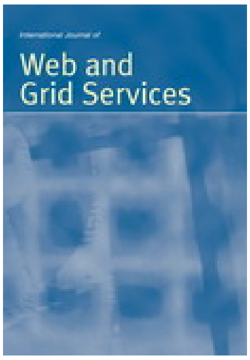


International Journal of Web and Grid Services

ISSN online: 1741-1114 - ISSN print: 1741-1106
https://www.inderscience.com/ijwgs

An augmented interpretive framework based on aspect sentiment words aggregation

Chao Li, Bo Shen, Yingsi Zhao, Qing-An Zeng

DOI: 10.1504/IJWGS.2024.10060552

Article History:

Received:	25 July 2023
Last revised:	26 August 2023
Accepted:	23 September 2023
Published online:	25 March 2024

An augmented interpretive framework based on aspect sentiment words aggregation

Chao Li

School of Electronic and Information Engineering,
Beijing Jiaotong University,
Beijing, 100044, China
Email: chaolixn@foxmail.com

Bo Shen

Key Laboratory of Communication and Information Systems,
Beijing Jiaotong University,
Beijing, 100044, China
Email: bshen@bjtu.edu.cn

Yingsi Zhao*

School of Economics and Management,
Beijing Jiaotong University,
Beijing, 100044, China
Email: yszhao@bjtu.edu.cn
*Corresponding author

Qing-An Zeng

Department of Computer Systems Technology,
North Carolina A&T State University,
North Carolina, 27695, USA
Email: qazeng@yahoo.com

Abstract: Given the mounting anxieties surrounding the interpretability of neural models, appraising interpretability remains an unsolved puzzle owing to the ineffectual performance of existing interpretation techniques and evaluation metrics. The architecture of neural network models varies depending on the task at hand, making it challenging to devise a universal method of explanation that can produce coherent justifications for each model. This paper proposes a framework to enhance the interpretability of text sentiment classification models using aspect sentiment words (ASW) aggregation, which can be applied to web services to improve transparency, accountability, and user trust. The proposed method extracts ASW from sentences and consolidates the token importance scores to provide more credible justifications. The paper also introduces new evaluation metrics for faithfulness, which assess whether interpretations accurately reflect the model's decision-making process. The proposed metrics are effective in evaluating the fidelity of rationales to models at the snippet-level.

Keywords: deep learning; text sentiment classification; interpretability; aspect sentiment words aggregation.

Reference to this paper should be made as follows: Li, C., Shen, B., Zhao, Y. and Zeng, Q-A. (2024) ‘An augmented interpretive framework based on aspect sentiment words aggregation’, *Int. J. Web and Grid Services*, Vol. 20, No. 1, pp.54–73.

Biographical notes: Chao Li is currently a Master’s student in the School of Electronic and Information Engineering at the Beijing Jiaotong University, holding a Bachelor’s in Engineering from the Beijing Jiaotong University in 2021. His research interests include language understanding, sentiment analysis, and information retrieval. He has a strong academic background in NLP and has actively contributed to research projects in the field. He aims to leverage his expertise to develop innovative techniques that enhance the accuracy and performance of NLP systems.

Bo Shen is an Associate Professor in the School of Electronic and Information Engineering at Beijing Jiaotong University. He received his PhD in The Communication and Information Systems, and his research interests include recommendation systems and computer communication.

Yingsi Zhao received his Bachelor’s in Engineering in 2007, and after completed the course work of Communication Engineering from the Beijing Jiaotong University, Beijing, China, in 2009, and she also received her Master’s in Engineering. Respectively, she received her Doctoral in Management in 2014 and currently works as a teacher in School of Economics and Management of Beijing Jiaotong University from 2014. Her research interests are in the area of enterprise management, including but not limited to human resources, complex network, innovation performance, blockchain technology and so on.

Qing-An Zeng received his PhD in Electronic Engineering from the Shizuoka University, Japan, in 1997. He is currently a faculty member with the Department of Computer Systems Technology, North Carolina A&T State University, USA. He has published over 150 books, book chapters, refereed journal papers, and conference proceeding papers. His research interests are in all areas of wireless and mobile networks, ad hoc and sensor networks, handoff, mobility management, heterogeneous networks, system modeling and performance analysis, simulations, QoS, security, NoC, smart grid, smart grid communications, PLC, social networks, deep learning, decision making, and queuing theory.

