies. While transformers outperform lexicons on Cohen's Kappa, lexicons offer comparable accuracy with advantages in simplicity, computational efficiency, and replicability—making them a viable choice. Future work can enhance lexicon performance to match transformers. Overall, combining transformers with XAI techniques enables robust polarization measurement and extends empirical research on socio-economic impacts

## 8 Limitations

We report three key limitations. First, since there are no standardized guidelines for the systematic evaluation of model explainability, we mostly relied on best practices from prior literature. Second, the selection of models was constrained to open-access and based on other architectures, such as recurrent neural networks (e.g., LSTM), as well as proprietary or closed-source models, were not included in the evaluation. Third, the models used in this study were not fine-tuned or trained on the target datasets; instead, openly available trained models were evaluated in an out-of-the-box setting. While this allows for fair comparison and reproducibility, it may limit performance and does not reflect the full potential. Fourth, our validation of SHAP features using SentiWordNet assumes that sentiment-bearing vocabulary should correspond to lexical sentiment scores. However, transformers may legitimately attend to contextual cues (such as negation markers or intensifiers) that do not carry independent sentiment valence but modify the sentiment of surrounding tokens.

## References


Shivaji Alaparthi and Manit Mishra. 2021. Bert: A sentiment analysis odyssey. *Journal of Marketing Analytics*, 9(2):118–126.

Salem Alghamdi and Yaser Alhasawi. 2024. Aspect-based sentiment analysis in smart devices: A comprehensive and specialized dataset. *Data in Brief*, 55:110642.

Wajdi Aljedaani, Furqan Rustam, Mohamed Wiem Mkaouer, Abdullatif Ghallab, Vaibhav Rupapara, Patrick Bernard Washington, Ernesto Lee, and Imran Ashraf. 2022. Sentiment analysis on twitter data integrating textblob and deep learning models: The case of us airline industry. *Knowledge-Based Systems*, 255:109780.

Ekim Arbatli and Dina Rosenberg. 2021. United we stand, divided we rule: how political polarization erodes democracy. *Democratization*, 28(2):285–307.

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC)*, pages 2200–2204.

Christopher A Bail, Lisa P Argyle, Taylor W Brown, John P Bumpus, Haohan Chen, MB Fallin Hunzaker, Jaemin Lee, Marcus Mann, Friedolin Merhout, and Alexander Volfovsky. 2018. Exposure to opposing views on social media can increase political polarization. *Proceedings of the National Academy of Sciences*, 115(37):9216–9221.

Cardiff NLP. 2022. twitter-roberta-base-sentiment-latest. https://huggingface.co/cardiffnlp/twitter-roberta-base-sentiment-latest. Machine learning model.

Rosario Catelli, Serena Pelosi, and Massimo Esposito. 2022. Lexicon-based vs. bert-based sentiment analysis: A comparative study in italian. *Electronics*, 11(3):374.

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.

A. Doewes, N. Kurdhi, and A. Saxena. 2023. Evaluating quadratic weighted kappa as the standard performance metric for automated essay scoring. In *Proceedings of the 16th International Conference on Educational Data Mining (EDM 2023)*, pages 103–113. International Educational Data Mining Society.

Joan-Maria Esteban and Debraj Ray. 1994. On the measurement of polarization. *Econometrica: Journal of the Econometric Society*, pages 819–851.

Daniel Fryer, Inga Strümke, and Hien Nguyen. 2021. Shapley values for feature selection: The good, the bad, and the axioms. *Ieee Access*, 9:144352–144360.

Diana Camila Garzón-Velandia and James W Pennebaker. 2025. A linguistic strategy to measure negative affective polarization through text content. *Journal of Language and Social Psychology*, page 0261927X251338360.

Yuting Guo, Sudeshna Das, Sahithi Lakamana, and Abeed Sarker. 2023. An aspect-level sentiment analysis dataset for therapies on twitter. *Data in Brief*, 50:109618.

Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. 2019. A benchmark for interpretability methods in deep neural networks. *Advances in neural information processing systems*, 32.

Ashlee Humphreys and Rebecca Jen-Hui Wang. 2018. Automated text analysis for consumer research. *Journal of Consumer Research*, 44(6):1274–1306.

Clayton Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media*, volume 8, pages 216–225.

Hohyun Hwang and Younghoon Lee. 2021. Semi-supervised learning based on auto-generated lexicon using xai in sentiment analysis. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 593–600.

Alon Jacovi and Yoav Goldberg. 2020. Towards faithfully interpretable nlp systems: How should we define and evaluate faithfulness? *arXiv preprint arXiv:2004.03685*.

Sarthak Jain and Byron C Wallace. 2019. Attention is not explanation. *arXiv preprint arXiv:1902.10186*.

Sunghoon Kang, Hyeoneui Kim, Hyewon Park, and Ricky Taira. 2024. Detecting redundant health survey questions using language-agnostic bert sentence embedding (labse). *arXiv preprint arXiv:2412.03817*.

Kian Kenyon-Dean, Eisha Ahmed, Scott Fujimoto, Jeremy Georges-Filteau, Christopher Glasz, Barleen Kaur, Auguste Lalande, Shruti Bhanderi, Robert Belfer, Nirmal Kanagasabai, and 1 others. 2018. Sentiment analysis: It's complicated! In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1886–1895.

Maikel Leon. 2025. From lexicons to transformers: An ai view of sentiment analysis. *Journal of Intelligent Communication*, 4(2):13–25.

Ro'ee Levy. 2021. Social media, news consumption, and polarization: Evidence from a field experiment. *American economic review*, 111(3):831–870.

Zachary C Lipton. 2018. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3):31–57.

Steven Loria, P Keen, M Honnibal, R Yankovsky, D Karesh, E Dempsey, W Childs, J Schnurr, A Qalieh, L Ragnarsson, and 1 others. 2019. Textblob: simplified text processing; 2018. *Online: https://textblob. readthedocs. io/en/dev/Accessed*, pages 08–02.

Hsiu-Chi Lu and Hsuan-wei Lee. 2025. Agents of discord: Modeling the impact of political bots on opinion polarization in social networks. *Social Science Computer Review*, 43(4):750–772.

Siwen Luo, Hamish Ivison, Soyeon Caren Han, and Josiah Poon. 2024. Local interpretations for explainable natural language processing: A survey. *ACM Computing Surveys*, 56(9):1–36.

Andrew Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pages 142–150.

Frank J. Massey Jr. 1951. The kolmogorov–smirnov test for goodness of fit. *Journal of the American Statistical Association*, 46(253):68–78.

Mary L. McHugh. 2012. Interrater reliability: The kappa statistic. *Biochemia Medica*, 22(3):276–282.

M. L. Menéndez, J. A. Pardo, L. Pardo, and M. D. C. Pardo. 1997. The jensen–shannon divergence. *Journal of the Franklin Institute*, 334(2):307–318.

Renáta Németh. 2023. A scoping review on the use of natural language processing in research on political polarization: trends and research prospects. *Journal of computational social science*, 6(1):289–313.

Finn Årup Nielsen. 2011. A new anew: Evaluation of a word list for sentiment analysis in microblogs. *arXiv preprint arXiv:1103.2903*.

Lata Pangtey, Anukriti Bhatnagar, Shubhi Bansal, Shahid Shafi Dar, and Nagendra Kumar. 2025. Large language models meet stance detection: A survey of tasks, methods, applications, challenges and future directions. *arXiv preprint arXiv:2505.08464*.

Sai Ram Aditya Parisineni and Mayukha Pal. 2024. Enhancing trust and interpretability of complex machine learning models using local interpretable model agnostic shap explanations. *International Journal of Data Science and Analytics*, 18(4):457–466.

