XAI における忠実性評価手法の考察

牧野 雅紘 [1]　浅妻 佑弥 [1,2]　佐々木 翔大 [2,1]　鈴木 潤 [1,2]
東北大学 [1]　理化学研究所 [2]
{masahiro.makino.r6、asazuma.yuya.r7}@dc.tohoku.ac.jp
shota.sasaki.yv@riken.jp　jun.suzuki@tohoku.ac.jp

概要

本稿では、XAI を評価する際に重要視されている忠実性において評価手法は複数存在しており、どの評価手法を使用して評価を行うのか同意が取れていないという課題について検証する。実験では複数ある忠実性評価手法間の相関を測定した。その結果、忠実性という 1 つの軸で論じられている評価手法であるにも関わらず、相関がない評価手法の組み合わせもあり、それぞれの評価手法が異なった観点で評価を行なっている可能性があることが判明した。本稿では、複数の忠実性評価手法を組み合わせて多角的に説明の忠実性を評価することを提案する。

1　はじめに

近年、深層学習を中心に人工知能技術の発展が著しい。しかしこの発展を支える深層学習の手法の多くは、内部の挙動が人間に理解できない複雑な数値計算により成立している。予測結果の理由が理解できないため，いくら正確な予測をしても利用者の信頼を得られず、高い信頼性を重要視する諸分野（金融・医療）における発展の妨げになりうる。

この問題を解消するために、モデル内部や予測の挙動を説明することのできる説明可能な AI (XAI) に関する研究が盛んに行われている。その中でもモデルの出力毎に事後的な説明を生成する手法は Feature Attribution と呼ばれる [1, 2, 3, 4, 5]。代表的な例として Local Interpretable Model-agnostic Explanations (LIME) [1] がある。LIME はモデルの出力が行われたのちに、入力の近傍で局所的に線形近似することで説明を生成する。

Feature Attribution によって生成された説明がモデルを正確に説明することができているかを表す指標が忠実性である。この忠実性が XAI の性能を測る代表的な指標の一つとなっており、忠実性評価について盛んに議論が行われている [6, 7]。実際に自然言語処理分野における代表的な会議である ACL、NAACL、EMNLP の論文を 2020-2022 年分調査した結果、説明を評価する際に忠実性が最も重要視されていることがわかった。しかしながら、この忠実性評価には、忠実性を評価する指標が複数存在し、どの指標を使用すれば良いのかの同意が取れていないという大きな課題がある。

そこで本稿では、異なる指標の評価結果間の相関を計算することで、評価手法間の関係性を分析する。その結果、忠実性という 1 つの軸で論じられている評価手法にも関わらず、結果に相関のない評価手法の組み合わせが存在することがわかった。ゆえに、実験結果から複数の評価手法を使用して忠実性を多角的に評価する必要があると結論づけた。本稿が説明の忠実性を評価する際、どの指標を用いて評価すればいいのかを選択する一助となり、忠実性の正しい評価へとつながることを期待する。

2　関連研究

モデルの出力毎に事後的な説明を生成する手法を Feature Attribution と呼ぶ。この手法には線形近似を利用した LIME [1] や勾配を利用した Integrated Gradients [2] などが存在する。

忠実性とは，Feature Attribution によって生成された説明がモデルを正確に説明することができているかを表す指標である。複数の既存研究が、忠実性による説明間の性能比較を行っている [8, 9, 10]。

説明評価の際に忠実性が重要視されているため、複数の忠実性評価手法が提案されてきた。しかしどの忠実性評価手法を使用するかは各論文で同意が取れておらず、著者に判断が委ねられている。そのため忠実性評価手法を分析し、評価手法を評価する試みが行われている [6, 7]。Chan ら [6] は、忠実性評価手法毎に評価結果が大きく異なる場合があると言及し、6 つの評価手法間において 2 つの評価軸から優劣を議論した結果 2 つの評価手法の使用を提案

表 1　論文本数調査

論文の種類	本数
XAI 手法の提案	46
XAI の評価	16
XAI のサーベイ	5

表 2　指標調査

指標	本数	割合
忠実性	11 本	50%
妥当性	4 本	18%
その他	3 本	14%
検証なし	4 本	18%

した。

我々は、評価手法毎に結果が大きく異なる理由を明らかにするために、結果間の相関の計算による分析を行った。その結果から，非類似性を持つ評価手法を組み合わせた多角的な忠実性評価を提案する。