1 Introduction

Deep neural networks (DNNs) have enabled deep learning to make significant advances in related fields such as natural language processing (NLP) (Yuan et al., 2020), image processing (Wang et al., 2020), and speech recognition (Ho et al., 2020). The success of deep neural networks is largely attributed to their deep structure, which allows for a complex combination of numerous nonlinear network layers to automatically extract features from raw data at various levels of abstraction, thus dramatically improving prediction performance. However, due to their high complexity, multitudinous parameters, and low transparency, these end-to-end models behave like black boxes, making it difficult to understand their decision-making mechanisms or to assess the reliability of their decisions.

The research on interpretability of large artificial intelligence (AI) models can be applied to web services to improve their transparency, accountability, and user trust. By understanding how these models make decisions and which factors are most important in producing their outputs, web service providers can better explain their actions to users and address potential biases or errors in the models. This can lead to increased user satisfaction and loyalty, as well as improved regulatory compliance and legal defensibility. Furthermore, interpretability research can inform the development of more explainable and trustworthy AI models that are optimised for deployment in web service applications.

Table 1 Saliency map of word importance (see online version for colours)

True label	*Predicted label (prob.)*	*Word importance*
		English
1	1 (1.00)	it ’s a charming and often affecting journey.
0	1 (0.93)	this one is definitely one to skip , even for horror movie fanatics.
		Chinese
1	1 (0.91)	交通 方便 ; 环境 很好 ; 服务态度 很好 房间 较小
0	1 (0.94)	风扇 确实 够 响 的 , 尤其 是 到 晚上 周围 安静 下来。风扇 频频 开启 , 发热量 有些 惊人

Note: The greener the colour of the word, the more verdant the hue of a word, the greater its significance in label prediction. Conversely, the more crimson its shade, the less crucial it is, and may even exert a negative effect.

Presently, there exist primarily two approaches to attain interpretable models:

1 interpreting existing models through post-hoc techniques

2 designing inherently interpretable models.

Compared to the latter, we prefer post-hoc interpretations, as they provide a balance between inner interpretability and model accuracy. And Jacovi and Goldberg (2020) had warned that the claim of a method being ‘inherently interpretable’ should be verified before it can be trusted. Generally, if a model has a simple structure and good interpretability, its fitting ability will be limited, resulting in low prediction accuracy, which may restrict the application scenarios of these algorithms.

Currently, there exist several deep neural network model architectures, such as convolutional neural networks (CNN) (LeCun et al., 1998), recurrent neural networks (RNN), long short-term memory (LSTM) (Hochreiter and Schmidhuber, 1997), transformer (Vaswani et al., 2017), and bidirectional encoder representation from transformers (BERT) (Kenton and Toutanova, 2019), among others. These models have yielded impressive results in tasks related to computer vision and NLP. Nonetheless, devising a universal explanation method for these models to provide a logical justification for their decision-making process is challenging. This is not only because of the diversity of these network models but also due to the requirement of defining distinct evaluation metrics for various processing tasks. For instance, the interpretability evaluation criteria for image recognition and text sentiment classification are evidently not universal. In this paper, we focus on the interpretability of text sentiment classification models, as sentiment analysis is one of the most representative tasks in NLP. Besides, it can be applied to various specific scenarios in web services, including but not limited to customer feedback analysis, content moderation, recommender systems and advertisement targeting.

The majority of researchers have employed post-hoc techniques to interpret text sentiment classification models. In the domain of post-hoc techniques, saliency methods are extensively utilised to explicate a model's decisions by apportioning relevance scores to the input tokens, thus representing their effects on predictions. As illustrated in Table 1, the saliency methods were utilised to generate a heatmap depicting the dependence of the model's prediction on each word within text sentiment classification tasks. Although these methods have advanced the performance of interpretation, there is still ample opportunity for further enhancement. Additionally, an increasing number of researchers are utilising faithfulness and plausibility as measures to evaluate the rationales extracted by saliency methods, albeit with differing methodologies for computing these metrics (Zhang et al., 2021; Madsen et al., 2022; Jacovi and Goldberg, 2020; Yin et al., 2022). Mathew et al. (2021) have employed the token F1-score to assess the credibility of token-level rationales. Differently, DeYoung et al. (2020) have proposed the intersection-over-union (IOU) F1-score and the area under the precision-recall curve (AUPRC) as measures to evaluate the plausibility of snippet-level rationales.

