Reversing Causal Assumptions: Explainability in Online Sports Dialogues

Asteria Kaeberlein and Malihe Alikanhi
Northeastern University
{kaeberlein.c, m.alikhani}@northeastern.edu

Abstract

Prior XAI research often assumes inputs must be ”causes” and outputs must be ”effects”, severely limiting applicability to analyzing behaviors that emerge as reactions or consequences. Many linguistic tasks, such as dialogues and conversations, involve such behaviors. To address this, we propose that the assumed causality from inputs to outputs can be reversed and still remain valid by using outputs that cause changes in features. We show how this enables analysis of complex feature sets through simpler metrics, propose a framework that is generalizable to most linguistic tasks, and highlight best practices for applying our framework. By training a predictive model from complex effects to simple causes, we apply feature attributions to estimate how the inputs change with the outputs. We demonstrate an application of this by studying sports fans' comments made during a game and compare those comments to a simpler metric, win probability. We also expand on a prior study of intergroup bias, demonstrating how our framework can uncover behaviors that other XAI methods may overlook. We discuss the implications of these findings for advancing interpretability in computational linguistics and improving data-driven-decision-making in social contexts.

1 Introduction

Explainable AI (XAI) techniques have proven a valuable tool for creating trustworthy and reliable models and have seen growing popularity as the need for transparency becomes more prevalent. Researchers across a range of disciplines (Kalasampath et al., 2025) have discovered that these techniques also have the potential to highlight important features and patterns (Fryer et al., 2021) in a dataset. This is most prevalent in biomedicine (Hossain et al., 2025) where finding important features can save lives. Such tasks assume a causal relationship (Doshi-Velez and Kim, 2017) where complex features directly result in a simple metric. In this paper, we show that the direction of causality in this assumption can be reversed by applying feature attribution techniques to a classifier that takes results as inputs and predicts the cause, as observed in 1 We show that these techniques are viable for finding causal assumptions in the opposite direction, even in difficult tasks.

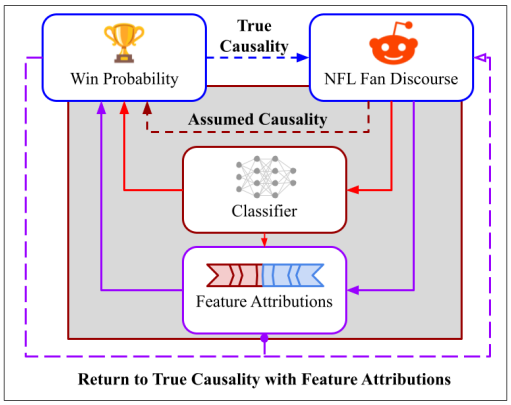

Figure 1: The proposed framework for evaluating the correlation between textual data and simple classifications like team performance. By assuming fan comments actually impact win probability, we can produce feature attributions for those comments to identify interesting behaviors. After we drop the assumed causality, we find that these attributions still accurately represent how win probability can influence fan behaviors.

Many forms of XAI share the perception that models must go from cause to effect (Carloni et al., 2025). The reasoning for this is clear in some applications, such as medical tools like risk analysis (Lundberg et al., 2018), where a limited set of features directly contribute to potential danger. However, models have been observed to learn correlation rather than causation (Molnar et al., 2020). This suggests a reversal of this framework is possible: Make a simple measurable task the 'predic-

Proceedings of Recent Advances in Natural Language Processing,pages 491–500
Varna, Sep 8–10, 2025
https://doi.org/10.26615/978-954-452-098-4-058

tion' and the complex results the 'features'. Unlike existing applications of XAI techniques, this provides a way to analyze reactions, which can enable deeper insights into decision making in discourse (Wu et al., 2024) and unintuitive linguistic patterns (Pennebaker et al., 2014).

There is substantial evidence that suggests this is feasible. When considering medical risk analysis (Lundberg et al., 2018), some features used to predict risk are part of the body's natural reaction to being put in danger, such as heart rate. While the abstraction of 'risk' could technically consider these as 'causes', they are a reaction to the ground truth rather than a contributing factor. Additionally, reversing causal assumptions have shown promising results for training models (Somerstep et al., 2024). This suggests that an XAI framework under a reversed assumption of causality should be effective.

To demonstrate the potential of reverse causality and XAI, we analyze the comments[1] NFL fans make on Reddit as their games progress (Govindarajan et al., 2022). The language of sports fans has been observed to be linguistically rich (Merullo et al., 2019) for analysis where even state-of-the-art models have been observed to struggle (Govindarajan et al., 2024). Despite this, our framework shows promising results for analyzing this data. We apply model agnostic feature attribution methods to show the potential for studying more complex tasks in the future. Overall, our paper provides the following contributions:

- We provide a generic framework for exploring textual data for a variety of potential tasks (§ 3).
- We demonstrate potential 'best practices' for applying our framework and explore the advantages and disadvantages of various feature attribution methods (§ 4).
- We highlight the use of our framework in a complex linguistic task (§ 5).

