High Dimensional Model Explanations: An Axiomatic Approach (Supplementary material)

Neel Patel
University of Southern California
neelbpat@usc.edu

Martin Strobel
National University of Singapore
mstrobel@comp.nus.edu.sg

Yair Zick
University of Massachusetts, Amherst
yzick@umass.edu

A USEFUL TECHNICAL RESULTS

Proposition A.1. *Given a primitive game p^R, for any $S \subseteq N$: if $S \not\subseteq R$ then $\forall T \subseteq N \setminus S$, $m_S(T, p^R) = 0$. In particular, $I^{p^R}_{\texttt{BII}}(S) = 0$.*

Proof. Suppose that $S \not\subseteq R$. We distinguish between two cases.

Case 1: $R \not\subseteq T \cup S$. In this case, $\forall L \subseteq S$, $T \cup L$ does not contain R, thus $p^R(T \cup L) = 0$ which implies $m_S(T, p^R) = 0$.

Case 2: $R \subseteq T \cup S$. In this case, $m_S(T, p^R)$ equals

$$\sum_{L \subseteq S} (-1)^{|S|-|L|} p^R(L \cup T) = \sum_{L \subseteq S : S \cap R \subseteq L} (-1)^{|S|-|L|} = \sum_{L \subseteq S \setminus R} (-1)^{|S|-|S \cap R|-|L|} = \sum_{L \subseteq S \setminus R} (-1)^{|S \setminus R|-|L|} = \sum_{k=0}^{|S \setminus R|} (-1)^k \binom{|S \setminus R|}{k} = 0$$

Thus in either case $m_S(T, p^R) = 0$, and we are done. □

We note that Proposition A.1 immediately holds for any game that is a scalar multiple of a primitive game, i.e. for any $v = c \times p^R$, and any $S \not\subseteq R$, $I^v_{\texttt{BII}}(S) = 0$.

Proposition A.2. *Given a primitive game p^R, for any $S \subseteq N$: If $v = c \times p^R$, and $S \subseteq R$ then*

$$I^v_{\texttt{BII}}(S) = \frac{c}{2^{|R|-|S|}}.$$

Proof. Since $S \subseteq R$, then for any $L \subset$ and any $T \subseteq N \setminus S$, $v(L \cup T) = 0$. Therefore,

$$I^v_{\texttt{BII}}(S) = \frac{1}{2^{n-|S|}} \sum_{T \subseteq N \setminus S} \sum_{L \subseteq S} (-1)^{|S|-|L|} v(L \cup T) = \frac{1}{2^{n-|S|}} \sum_{T \subseteq N \setminus S} v(S \cup T) \quad (1)$$

Next, if $T \subseteq N \setminus S$ does not contain $R \setminus S$, then $v(S \cup T) = 0$. Therefore, the summand in (1) equals

$$\sum_{T \subseteq N \setminus R} v(R \cup T) = c \times \sum_{T \subseteq N \setminus R} 1 = c \times 2^{n-|R|}$$

Plugging this back into (1), we obtain the desired result. □

Next, let us characterize how influence measures satisfying our axioms behave on primitive games. Note that Lemma A.3 offers a special case of Proposition A.1 for *any* influence measure, rather than just for BII .

Lemma A.3. *If I^v satisfies (S), (GE), (L) and (M) then for $v = c \times p^R$, $I^v(R) = c$*

Proof. We prove this lemma by inductively removing a feature $k \in N \setminus R$ and using the (GE) property at each step. Take any feature $i \in R$ and remove it from R and define $S := R \setminus \{i\}$. Now for any feature $j_1 \in N \setminus R \neq i$, by (GE) property we can write,

$$I^{v_{[ij_1]}}(S \cup [ij_1]) = I^v(S \cup i) + I^v(S \cup j_1)$$

Since $S \cup j_1 \not\subseteq R$, by Proposition A.1, $m_{S \cup j_1}(T, v) = 0$ for all $T \subseteq N \setminus \{S \cup j_1\}$. Therefore by Lemma **??**, $I^v(S \cup j_1) = 0$, which yields $I^v(R) = I^{v_{[ij_1]}}(S \cup [ij_1])$. We next remove $j_2 \neq j_1 \in N \setminus R$, and again invoke the (GE) property:

$$I^{v_{[ij_1j_2]}}(S \cup [ij_1j_2]) = I^{v_{[ij_1]}}(S \cup [ij_1]) + I^{v_{[ij_1]}}(S \cup j_2)$$