Catarina Pereira, Raquel da Silva, and Catarina Rosa. 2025. How to measure political polarization in text-as-data? a scoping review of computational social science approaches. *Journal of Information Technology & Politics*, 22(2):172–185.

JM Pérez. 2021. Finiteautomata/bertweet-base-sentiment-analysis. *Hugging Face*.

Nils Reimers, P Freire, G Becquin, O Espejel, and J Gante. 2024. Sentence-transformers/all-minilm-l6-v2 hugging face.

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.

Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, 1(5):206–215.

Andrew D Selbst, Danah Boyd, Sorelle A Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2019. Fairness and abstraction in sociotechnical systems. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 59–68.

Seongbum Seo, Sangbong Yoo, Hyelim Lee, Yun Jang, Ji Hwan Park, and Jeong-Nam Kim. 2025. A sentence-level visualization of attention in large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (System Demonstrations)*, pages 313–320.

Lucky Susanto, Musa Izzanardi Wijanarko, Prasetia Anugrah Pratama, Zilu Tang, Fariz Akyas, Traci Hong, Ika Karlina Idris, Alham Fikri Aji, and Derry Tanti Wijaya. 2025. A multi-labeled dataset for indonesian discourse: Examining toxicity, polarization, and demographics information. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 18863–18890.

Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. 2011. Lexicon-based methods for sentiment analysis. *Computational Linguistics*, 37(2):267–307.

Jesse Vig. 2019. A multiscale visualization of attention in the transformer model. *arXiv preprint arXiv:1906.05714*.

Xiaolong Wang, Jiahui Lyu, John D. Peter, Byung G. Kim, B. D. Parameshachari, Kai Li, and Wei Wei. 2025. Explaining sentiments: Improving explainability in sentiment analysis using local interpretable model-agnostic explanations and counterfactual explanations. *IEEE Transactions on Computational Social Systems*.

Sarah Wiegreffe and Yuval Pinter. 2019. Attention is not not explanation. *arXiv preprint arXiv:1908.04626*.

Bing Xiang and Liang Zhou. 2014. Improving twitter sentiment analysis with topic-based mixture modeling and semi-supervised training. In *Proceedings*

*of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 434–439.

Yuchen Zhang, Alex Warstadt, Ximing Li, and Samuel R. Bowman. 2021. When do you need billions of words of pretraining data? 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 1112–1125.

**Appendix 1.1: Meta-Data on Experimental Dataset-Sentiment prediction**

The present research extracted the unstructured textual data from Telegram using the application programming interface (API) provided by python's Telethon framework. This is big thick dataset which was used to undertake sentiment predicted using each of rule based-lexicon and transformer models. We extracted data from 108 Telegram groups and channels discussing various scientific topics which include discussion on various types of science. These groups were categorized as either anti-science or pro-science based on (1) their descriptions and (2) the initial discussions within the groups and channels. Table 2 presents the basic descriptions of the datasets developed for the PhD dissertation.

| S.No | Category | Number of Groups | No. of Posts | Date |
|---|---|---|---|---|
| 1 | Pro Science | 74 | 1006837 | 2016-2023 |
| 2 | Anti Science | 34 | 1025269 | 2016-2023 |

**Appendix 1.2: Meta-Data on Experimental Dataset-Sentiment prediction**

To experimenting with each of XAI techniques especially global techniques i.e. Shap and LRP, we create subset of dataset as presented in appendix 1.1. The decision to create subset was based on the fact that, techniques such as Shap required huge amount of computing and time resource. Thus, to optimize such, we undertake such decision.

Each Pro-Science and Anti-Science subset contains 5,000 samples. These 5,000 samples are balanced across sentiment classes, meaning each dataset includes 1,667 positive samples, 1,667 negative samples, and 1,666 neutral samples. We used a random state to randomly draw samples while maintaining balanced class distribution. Since we employed two models (RoBERTa-Base and BERTweet-Base), we created separate subsets for each model based on their label distributions.

## Appendix 2.1: Top 20 tokens predicting polarizing labels: Roberta-base Model

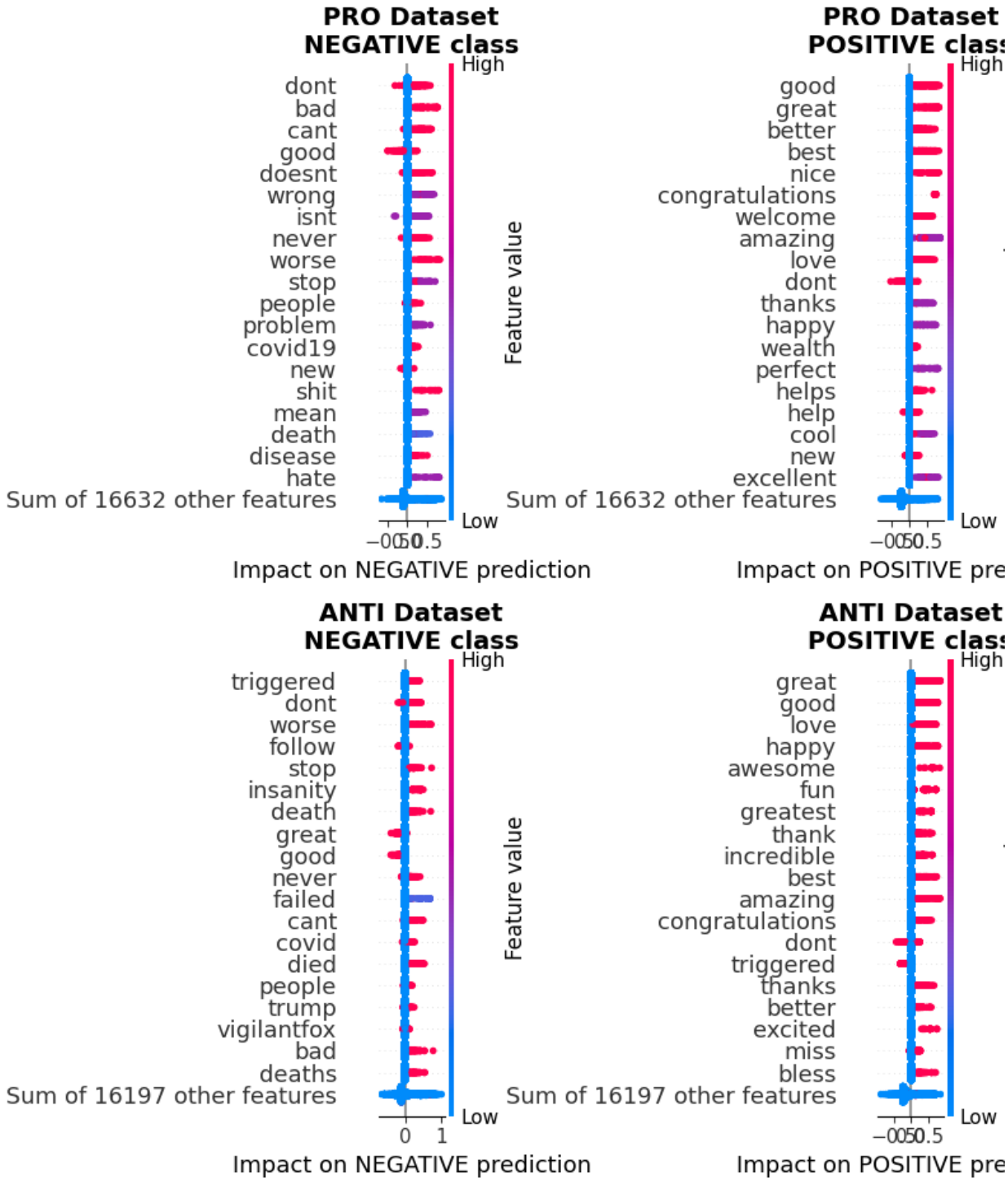
| Pro | | | | Anti | | | |
|---|---|---|---|---|---|---|---|
| **Positive Feature** | **Mean Shap** | **Negative Feature** | **Mean Shap** | **Positive Feature** | **Mean Shap** | **Negative Feature** | **Mean Shap** |
| gorgeous | 0.87038 | pissed | 0.801289 | proud | 0.7473 | unfuckingbelievable | 0.813057 |
| enjoyed | 0.866934 | cursed | 0.796791 | kindest | 0.674093 | sucks | 0.798774 |
| impressive | 0.82385 | garbage | 0.787841 | optimistic | 0.662681 | hell | 0.752039 |
| perfect | 0.79452 | rubbish | 0.784248 | congratulates | 0.660122 | trash | 0.702409 |
| awesomeness | 0.792379 | hells | 0.764458 | fascinating | 0.641905 | broken | 0.668941 |
| awesome | 0.79122 | bullshit | 0.762609 | enjoy | 0.636914 | ugh | 0.667406 |
| breathtaking | 0.761736 | useless | 0.741365 | impressive | 0.636283 | harassment | 0.661269 |
| fantastic | 0.749989 | suck | 0.739847 | shining | 0.624948 | corrupted | 0.638431 |
| brilliant | 0.739427 | hate | 0.735134 | awesome | 0.623924 | unimpressive | 0.621219 |
| nice | 0.725613 | fuck | 0.714404 | favourite | 0.618442 | fuckthenewworldqrder | 0.60716 |
| congratulations | 0.713669 | motherfucking | 0.711009 | good | 0.618346 | ominous | 0.599879 |
| adorable | 0.707263 | idiot | 0.710724 | excellent | 0.613992 | inexcusable | 0.593602 |
| appreciated | 0.707207 | killer | 0.70511 | positivequotes | 0.599544 | criminal | 0.593595 |
| superb | 0.699428 | holeefuck | 0.702435 | delighted | 0.587067 | criminals | 0.59062 |
| cutest | 0.686664 | regretted | 0.694385 | enjoyed | 0.572256 | laughable | 0.564509 |
| awesome | 0.673298 | lame | 0.687664 | welldone | 0.565207 | cuck | 0.555357 |
| amazingly | 0.671284 | hellish | 0.683414 | excited | 0.549781 | outrageous | 0.552988 |
| satisfying | 0.657758 | badworking | 0.680951 | glad | 0.531536 | shoddy | 0.552863 |
| amazing | 0.654619 | toxic | 0.679895 | fantastic | 0.530516 | cowardice | 0.53498 |
| yay | 0.647124 | cringe | 0.669696 | stunning | 0.530093 | suicide | 0.532912 |