3　忠実性評価手法

説明がモデルを正確に説明することができているかを表す忠実性について調査した結果、複数の仮説が生じた。この章では具体的な調査結果と仮説について表記する。

3.1　文献調査の結果

まずは 2020-2022 年の間の ACL、EMNLP、NAACL を対象に XAI に関連する論文を収集した。その結果を表 1 に示す。

次に収集した論文を Future Attribution 手法の XAI について論じられている論文に絞り、どのような観点で評価が行われているのかを調査した。その結果を表 2 に示す。この結果から実際に説明を評価する際には忠実性が重要視されている傾向にあることがわかった。

加えて、異なる忠実性評価手法がいくつ存在しているかの調査を行った。調査の結果、9 つの忠実性評価手法の存在を確認した。9 つの評価手法は、分類タスクにおいて、入力文の一部がモデルの予測ラベルの確率にどのように影響を与えるかを測定する点で共通していた。例えば、XAI が分類において重要度が高いと判断した入力文の一部をマスクした際にモデルの予測ラベルが大きく減少すれば、この XAI による説明はモデルを正しく説明していると言えるため高い評価を得る。

しかし、この 9 つの評価手法は、どの評価手法を使用すればいいのかの同意が取れていないことがわかった。論文ごとに研究者が独断で使用する評価手法を選択している状況である。この状況は、どの評価手法を使用するかで異なった評価結果を示す [6] 中で、XAI の忠実性を正しく評価できていない可能性を示唆している。

3.2　仮説

以上の調査結果から、XAI の説明を評価する際に忠実性は重要視されているものの、評価手法が複数存在し、どの手法を使用するかで大きく結果が異なるという状況にあるということがわかった。

しかしながら、複数の忠実性評価が異なった結果を示すことは先行研究で部分的に示唆されているものの、評価手法間の類似性や非類似性などの関係性の定量的に調査は不十分である。そこで本稿では、評価手法間の類似性や非類似性などの関係性を定量的に分析し、以下の仮説を検証する。

- 評価手法間の関係性を分析することで、手法を分類することができるのではないか。
- 分類により、結果が異なる原因や適切な評価手法の選択が可能となり、説明を正しく評価する一助になるのではないか。

4　実験方法

3.2 節で示した仮説を以下の方法 で検証する。まずは、文書分類タスクにおいてモデルに予測ラベルを出力させる。このモデルに対して XAI によって説明を行う。具体的には予測ラベルの結果に影響を与えた単語に対して重要度を付与する。

次に説明として得られた単語重要度に対して 9 つの忠実性評価手法を用いて忠実性を計算する。最後に各分類モデルごとに、それぞれの忠実性評価手法によって得られた結果間でピアソン相関を計算することで、評価手法間の関係性を分析する。

4.1　データと分類モデル

文書分類タスクのデータセットとして AG NEWS [11] を使用した。AG NEWS は、4 つのクラス (“World”、“Sports”、“Business”、“Sci/Tech”) の記事の見出しと説明フィールドを集めて構築したニュース記事のデータセットで、訓練データ 800 個、開

発データ 100 個、評価データ 100 個で構成される。このデータセットを用いて 3 つの分類モデル (BERT [12]、LSTM [13]、CNN [14]) を学習した。その結果、評価セットにおけるモデルの性能値 (F1) はを BERT が 0.919、LSTM が 0.904、CNN が 0.926 となった。

4.2 XAI

XAI は LIME [1]、Integrated Gradients [2]、Deep Lift [3]、InputXGradient [4]、Vanila Gradient [5] を使用した。LIME はモデルを入力の近傍で線形近似を利用することで説明を生成する。また Integrated Gradients、InputXGradient、Vanila Gradient は勾配を利用して説明を生成する手法である。

4.3 評価手法

忠実性評価手法は、Most Informative Token (Most) [15]、Decision Flip (Flip) [16]、Comprehensiveness (Com) [17]、Sufficiency (Suf) [17]、Correlation between Importance and Output Probability (Cor) [18]、Monotonicity (Mono) [18]、Logodds (Log) [19]、NAUC (NA) [20]、RAUC (RA) [21] の 9 つを使用した。実験では先行研究に習い、 式 (3) 、 (4) で $q \in B = \{1, 5, 10, 20, 50\}$ とし、式 (7) 、 (8) で $t = 20\%$ とした。詳しい数式は付録に記載した。