The remainder of the paper proceeds as follows: Section 2 reviews related work and describes the background of the feature attribution methods as well as intergroup bias in sports. Section 3 outlines our proposed framework for reversing the assumption of causality. Section 4 presents the dataset and preprocessing techniques used in this study. It also outlines the potential design decisions in our framework and highlights the advantages and disadvantages of each. Section 5 studies intergroup bias and explains what behaviors we were able to recreate from prior works.

[1]Disclaimer: This paper contains examples which some readers may find disturbing.

2 Related Work

Broadening Applications of Model Agnostic Feature Attribution There are a vast number of XAI techniques to calculate the significance of individual features (Adadi and Berrada, 2018). However, the 'gold standard' for many of these is SHAP (Mosca et al., 2022). This method of feature attribution is based on Shapley values, which have extensive theoretic background in game theory (Shapley, 1952). These techniques apply every permutation of input features and measure how those permutations affect the output. Since they only require inputs and outputs, they are particularly useful when analyzing black-box models. Similar methods have also been developed, with Leave-One-Out (LOO) (Maron and Moore, 1993) and Local Interpretable Model-Agnostic Explanation (LIME) (Ribeiro et al., 2016) being among the most prominent.

While there are many techniques that may perform better or faster for a variety of tasks, SHAP, LIME, and LOO are still widely applied in a variety of studies (Kalasampath et al., 2025). Additionally, they are generically applicable and have solid theoretical backing. However, a severe limitation is applicability. The logical background of game theory is believed to be less reliable if there isn't a clear sense of cause as input and effect as output (Fryer et al., 2021).

This paper proposes that this requirement can be expanded further, broadening which tasks feature attribution methods can be applied to. We demonstrate how our framework enables analysis of complex linguistic behavior, intergroup bias, that XAI is supposed to be inapplicable to.

Furthering Intergroup Bias Research Intergroup bias refers to the patterns in which stereotypes are communicated in linguistic asymmetry (Maass et al., 1989). This is represented by desirable in-group behaviors and undesirable out-group behaviors. Sports is a natural area for such groups to emerge, with both types of behaviors appearing in the supporters of various teams (Zhang et al., 2019).

Prior studies have contrasted win probability with intergroup bias (Govindarajan et al., 2024) to demonstrate in-group protection (Maass et al., 1989) during NFL games. We extend the observations of this research and focus more on identifying specific behaviors, rather than general trends. Through this, we show specific forms of intergroup bias that have been observed in other studies and demonstrate how introducing feature attribution helps highlight these patterns.

3 Explainability Framework

Most applications of 'XAI' rely on the original assumptions of game theory: The 'players' are to blame for the 'results' (Shapley, 1952). This is intuitive, but not a necessary requirement. As such, we propose a reversed structure: The 'results' can be blamed for the behaviors of the 'players'.

This is valuable, as exploring textual data and the usage of language involving particular events can not always be phrased as a reduction. Instead of moving from complex data to simple conclusions, it often requires identifying how simple data can have complex results. In our task, this is describing how people behave when their team is winning or losing. Feature attribution, however, is designed for causal correlation from the complex to the simple. This is highlighted by its use in the medical field: identifying which features contribute to potential risks (Lundberg et al., 2018).

Causality in feature attribution is important because it suggests the predicted change will occur in the real system, not just the the modeled one (Doshi-Velez and Kim, 2017). Unlike some other forms of XAI, feature attributions do not inherently provide causality (Zhou et al., 2022). This makes it very important for researchers to apply their own reasoning. Since humans have shown significant bias in causal reasoning (Leszczensky and Wolbring, 2022), most applications of XAI still use the default assumption of causality.

However, this order of causality is merely implied rather than built-in. Models measure correlation, not causation. Feature attribution determines the strength of that correlation. As such, we propose that reversing the assumed causality is an equally valid interpretation of observed results. In our task, there is both correlation and causation between fan reactions and probability of victory. Unlike in the medical field, however, it is overwhelmingly more reasonable that online speech is impacted by team performance rather than the other way around.

To demonstrate this, consider the feature attribution of Leave-One-Out (Maron and Moore, 1993) for a feature x_i in an input x.

$$\varphi_{LOO}(x_i) = f(x) - f(x \backslash x_i)$$

LOO estimates the importance of a feature by evaluating the change in output that occurs when removing it. For our task, the 'cause' x is the win probability. This is neither applicable for feature attribution nor helpful. We cannot perform attribution over a single feature and it will not provide us any interesting insight into the 'results' y, which is fan behavior. However, we can perform a similar operation to estimate this association in reverse.

$$\varphi_{LOO}(y_i) = f^{-1}(y) - f^{-1}(y \backslash y_i)$$

By estimating f^{-1}, we can approximate how each y_i 'affects' x. This demonstrates the connections the model makes between y_i and x to determine their association. In context of our case study, this would be the association between a word y_i and the probability of winning x. It also maintains the real-world grounding expected of feature attributions (Carloni et al., 2025), albeit applied in the opposite direction.

