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 has emerged as a critical challenge (Arbatli and Rosenberg, 2021). To study this phenomenon, researchers have leveraged social media as a proxy for broader societal dynamics, employing methodologies including surveys (Bail et al., 2018) and controlled experiments (Levy, 2021). Analyzing unstructured social media text data has also emerged as a prominent methodology (Pereira et al., 2025). Through automated text analysis, researchers measure polarization as constructs using both machine learning (ML) (Vaswani et al., 2017) and rule-based lexicon models (Garzón-Velandia and Pennebaker, 2025; Susanto et al., 2025). This line of research carries implications, as social media serves as a primary source of information and platform to express opinions and perspectives (Levy, 2021). The sentiment classification of text (using both model) as positive and negative to measure polarization is also grounded in the theory of polarization measurement (Esteban and Ray, 1994; Lu and Lee, 2025). However, methodological choice between rule-based and lexicons and ML model such as transformer involves a fundamental trade-off between interpretability and performance (Rudin, 2019). Rule-based and lexicon models, while inherently interpretable, measure polarization statically, failing to capture the contextual nuances in text data (Humphreys and Wang, 2018). On the other hand, transformer predict sentiment classes more effectively, leading to precise and nuanced polarization measurement (Alaparthi and Mishra, 2021). While interpretability is increasingly recognized as a context-dependent and multidimensional concept (Lipton, 2018), polarization researchers face a distinct challenge: transformer-based models lack the transparency and replicability that polarization research demands, where accurate predictions alone are insufficient and explanations must withstand public scrutiny and inform democratic deliberation.

Recent advances in explainable artificial intelligence (XAI) offer a principled solution to the opacity of transformer-based polarization models (Shrivastava et al., 2024). Post-hoc XAI methods yield both global and local explanations, elucidating overall model behavior and the reasoning behind individual predictions by identifying influential features, dominant layers, and critical input segments. Despite this promise, several gaps remain. First, systematic evaluation of XAI is needed to establish rigorous evidence of model interpretability. Second, a comparative analysis of local versus global techniques is warranted to guide researchers in selecting the appropriate approach. Third, a benchmarking of transformer-based against rule-based and lexicon models on a gold standard dataset remains needed. Finally, the perceived usefulness of XAI outputs from a human perspective remains underexplored in research and policy contexts. To address these gaps, this study examines three research questions: *RQ1*: How effectively can post-hoc XAI techniques establish model explainability, and do local or global explanations yield more reliable insights? *RQ2*: Do transformer-based sentiment models augmented with XAI techniques outperform rule-based and lexicon models in sentiment prediction and polarization measurement? *RQ3*:

Does perceived explainability derived from post-hoc XAI techniques enhance user trust, satisfaction, and perceived usefulness?

2 Review of Litreature

2.1 Sentiment Analysis and Polarization

Sentiment analysis has become a powerful tool for conducting polarization studies (Kenyon-Dean et al., 2018; Xiang and Zhou, 2014) grounded in classical conceptualizations of polarization as the clustering of opinions in vector of opposing extremes (Esteban and Ray, 1994).

2.2 Measuring polarization

In measuring polarization, Rule-based and lexicon models use 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, TextBlob, Valence Aware Dictionary and sEntiment Reasoner (VADER) (Loria et al., 2019; Aljedaani et al., 2022; Nielsen, 2011; Hutto and Gilbert, 2014). Consistently, transformer ML models have also increase in polarization studies as it overcome key limitations of lexicon by being trained on large corpora, capturing richer contextual nuances in language. BERT, for instance, employs an encoder-only architecture with self-attention mechanisms, pre-trained via masked language modeling and next sentence prediction (Pangtey et al., 2025; Kang et al., 2024; Devlin et al., 2019), and has been shown to outperform lexicon-based approaches in sentiment classification (Zhang et al., 2021). However, transformers demand substantial computational resources and lack transparency in their decision-making (Leon, 2025).

2.3 XAI-post-Hoc techniques

To enhance model interpretability, various post-hoc XAI techniques can be employed. It include Shapley Additive exPlanations (Shap), a game-theoretic approach (Fryer et al., 2021), where each token is treated as a player in a cooperative coalitional game and the payoff corresponds to the model's prediction (Parisineni and Pal, 2024) for global explainability. In contrast, local interpretability can be achieved using Local Interpretable Model-agnostic Explanations (LIME), which explains individual predictions by generating perturbed input samples and fitting an interpretable surrogate model to approximate local model behavior (Ribeiro et al., 2016; Wang et al., 2025). The other key techniques of XAI at global level include Layer-wise Relevance Propagation (LRP) that interpret model by propagating prediction backward through the network layers, decomposing the prediction into relevance scores for each input feature (Luo et al., 2024) and Attention Mechanism Visualization at local level which illustrate parts of the neural network that receive focus during prediction (Vig, 2019).The BertViz is commonly used attention visualization techniques (Vig, 2019). The use of attention weights as explanations remains contested. (Jain and Wallace, 2019) demonstrated that attention weights do not consistently correlate with gradient-based feature importance measures, questioning their faithfulness as explanations. While (Wiegreffe and Pinter, 2019) partially rebutted this, conceding that attention can be meaningful under certain conditions, attention weights should not be naively interpreted as definitive indicators of feature importance. More broadly, explanation faithfulness has emerged as a central concern in XAI research (Jacovi and Goldberg, 2020), as plausible explanations may not reflect the true computational mechanisms driving predictions and even faithful explanations are of limited value if end-users cannot meaningfully interpret or trust them, underscoring the need to assess perceived explainability, trust, and satisfaction from a human perspective

3 Method and Material

3.1 Selection of Sentiment Models

The present research has selected two open source transformer (blackbox) models and three explainable open source rule-based and lexicons. Table 1 shows details. These models are widely used open access for sentiment analysis and subsequent polarization measurement.

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

Table 1: Sentiment models for measuring polarization.

3.2 Experimental Dataset

This study examines polarization on social media and its measurement from unstructured text data.

Accordingly, data were collected from two distinct platforms of Twitter and Telegram. The Twitter dataset is openly sourced from (Ngo, 2022) consist of 1.6 million tweets. The Telegram datasets was collected firsthand via the Python Telethon application programming interface (API). Both datasets obtain from groups focusing on science-related discourse, covering vaccine and climate change and other topic. The dataset were manually labeled as *pro-science* consist of 1.002 million message or *anti-science* consist of 1.06 million message based on their group descriptions. For XAI and interpretability analysis, a subsample of 5,000 instances (1,667 instances per sentiment) class was drawn from each of three dataset because techniques such as Shap consume higher amount of computational resource.

3.3 XAI techniques and Analysis

This study employs two primary XAI techniques (a) Shap for global and LIME for local explainability. To assess interpretability, three (a) prediction stability via ablation (b) prediction consistency, and (c) prediction quality via perturbation are conduced. The analysis are complemented with LRP for global and attention mechanism visualization for local interpretation.

3.4 Trade-of analysis

We evaluated the prediction performance of black-box and Rule based and lexicon against two human annotated (Guo et al., 2023; Alghamdi and Alhasawi, 2024) and one sentiment benchmark (Maas et al., 2011) dataset using metric of accuracy and weighted Cohen's Kappa with quadratic weighting (Doewes et al., 2023)

3.5 Human Usefulness experiment

Finally, to examine whether XAI explanations influence user perceptions, a between-subjects experimental design was employed. Participants were assigned to either an explanation condition (experimental group) or a no-explanation condition (control group). The variables measured included perceived explainability (Wang and Benbasat, 2016), trust (Hoffman et al., 2023), usefulness (Gefen and Straub, 2000), and satisfaction (Hoffman et al., 2023) using 5-point Likert scale (appendix 8.3). Given the anticipated violation of multivariate normality assumptions, Permutational Multivariate Analysis of Variance (PERMANOVA) was selected over MANOVA to test whether group membership (explanation vs. no-explanation) produced significant differences across the four outcome variables simultaneously. Lastly, we investigated the effect of of perceived explainability on trust, satisfaction, and usefulness specified as separate outcome variables using regression models.

4 Data Analysis

Using each model, sentiment labels were predicted across both datasets. Appendix 1 shows that in the Telegram dataset, both pro-science (Pro-S) and anti-science (Anti-S) groups are dominated by neutral sentiment predictions. In contrast, in the Twitter dataset, rule-based and lexicon-based models more frequently predict positive sentiment, whereas transformer-based models produce a more balanced distribution across sentiment classes.

Inter-model agreement (see Table 2), measured using Cohen's Kappa via the scikit-learn Python library, shows mixed patterns across model types and datasets. In the Telegram dataset, lexicon-based models (VADER and Afinn) exhibit stronger agreement with each other, achieving higher Cohen's Kappa scores of 0.761 (pro-science) and 0.780 (anti-science), respectively. In contrast, in the Twitter dataset, transformer-based models show higher inter-model agreement, with a Cohen's Kappa of 0.761. This pattern is further supported by Fleiss' Kappa, where lexicon- and rule-based models show higher agreement in Telegram, whereas transformer-based models achieve higher agreement in Twitter.

Model 1	Model 2	Pro-S	Anti-S	twitter
Roberta	Bertweet	0.457	0.428	**0.760**
Vader	TextBlob	0.420	0.490	0.484
Vader	Afinn	**0.761**	**0.780**	0.752
TextBlob	Afinn	0.432	0.478	0.498
Fleiss-Kappa (Models)		0.446	0.399	**0.759**
Fleiss-Kappa (Lexicons)		0.539	**0.580**	0.575

Table 2: Inter-model agreement-Cohen Kappa.

4.1 Average Model Confidence

Prior to presenting the XAI interpretability results, Table 3 reports the average model confidence and token count required to predict positive and negative sentiment classes for each model. Token count serves as a relative indicator of computational efficiency, with results showing that Bertweet requires fewer tokens on average than Roberta. Furthermore, Bertweet exhibits higher model confidence

(>0.80) compared to Roberta (<0.80), further corroborating its superior predictive performance.

Dataset	Roberta			Bertweet		
	Conf.	Tok (+)	Tok (-)	Conf.	Tok (+)	Tok (-)
Pro-S	0.731	10.56	9.59	**0.822**	7.25	8.09
Anti-S	0.746	13.11	15.29	**0.828**	9.31	14.04
Twitter	0.760	10.09	9.62	**0.852**	7.14	8.29

Table 3: Model confidence and Average token.

4.2 Shap: Global Explainability

Global interpretability analysis was conducted by computing aggregated Shapley values to identify the most influential features driving model predictions across sentiment classes. The top 20 features for each dataset are reported in Appendices 2.1–2.3. To validate whether these influential features reflect a polarizing tone, we computed their polarity scores using SentiWordNet (Baccianella et al., 2010). This step serves as a lexical grounding check, ensuring that the Shap-identified features align with established sentiment lexicons (see Table 4).

We also present the Beeswarm plots in the appendix 3.1 (for Roberta-base) and 3.2 (for Bert-base) which is an special XAI visualization that helps to explain the model's interpretability by visualizing dominant (top-K) and other features that impact the model decision to predict a particular class of positive and negative sentiment

Token Group	Roberta	Bertweet
Pro-S Positive Shap	0.351	0.265
Anti-S Positive Shap	0.287	0.191
Pro-S Negative Shap	-0.202	-0.282
Anti-S Negative Shap	-0.219	-0.176
Twitter Positive	0.146	0.185
Twitter Negative	-0.250	-0.155

Table 4: Mean sentiment scores on Shap feature.

4.2.1 Stability: Ablation test

We assess interpretability via a stability-based ablation test. Specifically, we validate the importance of features identified by SHap as influential for each sentiment class. We remove these tokens and re-evaluate the model, measuring the resulting changes in predicted class labels and confidence scores. Table 6 presents the results, showing the marginal effect of removing Shap-predicted features on model predictions. Overall, the results suggest that both models are robust to feature removal.

These results are consistent with the established notion of transformer models' ability to understand and classify language contextually, such that removing these features does not significantly impact the model's overall confidence in understanding and classifying the text. This finding resonates with (Hooker et al., 2019), who demonstrated that removing ostensibly important features often fails to degrade model performance, a phenomenon termed the *masking problem*. Our stability analysis suggests that for transformer-based sentiment classification, holistic contextual representations are more meaningful than individual token-level importance, with direct implications for how XAI outputs should be interpreted in polarization research.

Model	Global	Class-Wise	TF-IDF	Flip
Roberta (Pro)	0.7296	0.7288	0.7312	0.0270
Roberta (Anti)	0.7448	0.7445	0.7464	0.0208
Roberta (Twitter)	0.7595	0.7592	0.7604	**0.0064**
Bertweet (Pro)	0.8297	0.8213	0.8220	0.0158
Bertweet (Anti)	0.8213	0.8260	0.8283	0.0322
Bertweet (Twitter)	0.8520	0.8518	0.8527	0.0142

Table 5: Ablation-based stability analysis

4.2.2 Consistency:

A consistency analysis was conducted to examine whether semantically similar inputs yield comparable Shap explanations. Sentence embeddings were generated using all-MiniLM-L6-v2 (Reimers et al., 2024) and using cosine similarity, 300 similar sentence were extracted. The Mean cosine similarity of embeddings was 0.7347 and 0.7383 for RoBERTa and BERTweet respectively on the Telegram dataset, and 0.6655 and 0.6814 on the Twitter dataset. To assess whether semantic similarity translates to consistency in SHAP explanations, Jensen-Shannon Divergence (JSD) (Menéndez et al., 1997) was computed and corroborated using the Kolmogorov-Smirnov (KS) test (Massey Jr., 1951). Results, presented in Table 6, indicate that semantically similar Telegram messages yield more stable Shap attribution patterns than Twitter data, where higher JSD and KS values reflect greater explanatory variability. Negative sentiment classes consistently exhibited lower JSD scores than positive classes, suggesting more coherent and stable explanations for negative content. Finally, marginal differences between Roberta and Bertweet indicate broadly comparable explanation consistency across both architectures.

Finally, we assessed whether embedding similarity also leads to similar Shap explanation vectors

Sentiment	Roberta		Bertweet	
	JSD	KS	JSD	KS
Overall-Telegram	0.3390	0.2870	0.3410	0.2820
Positive-Telegram	0.4060	0.2250	0.4250	0.2240
Negative-Telegram	0.2710	0.2950	0.2700	0.3000
Overall-Twitter	0.3626	0.3687	0.3797	0.3543
Positive-Twitter	0.4346	0.3643	0.4847	0.3541
Negative-Twitter	0.3116	0.3629	0.3089	0.3673

Table 6: Embedding and Explanation Similarity.

using Spearman correlation analysis. The results presented in Table 7 reveal mixed findings. Specifically, significant Spearman correlations were observed between embedding similarity and Shap explanation vectors for the Telegram dataset, suggesting that semantically similar Telegram messages tend to generate more consistent explanation patterns. However, the correlations for the Twitter dataset were found to be statistically insignificant, indicating that semantic similarity in tweets does not necessarily translate into stable or aligned Shap attribution vectors.

Models	Telegram-Dataset		Twitter-Dataset	
	p	r	p	r
Roberta	1.16e-05	0.251	0.2268	0.074
Bertweet	0.002225	0.176	0.1087	0.093

Table 7: Embedding and Shap Vectors Correlation.

4.2.3 Explanations Quality: Perturbation Test

Finally to assess the quality and robustness of Shap explanation, we undertaken perturbation test using random token deletion, shuffling, and synonym replacement. In random deletion, 10 percent 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.

Table 8 presents the results of model confidence under perturbation settings. The relatively small changes in confidence scores across all datasets suggest that both Roberta and Bertweet remain comparatively robust in maintaining their predictions despite perturbations, indicating stable and reliable predictive behavior.