5 実験結果/考察

得られた忠実性評価結果について相関を計算し、考察したことを以下に示す。高い正の相関を示した評価手法の組み合わせは Flip、Com、Log、NA であった。また、これらの指標の組み合わせはモデルと使用する XAI に依存することなく同じ結果を示す傾向にあることがわかった。この様子を図 1 に示した。なお箱ひげ図である図 1、2 は対象とする XAI を複数種類用意したときの値の分布をあらわしている。

Cor、Mono、Most、Suf、RA は、弱い正の相関を示した評価手法の組み合わせがあるものの、高い正の相関を示す評価手法の組み合わせは存在しなかった。これらの評価手法と Flip との相関を図 2 に示した。 ここでは (Flip、Com、Log、NA) と他の評価手法間の相関が弱いことを示すために、代表して Flip との相関を示している。また Cor、Mono、Most、Suf、RA 間の相関を図 2 に表した。

図 1 と図 2 の結果から、9 つの評価手法の中で

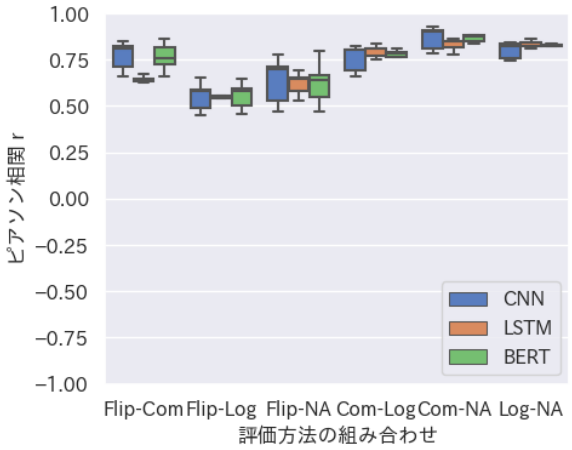

図 1 正の相関を示した評価手法

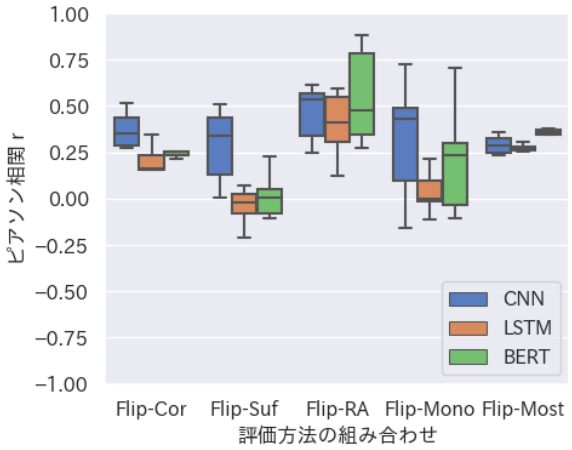

図 2 Flip と評価手法の相関

(Flip、Com、Log、NA) は同じ観点で評価を行なっている可能性が高いことがわかる。一方でその他の評価手法は、それぞれ異なった観点で説明の忠実性を測っている可能性が高いことがわかる。

この実験結果から 9 つの評価手法の分類についての考察を行う。各評価手法がどの程度の重要度を持つ単語に注目して忠実性を評価しているかに注目して分類を行なった。また各評価手法の測定している性質に注目して分類を行った。各評価手法が忠実性を評価する際に注目している単語重要度の範囲を示したのが図 4 であり、各評価手法が注目する単語と測定する性質を軸に分類した表が 3 である。

9 つの評価手法は評価の際に注目する単語によって 4 つに分類した。また評価で測定する性質でさらに分類した。単調増加性とは、入力に重要度の低い単語を順に一つずつ追加していった際に分類ラベルの確率が単調に増加するかを表す。相関性とは単語