We highlight this association in Figure 1. Even though the true cause is the win probability, we treat that as the output of f^{-1}. Similarly, the result of winning or losing, fan responses y, are treated as the input. This is effectively assuming a reverse of the true cause-and-effect relationship between them: How do individual words cause a team to win or lose? We then apply the original feature attribution methods under that false assumption. After doing so, we aggregate our results and return to the true causal direction. The feature attributions now point in the opposite direction: How is winning or losing associated with individual words?

In our experiments, we demonstrate that this reversal enables the exploration of datasets that would be impossible under existing assumptions.

4 Data & Experiments

4.1 Datasets & Preprocessing

Similar to Govindarajan et al. (2024), our dataset of posts comes from subreddits dedicated to different teams in the NFL. During the NFL season, each subreddit has posts for discussing each game

live. This enables a parallel intergroup language dataset, where we can observe the perspectives of the supporters of two teams actively in competition. It also makes it very easy to assume that the in-group of individual comments will be the team of that subreddit.

In contrast to the original study, we focus on games from 2023-2025 and their respective comments. In total, we assess 23,801 comments across twelve subreddits. To estimate the win probability, we use `nflFastR` (Carl et al., 2022), one of many tools developed for win prediction. By aligning comment timestamps with game events, we can accurately estimate the WP for a given comment to determine whether the team is winning or losing.

Overall, these existing tools and prior results (Govindarajan et al., 2024) provide solid grounding as a case study for how our framework can be applied to discover behaviors that were previously overlooked and highlight how it enables exploration of a complex topic: intergroup bias.

In order to assess our dataset, we first construct a correlation between win probabilities and online comments, rather than the game state, with a lightweight classifier. From this structure, we experiment with different feature attribution methods and ways to approach the data to highlight how it exposes various behaviors. Then, we reintroduce intergroup labels to determine whether our methods can provide more in-depth or subtle insights into the results of prior studies.

4.2 Exploring Feature Attribution Methods

In order to perform feature attribution, we must first have a function that converts inputs into outputs. While the prior study merely correlated win probability with comments (Govindarajan et al., 2024), that is insufficient. To achieve this, we use the embeddings of a small language model (AI@Meta, 2024) to convert text into an input space, then train a classifier to predict the win probability from the sentence embeddings. This network consists of 2 layers with 500 features each, and ReLU activation between them. We add one connected layer that outputs two features and apply a softmax to make an approximation of $f^{-}1(y)$ as described in the prior section. Our model achieves a Mean-Squared-Error of 0.074 on probability predictions. We then apply a variety of feature attribution methods to various inputs on that classifier.

For the purpose of defining these attribution methods, let x be the set of input features $\{x_0, x_1, ...x_n\}$ and have the score function for a feature x_i be $\varphi(x_i)$. Let $f(x)$ be the output of the model being explained. We do not use the language defined previously, since the definitions of feature attributions assume a sense of forward causality.

Leave-One-Out Leave-One-Out (Maron and Moore, 1993), or LOO, is the most simplistic form of feature attribution. It consists of removing an individual feature and observing how much a prediction changes. While this is efficient and intuitive, it lacks insight on the interplay between different features and is prone to inconsistency between different inputs.

$$\varphi_{LOO}(x_i) = f(x) - f(x\backslash i)$$

Shapley Values (Štrumbelj and Kononenko, 2014). In direct contrast to LOO, Shapley values or SHAP are expensive but effective. Instead of calculating individual influences, SHAP observes every subset of features and uses them to estimate how much influence can be ascribed to each individual. It is significantly more effective and has become commonplace in XAI. However, the cost of predicting every permutation of features is computationally expensive. NLP largely avoids this issue since the vast majority of features are not present. However, substantially long samples can make Shapley values unwieldy. Due to this, we filter out such samples while using Shapley feature attribution. To reduce the need for this as much as possible, we group every other word of each post together to reduce the number of features. This allows us to process longer posts with Shapley feature attribution.

$$\varphi_{SHP}(x_i) = \frac{\sum_{S \subseteq x\backslash\{i\}} \binom{n-1}{|S|}^{-1} (f(S \cup i) - f(S))}{n}$$

LIME(Ribeiro et al., 2016). The final form of feature attribution we consider is LIME. In contrast to Shapley and LOO, LIME does not measure the outputs of a model directly. Instead, it creates an interpretable version that can perform feature attributions through knowledge distillation. Similar to Shapley, LIME has been adopted for a wide range of applications in XAI. In contrast, however, it does not directly represent the impact of each feature, making it less accurate. As such, it is generally considered a decent balance between the instability of LOO and the cost of SHAP.