We compute the flip rate to quantify prediction stability under input perturbations. Specifically, for each model, we compare the original predicted label with the label obtained after applying a perturbation (i.e., deletion, shuffling, or synonym replacement). The results shown in table 9 suggest that a consistent pattern. First, In all settings, shuffle-based perturbations yield the highest flip rates (28–31 percent). In contrast, synonym replacement consistently produces the lowest flip rates (6–10 percent), while deletion-based perturbations show intermediate sensitivity. Importantly, this pattern remains stable across datasets and models, indicating that both architectures are primarily sensitive to syntactic structure rather than isolated lexical changes.

Model	Original	Deletion	Shuffle	Synonym
Roberta-(Pro-S)	0.731	0.731	0.726	0.727
Roberta-(Anti-S)	0.746	0.743	0.730	0.741
Roberta-(Twitter)	0.760	0.754	0.732	0.755
Bertweet-(Pro-S)	0.822	0.825	0.827	0.821
Bertweet-(Anti-S)	0.828	0.829	0.825	0.825
Bertweet-(Twitter)	0.852	0.850	0.849	0.850

Table 8: Model performance under perturbations.

	Roberta			Bertweet		
	Twitter	Pro-S	Anti-S	Twitter	Pro-S	Anti-S
Delete	12.00	13.34	10.84	12.02	13.88	10.28
Shuffle	28.80	30.32	28.28	30.54	31.36	31.44
Synonym	6.84	9.66	7.96	6.12	9.22	8.30

Table 9: Prediction Change: Perturbation.

Finally, we evaluated the stability of Shap values by computing explanations for each perturbed dataset and conducting 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 (especially for Telegram Pro and Anti-Science dataset), indicating that perturbations lead to significant changes in Shap value distributions across both models and datasets. Despite statistical significance, the corresponding t-statistics and Cohen's d effect sizes suggest that the magnitude of these changes is moderate. However, the relatively limited effect sizes also suggest that transformer-based models retain a degree of contextual robustness despite surface-level changes.

4.3 Local XAI Analysis

The present research further incorporates local XAI techniques to investigate and interpret model behavior at the individual prediction level. In this regard, we incorporated local explanations offered by both LIME, and Shap. These local XAI techniques are primarily visualization-based and enable the identification of the most influential features contributing to a specific class prediction at the

individual instance level

To generate local explanations, we selected a random subset of 15 instances that were already classified by both the Roberta and Bertweet models. The sample is balanced across sentiment classes, ensuring equal representation of positive. The results of the local-level analysis are presented in Appendices 5.1–5.16, where we provide both Shap and LIME visualizations for individual instances across sentiment classes and datasets, including Twitter and Telegram.

The key results suggest that both Shap and LIME accurately and consistently explain model's decision. For instance, appendices 5.1 (LIME) and 5.2 (Shap) identify features like *beauty*, *largest*, and *giants* as key drivers of positive sentiment predictions for Roberta model. Similarly appendix 5.15 (LIME) and 5.16 (Shap) for Bertweet model on Twitter dataset also suggest similar pattern where features such as *totally*, *gig*, and *another* are predicted as key feature leading model to predict positive sentiment.

These results show the effectiveness of both local techniques in enhancing model interpretability at the instance level. By providing consistent and converging explanations, local XAI analysis demonstrates that the models rely on handful of feature in the sentence to form predictions. Finally, alignment between both explanation methods across different datasets strengthens explanation claims.

4.4 complimentary Analysis

To further enhance the interpretability of the sentiment models, we complemented our global XAI with LRP with local through attention mechanism visualization.

4.4.1 LRP

A simplified LRP method was employed, focusing on the classification head of the model. Relevance scores are computed in three steps: (a) the predicted logit is calculated, (b) propagated backward to the linear classification head using the Z-LRP rule, and (c) relevance values are assigned to each feature proportional to its contribution to the final prediction. Dominant features and their relevance scores for each sentiment class are presented in Appendices 6.1-6.4. Three key findings emerge from the global class-wise LRP analysis. First, LRP identifies distinct and meaningful sentiment-bearing features, with words such as *poorly*, *burning*, *sink*, *detrimental*, *shooting*, and *spam* serving as strong negative stance indicators. Second, these features complement SHAP-identified features at the global level (Appendices 3.1-3.4), together providing a holistic understanding of model behavior in sentiment prediction. Third, three attribution metrics are computed (Table 10, Appendix 6.5): mean absolute relevance, assessing average token contribution strength; sparsity, measuring the fraction of highly relevant tokens; and entropy, capturing token concentration. Results indicate comparable token contribution strength across both models, with BERTweet marginally higher. Sparsity values close to zero indicate that relevance is distributed broadly across tokens rather than concentrated on a subset, while high entropy values corroborate this distributed attribution pattern across both models. Overall, LRP enhances global interpretability by quantifying token contribution strength, distribution, and concentration across datasets and models. These attribution metrics complement the SHAP analysis by providing additional insight into explanation dispersion, and together with consistency analysis, ablation testing, and perturbation experiments, constitute a comprehensive framework for evaluating sentiment model reasoning.

	Roberta			**Bertweet**		
	Pro-S	**Anti-S**	**Twitter**	**Pro-S**	**Anti-S**	**Twitter**
Relevance	2.01	1.77	1.8646	2.38	2.33	2.4608
Sparsity	0.00	0.00	1.8646	0.001	0.000	0.0009
Entropy	10.0250	11.50	1.864	11.34	11.78	11.1892

Table 10: Global explainability metrics: LRP.

4.4.2 Attention Mechanism Visualization

We complemented our local XAI analysis with attention mechanism visualization. This technique enables the visualization of layers and heads of the neural network process textual data during prediction. For example, in Appendix 7.1, we identify the dominant neural network layers activated while processing individual instances across different datasets and models. Similarly, Appendices 7.2–7.5 present the mean attention across neural network heads at both the overall and sentiment-specific levels. Finally, Appendix 7.6 provides a visualization of attention heads across all layers of the neural network. Collectively, these analyses provide additional insights into the model's local behavior and internal processing dynamics. The single-instance examples presented in these appendices demonstrate the relatively balanced and integrated contribution of multiple neural network

layers during prediction generation. Therefore, the attention-based analysis complements our local XAI investigation using LIME by offering an additional perspective on how the models internally distribute attention while processing sentiment-related information.

4.5 Tradeoff: Rule Based Vs ML Model

Following the XAI analysis, transformer were evaluated against lexicon-based models on two human-annotated datasets (Guo et al., 2023; Alghamdi and Alhasawi, 2024), and the benchmark sentiment dataset (Maas et al., 2011) using accuracy and Cohen's Kappa.

The Accuracy measures the proportion of instances correctly predicted by each rule-based and transformer-based model in agreement with the labels in human-annotated and benchmark datasets. The results reported in Table 11 yield three key insights. First, both transformer-based models consistently outperform rule-based and lexicon models in terms of classification accuracy, confirming their superior performance. Second, Bertweet demonstrates stronger performance on the aspect-based (Guo et al., 2023) and the Amazon product review dataset sentiment dataset (Alghamdi and Alhasawi, 2024), while it slightly underperforms compared to Roberta on the benchmark sentiment dataset (Maas et al., 2011). This suggests that Bertweet is particularly well-suited for domain-specific and social media–oriented text, whereas Roberta exhibits greater robustness on general benchmark data. Finally, despite their inherent limitations, rule based models remain competitive in certain settings. Notably, all three lexicons outperform the Bertweet-based model on the benchmark sentiment dataset, while also achieving reasonable performance on the Amazon review dataset, with accuracy levels exceeding 50 percent 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 11: Model performance: Accuracy.

Secondly, we have assessed the performance 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). The results in table 12 show 4 key insights First, transformers consistently achieve higher agreement with human annotations. Second, performance varies across datasets where highest is reported on the Amazon Review, moderate on the Benchmark Sentiment dataset, and lowest on the Aspect-Based sentiment dataset. Third, among the transformer models, Bertweet-base slightly outperforms Roberta across all three datasets. Finally, lexicon-based models (VADER, TextBlob, and Afinn) show substantially lower agreement with human annotations compared to transformer-based models, with Cohen's Kappa values generally below 0.50. Among the 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 12: Model performance: Cohen Kappa

4.6 Human Evaluation Experimental Results

Data were collected from 75 participants (experimental group: 38, control group: 37; male: 31, female: 44). The sample was technically proficient, with 96% having used ML models, 94.66% having trained such models for academic or professional purposes, 93% familiar with sentiment analysis, and 92% familiar with the black-box nature of ML models. Both groups evaluated 10 sentences annotated with sentiment labels. Agreement with assigned labels was comparable across groups, at 63.16% in the control group and 63.36% in the experimental group. In the experimental group, participants were additionally provided with Shap-generated explanations for model predictions, and their mean agreement with these explanations was 55.29%.

We employed a permutation-MANOVA (due to non-normality of data; See appendices 8.1-8.2) to test group difference on group difference across outcome variable (Explainability, Usefulness, Trust and Satisfaction). The PERMANOVA revealed no statistically significant difference between the explanation and no-explanation conditions across the four constructs (F-Statistics = 0.51, p = .608, permutations = 999), suggesting that the mere presence of XAI explanations did not produce a mean-

ingful shift in users' perceived explainability, trust, satisfaction, or usefulness at the multivariate level.

Item	Test	F-Statistic	p-value	Permutations
Value	pseudo-F	0.514051	0.608	999

Table 13: PERMANOVA test results.

The table 14 presents the robust regression (due to non-normality) results for the effect of perceived explainability on trust, satisfaction, and usefulness, controlling for group membership. Perceived explainability emerged as a significant positive predictor across all three outcomes ($p < .001$). The strongest effect was observed for usefulness ($B = 0.62$, $R^2 = 0.44$), followed by trust ($B = 0.44$, $R^2 = 0.34$) and satisfaction ($B = 0.21$, $R^2 = .18$), indicating that users who perceived the XAI explainable consistently reported higher levels of trust, satisfaction, and usefulness.

Outcome	Pseudo R^2	β (Explainability)	p-value
Trust	0.3479	0.4469	0.0000
Satisfaction	0.1816	0.2176	0.0001
Usefulness	0.4410	0.6215	0.0000

Table 14: Regression results.

5 Conclusion

The measurement of polarization from text data using NLP methods (especially using ML) offers opportunities for policy-relevant research. However, the black-box nature of ML models raises concerns of reproducibility. The experimental results of this study demonstrate that XAI techniques can effectively be employed to interpret model behavior and ultimately strengthening the reproducibility of research outcomes. Our results suggest that, global XAI analysis is more robust in inferring the model behavior. Although in absence of standard for inferring, our experimental techniques at global and local level attempted to understand model behavior.

Our results have confirmed that, Shap extracted features (complimented with LRP) validated using SentiWordNet correspond to true sentiment of polarizing positive and negative sentiment. The ablation test demonstrate their importance for model processing of the data. Our consistency analysis demonstrates that Shap-identified features are not isolated signals; rather, they exhibit significant correlations with contextual embeddings, suggesting that model explanations are embedded within broader semantic representations. Finally, perturbation further strengthen importance of features and their contextual position. In contrast, local explanation effective at individual instance may not optimal way to infer the model interpretability as compared to the global explainability. In this regard, Our results suggest a 3 practical workflow for polarization user using transformer to measure polarization (1) report global feature importance using Shap or LRP to demonstrate that the model attends to sentiment-relevant vocabulary, (2) conduct stability, consistency and quality 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).

Responding to needs to human understanding and usefulness, an experiment was conducted. Although the explanation and no-explanation conditions did not differ significantly at the multivariate level, perceived explainability emerged as a significant positive predictor of usefulness, trust, and satisfaction, suggesting that explanation quality matters more than explanation exposure alone in XAI system design. This demonstrate the importance of XAI for stakeholder in decision making.

6 Limitations

We report four key limitations. First, due to lack of guidelines for the systematic evaluation of model explainability, we mostly relied on best practices from prior literature. Second, the selection of black-box models was limited to open-access approaches; therefore, architectures such as recurrent neural networks, as well as proprietary or closed-source models, were excluded from 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. 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 that do not carry independent sentiment valence.

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.

David Gefen and Detmar W Straub. 2000. The relative importance of perceived ease of use in is adoption: A study of e-commerce adoption. *Journal of the association for Information Systems*, 1(1):8.

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.

Robert R Hoffman, Shane T Mueller, Gary Klein, and Jordan Litman. 2023. Measures for explainable ai: Explanation goodness, user satisfaction, mental models, curiosity, trust, and human-ai performance. *Frontiers in Computer Science*, 5:1096257.

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.

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.

Vu Minh Ngo. 2022. Dataset of discourses about covid-19 and financial markets from twitter. *Data in Brief*, 43:108428.

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.

Gitanjali Shrivastava, Vivek Veeraiah, S Praveenkumar, Punit Pathak, Tripti Sharma, and Ankur Gupta. 2024. Explainable ai in nlp: interpretable models for transparent decision making. In *2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, pages 1–6. IEEE.

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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

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

Weiquan Wang and Izak Benbasat. 2016. Empirical assessment of alternative designs for enhancing different types of trusting beliefs in online recommendation agents. *Journal of management information systems*, 33(3):744–775.

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.

A Ethics in Involving the Human Subjects

The human subjects have been involved in the current study. In this regard, the opinion of the ethics committee of researcher's associate university have been obtained. The data involving human subject is collected after positive opinion of the ethics committee.

Appendix 1.1: Result of Sentiment Model Predictions across datasets

Model	Pro-Science Telegram			Anti-Science Telegram			Twitter		
	Negative	**Neutral**	**Positive**	**Negative**	**Neutral**	**Positive**	**Negative**	**Neutral**	**Positive**
Bertweet	12.00	**77.91**	10.08	31.05	**61.25**	7.68	31.58	**35.39**	33.02
Roberta	13.26	**63.40**	23.32	32.25	**35.66**	32.08	**34.09**	32.70	33.19
Afinn	16.67	**57.39**	25.92	33.93	**42.39**	23.66	25.67	31.74	**42.57**
Vader	16.18	**50.63**	33.17	30.97	**38.39**	30.63	26.92	25.13	**47.93**
TextBlob	11.57	**56.67**	31.74	18.65	**44.34**	37.00	21.19	35.15	**43.65**

Appendix 2.1: Pro Science-Telegram Dataset

Roberta				Bertweet-Base			
Positive Feature	**μ Abs Shap**	**Negative Feature**	**μ Abs Shap**	**Positive Feature**	**μ Abs Shap**	**Negative Feature**	**μ Abs Shap**
gorgeous	0.87038	pissed	0.801289	lucky	0.937570	crapshoot	0.949790
enjoyed	0.866934	cursed	0.796791	congratulations	0.936437	sucks	0.949719
impressive	0.82385	garbage	0.787841	impressed	0.924642	fuck	0.946441
perfect	0.79452	rubbish	0.784248	proud	0.921791	shitting	0.946339
awesomeness	0.792379	hells	0.764458	sweet	0.904331	miserable	0.933623
awesome	0.79122	bullshit	0.762609	amazing	0.899627	terrible	0.929947
breathtaking	0.761736	useless	0.741365	congrats	0.896143	suck	0.924983
fantastic	0.749989	suck	0.739847	congrats	0.892491	awful	0.921612
brilliant	0.739427	hate	0.735134	favorite	0.880586	lousy	0.921178
nice	0.725613	fuck	0.714404	adorable	0.875280	fieldworkfail	0.919701
congratulations	0.713669	motherfucking	0.711009	grateful	0.874395	unsustainable	0.916284
adorable	0.707263	idiot	0.710724	banggood	0.874294	disgusting	0.914893
appreciated	0.707207	killer	0.70511	miraculous	0.862329	horrifying	0.914520
superb	0.699428	holeefuck	0.702435	cute	0.861257	dumbass	0.913687
cutest	0.686664	regretted	0.694385	interesting	0.855965	fucked	0.912953
awesome	0.673298	lame	0.687664	attractive	0.841380	whack	0.911274
amazingly	0.671284	hellish	0.683414	welcomed	0.840065	unfortunately	0.908462
satisfying	0.657758	badworking	0.680951	mathisfun	0.830184	miseries	0.899782
amazing	0.654619	toxic	0.679895	fantastic	0.809544	abusing	0.895160
yay	0.647124	cringe	0.669696	perfectly	0.798457	shite	0.894095