It is easy to check that $I^{v_{[ij_1]}}(S \cup j_2) = 0$ by Proposition A.1 and Lemma **??** for the reduced game $v_{[ij_1]}$.

$$v_{[ij_1]}(S') = \begin{cases} c, & \text{if } S \cup [ij_1] \subseteq S' \\ 0, & \text{else} \end{cases}$$

$m_{S \cup j_2}(T, v_{[ij_1]}) = 0$ for any $T \subseteq \{N \setminus \{i, j_1\} \cup [ij_1]\} \setminus \{S \cup j_2\}$. This implies that $I^v(R) = I^{v_{[ij_1j_2]}}(S \cup [ij_1j_2])$. By repeating this argument for all $j \in N \setminus R$, we will have $I^v(R) = I^{v_{[N \setminus S]}}(S \cup [N \setminus S])$. We can write the reduced game $v_{[N \setminus S]}$ as

$$v_{[N \setminus S]}(S') = \begin{cases} c, & \text{if } S' = S \cup [N \setminus S] \\ 0, & \text{else.} \end{cases}$$

By the Limit (L) property,

$$I^{v_{[N \setminus S]}}(S \cup [N \setminus S]) = m_{S \cup [N \setminus S]}(\emptyset, v_{[N \setminus S]}) = c,$$

which concludes the proof □

B ADDITIONAL EXAMPLE

In the following example, we demonstrate that the Shapley interaction index can be misleading in simple situations. We consider the general majority classification function which exhibits pairwise feature interaction. Shapley interaction indices fail to capture these interactions. Moreover, Shapley-Taylor interaction indices fail to capture the sign of pairwise interactions for the same function.

Example B.1. [2] Consider a classification function whose input space is binary. Let the classification function f be: $f(x_1, \ldots, x_n) = 1$ iff $\sum_i x_i \geq k$ and 0 elsewhere with the baseline vector $\vec{x}' = \vec{0}$. Thus, $v(S, \vec{1}, \vec{x}') = 1$ iff $|S| \geq k$ and 0 otherwise. For $k = \frac{n}{2}$, the function coincides with the majority function discussed in the cooperative game theory literature. Clearly, there exists pairwise interaction among features, however, the Shapley interaction value for each pairwise feature is 0. In contrast, the pairwise Banzhaf interaction index for any feature pair $\{i, j\}$ is $c_k(2k - (n + 3))$ where $c_k > 0$. Pairwise interaction is negative when $2k < n + 3$. This can be explained by the following argument: the number of winning coalitions containing $\{i, j\}$ is $\binom{n-2}{k-2}$, and the number of

winning coalitions that do not contain $\{i, j\}$ $\binom{n-2}{k}$, which is higher for smaller k. This shows that $\{i, j\}$ has more interaction effect for the output 0. On the other hand, the Shapley-Taylor pairwise interaction index is $\frac{2}{n(n-1)}$, and fails to capture the sign of the interaction index.

C COMPUTATIONAL COMPLEXITY

Computing BII exactly — even for individual features, where it coincides with the Banzhaf index — is intractable (see [4, Chapter 4] for a general overview); however, due to its probabilistic nature, BII can be well-approximated via random sampling for individual features [3]. The sampling approach is widely used for approximating cooperative game theoretic influence measures for the individual features [5, 10].

To approximate the pairwise interaction for BII , we can decompose pairwise BII as following:

$$I^{v}_{\texttt{BII}}(ij) = \frac{1}{2^{n-2}} \sum_{T \subseteq N \setminus ij} (v(T \cup ij) - v(T \cup i) - v(T \cup j) + v(T))$$
$$= I^{v_{[ij]}}_{\texttt{BII}}([ij]) - I^{v(N\setminus i)}_{\texttt{BII}}(j) - I^{v(N\setminus j)}_{\texttt{BII}}(i) \quad (2)$$

This shows that there exists an efficient sampling scheme which approximates the pairwise BII : each term in Equation 2 can be approximated by polynomially many samples similar to the [5, 10]. Moreover, we further note that the approximation algorithms discussed for approximating Shapley interaction index in Lundberg et al. [9] and polynomial algorithms for the tree-based decision model can be implemented for BII as well.