## Appendix 2.2: Top 20 tokens predicting polarizing labels: Bertweet-base Model

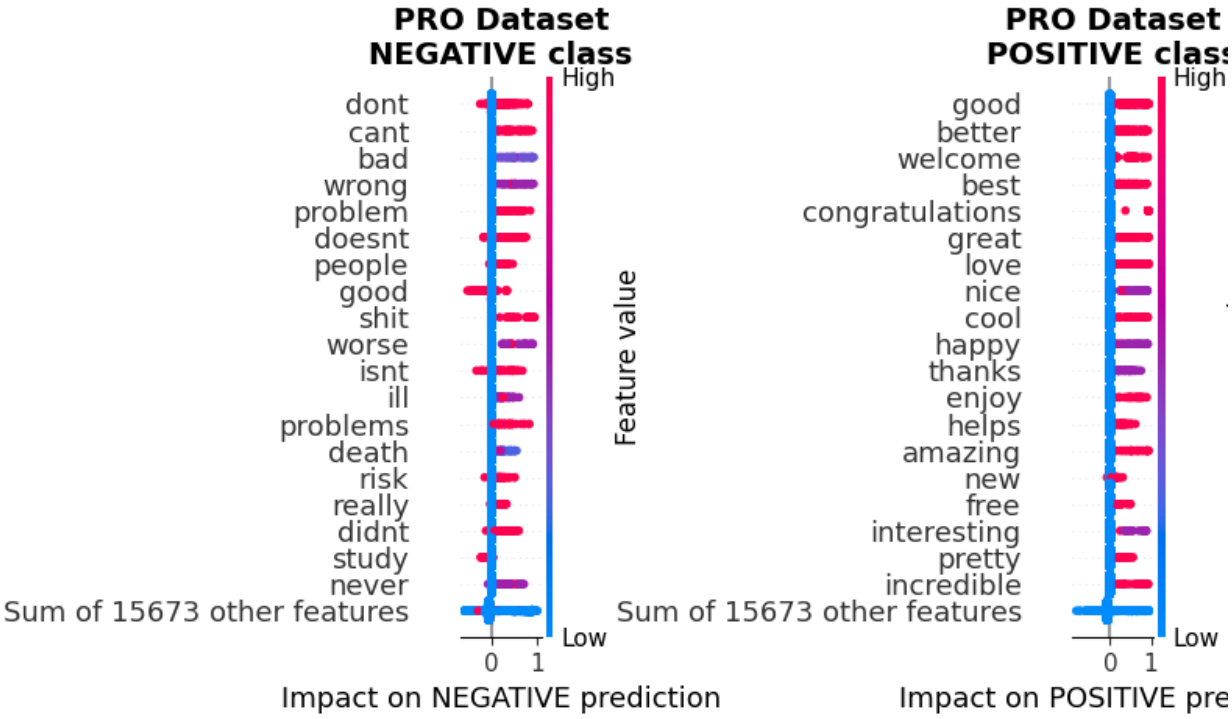
| **Positive feature** | **Mean Shap** | **Negative Feature** | **Mean Shap** | **Positive Feature** | **Mean Shap** | **Negative Feature** | **Mean Shap** |
|---|---|---|---|---|---|---|---|
| lucky | 0.937570 | crapshoot | 0.949790 | enjoy | 0.911395 | terrible | 0.951067 |
| congratulations | 0.936437 | sucks | 0.949719 | pleased | 0.910270 | retard | 0.929044 |
| impressed | 0.924642 | fuck | 0.946441 | loving | 0.900675 | dumbest | 0.923225 |
| proud | 0.921791 | shitting | 0.946339 | better | 0.858524 | disastrous | 0.913929 |
| sweet | 0.904331 | miserable | 0.933623 | smiles | 0.847337 | genocide | 0.913806 |
| amazingt | 0.899627 | terrible | 0.929947 | welcomes | 0.845239 | babykillers | 0.896380 |
| congrats | 0.896143 | suck | 0.924983 | rejoiced | 0.743534 | fraud | 0.880659 |
| congrats | 0.892491 | awful | 0.921612 | powerful | 0.742407 | alarming | 0.856069 |
| favorite | 0.880586 | lousy | 0.921178 | impressive | 0.732446 | diedsuddenly | 0.848860 |
| adorable | 0.875280 | fieldworkfail | 0.919701 | yay | 0.731197 | fiasco | 0.833591 |
| grateful | 0.874395 | unsustainable | 0.916284 | loooove | 0.727871 | pranksters | 0.832084 |
| banggood | 0.874294 | disgusting | 0.914893 | interesting | 0.723045 | laughable | 0.828146 |
| miraculous | 0.862329 | horrifying | 0.914520 | bravely | 0.706284 | abomination | 0.827135 |
| cute | 0.861257 | dumbass | 0.913687 | holyyyy | 0.699404 | illegal | 0.813145 |
| interesting | 0.855965 | fucked | 0.912953 | excited | 0.694914 | destroy | 0.793587 |
| attractive | 0.841380 | whack | 0.911274 | invencible | 0.688607 | inaccurate | 0.792296 |
| welcomed | 0.840065 | unfortunately | 0.908462 | woohoo | 0.685652 | satanists | 0.791240 |
| mathisfun | 0.830184 | miseries | 0.899782 | fun | 0.679726 | jerks | 0.782352 |