Appendix 2.2 Anti-Science Telegram Dataset

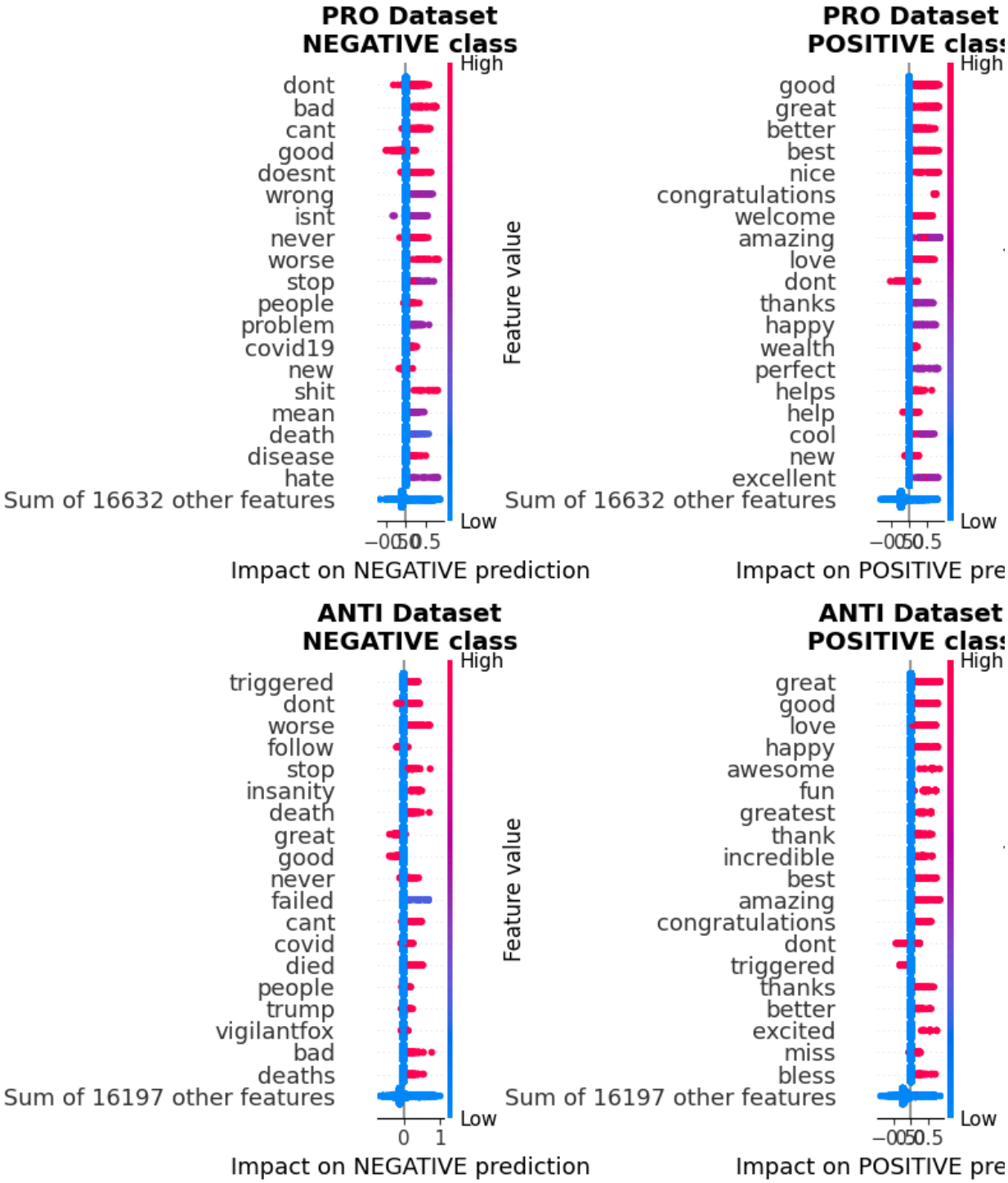
Roberta				Bertweet-Base			
Positive Feature	**μ Abs Shap**	**Negative Feature**	**μ Abs Shap**	**Positive Feature**	**μ Abs Shap**	**Negative Feature**	**μ Abs Shap**
proud	0.7473	unfuckingbelievable	0.81305	enjoy	0.911395	terrible	0.951067
kindest	0.674093	sucks	0.79877	pleased	0.910270	retard	0.929044
optimistic	0.662681	hell	0.75203	loving	0.900675	dumbest	0.923225

congratulates	0.660122	trash	0.70240	better	0.858524	disastrous	0.913929
fascinating	0.641905	broken	0.66894	smiles	0.847337	genocide	0.913806
enjoy	0.636914	ugh	0.66740	welcomes	0.845239	babykillers	0.896380
impressive	0.636283	harassment	0.66126	rejoiced	0.743534	fraud	0.880659
shining	0.624948	corrupted	0.63843	powerful	0.742407	alarming	0.856069
awesome	0.623924	unimpressive	0.62121	impressive	0.732446	diedsuddenly	0.848860
favourite	0.618442	fuckthenewworldqrder	0.60716	yay	0.731197	fiasco	0.833591
good	0.618346	ominous	0.59989	loooove	0.727871	pranksters	0.832084
excellent	0.613992	inexcusable	0.59360	interesting	0.723045	laughable	0.828146
positivequotes	0.599544	criminal	0.59359	bravely	0.706284	abomination	0.827135
delighted	0.587067	criminals	0.59062	holyyyy	0.699404	illegal	0.813145
enjoyed	0.572256	laughable	0.56450	excited	0.694914	destroy	0.793587
welldone	0.565207	cuck	0.55535	invencible	0.688607	inaccurate	0.792296
excited	0.549781	outrageous	0.55298	woohoo	0.685652	satanists	0.791240
glad	0.531536	shoddy	0.55286	fun	0.679726	jerks	0.782352
fantastic	0.530516	cowardice	0.53498	happy	0.676081	worse	0.779124
stunning	0.530093	suicide	0.53291	beginningbrighter	0.665395	ruining	0.773678

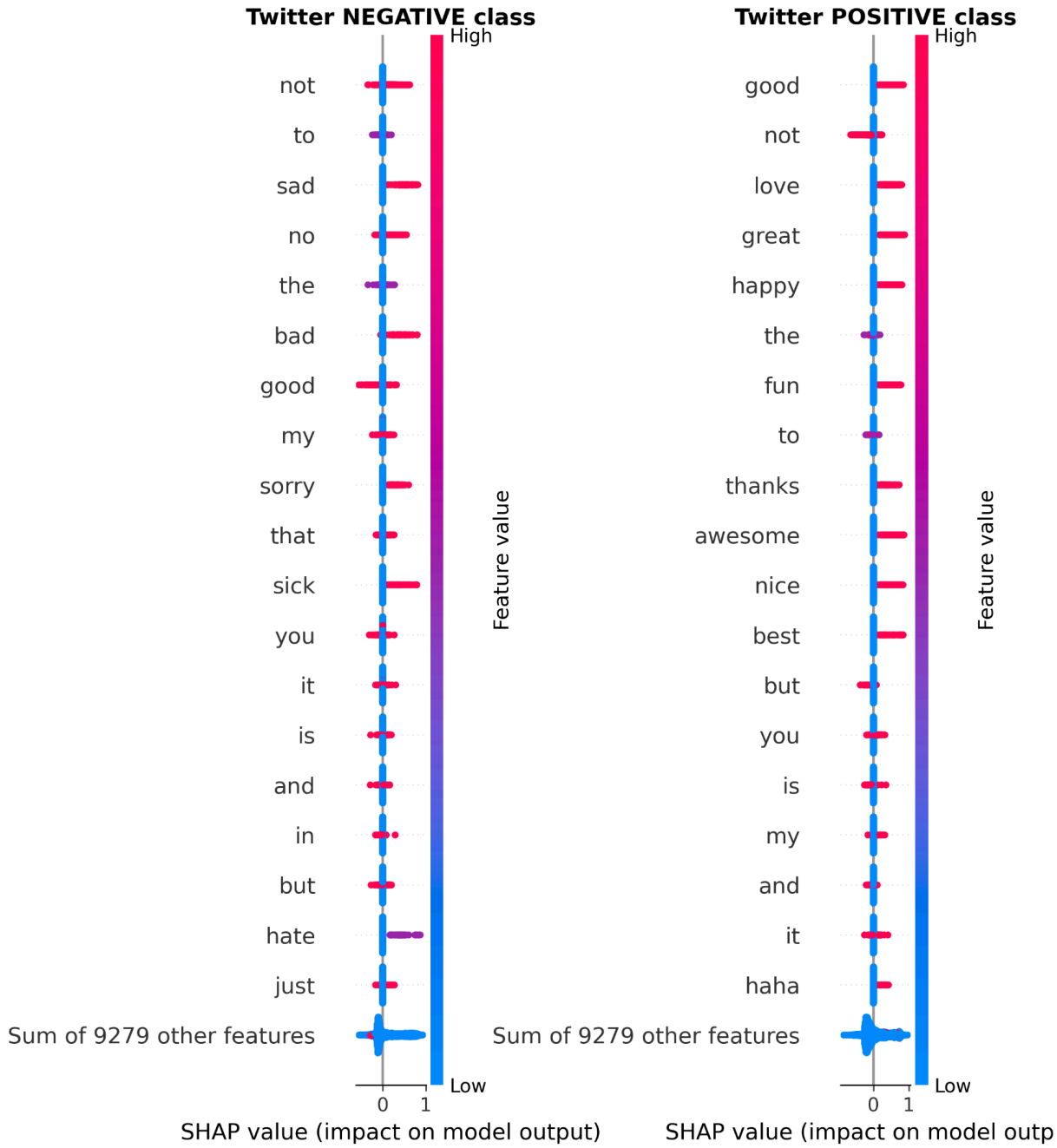
Appendix 2.3 Twitter-Global Shap from Roberta-Base and Bertweet Base Model

Roberta-Base Model				**Bertweet Base Model**			
Positive Feature	**μ Abs Shap**	**Negative Feature**	**μ Abs Shap**	**Positive Feature**	**μ Abs Shap**	**Negative Feature**	**μ Abs Shap**
excitedddd	0.932380	suck	0.941906	congratulations	0.830842	suckish	0.886848
yaaaaaaay	0.930996	scoundrel	0.935006	coolest	0.813121	suckss	0.829175
happier	0.930442	douche	0.932774	loooove	0.794564	shite	0.788138
gooood	0.926852	stupid	0.929179	coolest	0.778070	disgusting	0.786887
exited	0.915331	unfortunetly	0.925747	brilliant	0.777448	pathetic	0.761138
nicest	0.904545	nasty	0.924212	enjoyment	0.777165	trainwrecked	0.756053
funnnn	0.896356	failed	0.923488	happier	0.771090	racist	0.748370
niiceee	0.895981	cancelations	0.921812	fabulous	0.768051	cba	0.728651
yayyyy	0.886554	frustration	0.901531	excited	0.763403	shitting	0.724147
encouraging	0.868507	booooring	0.899318	happiest	0.753301	noooooooo	0.700437
niiice	0.859181	died	0.897745	yummm	0.751458	unfortunate	0.688550
yay	0.857896	despise	0.897344	thankful	0.746313	disapointed	0.670943
refreshed	0.854752	losers	0.891094	goooood	0.744071	devastated	0.664887
awesomeee	0.853086	murderer	0.888871	funnn	0.740658	unfair	0.659426
goooooodmorning	0.851313	enemies	0.887477	bestt	0.736219	suck	0.654513
yaay	0.844401	cruel	0.884766	ahhhblessyou	0.720343	hates	0.653045
perfect	0.842135	unfit	0.884746	enjoyin	0.716762	terribly	0.651519
woohoo	0.841700	fail	0.884601	yummy	0.693023	boredd	0.644093
fortunate	0.836793	dumb	0.880319	merry	0.689510	trash	0.637092
delicious	0.836441	unpleasant	0.877313	delish	0.682175	hurt	0.632767

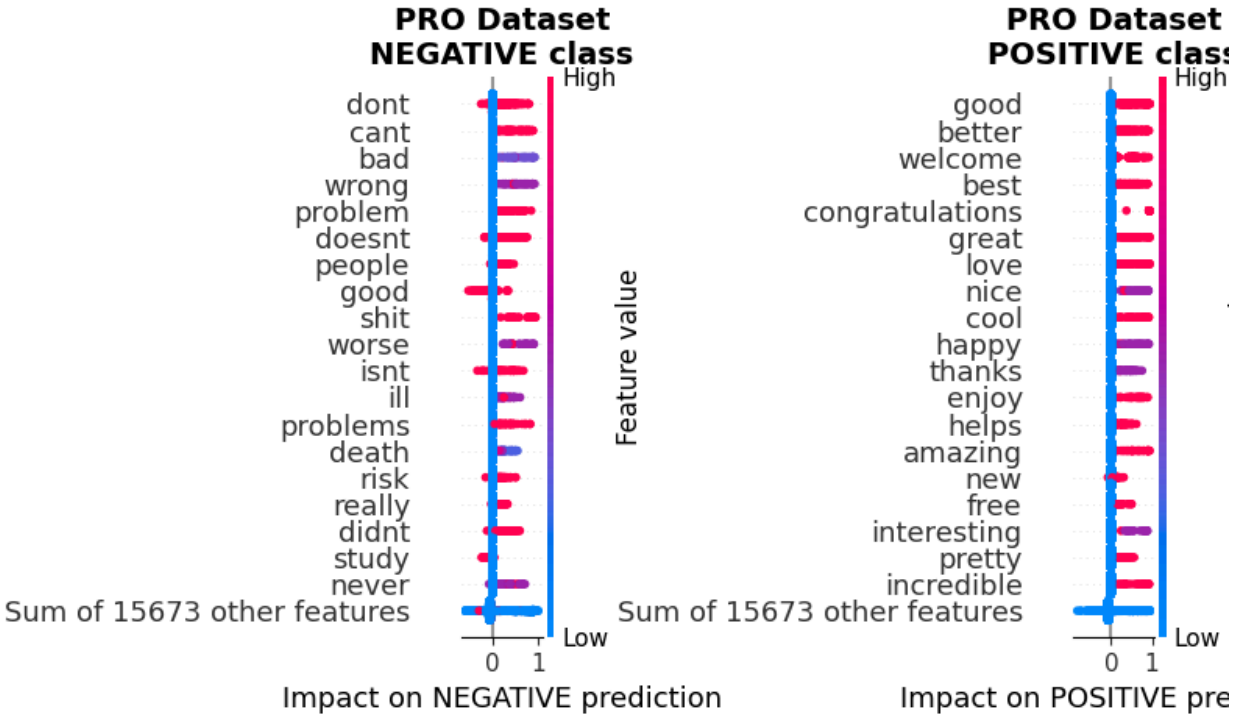
Appendix 3.1: Beeswarm plots: Roberta-Base Model-Telegram