D BROADER IMPACT

The recent wave of research on explainable machine learning is motivated by the disparity between the success of machine learning, and its pronounced lack of public accountability. This work takes the next natural step towards a more general theory of model explanations. Our work treats explanation complexity as a *tunable parameter*: we can generate explanations of varying degree of complexity, as is appropriate for different contexts and stakeholders. The authors view this as a *positive* effect on society.

Machine Learning models have been shown to potentially leak information about the training data [12] and there is some recent work that shows how explanations can be used to reconstruct models [11] or recover user data [13]. There is a growing body of evidence that explanations make models more prone to leak private training data; such leakage could undermine public trust in machine learning models, the very same trust that model explanations try to establish. It is only natural to assume that our high-dimensional model explanations are more vulnerable to adversarial exploitation, as they release more information per query. This is a potential societal *risk*. Advances in developing differentially private machine learning models [1] offer some hope that privacy preserving model explanations can be developed. In fact, QII is shown to be differentially private (with respect to the sample set) [5].

Finally, our work offers several *mathematical* guarantees of explanation quality, but does not focus on its *actual adoptability*. There is currently little evidence that mathematically justified model explanations, as elegant as they may be, are acceptable by human stakeholders, or actually help humans understand model decisions. In fact, recent work suggests that model explanations can result in overconfidence of software developers in their understanding of models [7]. We believe that an axiomatic approach, like the one chosen in the paper helps ensure that stakeholders understand the extent to which an explanation can be useful. This makes explanation methods such as ours more *trustworthy*. However, we also believe that before model explanation methods are deployed, the research community needs to clearly specify which tools can be used for which tasks, and that end users understand their limitations. Some parts of the community are already moving towards a more user-centric direction and believe that this can help overcome existing shortcomings [6, 8, 14].

E ADDITIONAL EXAMPLE EXPLANATIONS FOR BERT

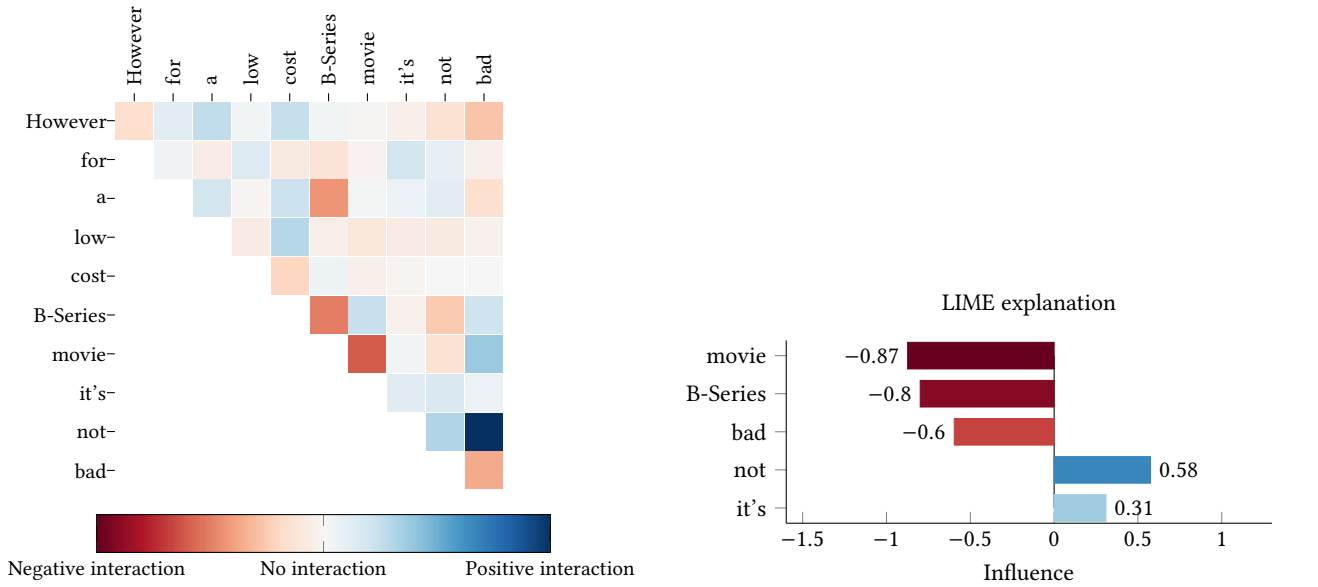

Figure 1: While most individual features have negative influence for both `BII` and `LIME` , `BII` is able to pick up on the strong positive interaction between "not" and "bad" that leads to a positive prediction of this review.