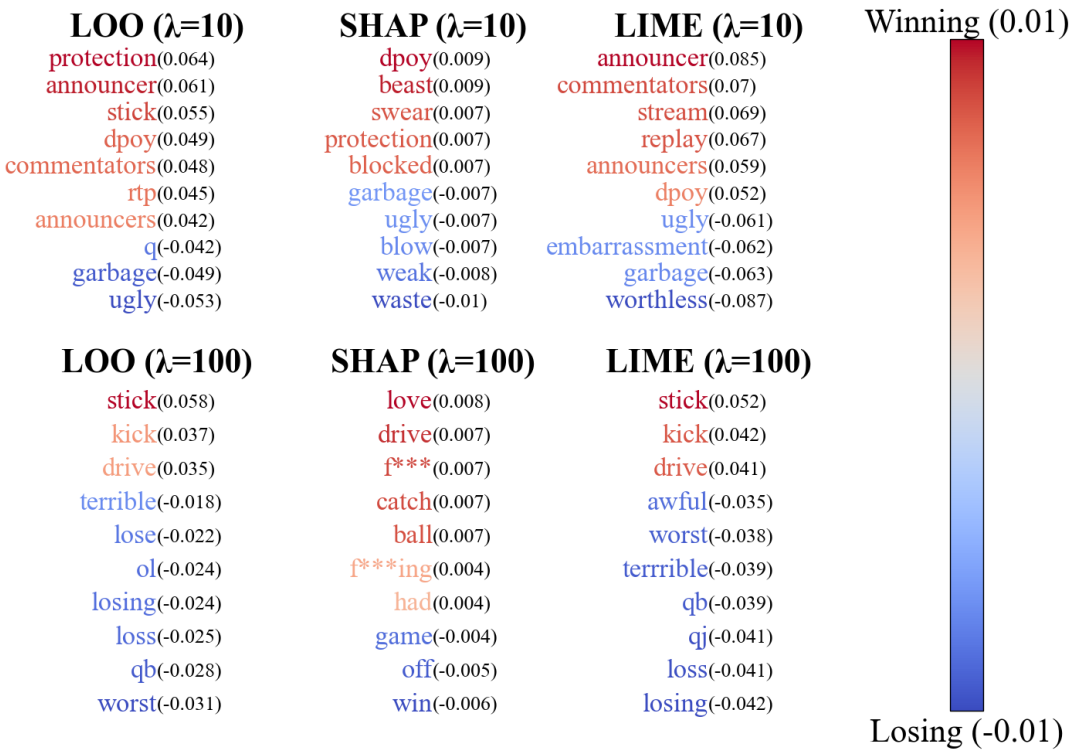

Figure 2: Words with the 10 highest attribution scores for LOO, SHAP, and LIME with $\lambda = 100$ and $\lambda = 10$. Red words are associated with winning, while blue words are associated with losing. The intensity of the color describes how strongly the association is. Some patterns that emerge are that words referring to game actions are observed more often while winning (such as 'kick' and 'drive'), and players are more often referred to while losing. This is observed in terms like 'qb' (quarterback) or 'ol' (offensive line).

4.3 Analyzing Patterns Highlighted by Attribution

For the purpose of visualization, we define a λ that represents the minimum number of samples that have to include a feature in order to have that feature represented in the visualization. This allows us to focus on features with consistent use and meaning across multiple posts, rather than those that only showed up once or twice. For each attribution method and each λ, Figure 2 shows these top 10 features with the most positive or negative median attributions.

Positive attributions represent features that are used by the classifier to predict 'winning' while negative attributions means a term was used to predict 'losing'. Our filtering restricts the presented results to those that are consistently associated with a certain outcome.

λ = 10 Analysis The top row of Figure 2 represents $\lambda = 10$, requiring each feature to have been observed at least 10 times.

We can see the emergence of some interesting patterns. First, we see that LIME ($\lambda = 10$) finds that 'announcers' and 'commentators' both have a strong association with winning. This is because when a team starts performing well, their fans become defensive towards criticism. One example reads as "I love how the announcers are mocking... Looks pretty pathetic when he's making some damn good plays." In contrast to this defensiveness, we also see the introduction of new insults towards their team's performance: 'weak', 'garbage', and 'ugly'. These tend to have more extreme associations with losing, with LIME scores around 0.067. Additionally, we see shorthand of sports terms emerge in these posts. For example, 'q' is shorthand for 'quarter'. The negative association (-0.04) emerges when people are asking for their team to pull through in the fourth quarter. In contrast to that, we have the 'DPOY', or 'Defensive Player of the Year'. This has a strong positive association (0.06) as fans associate their 'DPOY' with a strong defense and better performance in the game.

λ = 100 Analysis The bottom row of Figure 2 shows the results of LOO, SHAP, and LIME while requiring at least 100 occurrences of a word.