Compared to plausibility, faithfulness is more challenging to define and compute owing to the limitations of human cognition regarding deep learning models. DeYoung et al. (2020) offered explicit computations for faithfulness from the standpoints of sufficiency and comprehensiveness. This definition solely considers the impact of the sequential arrangement of multiple words within a single sentence on the prediction; however, it may not precisely mirror the underlying semantics, potentially yielding inaccuracies. Wang et al. (2022) utilised mean average precision (MAP) to compute the coherence of rationales under perturbation, in order to assess faithfulness. Nonetheless, there exists a degree of negative correlation between MAP and the plausibility metric, and MAP breaches the faithfulness criterion (Jacovi and Goldberg, 2020). This implies that it may not accurately evaluate interpretability.

In order to address the aforementioned issues, we put forth an interpretive framework based on ASW aggregation, which extends current saliency methods by consolidating individual tokens into ASW. This approach can enhance the effectiveness of saliency methods and generate more plausible and accurate explanations. Additionally, a new evaluation metric for interpretability is proposed to measure the faithfulness of generated explanations.

Specifically, our contributions in this paper are as follows:

- An augmented interpretive framework of ASW aggregation has been proposed, thereby elevating both the rationality and fidelity of interpretations derived from saliency methods within text sentiment classification models.
- To more accurately assess the interpretability of sentiment analysis models, we suggest a more sound evaluation metric for fidelity, namely the RBO between two rankings: one involving the sorting of model output based on rationales, and the other pertaining to the sorting of token importance scores.
- We conducted experiments and provided a comprehensive analysis of our framework, employing three standard models in conjunction with three commonly used saliency methods. Furthermore, we present a comparative analysis between the metrics we propose and those that currently exist.

2 Related work

In this section, our primary focus will be on saliency methods, evaluation metrics for interpretability, and ASW extraction models. The crux of our work involves presenting an augmented interpretive framework centred on ASW aggregation, which heightens the interpretability of text sentiment analysis models. Additionally, we also introduce novel metrics for measuring fidelity.

2.1 Saliency methods

In the domain of post-hoc explanation technique, saliency methods are commonly utilised to comprehend the decision-making process of models. These methods distribute importance scores across input tokens as a means of illuminating their impact on model predictions (Simonyan et al., 2014; Murdoch et al., 2018; Ribeiro et al., 2016). Interpretability can be classified into two categories based on the target of interpretation: those grounded in input features (Zeiler and Fergus, 2014; Lundberg and Lee, 2017; Ahern et al., 2020; Smilkov et al., 2017; Ribeiro et al., 2016; Sundararajan et al., 2017) and those anchored in intermediate process features (Selvaraju et al., 2017; Wang et al., 2019; Abnar and Zuidema, 2020; Yuan et al., 2021). Furthermore, interpretability can also be divided into four categories based on feature attribute methods: fit-based, attention-based, removal-based, and gradient-based. Fit-based methods utilise a straightforward and comprehensible model to partially align the consequence of the evaluated model (Alvarez-Melis and Jaakkola, 2017; Ribeiro et al., 2016; Ahern et al., 2020). Attention-based methods are particularly suitable for models that employ attention mechanisms, as they utilise attention weights as a means of providing explanations (Jain and Wallace, 2019; Wiegreffe and Pinter, 2019). Removal-based methods provide explanations for the behaviour of the model by comparing the changes in model output before and after the removal of tokens (Li et al., 2016; Feng et al., 2018). Gradient-based methods reveal the decision-making behaviour of the model by calculating the gradient of the model during the training process (Simonyan et al., 2014; Lundberg and Lee, 2017; Sundararajan et al., 2017; Smilkov et al., 2017; Selvaraju et al., 2017). Each of these methods has its own unique advantages and disadvantages in terms of factors such as computational efficiency, interpretability performance, and so forth (De Cao et al., 2020; Sixt et al., 2020).