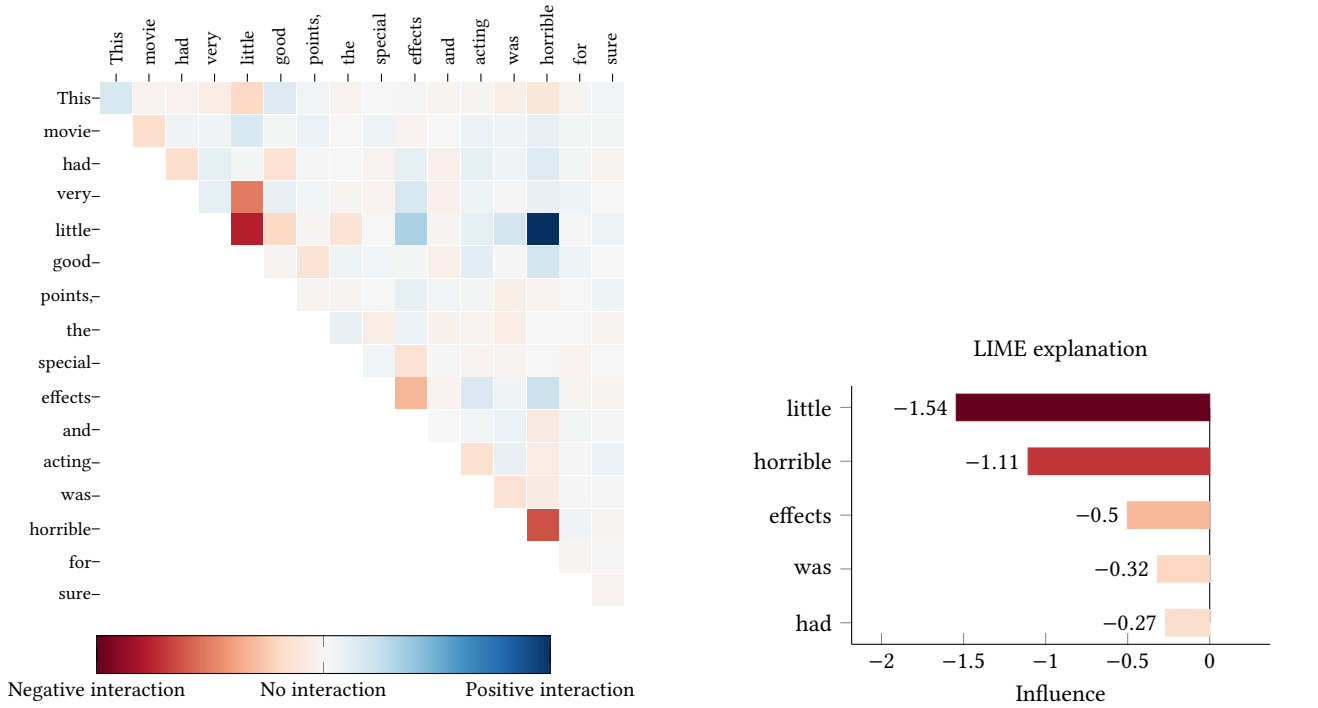

Figure 2: While most individual features have negative influence suggested by `BII` and `LIME` , `BII` is able to pick up on the strong positive interaction between "little" and "horrible" that leads to less confidence on the negative prediction of the review.