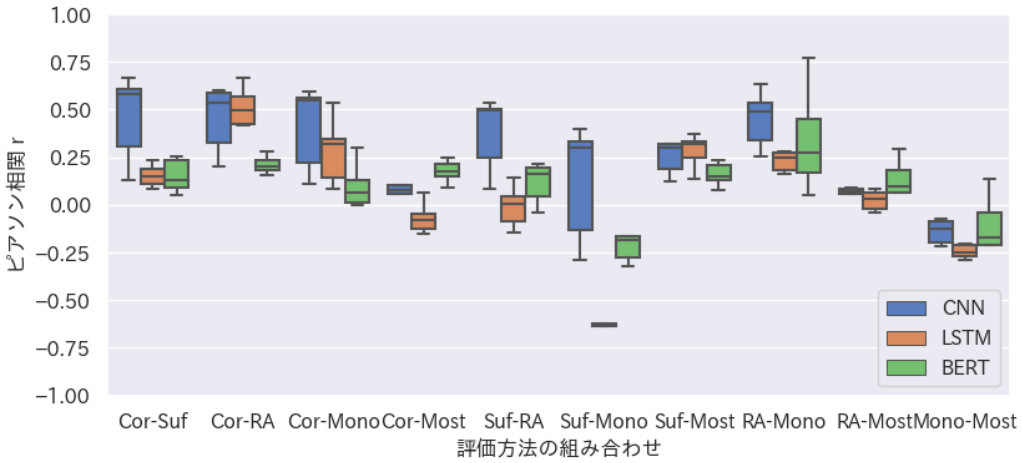

図 3　Cor、Mono、Most、Suf、RA 間の相関

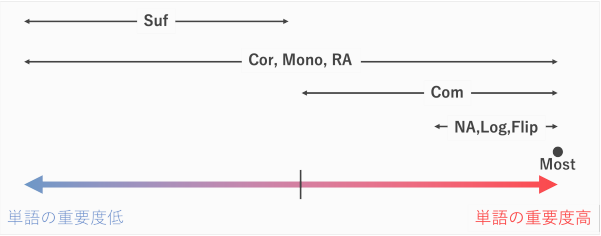

図 4　各評価手法の注目している単語

表 3　忠実性評価手法の分類

	注目する単語	測定する性質	評価指標
忠実性	全ての単語	単調増加性	**Mono**
		相関性	**Cor**
		ランダム優位性	**RA**
	最も重要度の高い単語	確率変動性	**Most**
	重要度の高い単語	確率変動性	**Flip, Com, NA, Log**
	重要度の低い単語	確率変動性	**Suf**

の重要度と分類ラベルの確率変動の間の相関を表す。ランダム優位性とはランダムに単語をマスクした際と重要度の高い単語順にマスクした際の分類ラベルの確率変動の違いを表す。確率変動性とは単語をマスクした際に分類ラベルの確率がどの程度変動するかを示す。つまり忠実性をはかる評価手法の中で、測定している性質が異なるということである。

そこで本稿では、複数の評価手法を用いて、さまざまな観点から忠実性評価を行う必要があると主張する。ある評価手法で、高い評価を得たとしても違う評価手法では低い評価を得る可能性があるからである。特に評価手法間で優劣がはっきりしていないからこそ複数の評価手法を用いて多角的に評価する必要がある。

6　おわりに

本稿では、忠実性が XAI 評価において重要視されていることを示した。しかしその忠実性を測る指標は乱立し、どの評価手法を使用するのか同意が取れていないことがわかった。

そこで本稿の実験で、9 つの忠実性評価手法を対象にそれぞれの相関を計算し、9 つの評価手法間の関係を分析した。その 4 つの評価手法間で高い相関が見られた一方で他の 5 つの評価手法間では相関がなかった。つまり多くの評価方法間で、異なった結果を示すことが定量的に示された。

そこで本稿では、複数の評価手法を用いて、さまざまな観点から忠実性評価を行う必要があると主張する。ある評価手法で、高い評価を得たとしても違う評価手法では低い評価を得る可能性があるからである。特に評価方法間で優劣がはっきりとしていないからこそ複数の評価手法を用いて多角的に評価する必要がある。

本研究では、複数の評価手法のうち、どの手法を用いて多角的に評価すべきかまでは分析することができなかった。そのため、各評価手法が何を評価しているのか、そして各評価手法の使用に優劣をつけることができるのかを分析することが今後の方針として考えられる。この分析により具体的にどの評価手法を用いて説明を評価するべきかを特定していきたい。

謝辞

本研究は、JSPS 科研費 JP21H04901、JST ムーンショット型研究開発事業 JPMJMS2011 (fundamental research) の助成を受けて実施されたものである。

参考文献

[1] Marco Ribeiro, Sameer Singh, and Carlos Guestrin. "why should I trust you?": Explaining the predictions of any classifier. In **Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations**, pp. 97–101, San Diego, California, June 2016. Association for Computational Linguistics.

[2] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic Attribution for Deep Networks. **In International Conference on Machine Learning**, pp. 3319–3328, 2017.