2.2 Interpretability metrics

When it comes to highlight-based rationales, interpretability is frequently evaluated in terms of plausibility and faithfulness from the standpoint of both humanity comprehension and the actual decision-making process of models. DeYoung et al. (2020) suggested employing IOU F1 score and AUPRC to gauge the plausibility of snippet-level rationales. DeYoung et al. (2020) also offered specific equations for evaluating faithfulness in terms of both the sufficiency and comprehensiveness of rationales. Nevertheless, this evaluation metric carries with it uncontrollable factors that could impact interpretability evaluation. Jacovi and Goldberg (2020) proposed several criteria for the definition and assessment of fidelity. Ding and Koehn (2021) assessed the trustiness of saliency methods on NLP models by measuring the consistency of rationales under perturbations. Wang et al. (2022) employed the correspondence between the rationales offered on instances prior to and following perturbation, which are skilfully designed to preserve the model's decision-making process, as a metric for evaluating faithfulness.

2.3 ASW extraction model

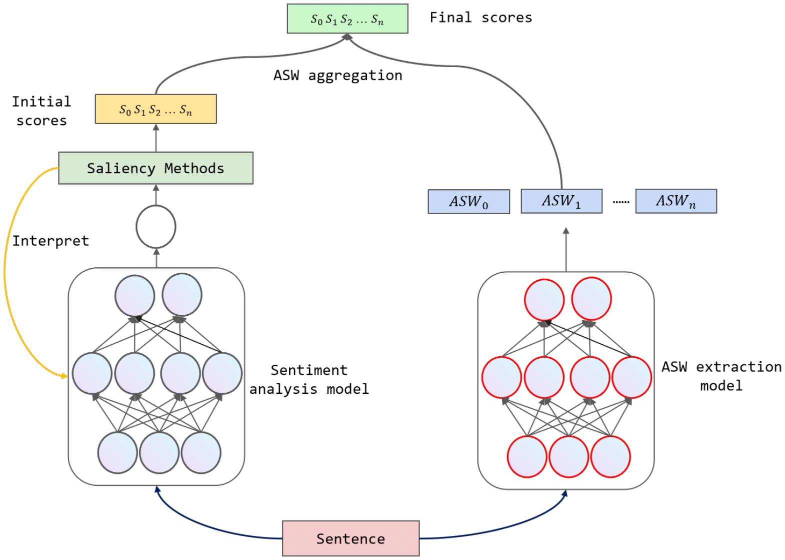
The task of aspect term extraction (ATE) (Yin et al., 2016; Li et al., 2018; Ma et al., 2019) centres on recognising aspect targets, whereas opinion term extraction (OTE) (Yang and Cardie, 2012; Klinger and Cimiano, 2013; Yang and Cardie, 2013) aims to extract opinion words or phrases that chiefly influence the sentiment polarity of the sentence or the corresponding target term. The most recently proposed subtask of aspect-based sentiment analysis (ABSA) (Zhang and Liu, 2012; Pontiki et al., 2014) is aspect sentiment triplet extraction (ASTE) (Peng et al., 2020), which forms a more complete picture of sentiment information through a triplet consisting of an aspect target term, its corresponding opinion term, and the expressed sentiment. Xu et al. (2021) proposed a span-based approach for learning the interaction between target words and opinion words, and introduced a dual-channel span pruning strategy to reduce the computational cost brought by span enumeration, which we use for English ASW extraction.

Furthermore, semantic dependency analysis can be leveraged for the extraction of ASW, as there frequently exists a dependency relation between opinion words and aspect words. Zhuang et al. (2006), Kobayashi et al. (2006), Somasundaran and Wiebe (2010) and Kessler and Nicolov (2009) utilise the parsing of sentence dependency relationships to identify the targets modified by sentiment words. Qiu et al. (2011) further generalise this approach using a double-propagation method to simultaneously extract both targets and opinion words. Recognising that evaluative targets may comprise of noun or verb phrases, rather than just single words, Wu et al. (2009) employ the semantic dependencies between phrases in a sentence to identify candidate targets, which are then filtered using a language model.

3 Methods

When individuals assess the emotional orientation of a sentence, they frequently base their judgment on the presence of ASW within the sentence. Therefore, this paper proposes an interpretation framework based on ASW aggregation, which takes into account human cognitive tendencies when assessing the sentiment of a sentence. Figure 1 provides an overview of our augmented interpretive framework, which initially extracts ASW using either syntactic dependency analysis or deep learning models, and then consolidates the token scores obtained by saliency methods based on the ASW. Subsequently, we will provide a detailed account of the implementation of the framework and the proposed metrics that are more suitable for evaluating loyalty.