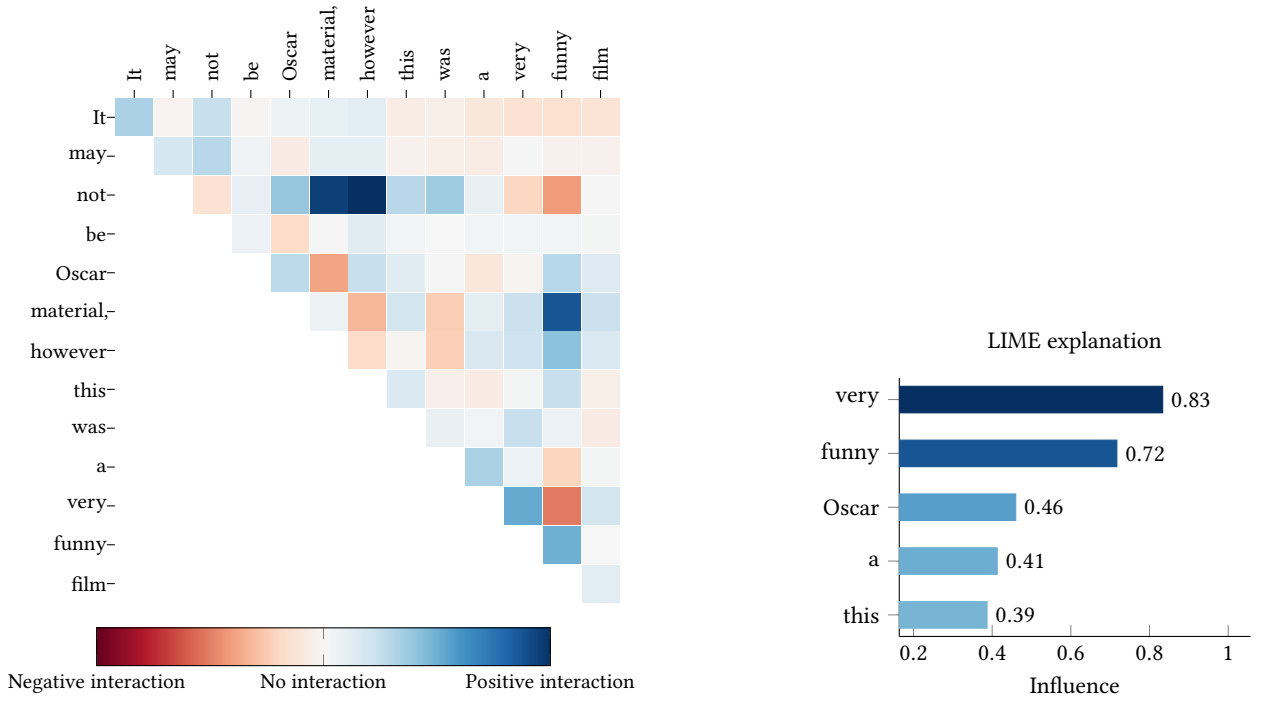

Figure 3: While most individual features have positive influence for `BII` and `LIME` both, `BII` is able to pick up some interesting interactions among words. For example the strong negative interaction between "not" and "funny" and "very" and "funny" that leads to less confidence on the negative prediction of the review.