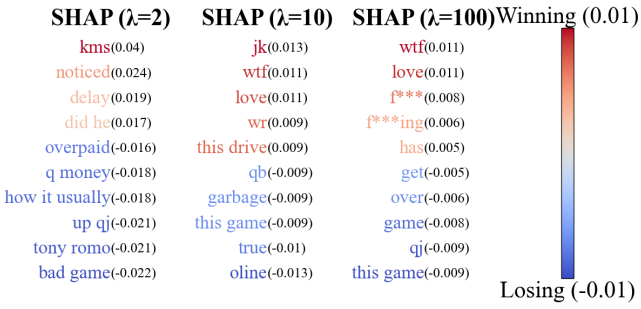

Figure 3: Top 10 word attributions for Clustered SHAP with $\lambda = 2$, $\lambda = 10$, and $\lambda = 100$. We observe multi-word phrases showing up, though often being overtaken by simple and fairly independent words. This suggests NLTK's POS tagging is effective, but imperfect.

With $\lambda = 100$, common and significantly hostile terms like 'terrible', 'worst', or 'losing' are strongly associated with negative attributions. These can be observed throughout the bottom row of Figure 2 and had LOO scores around -0.03, SHAP near -0.01, and LIME scores of -0.04, indicating that they are most used while a team is losing. Interestingly, positive terms align better with a team's actions in the game, rather than complimenting the team's members. When a team is winning, the fans focus more on the most recent 'catch', 'drive', or 'kick', with LOO scores around 0.03, SHAP around 0.09, and LIME scores around 0.04.

Even before introducing intergroup labeling, we can see how feature attributions are able to highlight both known behaviors and unintuitive patterns. In addition to the common sense associations of negative words being used while losing, we can already observe a bias emerge against outside commentators. Additionally, we highlight an interesting behavior where fans use subjective terms while losing ('terrible', 'worst', 'garbage') but focus on objective actions while winning ('catch', 'blocked', 'drive').

4.4 Reducing Computational Complexity of SHAP

One of the major downsides to SHAP is the computational cost. To reduce the number of features contained in each comment, we apply the Natural Language Toolkit's Regular Expression Parser (Bird et al., 2009). This process let us track the structure of a sentence and separate it into the shorter phrases that compose it, using Part-of-Speech tagging to create a tree of sentence structure. For the purpose of clustering, we elect to group leaves together with their root. This ensures only short phrases are captured and guarantees relevant connections. However, a more advanced speech clustering method could potentially provide stronger insights than discovered here.

In addition to the increased efficiency of Shapley Attribution, grouping words together also allows us to gather additional insights into common phrases, instead of individual words. Since these features are composed of meaningfully associated words, the phrases that emerged were only those with significance. Additionally, this structure maintains many of the words that were independently important.

4.5 Novel Behaviors Observed with Clustering

The advantages of clustering can be observed in Figure 3 where requiring only a small number of occurrences, such as $\lambda = 2$ (the first graph) and $\lambda = 10$ (the second graph) creates a number of two or three word phrases. The phrase 'how it usually' can be observed in $\lambda = 2$ and is associated with losing. It also provides greater explanations for prior results. The term 'game' emerged in Figure 2 with a negative attribution of -0.006 and $\lambda = 2$ in Figure 3 shows that this may be due to the phrase 'bad game' which has a much stronger relation to losing, -0.022. Similarly, 'these commentators' is also related to 'commentators'.

Despite the insightful results, this method also severely limits the number of features that emerge. Limiting attributions to $\lambda = 100$, as seen in the third graph, produces only single-word attributions. This is because two-word features are inherently less common. Even reducing to $\lambda = 10$, only provides us with a few additions such as 'this drive' or 'this game'.

Clustering based on part-of-speech tagging provides some additional insights and improved efficiency. It also better highlights features that LOO and LIME originally mentioned that SHAP does not, such as the 'commentators'. These benefits come at a trade-off of more human effort and inconsistency, requiring small λ to identify multi-word terms and phrases. However, it remains significantly faster than trying to observe these behaviors directly from Reddit posts, making this framework a potentially valuable tool for efficiently exploring data in depth.

5 Case Study: Intergroup Bias

5.1 Implementing Intergroup Bias

Introducing intergroup tagging allows us to expand the insights from Govindarajan et al. (2024) as a case study for our framework. In prior research, GPT-4o (Hurst et al., 2024) performed accurate tagging of in-group and out-group labels. These tags refer to whether the commenter refers to part of the team they're supporting (IN) or part of a team they're opposed to (OUT). We acquire tags for each comment through few-shot prompting. These tags are matched to the feature attributions acquired previously to further explore how intergroup bias plays a role in Reddit comments.

For the purpose of exploring intergroup bias, we focus largely on LIME. While SHAP Clusters are able to provide more details, they occur far less often. This means inconsistency in GPT-4o based labels can be significantly more impactful. Since the prior study found GPT-4o had an F1-score of 65.3 (Govindarajan et al., 2024), this is a significant concern.