[3] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning Important Features Through Propagating Activation Differences, 2019.

[4] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. **In Proceedings of the 34th International Conference on Machine Learning**.

[5] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. 2014.

[6] Chun Sik Chan, Huanqi Kong, and Liang Guanqing. A Comparative Study of Faithfulness Metrics for Model Interpretability methods. **Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Long Papers)**, Vol. 1, pp. 5029–5038, 2022.

[7] Alon Jacovi and Yoav Goldberg. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 4198–4205, Online, July 2020. Association for Computational Linguistics.

[8] Giuseppe Attanasio, Debora Nozza, Eliana Pastor, and Dirk Hovy. Benchmarking post-hoc interpretability approaches for transformer-based misogyny detection. In **Proceedings of NLP Power! The First Workshop on Efficient Benchmarking in NLP**, pp. 100–112, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[9] Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. A diagnostic study of explainability techniques for text classification. In **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, pp. 3256–3274, Online, November 2020. Association for Computational Linguistics.

[10] George Chrysostomou and Nikolaos Aletras. Improving the Faithfulness of Attention-based Explanations with Task-specific Information for Text Classification. **Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing**, pp. 477–488, 2021.

[11] Ag's corpus of news articles, 2004. http://groups.di.unipi.it/~gulli/AG_corpus_of_news_articles.html.

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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)**, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[13] Felix A Gers, Jurgen Schmidhuber, and Fred Cummins. Learning to forget: Continual prediction with LSTM. **Neural computation**, pp. 2451–2471, 2000.

[14] Krizhevsky, Sutskever, and Hinton. Imagenet classification with deep convolutional neural networks. **Advances in Neural Information Processing Systems 25**, pp. 1097–1105, 2012.

[15] George Chrysostomou and Nikolaos Aletras. Improving the faithfulness of attention-based explanations with task-specific information for text classification. 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)**, pp. 477–488, Online, August 2021. Association for Computational Linguistics.

[16] Sofia Serrano and Noah A. Smith. Is attention interpretable? In **Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics**, pp. 2931–2951, Florence, Italy, July 2019. Association for Computational Linguistics.

[17] Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. ERASER: A benchmark to evaluate rationalized NLP models. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 4443–4458, Online, July 2020. Association for Computational Linguistics.

[18] Vijay Arya, Rachel K. E. Bellamy, Pin-Yu Chen, Amit Dhurandhar, Michael Hind Samuel C. Hoffman, Stephanie Houde, Q. Vera Liao, Ronny Luss, Aleksandra Mojsilović Sami Mourad, Pablo Pedemonte, Ramya Raghavendra, John Richards, Prasanna Sattigeri Karthikeyan Shanmugam, Moninder Singh, Kush R. Varshney, Dennis Wei, and Yunfeng Zhang. One Explanation Does Not Fit All: A Toolkit and Taxonomy of AI Explainability Techniques.

[19] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje, editors. **Learning important features through propagating activation differences**.

[20] Hillary Ngai and Frank Rudzicz. Doctor XAvIer: Explainable diagnosis on physician-patient dialogues and XAI evaluation. In **Proceedings of the 21st Workshop on Biomedical Language Processing**, pp. 337–344, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[21] Andreas Madsen, Vaibhav Adlakha Nicholas Meade, and Siva Reddy. Evaluating the Faithfulness of Importance Measures in NLP by Recursively Masking Allegedly Important Tokens and Retraining. 2021.

付録

忠実性評価手法

入力文を x とする。x の単語の数を l_x とする。x の予測されるクラスは $c(x)$ とし, クラス j に対応する予測確率を $p_{j(x)}$ とする。説明が与えられたとき k 番目の重要な単語を x_k とする。上位 k 個 (または上位 q%) の重要な単語のみを含む入力列を $x_{:k}$ (または $x_{:q\%}$) とする。修正された単語部分列が含まれる修正された入力列を x' をマスクした修正入力系列を x/x' とする。

また Area Under the Curve (AUC) は分類ラベルの確率 を、重要度の降順でマスクした単語の割合に対してプロットすることで 得られる曲線と x 軸間の面積である。y を、x の重要度で降順に並べ替えた単語列とすると図 5 の AUC ($y_{10:20\%}$) は y を上位 0% – 20% の単語をマスクした際に計算される AUC を表す。