Figure 1 Overview of our augmented interpretive framework (see online version for colours)



3.1 *An augmented interpretive framework for enhancing saliency methods based on ASW aggregation*

For the extraction of ASW, this paper utilises two methods:

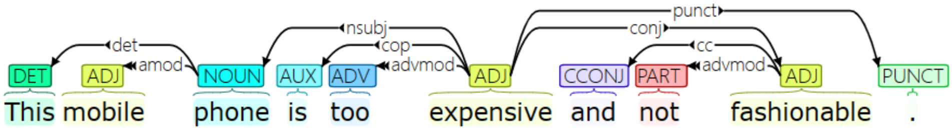
1 syntactic dependency analysis for both English and Chinese

2 ASTE (Xu et al., 2021) for English and sentiment knowledge enhanced pre-training for sentiment analysis (SKEP) (Tian et al., 2020) for Chinese.

Syntactic analysis is of paramount importance in NLP tasks such as opinion extraction. It has the capability to unearth the remote lexical dependencies in sentences that are arduous to acquire through lexical analysis, and aids in extracting the semantic information that is obscured in a more profound hierarchical stratum. For instance, in the sentence “this mobile phone is too expensive and not fashionable”, the intricate annotation details of the dependency relationship within the sentence are illustrated in Figure 2.

In the dependency analysis of a sentence, every dependency is portrayed as directed edges, where the origin of the arrow denotes the modified dominant word, and the destination indicates the subordinate word that functions as the modifier. Ultimately, the semantic dependencies of the sentence are derived through the edge annotations, revealing the interrelatedness between the words. Based on the outcomes of dependency parsing, a definite relationship between aspect words and opinion words can be discerned. Hence, it can be employed to extract aspect words and their corresponding sentiment words.

Figure 2 Dependency parsing tree (see online version for colours)



As for the alternative technique for extracting ASW, we utilise SKEP (Tian et al., 2020) to formulate an ASW extraction model for Chinese, and Span-ASTE (Xu et al., 2021) for English. With respect to SKEP, we can extract the aspects in sentences and their corresponding opinions through sequence labelling. To achieve this objective, we have expanded the labelling system founded on the begin inside outside (BIO) sequence labelling system, encompassing B-aspect, I-aspect, B-opinion, I-opinion, and O. The initial two labels are employed to indicate aspects, the subsequent two labels are employed to indicate the related opinions, while the last label represents neither aspect nor opinion. Regarding Span-ASTE, further details on extracting ASW from English sentences can be found in Lu's publication (Xu et al., 2021).

In summary, the framework for enhancing saliency methods through ASW aggregation involves the following steps:

Input: an input sentence X, the text sentiment classification model $f(x)$, and the specific feature attribution method Ω.

Output: the importance score w of each token in the sentence X.

1 Given a saliency method Ω, we first compute the original token importance score $w_0 = \Omega(X, f)$ by Ω

2 Use Arc-Standard (Nivre, 2003) to perform dependency analysis on the input sentence and extract ASW sets $S = \{[(a_0, ..., a_i), (o_0, ..., o_j)], ...\}$ according to the defined rules or use deep learning models like Span-ASTE and SKEP to extract them, where a_i is the token that makes up aspect words and o_j is the composition of sentiment section.

3 Traverse the set of ASW in S and accumulate their scores as follows: for each set $S_k = [(a_0, ..., a_i), (o_0, ..., o_j)]_k$, calculate the score s_k, and assign the score $t_{ik} = s_k + h * l$ for each token i belonging to S_k. Here, h is a small constant which we set to 0.001 and l is the reverse position of the token in the ASW.

4 Store these scores in each token's list $t_i = (t_{i0}, ..., t_{ik})$, where t_{i0} represents the original token score and t_{ik} represents the cumulative value calculated above if the token belongs to S_k.

5 Set the ultimate feature score of token i in the sentence as $w_i = \max(t_i)$, where $\max(t_i)$ represents the maximum value in t_i.

3.2 Interpretability metrics

In line with prior research (DeYoung et al., 2020; Ding and Koehn, 2021; Mathew et al., 2021), we assess interpretability based on the criteria of plausibility and faithfulness. In regards to models for analysing sentiment at the sentence-level, plausibility is determined by the degree to which the justifications produced by saliency techniques correspond with those annotated by human. Fidelity refers to whether the token importance scores generated by saliency methods accurately reflect the decision-making process of the model, meaning whether the model's prediction of the dependency on each part of the input is consistent with the token importance scores obtained.