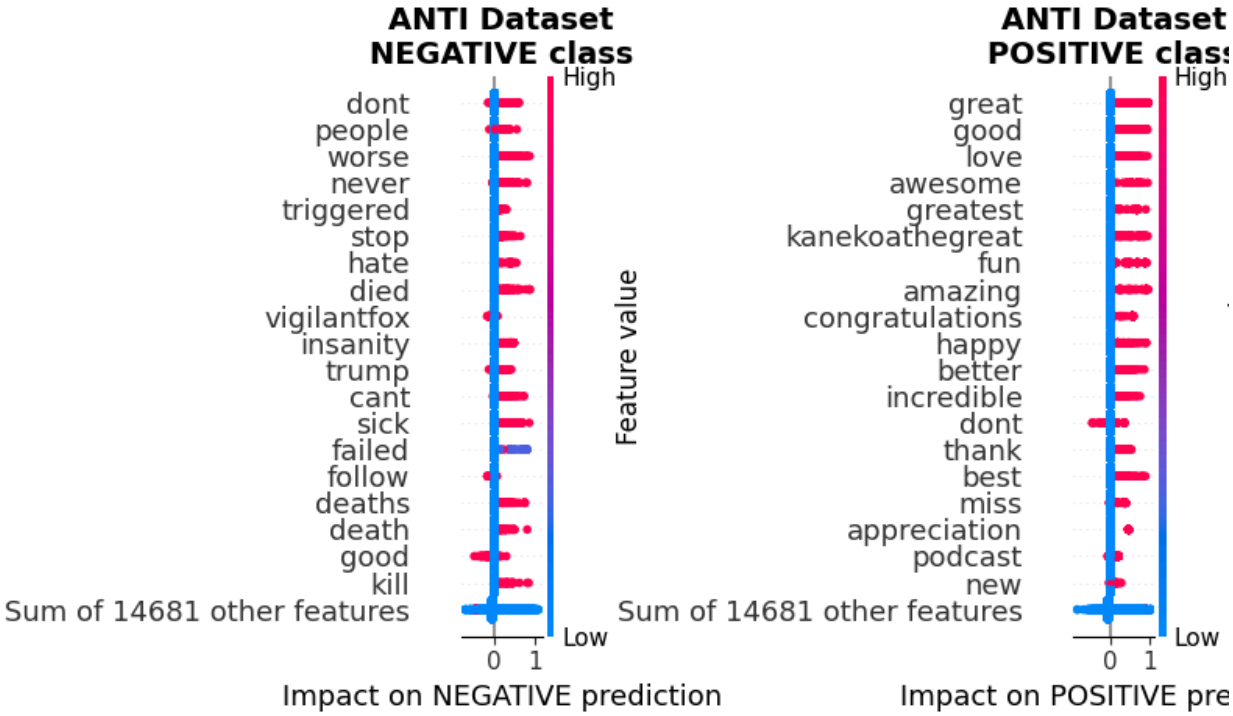


Appendix 3.2: Beeswarm plots: Roberta-Base Model-Twitter

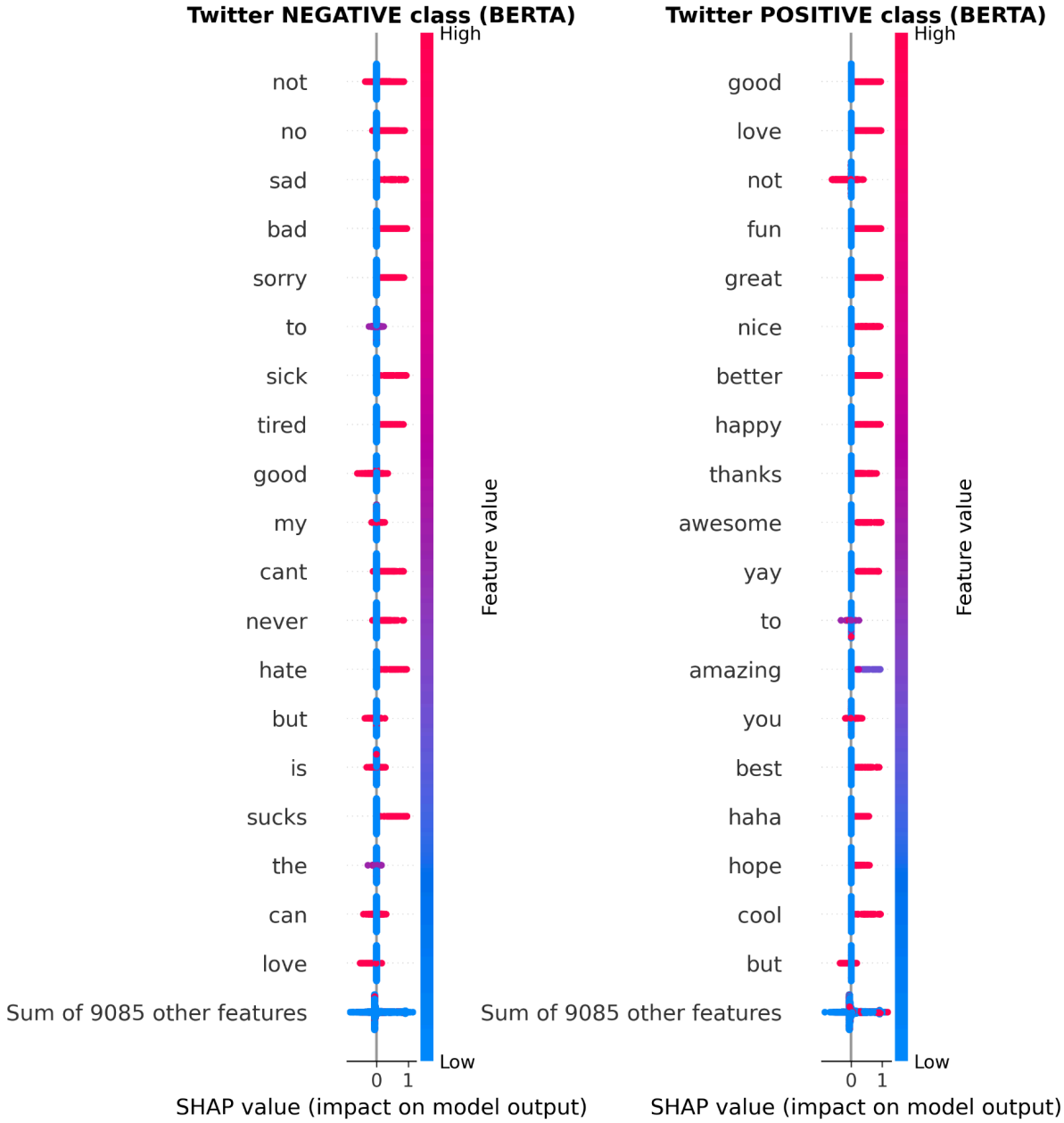


Appendix 3.3: Beeswarm plots: Berta-Base Model-Telegram





Appendix 3.4: Beeswarm plots: Berta-Base Model-Telegram



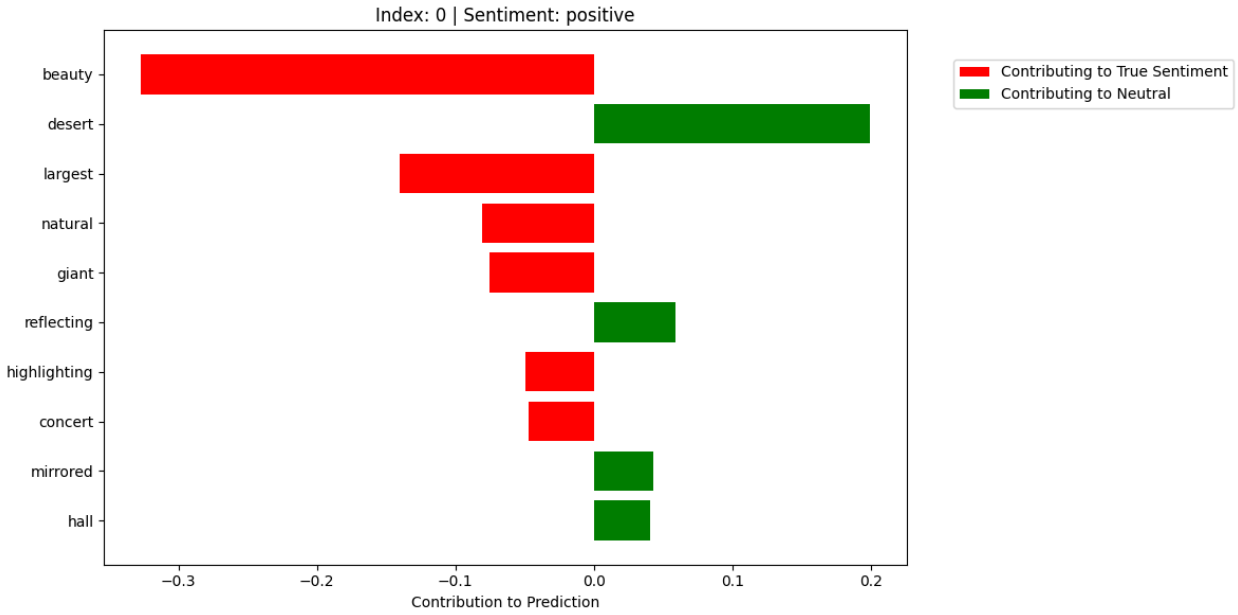
Appendix 4.1: pairwise comparison in between Original Shap vector and each of perturbed data Shap vectors-Roberta-Base Model

Perturbation	Pro-Science			Anti-Science			Twitter		
	t-value	p-value	Cohen's d	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	0.448	0.6544	0.006
Shuffle	2.364	0.0181	0.033	11.021	0.0000	0.156	-2.135	0.0328	-0.030
Synonym	4.479	0.0000	0.063	6.307	0.0000	0.089	2.280	0.0226	0.032

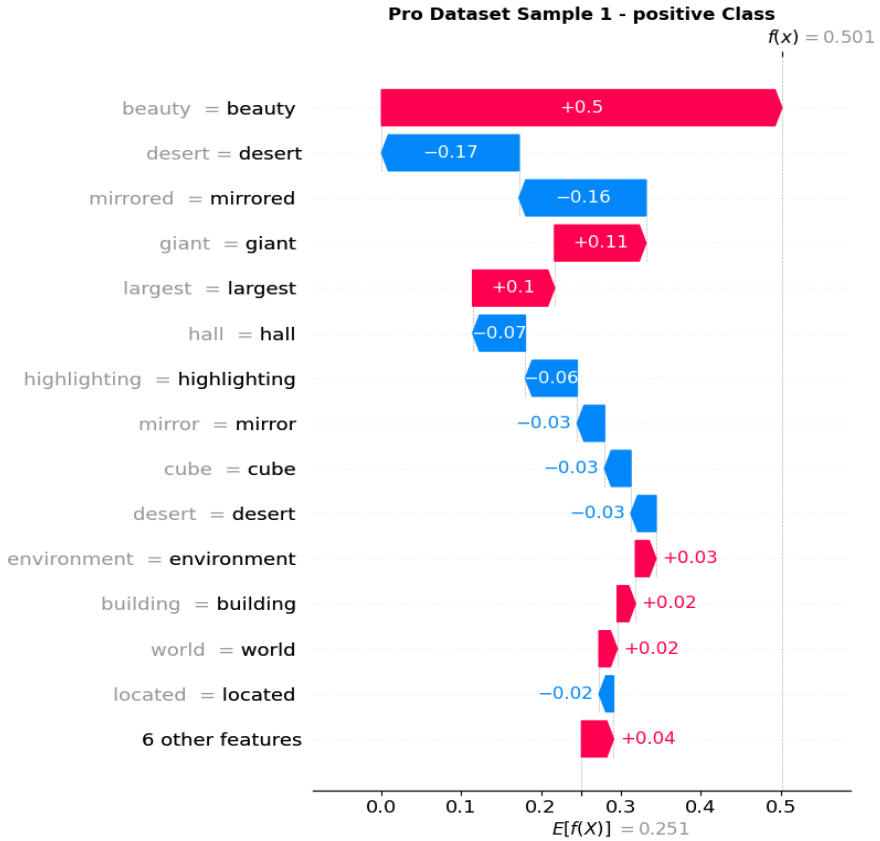
Appendix 4.2: Pairwise comparison in between Original Shap vector and each of perturbed data Shap vectors-Bertweet-Base Model

Perturbation	Pro-Science			Anti-Science			Twitter		
	t-value	p-value	Cohen's d	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	-1.746	0.0808	-0.025
Shuffle	2.074	0.0381	0.029	8.625	0.0000	0.122	0.864	0.3877	0.012
Synonym	3.415	0.0006	0.048	5.309	0.0000	0.075	3.685	0.0002	0.052

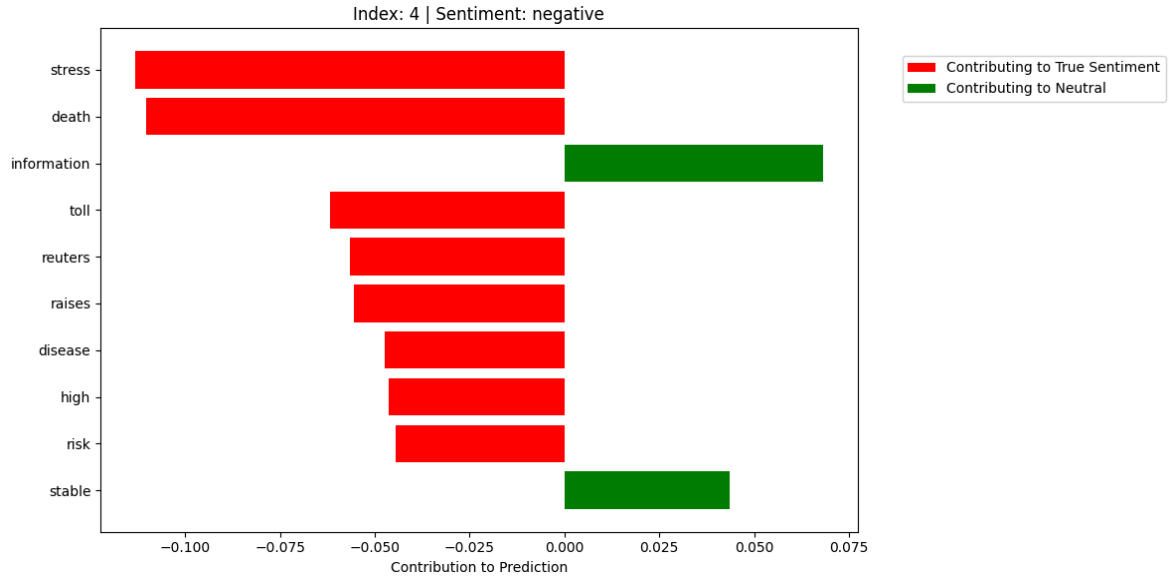
Appendix 5.1: LIME explanation on Roberta Model-Positive Sentiment-Telegram Dataset



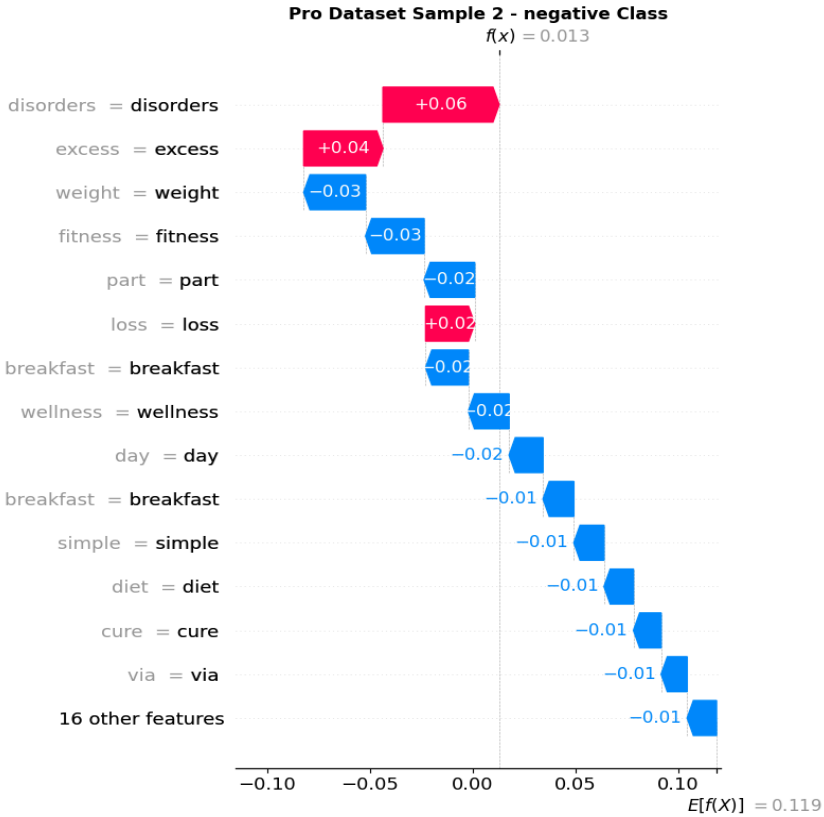
Appendix 5.2: Shap explanation on Roberta-Base Model-Positive Sentiment-Telegram Dataset