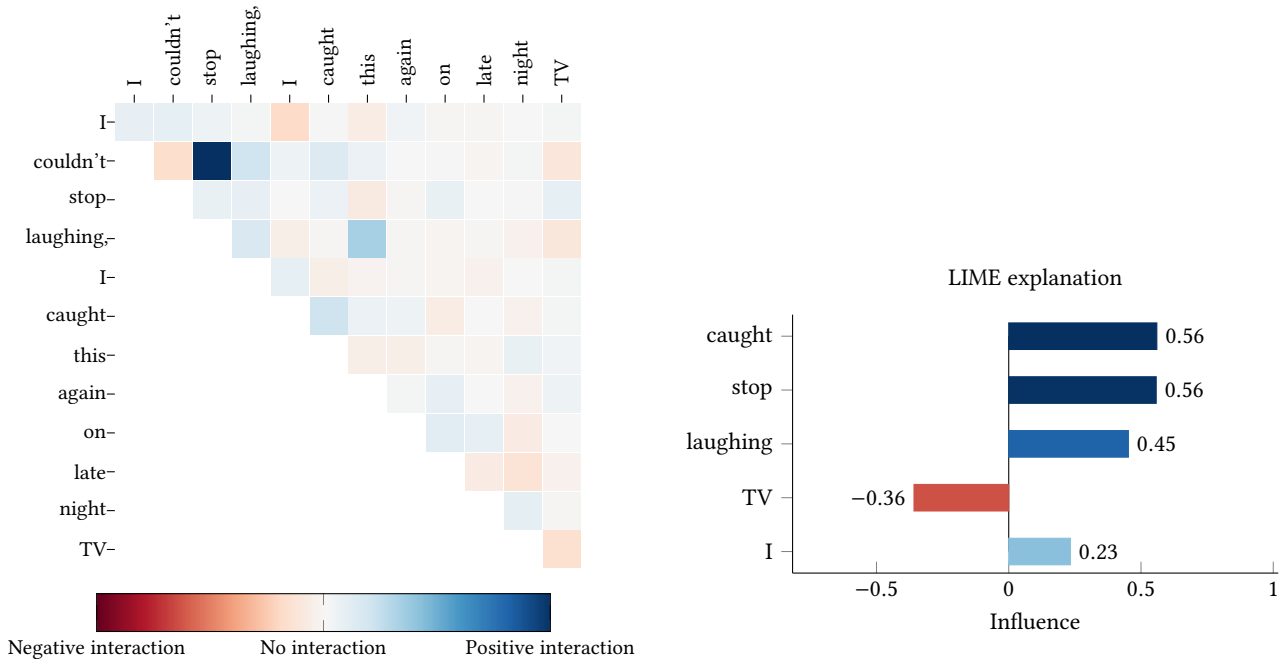

Figure 4: `BII` assigns almost negligible importance to individual word and picks up on the strong positive interaction between "couldn't" and "stop" that leads to a positive prediction of this review. However, `LIME` or any other feature based model explanation fails to capture such interactions.