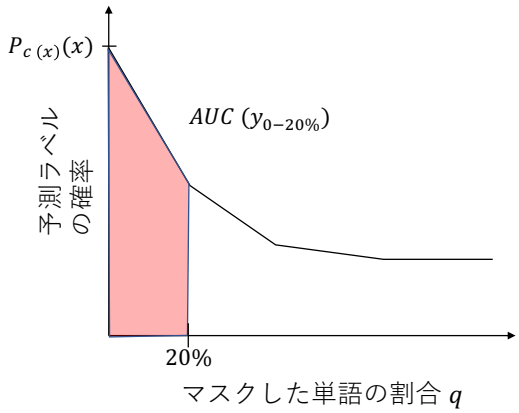

図 5 AUC の略図

Most Informative Token (Most) 入力文の中の最も重要度の高い単語をマスクした際に分類ラベルが変化した時、忠実と判断される指標である。

$$\text{Most} = \begin{cases} 1 & \text{if } c(x) \neq c(x/x_{:1}) \\ 0 & \text{if } c(x) = c(x/x_{:1}) \end{cases} \tag{1}$$

Decision Flip (Flip) 分類ラベルが変化するまで、重要度の高い順に単語をマスクしていき、分類ラベルを変化させるのに必要とした単語のインスランス内全ての単語に対する割合を計算する。

$$\text{Flip} = \begin{cases} -\min \dfrac{k}{l_x} & \text{s.t. } c(x) \neq c(x/x_{:k}) \\ -1 & \text{if } c(x) = c(x/x_{:k}) \text{ for any } k \end{cases} \tag{2}$$

Comprehensiveness (Com) 分類ラベルの確率と重要度の高い単語をマスクしていった際の分類ラベルの確率の差を計算する。

$$\text{Com} = \frac{1}{|B|} \sum_{q \in B} (p_{c(x)}(x) - p_{c(x)}(x/x_{:q\%})) \tag{3}$$

Sufficiency (Suf) 分類ラベルの確率と重要度の低い単語をマスクした後の分類ラベルの確率の差を計算する。

$$\text{Suf} = -\frac{1}{|B|} \sum_{q \in B} (p_{c(x)}(x) - p_{c(x)}(x_{:q\%})) \tag{4}$$

Correlation between Importance and Output Probability (Cor) 重要度の高い単語を一つずつマスクしていき、その際の分類ラベルの確率との相関を計算する。

$$\text{Cor} = -\rho(\boldsymbol{u}、\boldsymbol{p}) \tag{5}$$

$\boldsymbol{u}$ は単語の重要度を降順に並べたベクトルである。$\boldsymbol{p} = [p_{c(x)}(x)、p_{c(x)}(x/x_1), p_{c(x)}(x/x_2), \cdots, p_{c(x)}(x/x_{l_x})$.$\rho$ はピアソン相関係数を表している。

Monotonicity (Mono) 単語の重要度とその単語を追加した後の分類ラベルの確率を計算する。

$$\text{Mono} = \rho(\boldsymbol{u}、\boldsymbol{p}) \tag{6}$$

$\boldsymbol{u}$ は単語の重要度を降順に並べたベクトルである。$\boldsymbol{p} = [p_{c(x)}(x)、p_{c(x)}(x/x_{:1}), p_{c(x)}(x/x_{:2}), \cdots, p_{c(x)}(x/x_{:l_{x-1}})$.$\rho$ はピアソン相関係数を表している。

Logodds (Log) 単語重要度の高い単語をマスクする前と後の予測されたクラスに関する負の対数確率を計算する。

$$\text{Log} = \log \frac{p_{c(x)}(x/x_{:l_x})}{p_{c(x)}(x/x_{:t\%})} \tag{7}$$

NAUC (NA) 縦軸を予測ラベルの確率、横軸をマスクする単語の割合とする。予測ラベルの確率とマスクした単語の割合の乗算 (つまり四角形の面積) に対して AUC が占める割合を計算する。

$$\text{NA} = \frac{\text{AUC}(y_{0\%:t\%})}{q \cdot p_{c(x)}(x)} \tag{8}$$

RAUC (RA) 単語をランダム順でマスクした際の AUC に対して単語を重要度の降順でマスクした際の AUC が占める割合を計算する。z を x をランダムに並び替えた単語列とする。

$$\text{RA} = \sum_{k=0}^{9} \frac{\text{AUC}(y_{k*10\%:(k+1)*10\%})}{\text{AUC}(z_{k*10\%:(k+1)*10\%})} \tag{9}$$