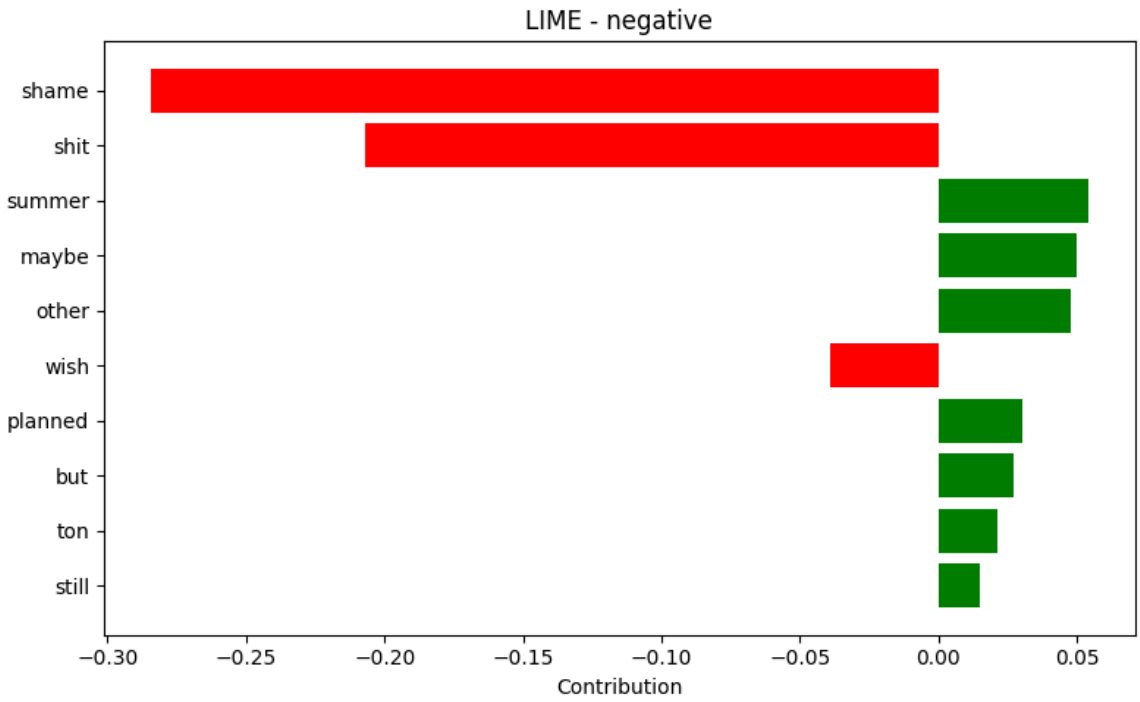
Appendix 5.3: LIME explanation on Roberta-Base Model-Negative Sentiment-Telegram Dataset



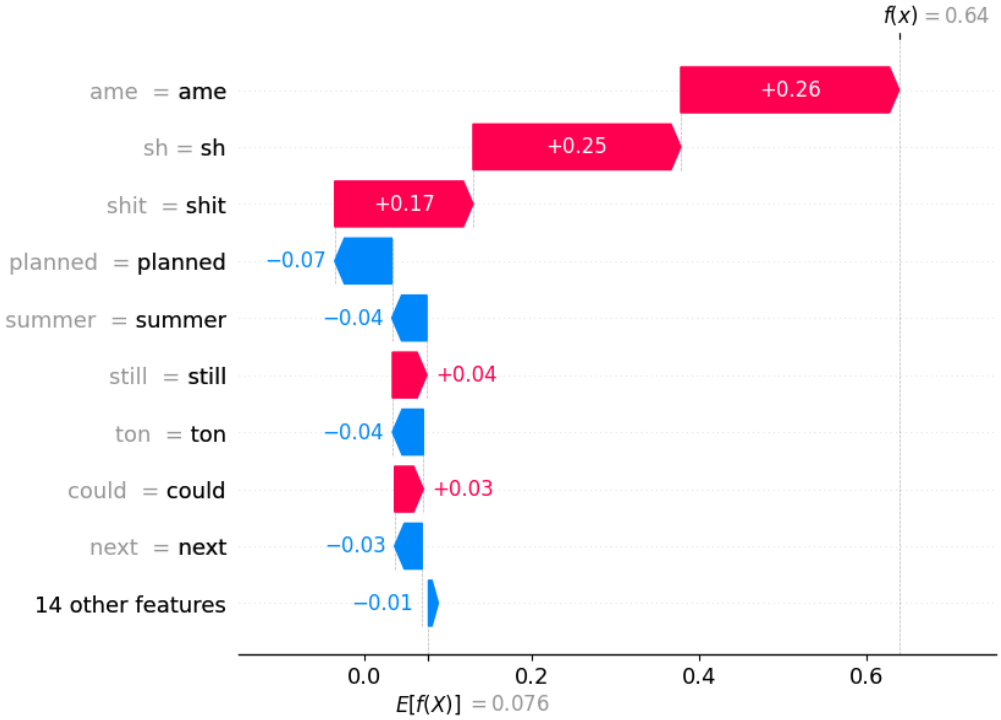
Appendix 5.4: Shap explanation on Roberta Model-Positive Sentiment-Telegram Dataset



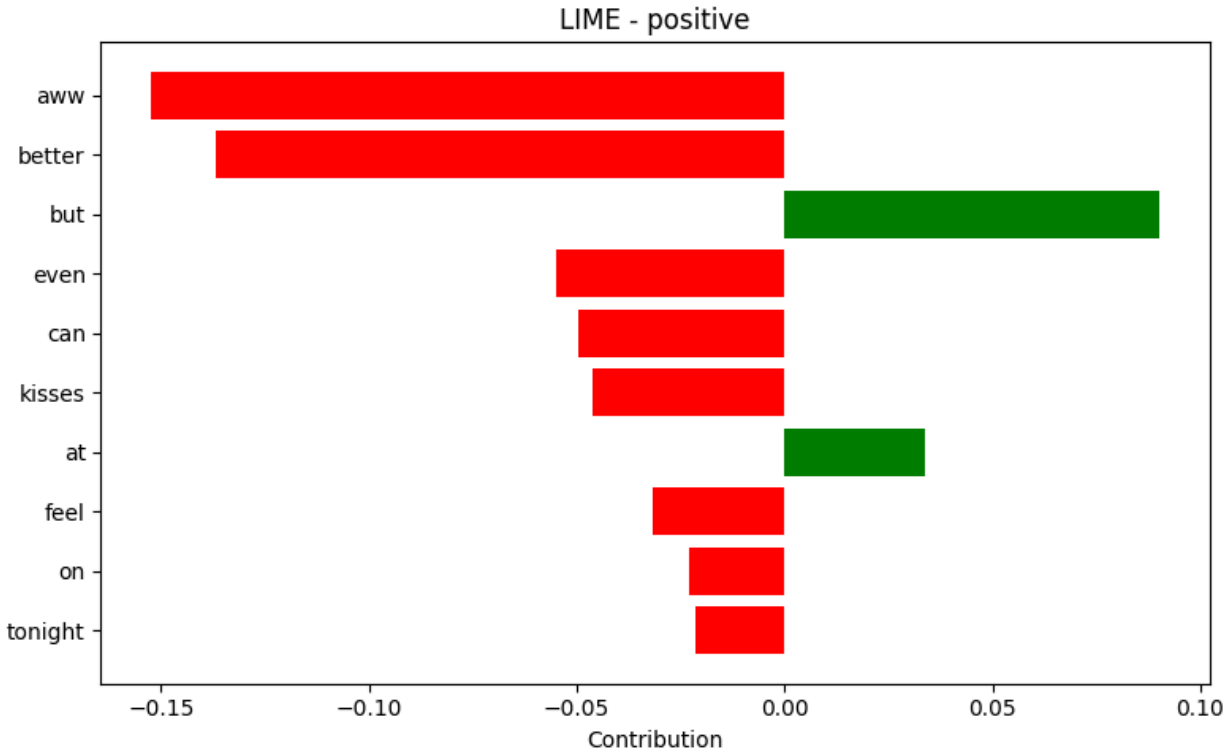
Appendix 5.5 LIME explanation on Roberta- Model-Negative Sentiment-Twitter Dataset



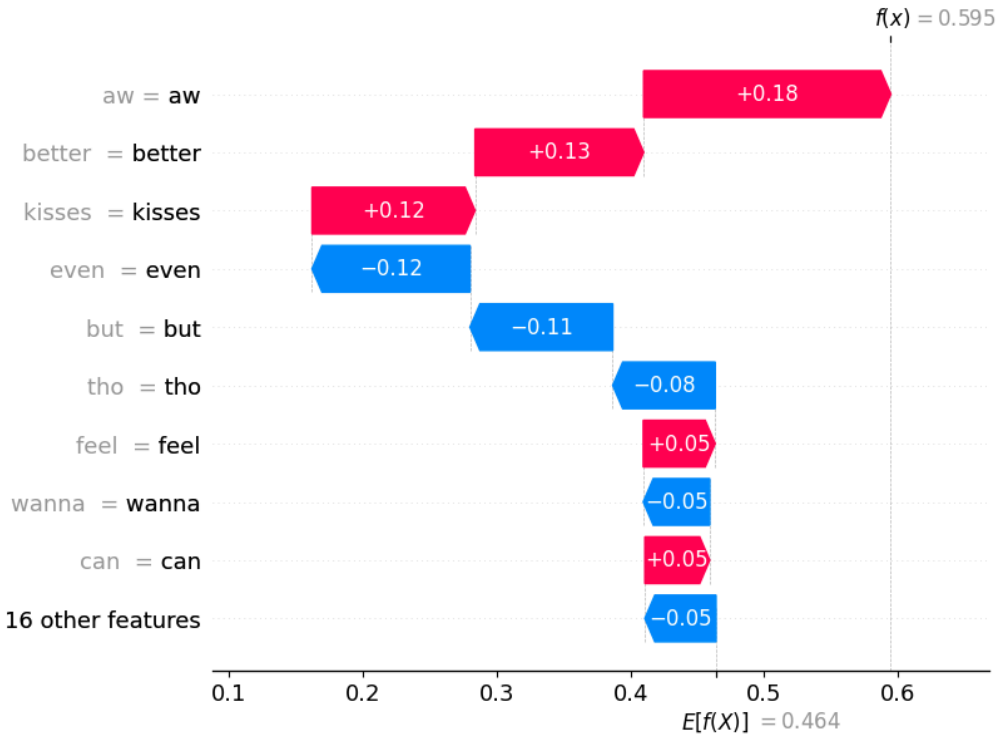
Appendix 5.6 Shap explanation on Roberta- Model-Negative Sentiment-Twitter Dataset



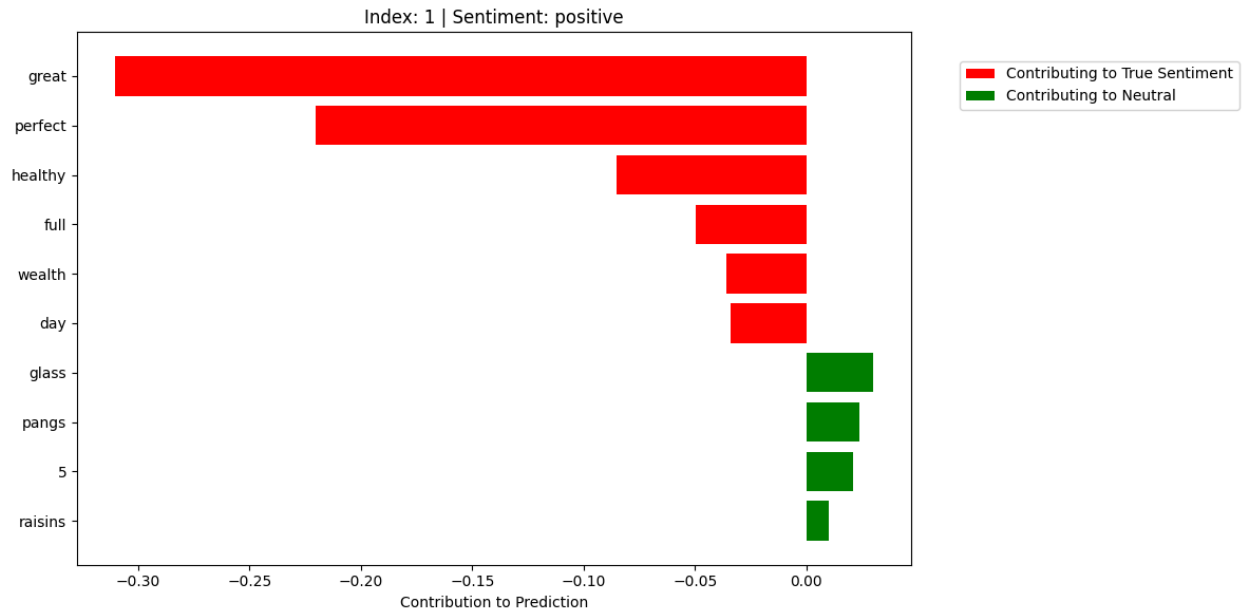
Appendix 5.7: LIME explanation on Roberta Model-Positive Sentiment-Twitter Dataset



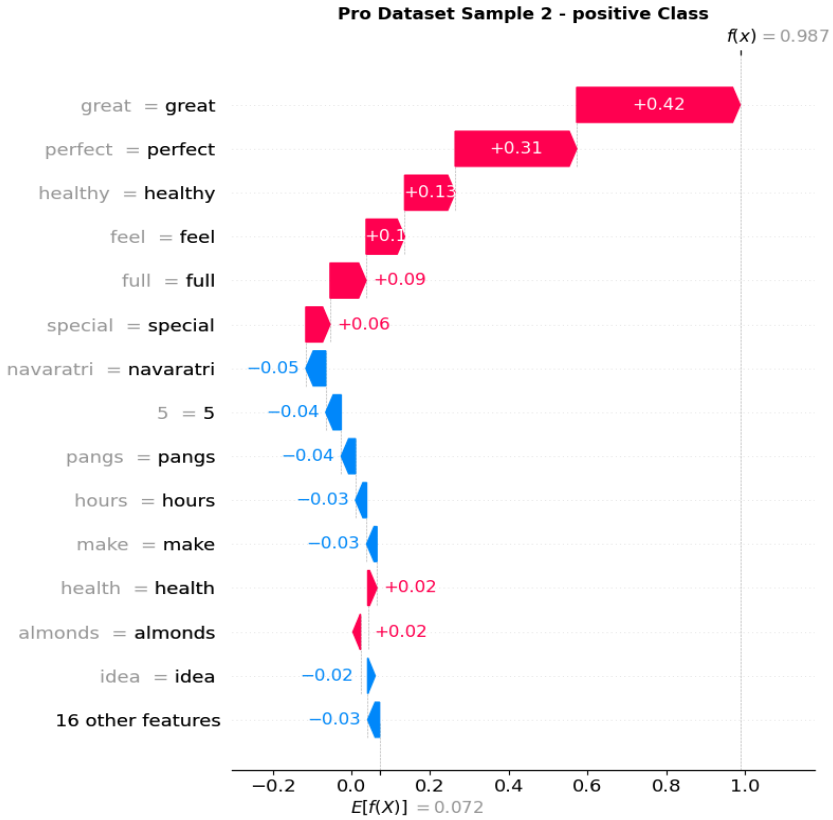
Appendix 5.8: Shap explanation on Roberta Model-Positive Sentiment-Twitter Dataset