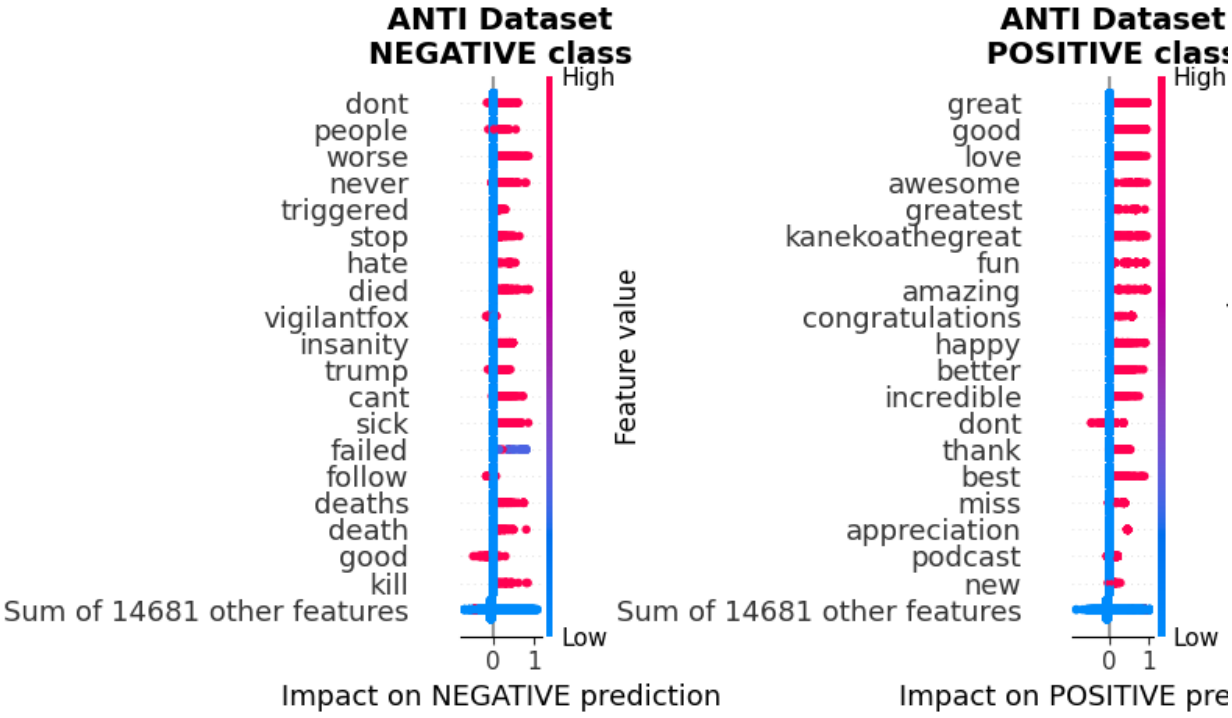
| | | | | | | | |
|---|---|---|---|---|---|---|---|
| fantastic | 0.809544 | abusing | 0.895160 | happy | 0.676081 | worse | 0.779124 |
| perfectly | 0.798457 | shite | 0.894095 | beginningbrighter | 0.665395 | ruining | 0.773678 |

## Appendix 3.1: Beeswarm plots: Roberta-Base Model



**Appendix 3.2: Beeswarm plots: Berta-Base Model**



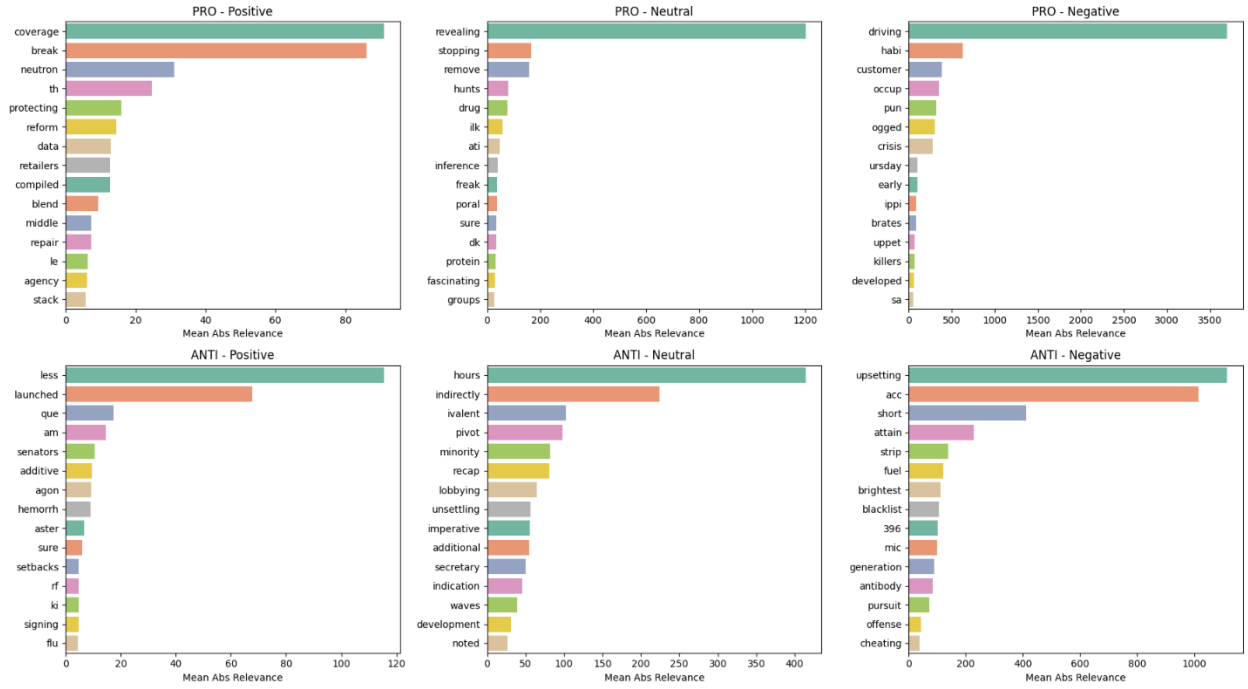
**Appendix 4.1: pairwise comparison in between Original Shap vector and each of perturbed data Shap vectors-Roberta-Base Model**

| Perturbation | Pro | | | Anti | | |
|---|---|---|---|---|---|---|
| | t-value | p-value | Cohen's d | t-value | p-value | Cohen's d |
| Deletion | 3.343 | 0.0008 | 0.047 | 6.973 | 0.0000 | 0.099 |
| Shuffle | 2.364 | 0.0181 | 0.033 | 11.021 | 0.0000 | 0.156 |
| Synonym | 4.479 | 0.0000 | 0.063 | 6.307 | 0.0000 | 0.089 |