Features in LIME occur more frequently, making the unreliability of the model's performance less of a concern and our results more stable. However, a more accurate means of intergroup labeling could likely achieve even more complex insights than we discuss here.

5.2 Expanding Prior Research With Specific Behaviors

Demonstrations of Outgroup-Animosity Intergroup bias can manifest in a number of ways, but one of the most consistent is outgroup-animosity, which refers to consistent negativity towards those not in the ingroup (Rathje et al., 2021). Outgroup-animosity has been observed in sports (Cobbs et al., 2017), even for adjacent groups such as sponsors (Lin and Bruning, 2020).

We observe such behaviors in our study as well. This can be seen in Figure 5, where the terms used while referring to the outgroup are generally more negative than those referring to the in-group. We achieve these results by restricting features to the group they were used to refer to more often. This helps us avoid the low accuracy of intergroup labeling that we observed prior. As Figure 5 shows the top 10 features most associated with certain outcomes for both groups, we observe that in-group features are generally more positive. Multiple terms like 'mvp', 'love', 'special', and 'swear' emerge as examples of this. In contrast, terms used most often to refer to the out-group appear negative - 'garbage', 'weak', 'horrible'.

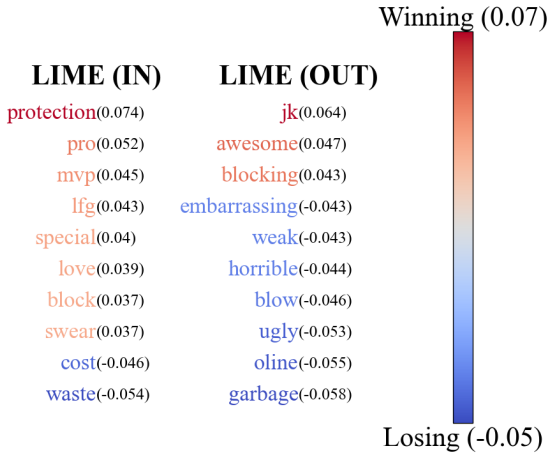

Figure 4: The most associated and commonly used terms for the In-group and the Out-group. While the in group does have some unfavorable terms, the terms are generally positive in meaning. In contrast, terms labeled part of the out group are almost overwhelmingly negative, both in meaning and in relation to win probability. This is likely a form of outgroup-animosity. When a team is losing, their fans are more likely to be negative towards the opposing team.

Notably, we observe that this behavior is enhanced by competition, matching prior studies. Animosity towards their opponents emerges most when the in-group's team is losing. However, it is worth noting once more that the labeling mechanism is heavily flawed. 'Embarassing' emerges as an out-group term, which seems intuitively inaccurate. Despite this, the results appear to support a behavior that has been observed by prior studies, suggesting our framework is likely effective.

Leader-Oriented Animosity Against Other Coaches By filtering our resultant data to the names of 'leaders' of individual teams, we see intense bias for and against individuals. Figure 5 shows that while there's a variance of attitude towards individual coaches, the term 'coach' itself has stronger connotations with losing. This suggests that fans place heavy emphasis on their leaders, blaming them for the team's success or failures.

Most notably, 'coach' being negative demonstrates out-group animosity directed toward the leader of opposing teams. While we manually verify that this holds true within the dataset, it is also

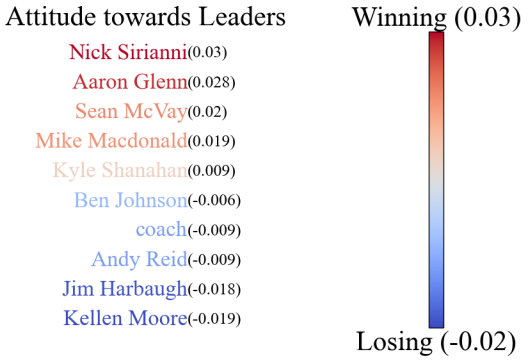

Figure 5: The NFL coaches that appeared most often and had strong association with winning or losing. While individual coaches cover a wide range of attributions, the term 'coach' is negative. This is most likely another form of outgroup-animosity targeted at the leaders of the outside group. This is because the in-group is far more likely to know and use their coach's name.

intuitive. The fans of opposing teams may only know the name of their own coach. As such, they use a less-specific term, 'coach', to refer to their opponent's coach instead. These results match prior research (Merten et al., 2023) that suggest this is another manifestation of intergroup bias.

Role in XAI Research Intergroup bias is a complex linguistic task that requires analysis of textual datasets with large vocabularies. While we can approximate similar results using classifiers and intuitively reasonable words (Govindarajan et al., 2024), these do not enable analysis of more complex behaviors.

Additionally, existing XAI research would assume that the majority of techniques cannot be used to analyze reactive behaviors (Lundberg et al., 2018). However, our results show that this limitation does not hold. By applying feature attribution on fan reactions rather than the cause of team performance, we are able to analyze intergroup bias to provide novel insights into human behavior. This demonstrates both the effectiveness of our framework and the potential for future research.