Unlike previous studies, we have chosen to use Token-F1 (Wang et al., 2022) to evaluate plausibility. However, for faithfulness, we have proposed new metrics, *RBO sufficiency (RBO-Suf)* and *RBO comprehensiveness (RBO-Com)*. These metrics eliminate the negative correlation with Token-F1, unlike MAP, and adhere to the rules proposed by Jacovi and Goldberg (2020). Compared to DeYoung et al. (2020), calculating the confidence score based on all rationales selected from the ranking of token importance scores is more reasonable. This approach provides an overall assessment of the results obtained by saliency methods, rather than relying on a single explanation.

RBO (Webber et al., 2010), as defined in equation (1), is a metric used to evaluate the similarity between two lists. In equation (1), S and T represent two arbitrary lists. p is a parameter that can be specified. $S_{c:d}$ represents the set of all elements from position c to position d in the list. The value of this metric ranges between 0 and 1, with a value closer to 1 indicating a higher degree of similarity between the two lists.

$$\text{RBO}(S, T, p) = (1-p)\sum_{d=1}^{\infty} p^{d-1} \cdot A_d \tag{1}$$

where

$$A_d = \frac{|I_d|}{d} = \frac{S_{1:d} \cap T_{1:d}}{d}$$

RBO-Com: inspired by DeYoung et al. (2020), we construct a contrast example for x_i, denoted as $\tilde{x_i}$, which is obtained by removing the predicted rationals r_i from x_i. r_i represents the rationale of x_i, and $x_i \backslash r_i$ represents its non-rationale. First, we calculate the output $F(x_i \backslash r_i)$ provided by a specified deep learning model F according to the $x_i \backslash r_i$. Then we calculate whether $F(x_i \backslash r_i)$ rankings is consistent with the rationals r_i rankings. We can measure this through calculating RBO between the two rankings as equation (2), where x_i is the i^{th} sentence, r_{ij} is the top-j elements of r_i and k_i represents the length of r_i.

$$\text{RBO-Com} = \frac{1}{N}\sum_{i=1}^{N}\{\text{RBO}[argsort(L_0), L]_i\} \tag{2}$$

where

$$L_0 = [1 - F(x_i \backslash r_{i1}), 1 - F(x_i \backslash r_{i2}), ..., 1 - F(x_i \backslash r_{ik_i})]$$

$$L = [1, 2, ..., k_i]$$

RBO-Suf: this captures the degree to which the snippets within the extracted rationales are adequate for a model to make a prediction. We let $F(r_i)$ be the prediction provided by a model F according to rationales from the token importance scores ranking. Then we calculate whether $F(r_i)$ rankings is consistent with the rationals r_i rankings. We also measure this through calculating RBO between the two rankings as equation (3).

$$\text{RBO-Suf} = \frac{1}{N}\sum_{i=1}^{N}\{\text{RBO}[argsort(L_0), L]_i\} \tag{3}$$

where

$$L_0 = [F(r_i1), F(r_i2), ..., F(r_ik_i)]$$

$$L = [1, 2, ..., k_i]$$

4 Experiments

4.1 Datasets

Wang et al. (2022) proposed interpretable evaluation benchmark datasets where our experiment is conducted. The dataset comprises 1,500 randomly selected instances from the dev/test sets of the Stanford Sentiment Treebank (SST) (Socher et al., 2013), along with 400 instances from the test set of the movie reviews (Zaidan and Eisner, 2008) dataset for English. Additionally, the dataset includes 60,000 randomly sampled instances from the logs of an open sentiment analysis (SA) application programming interface (API) for Chinese, with the permission of its users.

Table 2 shows the size of the original and perturbed pairs, as well as the average ratio of rationale length to input length (RRL) and the number of rationale sets in an input (NRS) across all data.