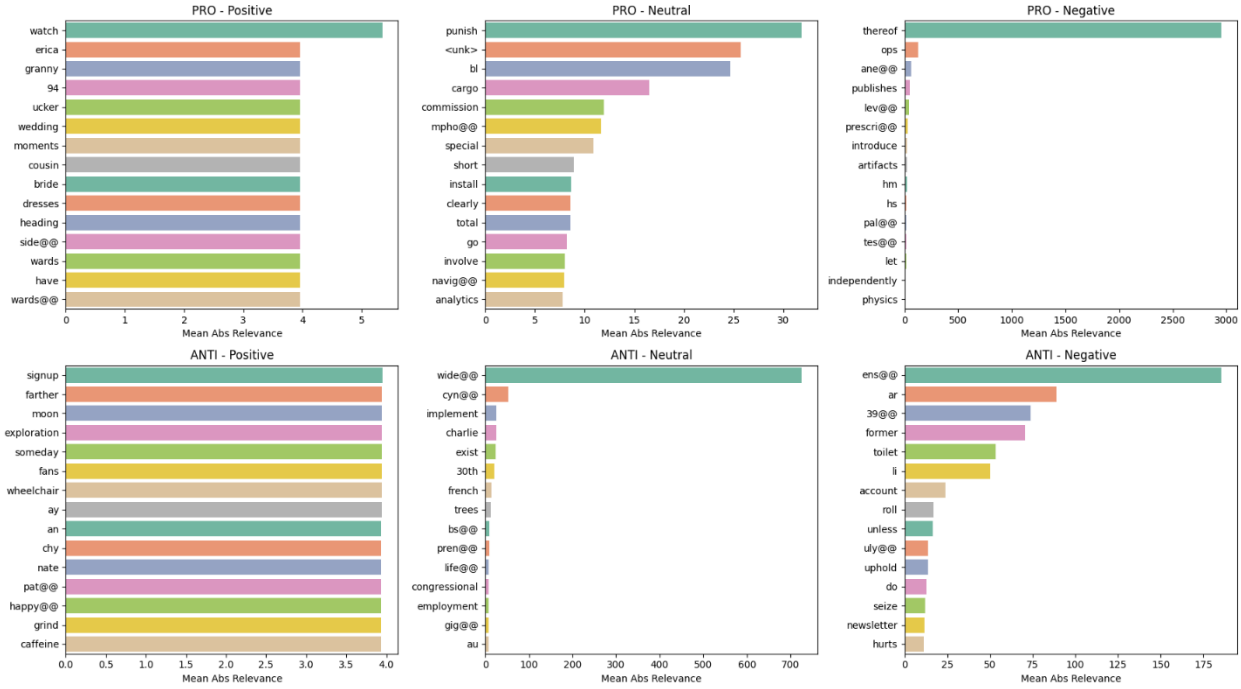
**Appendix 4.2: Pairwise comparison in between Original Shap vector and each of perturbed data Shap vectors-Bertweet-Base Model**

| Perturbation | Pro | | | Anti | | |
|---|---|---|---|---|---|---|
| | t-value | p-value | Cohen's d | t-value | p-value | Cohen's d |
| Deletion | 0.016 | 0.9870 | 0.000 | 3.198 | 0.0014 | 0.045 |
| Shuffle | 2.074 | 0.0381 | 0.029 | 8.625 | 0.0000 | 0.122 |
| Synonym | 3.415 | 0.0006 | 0.048 | 5.309 | 0.0000 | 0.075 |

**Appendix 5.1: Dominant features by the LRP simplified rule-Roberta-Base**



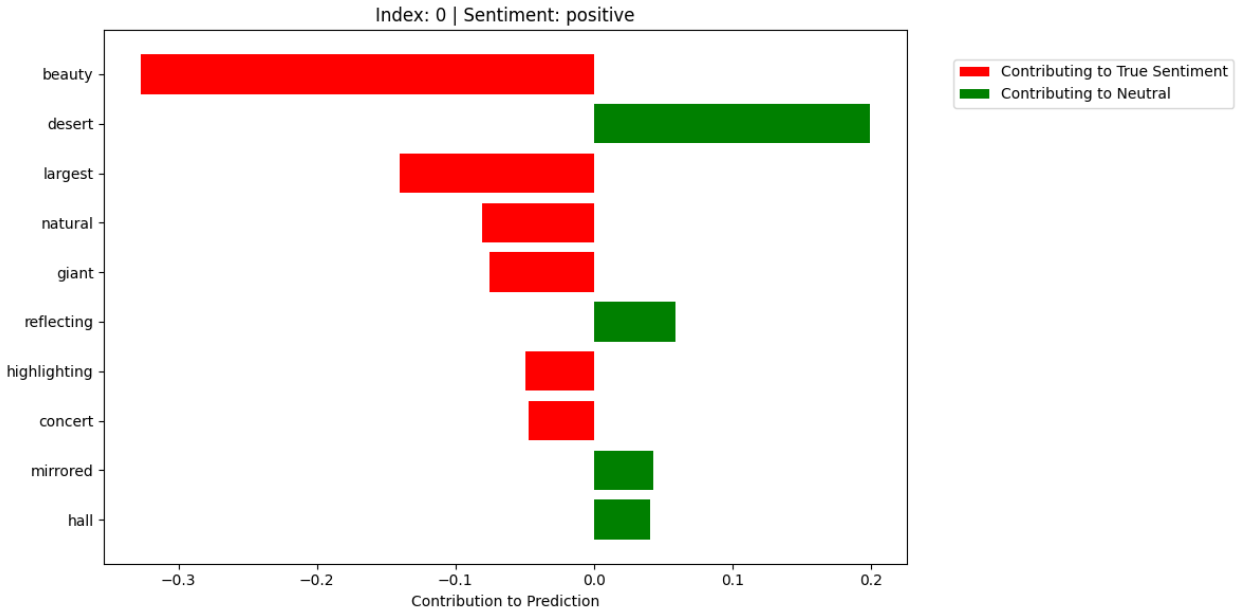
## Appendix 5.2: Dominant features by the LRP simplified Bertweet-Base