6 Conclusion

In this work, we propose a framework that adapts feature exploration methods to textual datasets with significant size. We demonstrate how this can be used by expanding the analysis of an existing research topic (Govindarajan et al., 2024), intergroup bias. Additionally, we show that feature attributions methods are reliable if we have many dimensions in the 'effect' and few for the 'cause', reversing previous assumptions of the required causality for XAI techniques. We show how this framework allows us to observe forms of behavior that previously rely on more in-depth studies. We compare attributions methods such as SHAP and LIME to determine what patterns are highlighted and propose text clustering as a way to modify the explored set of features. We also highlight numerous behavioral patterns observed through these methods, including a tendency towards more concrete observations while winning and more emotional outbursts while losing.

We demonstrate how our method allows us to explore intergroup bias and how fans treat various groups unequally. Our results show that the use of XAI for feature exploration and discovery can be expanded to reactive contexts. This has notable applications in a variety of other textual contexts, including AI discourse (Atwell et al., 2024), media parsing (Lei et al., 2022), and political bias (Wang et al., 2024). In future works, we plan to explore how other forms of bias manifest in different mediums, and how XAI can be used to identify significant features for bias evaluation.

7 Limitations

Our work relies on inconsistent and often unreliable models. There are two models that predict win-probability, one from game state and another from text embeddings. GPT-4o is also black box. Involving AI in so many steps of the process limits both reproducibility and accuracy.

Feature attribution methods are also imperfect. These methods can be unreliable (Shen et al., 2025) and may not align with human interpretations (Nguyen et al., 2021). Despite this, our results show alignment with known and expected behaviors, suggesting generalization.

Similarly, our study heavily relies on human interpretation. While prior studies have similar practices (Lundberg et al., 2018), they are often shortened sets of features and more aligned with the general understanding of XAI as a tool to help humans interpret models. Additionally, we only consider data from NFL fans on Reddit. Since we only show the results of a single case study, we cannot guarantee generalization. However, our framework is developed from reliable XAI methods that have proven consistent across fields.

Ethics Statement

This research presents valuable insights into expanding the applications of XAI. While it relies heavily on well-supported attribution methods, the proposed application steps beyond their expected use case. Our results suggest that this remains effective, but it could encourage further misuse of such tools. Further research should be performed to ensure that attribution methods appear to consistently follow causality in the ways described here, even across different tasks and datasets.

Additionally, this could potentially further encourage companies to introduce flawed AI systems to tasks they aren't ready for yet. This paper suggests a new application of XAI techniques that is not thoroughly explored, but has significant motivation for application, particularly in moderation. This emphasizes the importance of validating the presented results and assessing their generalizability to other tasks.

References

Amina Adadi and Mohammed Berrada. 2018. Peeking inside the black-box: a survey on explainable artificial intelligence (xai). *IEEE access*, 6:52138–52160.

AI@Meta. 2024. [link].

Katherine Atwell, Mert Inan, Anthony B. Sicilia, and Malihe Alikhani. 2024. Combining discourse coherence with large language models for more inclusive, equitable, and robust task-oriented dialogue. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3538–3552, Torino, Italia. ELRA and ICCL.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

Sebastian Carl, Ben Baldwin, and L Sharpe. 2022. nflfastr: Functions to efficiently access nfl play by play data. *R package version*, 4(1).

Gianluca Carloni, Andrea Berti, and Sara Colantonio. 2025. The role of causality in explainable artificial intelligence. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 15(2):e70015.

Joe Cobbs, B Daniel Sparks, and B David Tyler. 2017. Comparing rivalry effects across professional sports: National football league fans exhibit most animosity. *Sport Marketing Quarterly*, 26(4):235–246.

Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

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.

Venkata S Govindarajan, Katherine Atwell, Barea Sinno, Malihe Alikhani, David I Beaver, and Junyi Jessy Li. 2022. How people talk about each other: Modeling generalized intergroup bias and emotion. *arXiv preprint arXiv:2209.06687*.

Venkata S Govindarajan, Matianyu Zang, Kyle Mahowald, David Beaver, and Junyi Jessy Li. 2024. Do they mean'us'? interpreting referring expressions in intergroup bias. *arXiv preprint arXiv:2406.17947*.

MD Imran Hossain, Ghada Zamzmi, Peter R Mouton, MD Sirajus Salekin, Yu Sun, and Dmitry Goldgof. 2025. Explainable ai for medical data: current methods, limitations, and future directions. *ACM Computing Surveys*, 57(6):1–46.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Khushi Kalasampath, K. N. Spoorthi, Sreeparvathy Sajeev, Sahil Sarma Kuppa, Kavya Ajay, and Angulakshmi Maruthamuthu. 2025. A literature review on applications of explainable artificial intelligence (xai). *IEEE Access*, 13:41111–41140.