Table 2 Overview of datasets

English			*Chinese*		
Size	*RRL*	*NRS*	*Size*	*RRL*	*NRS*
1,999	20.1%	2.1	2,160	27.6%	1.4

Source: Wang et al. (2022)

4.2 Models

Similar to Wang et al. (2022), we use a robustly optimised BERT pretraining approach (RoBERTa-base), RoBERTa-large (Zhuang et al., 2021) and LSTM (Hochreiter and Schmidhuber, 1997) to evaluate interpretability on these two datasets. To fine-tune our English and Chinese models, we used the training sets of SST and ChnSentiCorp, respectively. However, due to the low accuracy of the Chinese model on the evaluation dataset, we artificially marked 1/4 of the test set and retrained the model to prevent impact on the evaluation of model interpretation. Table 3 presents the model's performance on the original dataset and the evaluation dataset.

Table 3 Accuracy of models

Models	*LSTM*	*RoBERTa-base*	*RoBERTa-large*
		English	
Acc^t	88.6	94.3	*95.6*
Acc^o	80.4	92.3	*92.8*
		Chinese	
Acc^t	90.4	93.2	*94.7*
Acc^o	61.2	60.4	*66.8*
Acc^r	84.8	92.9	*93.8*

Note: Acc^t represents the accuracy of the models on the training sets, Acc^o represents the accuracy on the original test sets, and Acc^r represents the accuracy of the retrained models after labelling 1/4 of the test dataset.

4.3 Saliency methods

- Attention (ATT) (Jain and Wallace, 2019; Wiegreffe and Pinter, 2019): attention-based methods utilising attention weights to measure the importance of each token in the input sequence.
- Integrated gradient (IG) (Sundararajan et al., 2017): integrating the gradients from a baseline input (zero embedding) to the original input taken along a direct route and utilising the gradients as indicators of token significance.
- Local interpretable model-agnostic explanations (LIME) (Ribeiro et al., 2016): adding slight perturbations to the input samples and observing the changes in the output of the model to be explained, training a linear model on the original input based on these changes, and using the linear model to locally approximate the predictions of the model to be explained, representing the weights of the linear model as importance scores of the tokens in the original input.

4.4 Metrics

For evaluation metrics, we employ Token-F1 (Wang et al., 2022) to evaluate the rationality of interpretations. We also utilise MAP (Wang et al., 2022), Score-Suf (DeYoung et al., 2020), and Score-Com (DeYoung et al., 2020) compared with RBO-Suf and RBO-Com to collectively assess the faithfulness of the explanation.

Token-F1, as defined in equation (4), is computed by assessing the degree of overlap between the predicted and ground truth rationale tokens.

$$\text{Token-F1} = \frac{1}{N}\sum_{i=1}^{N}\left(2 \times \frac{P_i \times R_i}{P_i + R_i}\right) \tag{4}$$

where

$$P_i = \frac{|S_i^p \cap S_i^g|}{|S_i^p|} \quad \text{and} \quad R_i = \frac{|S_i^p \cap S_i^g|}{|S_i^g|}$$

where S_i^p and S_i^g represent the rationale set of i^{th} instance provided by models and human respectively; N is the total number of instances.

Mean average precision (MAP), as defined in equation (5), is utilised to gauge the coherence of rationales under perturbations, thereby serving as a metric to evaluate the faithfulness.

$$\text{MAP} = \frac{\sum_{i=1}^{|X^p|} \sum_{j=1}^{i} \frac{1}{i} G(x_j^p, X_{1:i}^o)}{X^p} \tag{5}$$

where X^o and X^p represent the sorted rationale token list of the original and perturbed inputs, according to the token important scores assigned by a certain saliency method. $|X^p|$ represents the number of tokens in X^p. $X_{1:i}^o$ consists of top-i important tokens of X^o. The function $G(x, Y)$ is to determine whether the token x belongs to the list Y, where $G(x, Y) = 1 \; iff \; x \in Y$.

Equation (6) demonstrates the *Score-Suf* and *Score-Com*. A lower Score-Suf indicates that the rationale is more than sufficient, while a higher Score-Com signifies that the rationale has a greater impact on the output. For a rationale to be considered faithful, it should possess a low Score-Suf and a high Score-Com.

$$\begin{aligned} \text{Score-Suf} &= \frac{1}{N} \sum_{i=1}^{N} \left(F(x_i)_j - F(r_i)_j\right) \\ \text{Score-Com} &= \frac{1}{N} \sum_{i=1}^{N} \left(F(x_i)_j - F(x_i \backslash r_i)_j\right) \end{aligned} \tag{6}$$

where $F(x_i)_j$ represents the prediction probability provided by the model F for class j on the inputx_i; r_i represents the rationale of x_i, and $x_i \backslash r_i$ represents its non-rationale; N is the total number of datasets.