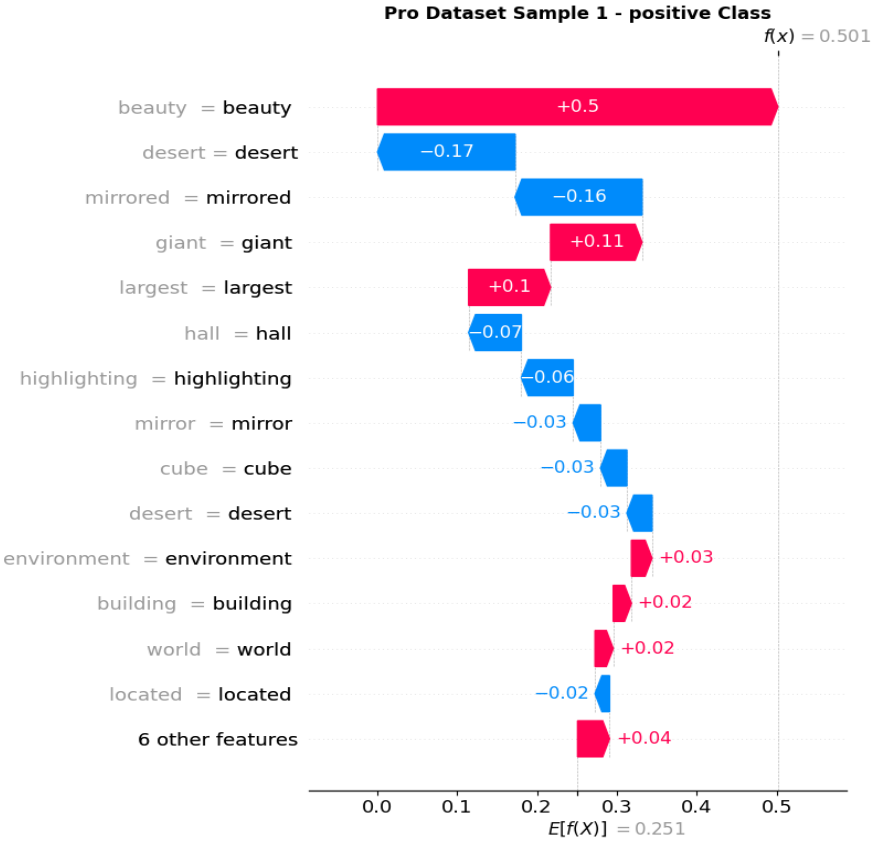

## Appendix 5.3: A comparative LRP global metrics of Sentiment classes

| Sentiment | Dataset | Mean Absolute Relevance | Sparsity Score | Entropy Score |
|---|---|---|---|---|
| **Roberta-Base** | | | | |
| Positive | Pro | 2.8449 | 0.0001 | 7.1882 |
| Negative | Pro | 1.6997 | 0.0004 | 10.5638 |
| Positive | Anti | 1.9799 | 0.0004 | 10.3483 |
| Negative | Anti | 1.9543 | 0.0002 | 10.7689 |
| **Bertweet-Base** | | | | |
| **Sentiment** | **Dataset** | **Mean Absolute Relevance** | **Sparsity Score** | **Entropy Score** |
| Positive | Pro | 2.3552 | 0.0001 | 9.9265 |
| Negative | Pro | 2.5380 | 0.9820 | 10.4642 |
| Positive | Anti | 2.4930 | 0.0005 | 10.7850 |
| Negative | Anti | 2.5193 | 0.9800 | 10.7759 |

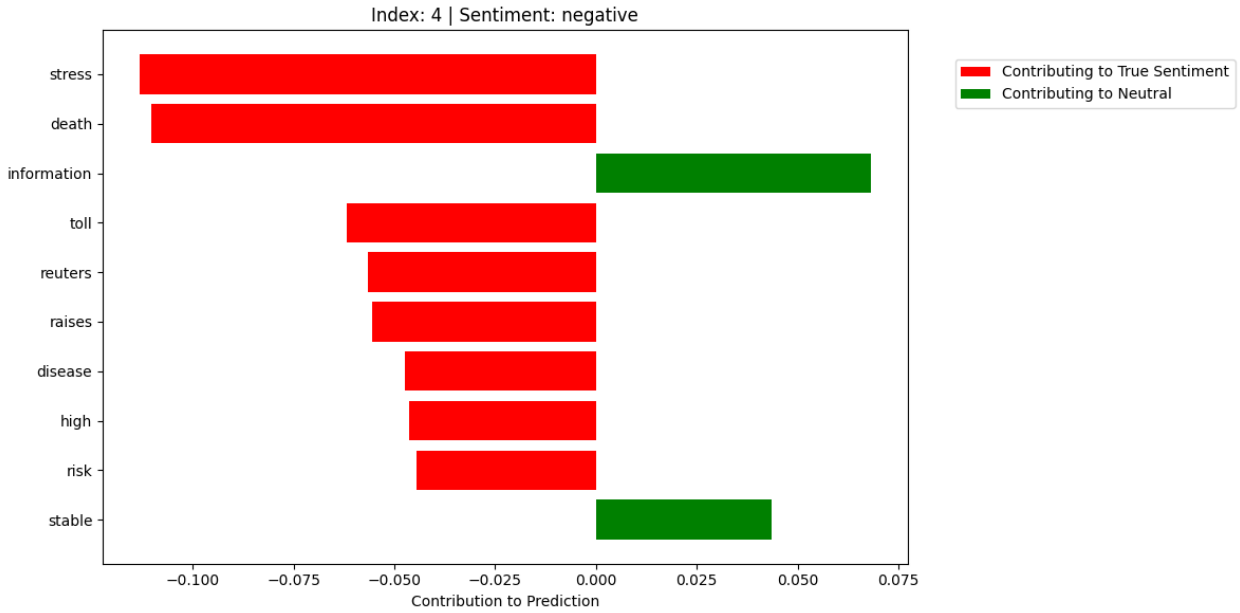
## Appendix 6.1: LIME local explanation on Roberta-Base Model-Positive Sentiment



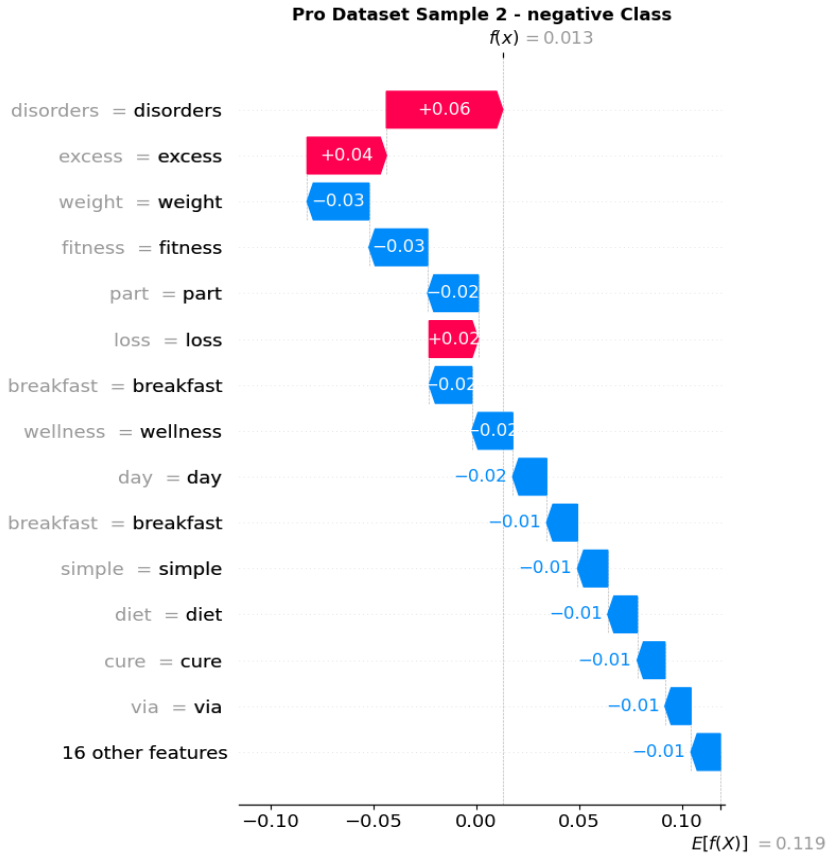

## Appendix 6.2: Shap local explanation on Roberta-Base Model- Positive Sentiment



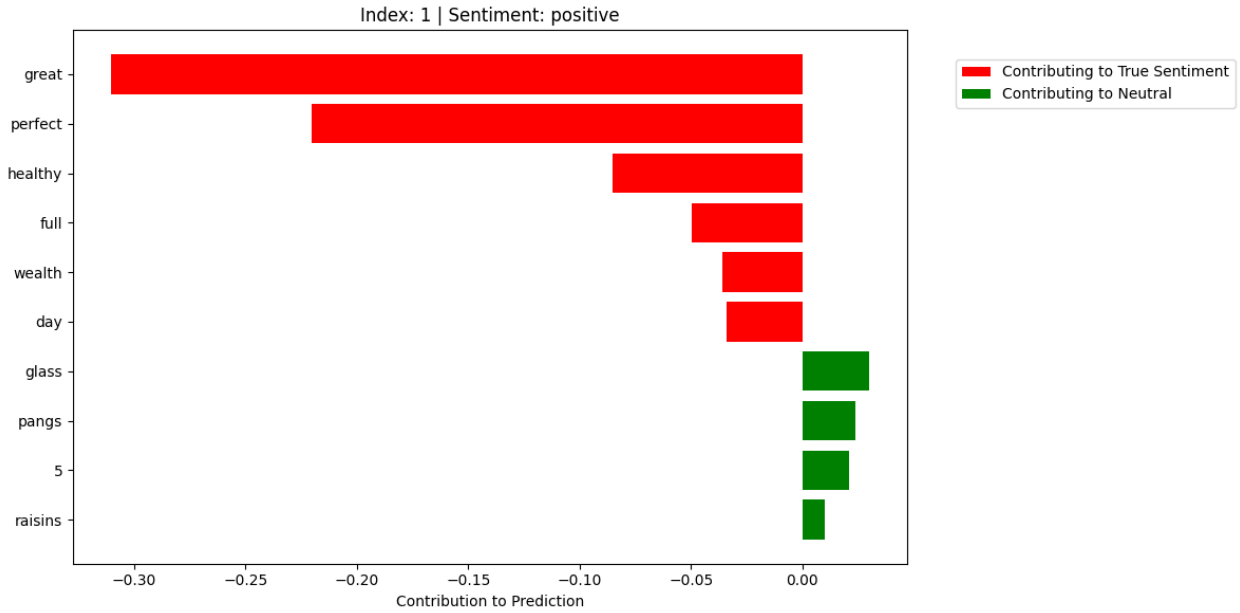
## Appendix 6.3: LIME local explanation on Roberta-Base Model-Negative Sentiment