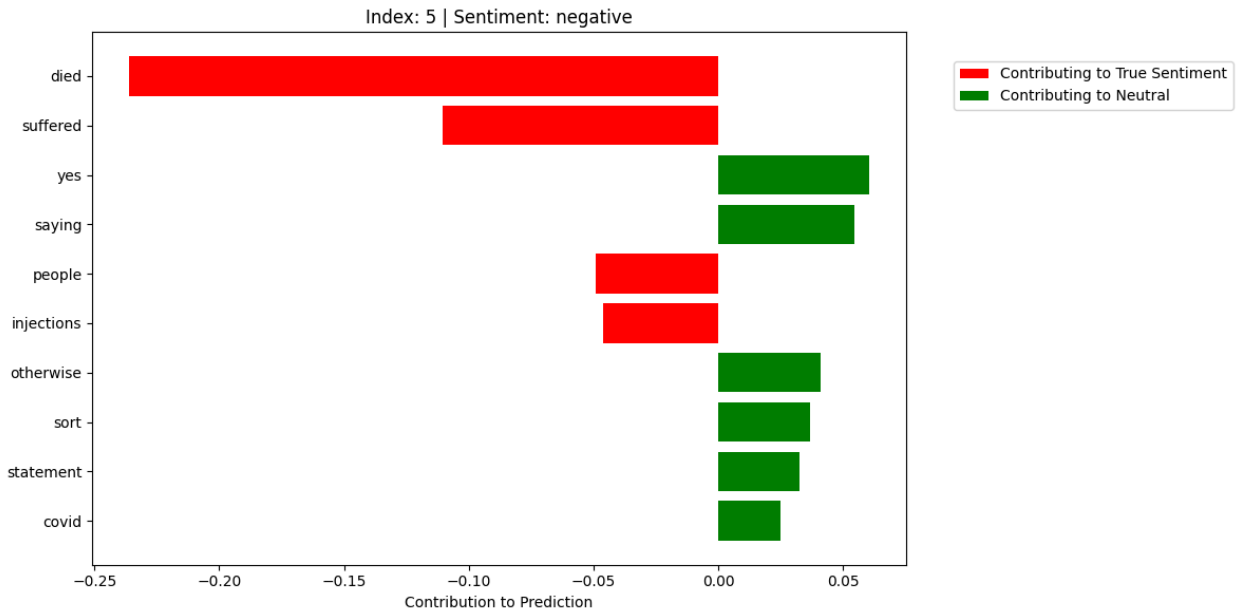
Appendix 5.9: LIME explanation on Bertweet- Model-Positive Sentiment-Telegram Dataset



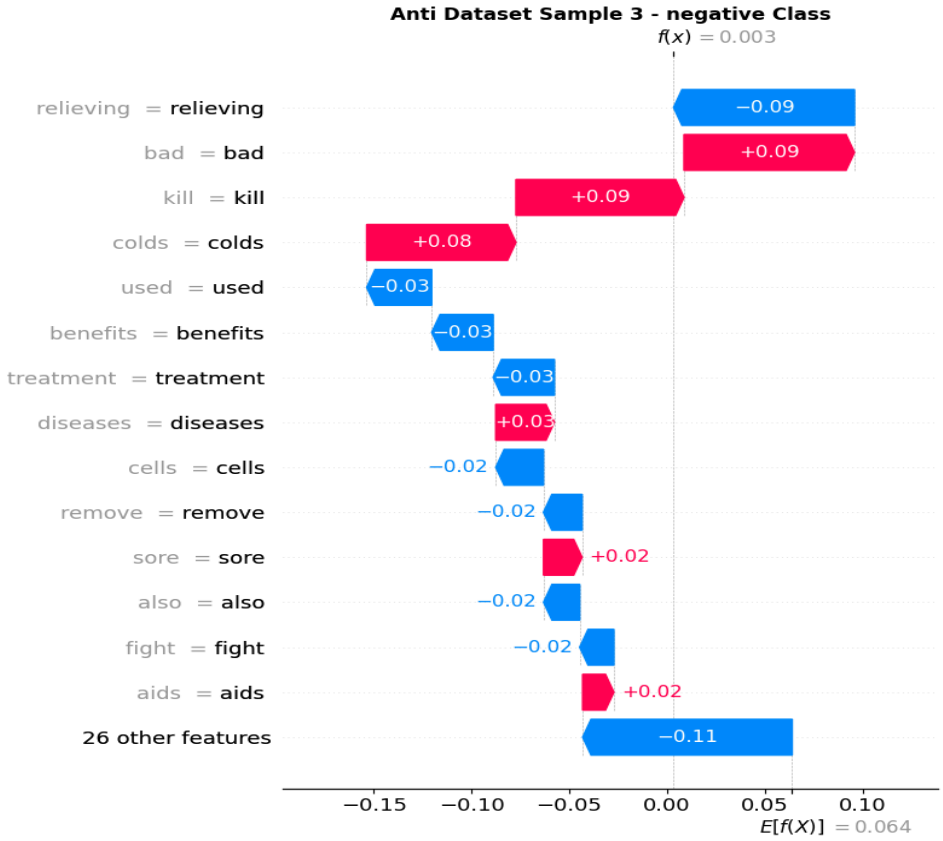
Appendix 5.10: Shap explanation on Bertweet Model-Positive Sentiment--Telegram Dataset



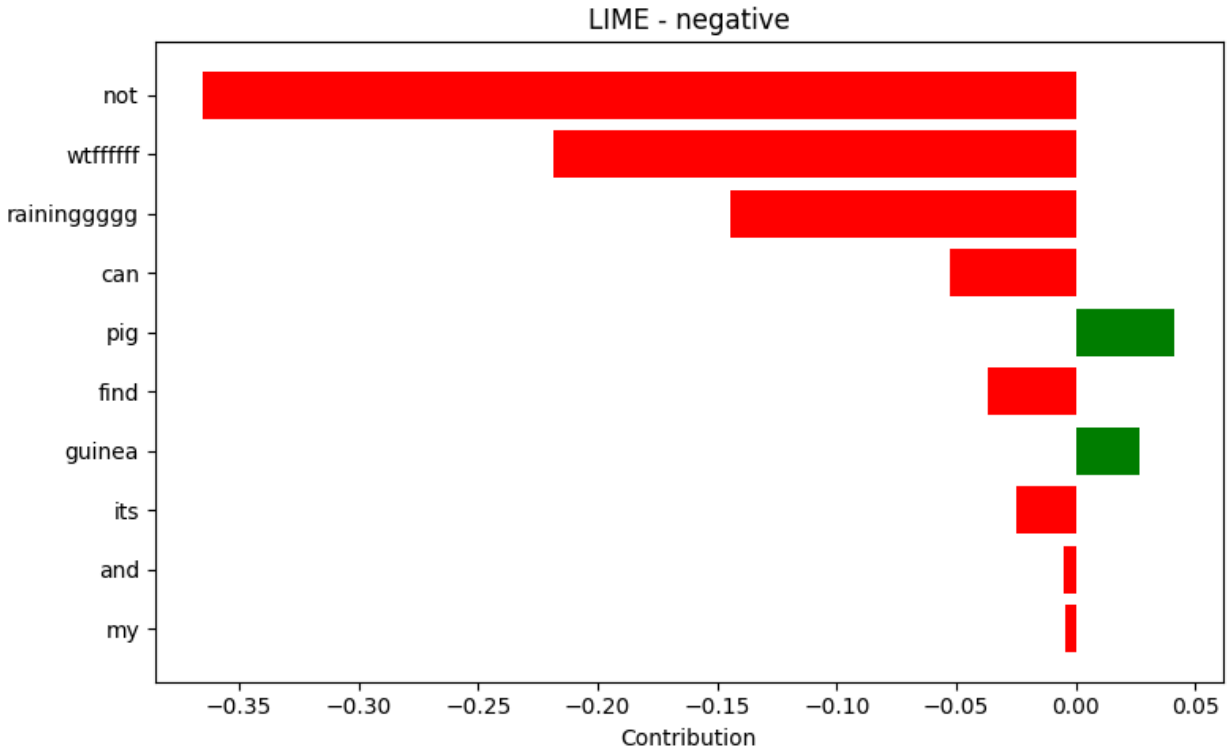
Appendix 5.11: LIME explanation on Bertweet Model-Negative Sentiment-Telegram



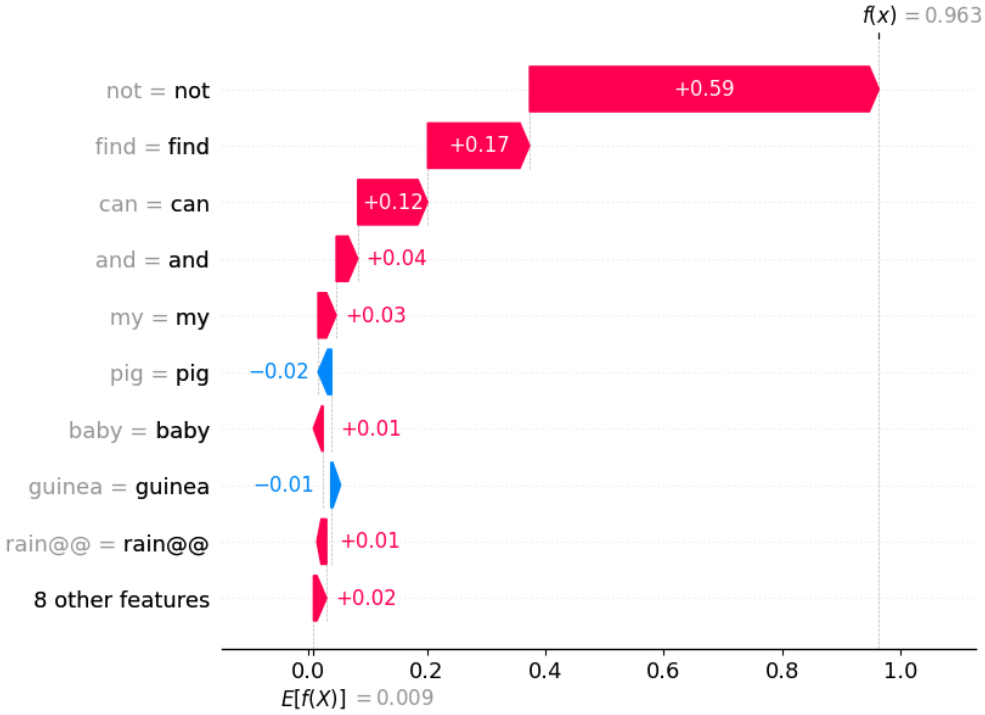
Appendix 5.12: Shap explanation on Bertweet Model-Negative Sentiment-Telegram



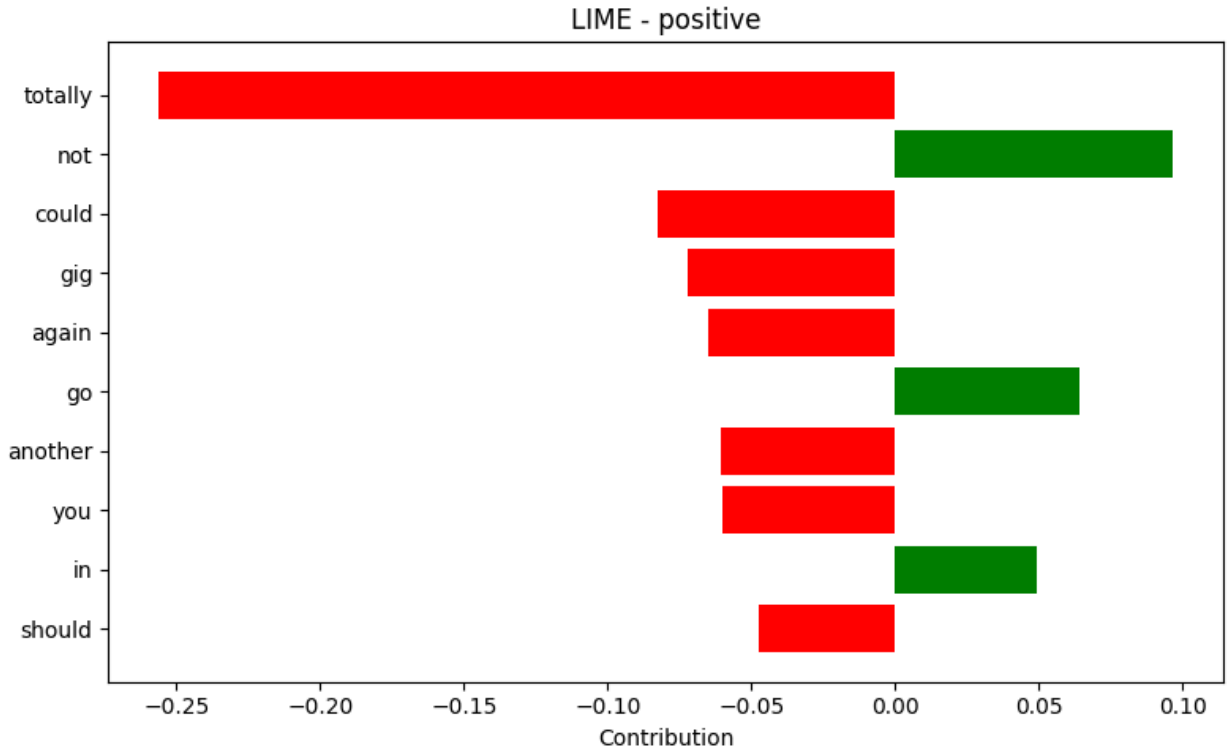
Appendix 5.13: LIME explanation on Bertweet Model-Negative Sentiment-Twitter



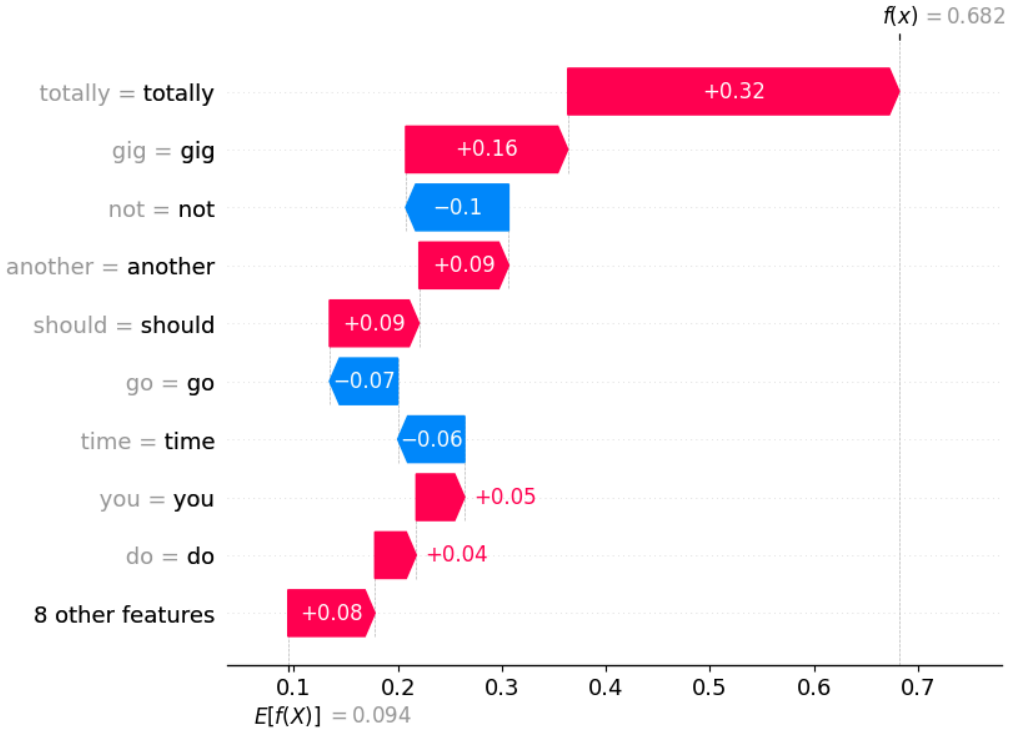
Appendix 5.14: Shap explanation on Bertweet Model-Negative Sentiment-Twitter