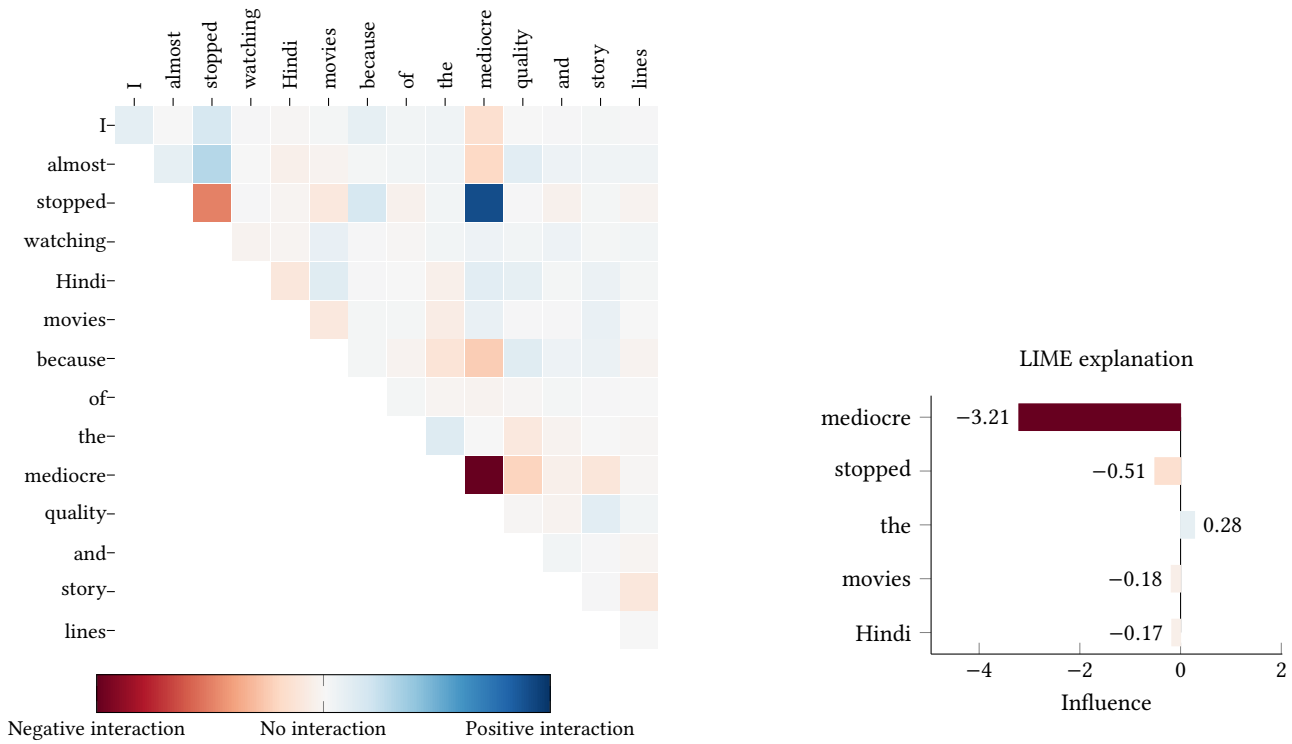

Figure 5: While `BII` and `LIME` agrees on the feature importance level and suggests that "stopped" and "mediocre" has high negative influence. However, `BII` points out that only one of the words "stopped" and "mediocre" has more negative influnce and they both together has positive interactions that leads to less confidence on the negative prediction of the review.

REFERENCES

[1] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. 2016. Deep learning with differential privacy. In *Proceedings of the 23rd ACM SIGSAC Conference on Computer and Communications Security (CCS)*. 308–318.

[2] Ashish Agarwal, Kedar Dhamdhere, and Mukund Sundararajan. 2019. A New Interaction Index inspired by the Taylor Series. *arXiv preprint arXiv:1902.05622* (2019). arXiv:1902.05622

[3] Yoram Bachrach, Evangelos Markakis, Ariel D Procaccia, Jeffrey S Rosenschein, and Amin Saberi. 2008. Approximating power indices. In *Proceedings of the 7th International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*. 943–950.

[4] Georgios Chalkiadakis, Edith Elkind, and Michael Wooldridge. 2011. Computational aspects of cooperative game theory. *Synthesis Lectures on Artificial Intelligence and Machine Learning* 5, 6 (2011), 1–168.

[5] Anupam Datta, Shayak Sen, and Yair Zick. 2016. Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems. In *Proceedings of the 37th IEEE Symposium on Security and Privacy (Oakland)*. 598–617.

[6] Finale Doshi-Velez and Been Kim. 2017. Towards A Rigorous Science of Interpretable Machine Learning. *arXiv preprint arXiv:1702.08608* (2017).

[7] Harmanpree Kaur, Harsha Nori, Samuel Jenkins, Rich Caruana, Hanna Wallach, and Jennifer Wortman Vaughan. 2020. Interpreting Interpretability: Understanding Data Scientists' Use of Interpretability Tools for Machine Learning. In *Proceedings of the 39th CHI Conference on Human Factors in Computing Systems (CHI)*. 1–14.

[8] Zachary C. Lipton. 2016. The Mythos of Model Interpretability. *arXiv preprint arXiv:1606.03490* (2016). arXiv:1606.03490

[9] Scott M. Lundberg, Gabriel G. Erion, and Su-In Lee. 2018. Consistent Individualized Feature Attribution for Tree Ensembles. *arXiv preprint arXiv:1802.03888* (2018). arXiv:1802.03888

[10] Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Proceedings of the 31st Annual Conference on Neural Information Processing Systems (NIPS)*. 4765–4774.

[11] Smitha Milli, Ludwig Schmidt, Anca D. Dragan, and Moritz Hardt. 2019. Model Reconstruction from Model Explanations. In *Proceedings of the 2nd Conference on Fairness, Accountability, and Transparency (FAT*)*. 1–9.

[12] Milad Nasr, Reza Shokri, and Amir Houmansadr. 2019. Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning. In *Proceedings of the 40th IEEE Symposium on Security and Privacy (Oakland)*. 739–753.

[13] Reza Shokri, Martin Strobel, and Yair Zick. 2019. Privacy Risks of Explaining Machine Learning Models. *arXiv preprint arXiv:1907.00164* (2019). arXiv:1907.00164

[14] Danding Wang, Qian Yang, Ashraf Abdul, and Brian Y. Lim. 2019. Designing Theory-Driven User-Centric Explainable AI. In *Proceedings of the 38th CHI Conference on Human Factors in Computing Systems (CHI)*. 1–15.