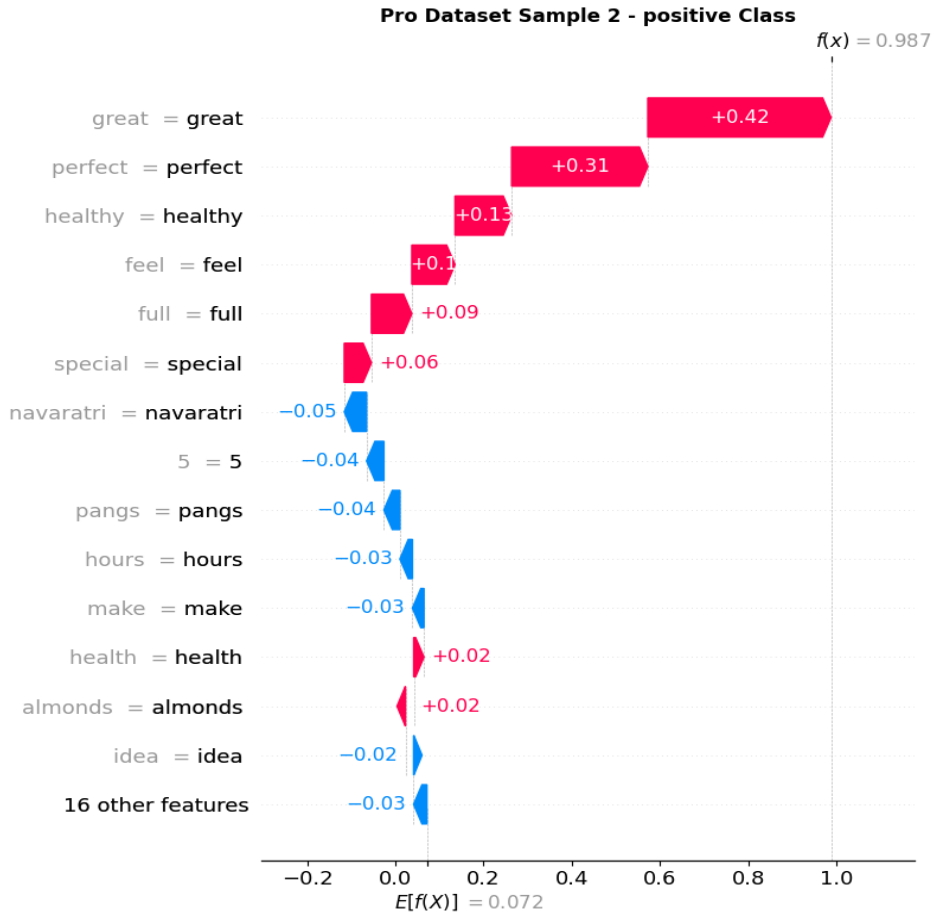

## Appendix 6.4: Shap local explanation on Roberta-Base Model-Negative Sentiment



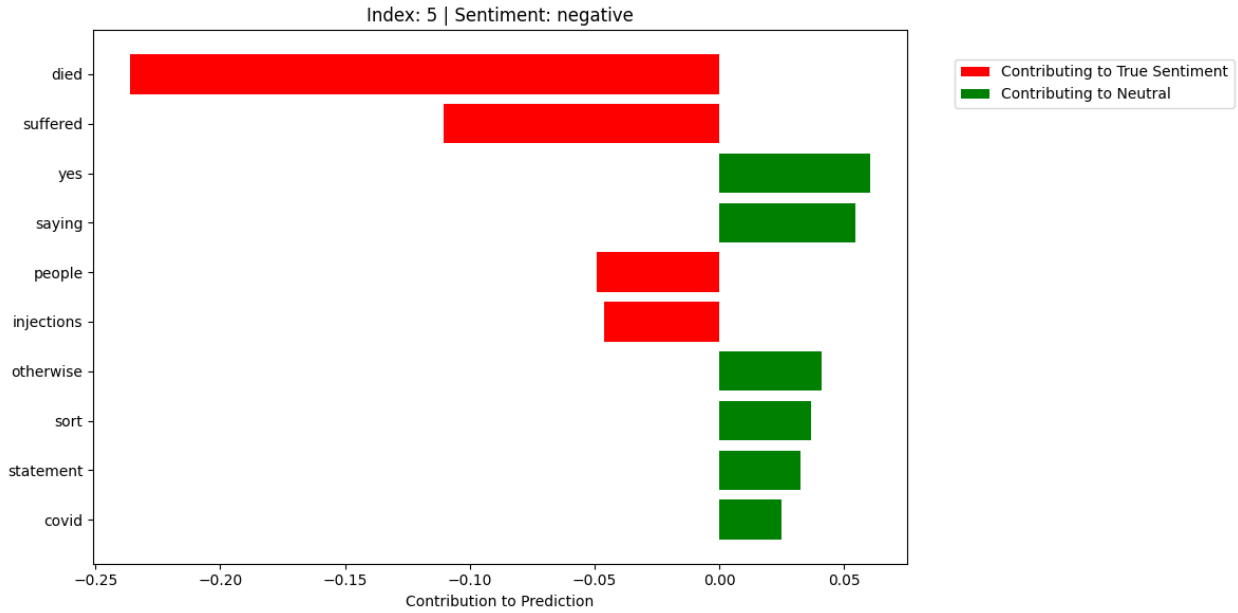
## Appendix 6.5: LIME local explanation on Bertweet-Base Model-Positive Sentiment