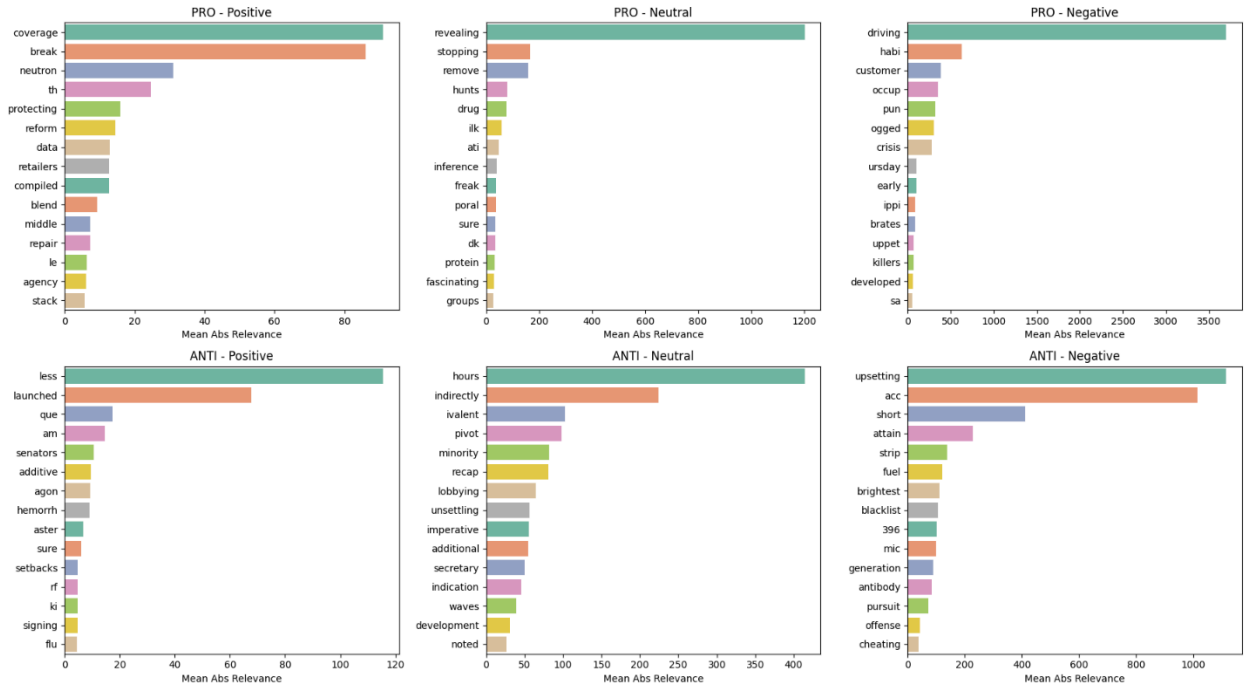
Appendix 5.15: LIME explanation on Bertweet Model-Positive Sentiment-Twitter Dataset



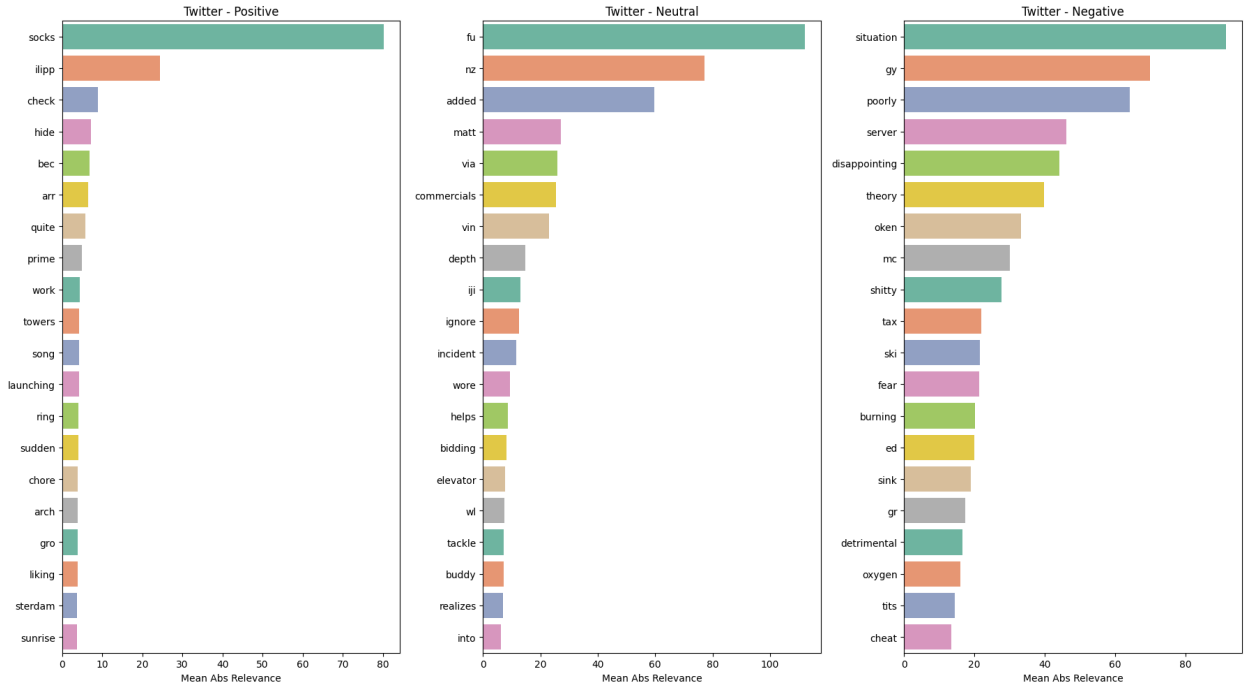
Appendix 5.16: Shap explanation on Bertweet Model-Positive Sentiment-Twitter Dataset



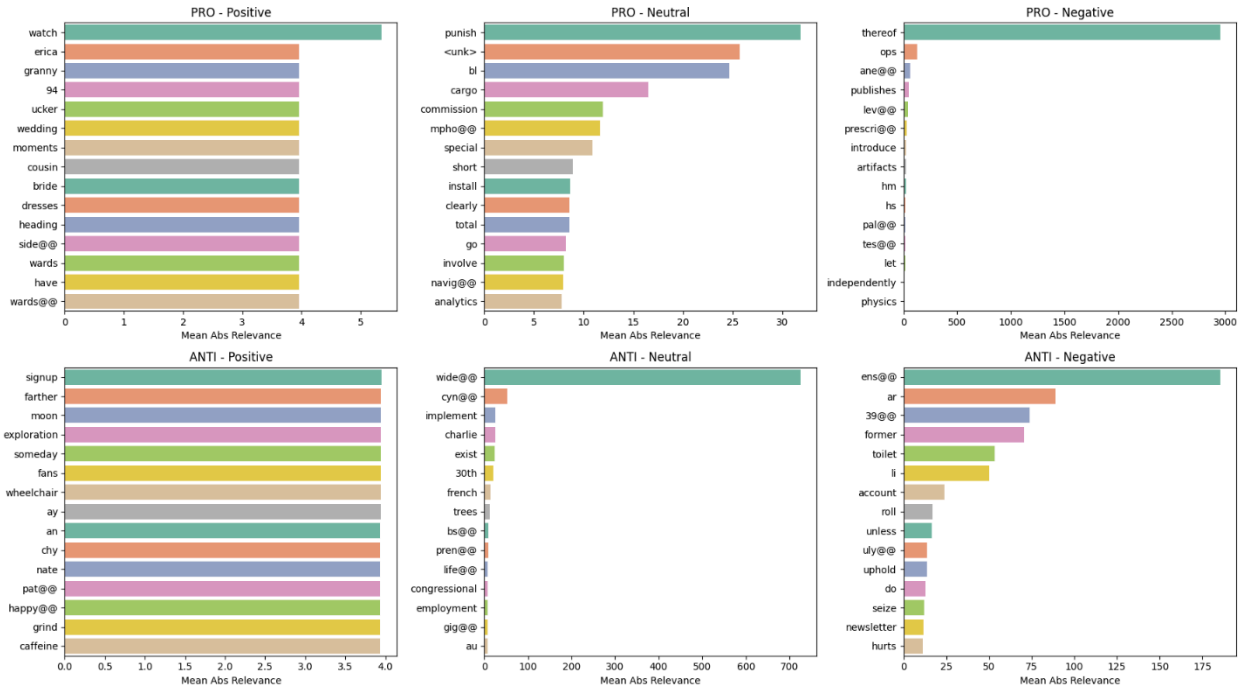
Appendix 6.1: Dominant features by the LRP simplified rule-Roberta-Base-Telegram



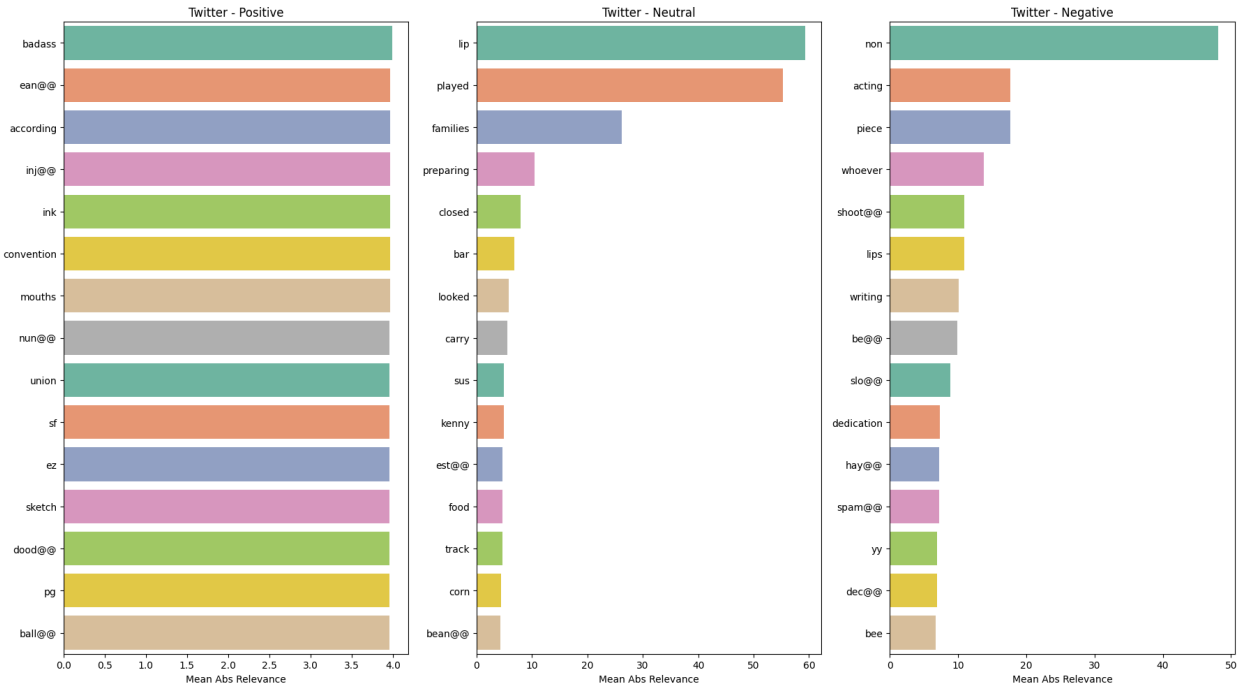
Appendix 6.2: Dominant features by the LRP simplified rule-Roberta-Base-Twitter



Appendix 6.3: Dominant features by the LRP simplified Bertweet-Base-Twitter



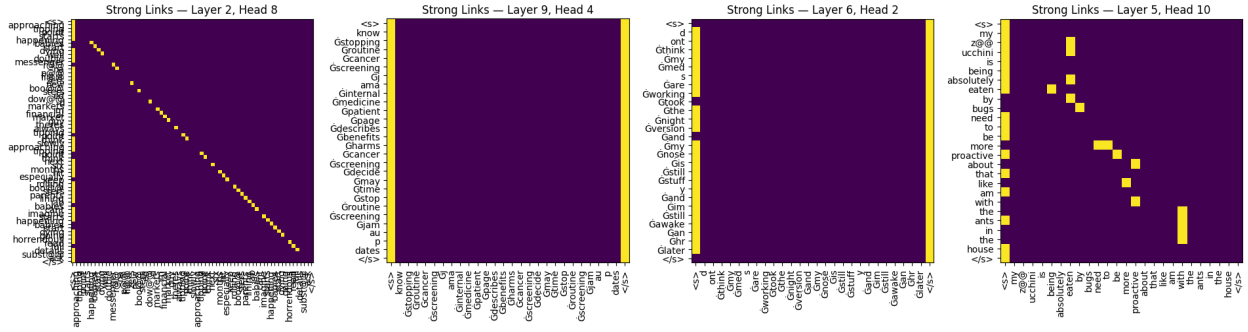
Appendix 6.4: Dominant features by the LRP simplified Bertweet-Base-Telegram-



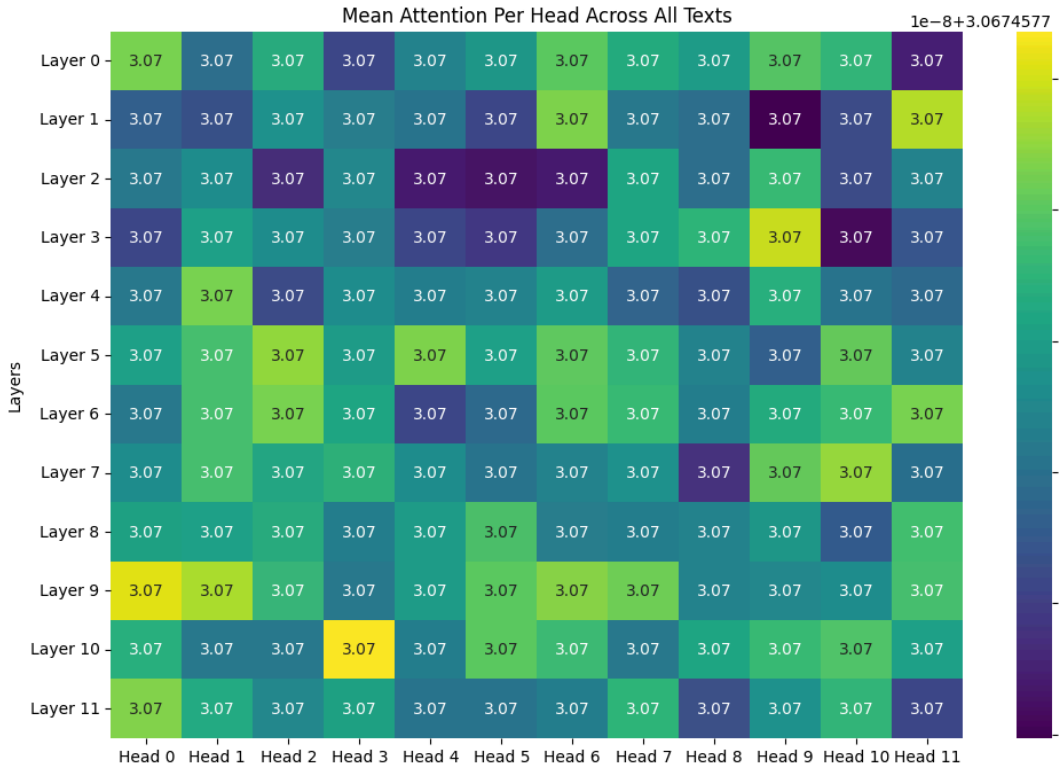
Appendix 6.5: A comparative LRP global metrics of Sentiment classes