Yuanyuan Lei, Ruihong Huang, Lu Wang, and Nick Beauchamp. 2022. Sentence-level media bias analysis informed by discourse structures. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10040–10050, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Lars Leszczensky and Tobias Wolbring. 2022. How to deal with reverse causality using panel data? recommendations for researchers based on a simulation study. *Sociological Methods & Research*, 51(2):837–865.

Hsin-Chen Lin and Patrick F Bruning. 2020. Comparing consumers' in-group-favor and out-group-animosity processes within sports sponsorship. *European Journal of Marketing*, 54(4):791–824.

Scott M Lundberg, Bala Nair, Monica S Vavilala, Mayumi Horibe, Michael J Eisses, Trevor Adams, David E Liston, Daniel King-Wai Low, Shu-Fang Newman, Jerry Kim, et al. 2018. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nature biomedical engineering*, 2(10):749–760.

Anne Maass, Daniela Salvi, Luciano Arcuri, and Gün R Semin. 1989. Language use in intergroup contexts: The linguistic intergroup bias. *Journal of personality and social psychology*, 57(6):981.

Oded Maron and Andrew Moore. 1993. Hoeffding races: Accelerating model selection search for classification and function approximation. *Advances in neural information processing systems*, 6.

Sebastian Merten, Nicolas Reuland, Mathieu Winand, and Mathieu Marlier. 2023. Fan identification in football: professional football players and clubs competing for fan loyalty. *Sport, Business and Management*, 14(2):169–187.

Jack Merullo, Luke Yeh, Abram Handler, Alvin Grissom II, Brendan O'Connor, and Mohit Iyyer. 2019. Investigating sports commentator bias within a large corpus of American football broadcasts. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6355–6361, Hong Kong, China. Association for Computational Linguistics.

Christoph Molnar, Gunnar König, Julia Herbinger, Timo Freiesleben, Susanne Dandl, Christian A Scholbeck, Giuseppe Casalicchio, Moritz Grosse-Wentrup, and Bernd Bischl. 2020. General pitfalls of model-agnostic interpretation methods for machine learning models. In *International Workshop on Extending Explainable AI Beyond Deep Models and Classifiers*, pages 39–68. Springer.

Edoardo Mosca, Ferenc Szigeti, Stella Tragianni, Daniel Gallagher, and Georg Groh. 2022. Shap-based explanation methods: a review for nlp interpretability. In *Proceedings of the 29th international conference on computational linguistics*, pages 4593–4603.

Giang Nguyen, Daeyoung Kim, and Anh Nguyen. 2021. The effectiveness of feature attribution methods and its correlation with automatic evaluation scores. *Advances in Neural Information Processing Systems*, 34:26422–26436.

James W Pennebaker, Cindy K Chung, Joey Frazee, Gary M Lavergne, and David I Beaver. 2014. When small words foretell academic success: The case of college admissions essays. *PloS one*, 9(12):e115844.

Steve Rathje, Jay J Van Bavel, and Sander Van Der Linden. 2021. Out-group animosity drives engagement on social media. *Proceedings of the national academy of sciences*, 118(26):e2024292118.

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.

Lloyd S. Shapley. 1952. *A Value for N-Person Games*. RAND Corporation, Santa Monica, CA.

Gaofei Shen, Hosein Mohebbi, Arianna Bisazza, Afra Alishahi, et al. 2025. On the reliability of feature attribution methods for speech classification. *arXiv preprint arXiv:2505.16406*.

Seamus Somerstep, Yuekai Sun, and Ya'acov Ritov. 2024. Learning in reverse causal strategic environments with ramifications on two sided markets. *arXiv preprint arXiv:2404.13240*.

Erik Štrumbelj and Igor Kononenko. 2014. Explaining prediction models and individual predictions with feature contributions. *Knowledge and information systems*, 41:647–665.

Sidi Wang, Gustav Eggers, Alexia de Roode Torres Georgiadis, Tuan Anh o, Léa Gontard, Ruth Carlitz, and Jelke Bloem. 2024. Towards quantifying politicization in foreign aid project reports. In *Proceedings of the Second Workshop on Natural Language Processing for Political Sciences@ LREC-COLING 2024*, pages 85–90.

Yating Wu, Ritika Rajesh Mangla, Alex Dimakis, Greg Durrett, and Junyi Jessy Li. 2024. Which questions should I answer? salience prediction of inquisitive questions. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 19969–19987, Miami, Florida, USA. Association for Computational Linguistics.

Jason Shuo Zhang, Chenhao Tan, and Qin Lv. 2019. Intergroup contact in the wild: Characterizing language differences between intergroup and single-group members in nba-related discussion forums. *Proceedings of the ACM on Human-Computer interaction*, 3(CSCW):1–35.

Yilun Zhou, Serena Booth, Marco Tulio Ribeiro, and Julie Shah. 2022. Do feature attribution methods correctly attribute features? In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 9623–9633.