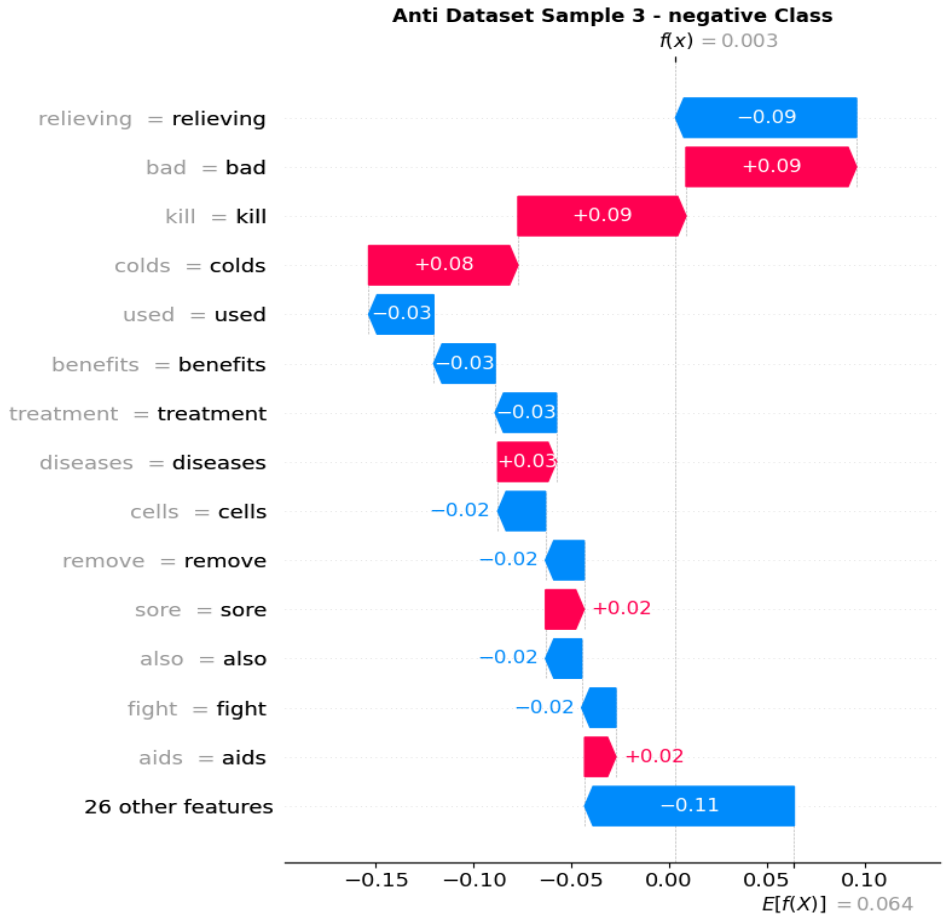

## Appendix 6.6: Shap local explanation on Bertweet-Base Model-Positive Sentiment



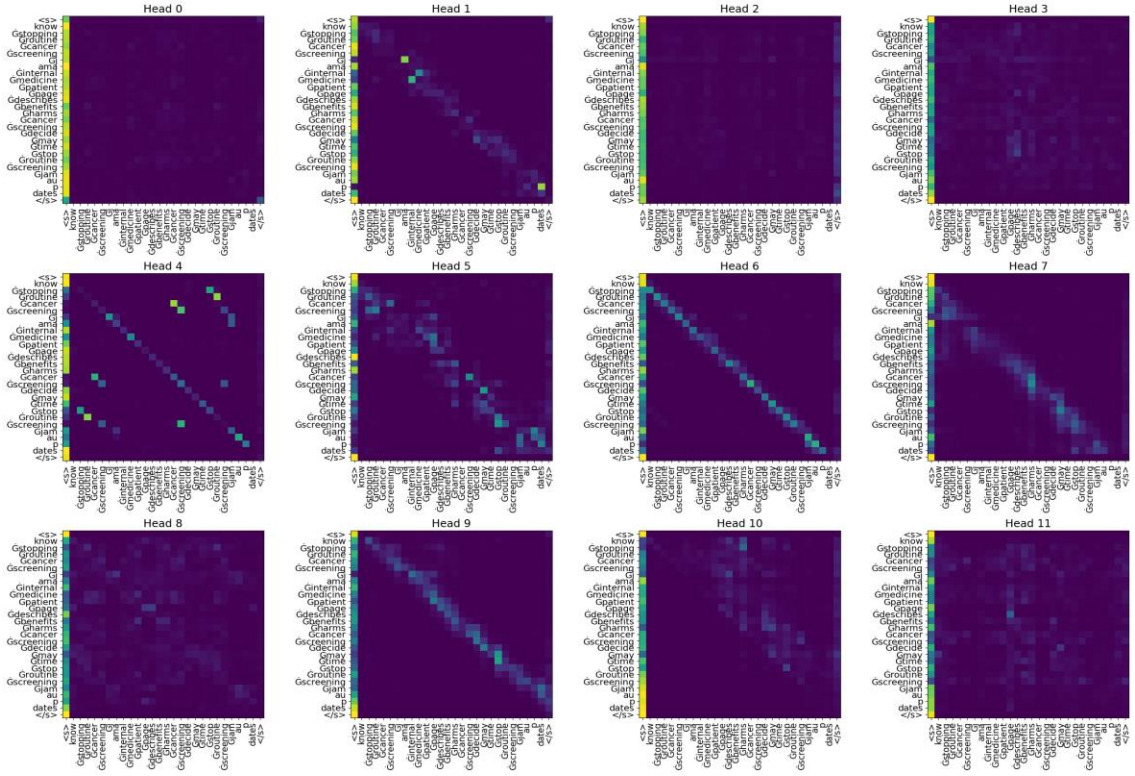
## Appendix 6.7: LIME local explanation on Bertweet-Base Model-Negative Sentiment



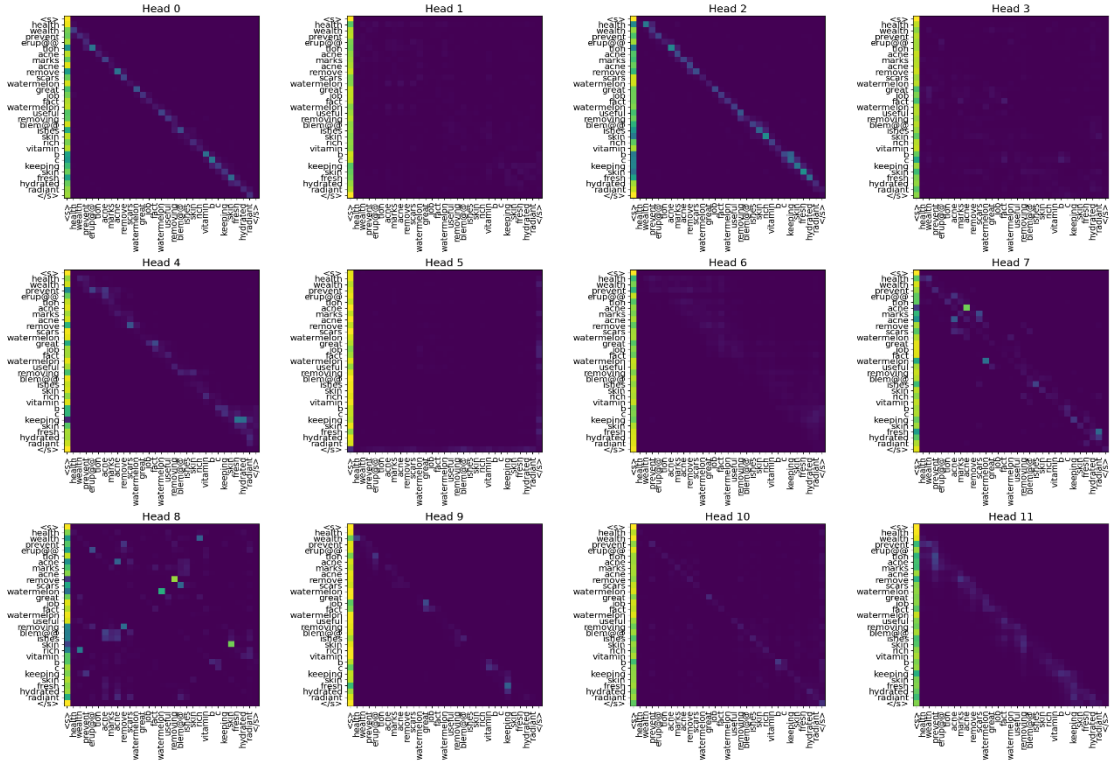

## Appendix 6.8: Shap local explanation on Bertweet-Base Model-Negative Sentiment



## Appendix 7.1: Heads of Layer 0: Roberta Base




## Appendix 7.2: Heads of Layer 0: Bertweet Base Example 1



## Appendix 7.3: Dominant layers