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
Positive	Twitter	1.9203	0.0019	9.8385
Negative	Twitter	2.3224	0.0031	10.1686
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
Positive	Twitter	2.3835	0.0015	10.1267
Negative	Twitter	3.0495	0.9934	10.1387

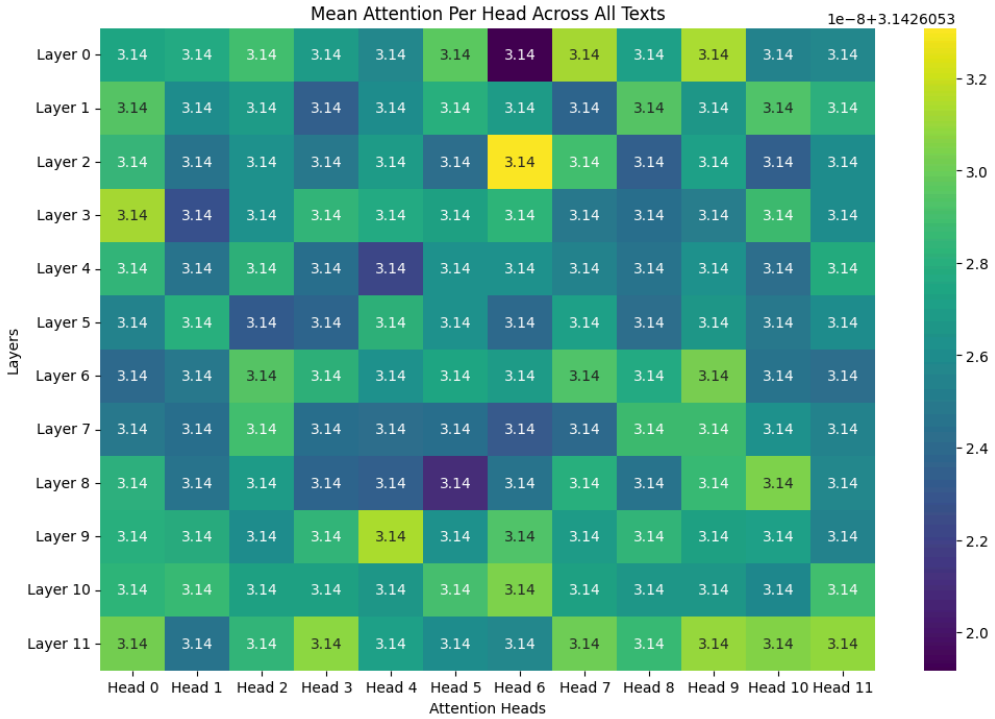
Appendix 7.1: Dominant Layers and head in the neural network



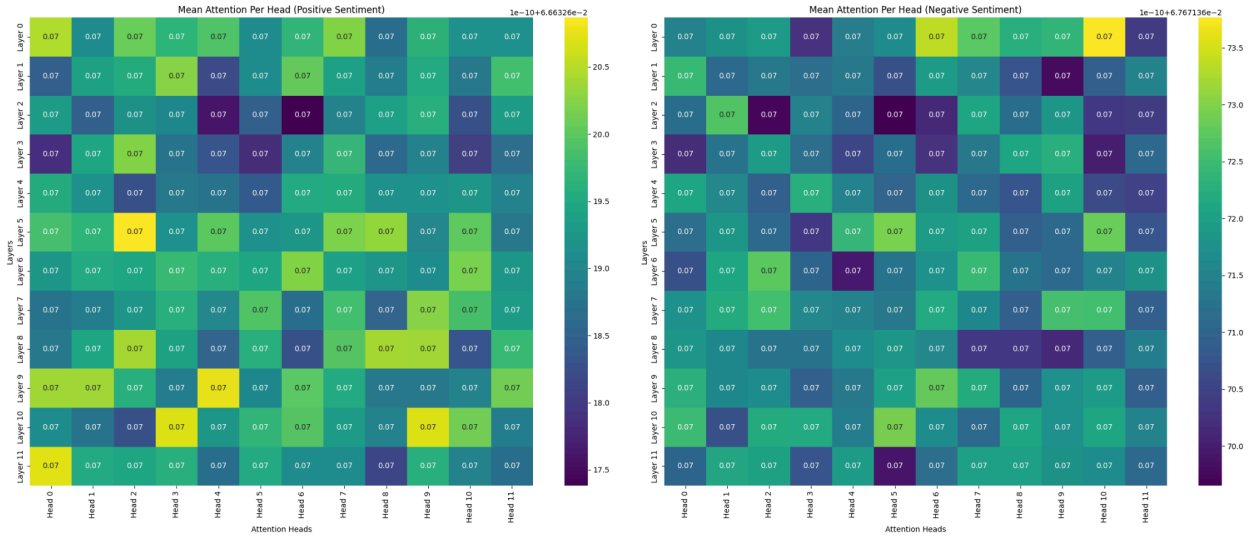
Appendix 7.2: Mean attention per head-Roberta-Twitter



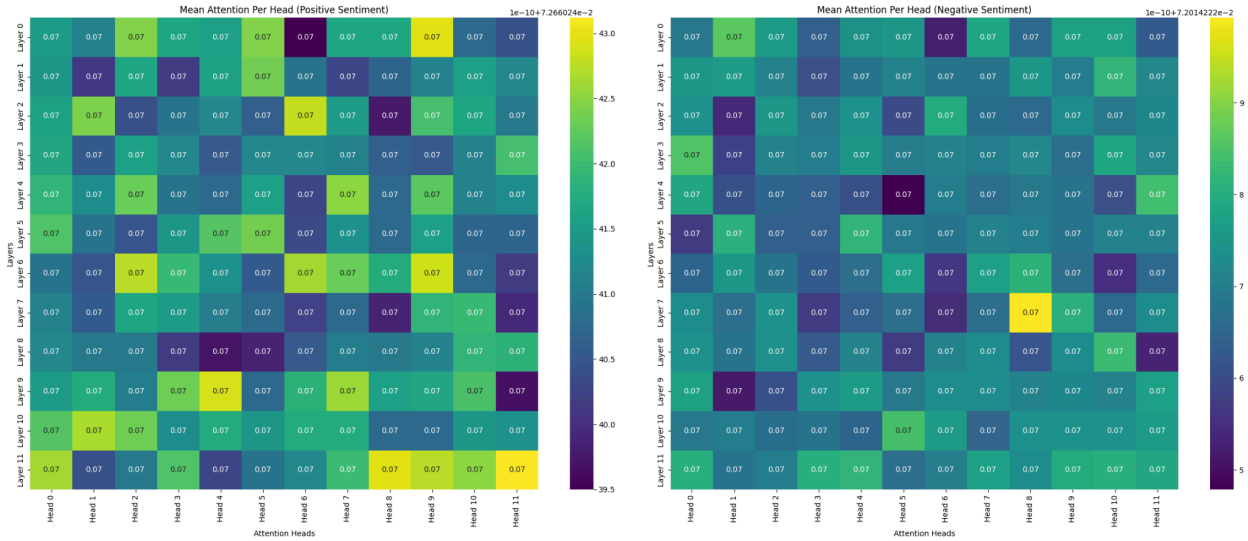
Appendix 7.3: Mean attention per head-Bertweet



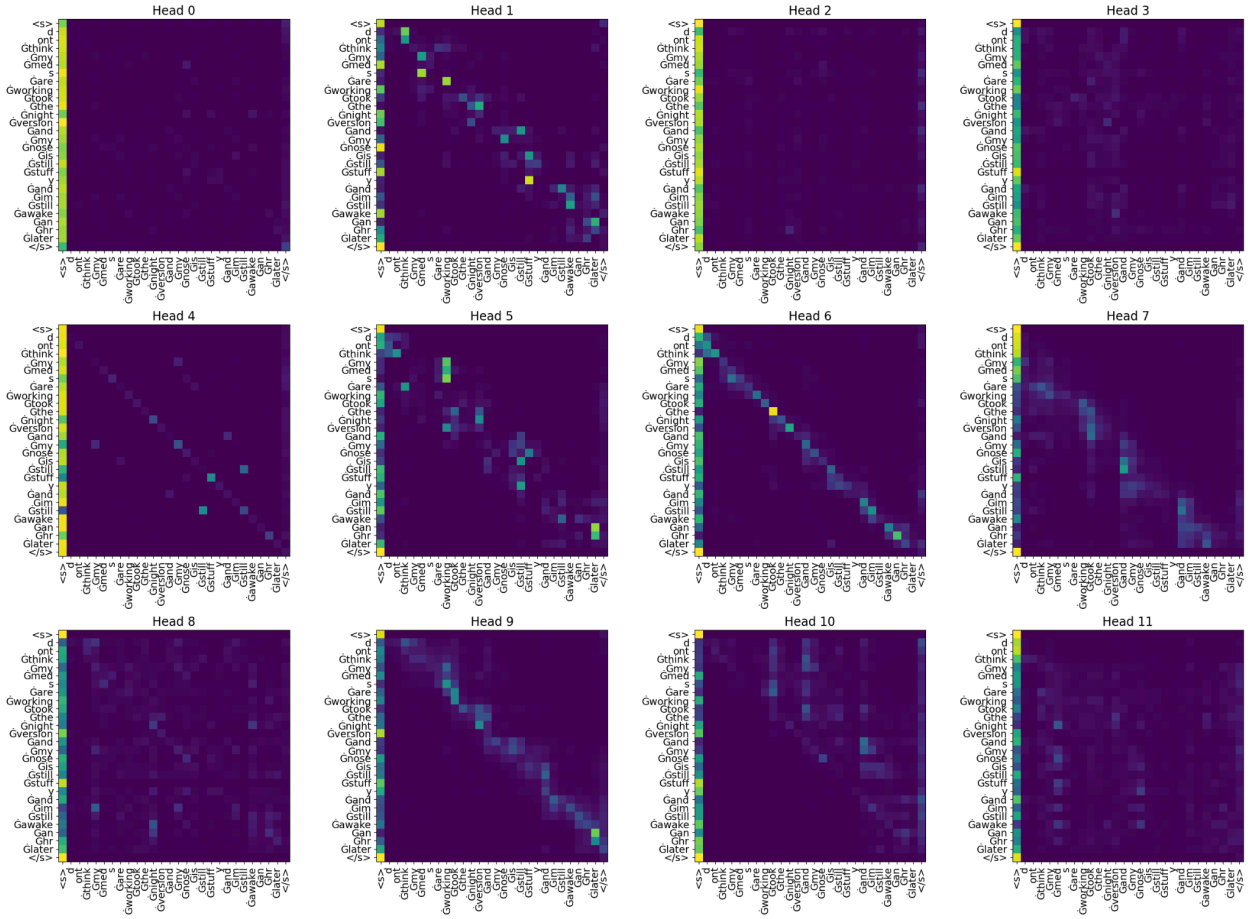
Appendix 7.4: Mean attention by sentiment class wise-Roberta-Twitter



Appendix 7.5: Mean attention by sentiment class wise-Bertweet-Twitter



Appendix 7.6: Head of Layer 0 for neural network



Appendix 8.1 Normality test using Shipro-Wilk: Experiment Group

Variable	Shapiro p-value	Skewness	Kurtosis	Normality
Perceived Usefulness	0.0731	-0.6716	0.6453	Approximately normal
Perceived explainability	0.0000	-3.2076	12.9595	Not normally distributed
Satisfaction	0.0300	-0.6037	0.7887	Not normally distributed
Trust	0.0216	-0.7576	0.5359	Not normally distributed

Appendix 8.2 Normality test using Shipro-Wilk: Control Group

Variable	Shapiro p-value	Skewness	Kurtosis	Normality
Perceived Usefulness	0.0000	-2.6051	7.6675	Not normally distributed
Perceived explainability	0.0000	-3.2739	14.4069	Not normally distributed
Satisfaction	0.0000	-3.0947	9.4059	Not normally distributed
Trust	0.0000	-2.8371	9.7178	Not normally distributed

Appendix 8.3: Scale used to collect the data using between the subject design experiment

Usefulness -Experimental	Usefulness -Control
The explanation on the model's prediction improves my performance in identifying sentiment of sentences.	The used sentiment model's prediction improves my performance in identifying sentiment of sentences shown to me.
The explanation sentiment model's prediction enables me to quickly identify the sentiment of a sentence**.**	The used sentiment model's prediction enables me to quickly identify the sentiment of a sentence shown to me**.**
The explanation on the sentiment model's prediction enhances my effectiveness in identifying sentiment of text.	The used sentiment model's prediction enhances my effectiveness in identifying sentiment of text
The explanation on the sentiment model's prediction makes it easier to assign the sentiment label.	The used sentiment model's prediction makes it easier to assign the sentiment label.
The explanation on sentiment model's prediction increases my productivity in sentiment identification and labelling to text.	The used sentiment model's prediction increases my productivity in sentiment identification and labelling to text.
Explainability-Experimental	**Explainability-Control**
The explanation on the sentiment labels is clear to me.	The sentiment model decision to predict all of the 10 statement sentiment labels is clear to me.

By reading the text and explanation, it was easier for me to see how sentiment models have predicted the sentiment label for each text	By reading the text, it was easier for me to see how sentiment model have predicted the sentiment label for each text
From explanation it was easier to understand the inner workings of the sentiment model.	It was easier to understand the inner workings of the sentiment model by looking at their prediction of text.
By reading explanation, I could understand why and how sentiment models used to predict the label.	I could understand why and how sentiment models used to predict the label for each of sentence.
Satisfaction-Experimental	**Satisfaction-Control**
From explanation, I know how the machine learning model works and predict the sentiment.	From the prediction of sentiment of text, I know how the machine learning model works.
The explanation on sentiment model's prediction satisfying	The prediction of sentiment model of text is explainably satisfying
The explanation of sentiment model has sufficient detail.	The prediction of sentiment model of texts has sufficient detail.
The explanation in sentiment model prediction seems complete	The prediction of sentiment model of texts seems complete
The explanation tells how easy it is to predict the sentiment of the text	The prediction of sentiment model of texts tells how easy it is to predict the sentiment of the text
The explanation on the sentiment model of texts seems useful.	The prediction of the sentiment model of texts seems useful.
Trust-Experimental	**Satisfaction-Control**
I am confident that, explanation on sentiment analysis machine learning model work well in predicting sentiment of text	I am confident that, used sentiment analysis machine learning model work well in predicting sentiment of text
From explanation, output of sentiment model seems very predictable	The out of sentiment model seems very predictable

The explanation on sentiment models seems reliable and can be counted at all the time.	The sentiment analysis machine learning model seems reliable and can be counted at all the time.
The explanation on sentiment models seems efficient.	The sentiment analysis machine learning model seems efficient.
With explanation, It is clear that, sentiment model can perform better than novice human in sentiment labelling task.	The sentiment analysis machine learning model model can perform better than novice human in sentiment labelling task.