5 Results and discussions

Tables 4 and 5 presents a comparison of the experimental results between the augmented interpretive framework based on ASW aggregation and the original interpretation method. The metric denoted by the upward arrow signifies that a higher score corresponds to superior performance. Conversely, a lower score indicated by the downward arrow is indicative of commendable performance. The acronym 'ASW-DA (dependency analysis)' denotes the ASW aggregation framework, which is founded upon dependency analysis. On the other hand, 'ASW-DL (deep learning)' relies on deep learning models. Moreover, 'Pla.' is a succinct representation of 'plausibility', whereas 'Fai.' abbreviates 'faithfulness'. 'Suf' serves as an abbreviation for 'Score-Suf', 'Com" stands for 'Score-Com', 'R-Suf' represents 'RBO-Suf', and 'R-Com' signifies 'RBO-Com'.

Table 4 Interpretability evaluation results on English datasets

Model + method	*Original*						*ASW-DA*						*ASW-DL*					
	Pla.	*Fai.*					*Pla.*	*Fai.*					*Pla.*	*Fai.*				
	Token-F1↑	*MAP*↑	*Suf*↓	*Com*↑	*R-Suf*↑	*R-Com*↑	*Token-F1*↑	*MAP*↑	*Suf*↓	*Com*↑	*R-Suf*↑	*R-Com*↑	*Token-F1*↑	*MAP*↑	*Suf*↓	*Com*↑	*R-Suf*↑	*R-Com*↑
LSTM + IG	31.3	*71.3*	30.9	45.1	88.2	73.9	35.4	70.1	29.8	44.7	88.1	74.1	35.6	70.0	29.4	44.6	88.1	74.1
RoBERTa-base + IG	35.2	59.9	11.7	63.0	89.3	80.2	39.9	59.5	11.9	61.3	89.2	80.1	40.4	60.2	13.4	60.2	89.1	80.1
RoBERTa-large + IG	34.6	40.2	12.4	47.8	89.1	79.8	38.6	41.6	11.9	48.0	89.1	79.8	39.6	43.1	12.2	47.0	89.0	79.9
LSTM + ATT	36.7	*71.3*	28.2	37.9	88.0	74.9	39.8	*70.5*	27.2	37.1	88.1	75.0	41.0	*71.7*	27.6	36.4	87.9	75.0
RoBERTa-base + ATT	25.7	69.0	22.9	45.8	85.9	74.4	33.2	66.7	11.9	61.3	89.2	80.1	35.4	68.2	26.0	37.0	86.5	75.8
RoBERTa-large + ATT	22.8	69.3	17.0	41.1	85.5	74.8	33.1	66.8	21.8	32.0	86.8	76.4	35.3	67.9	21.9	30.8	86.6	76.4
LSTM + LIME	30.2	68.0	33.0	45.8	87.8	73.8	30.9	59.5	31.5	45.6	87.8	74.0	31.0	60.1	31.7	45.2	87.8	73.9
RoBERTa-base + LIME	*41.0*	61.6	*9.9*	*87.1*	*90.5*	80.7	*45.5*	62.3	*9.9*	*84.0*	*90.3*	80.6	*45.9*	62.4	11.5	*79.3*	*90.2*	80.4
RoBERTa-large + LIME	39.2	60.1	10.3	82.3	90.3	*80.8*	43.2	60.5	10.2	78.0	90.1	*80.7*	44.1	61.5	*10.9*	74.7	90.0	*80.8*

Table 5 Interpretability evaluation results on Chinese datasets

Model + method	*Original*						*ASW-DA*						*ASW-DL*					
	Pla.	*Fai.*					*Pla.*	*Fai.*					*Pla.*	*Fai.*				
	Token-F1↑	*MAP*↑	*Suf*↓	*Com*↑	*R-Suf*↑	*R-Com*↑	*Token-F1*↑	*MAP*↑	*Suf*↓	*Com*↑	*R-Suf*↑	*R-Com*↑	*Token-F1*↑	*MAP*↑	*Suf*↓	*Com*↑	*R-Suf*↑	*R-Com*↑
LSTM + IG	43.6	*57.9*	69.0	73.7	88.4	73.9	52.0	*57.6*	48.0	81.6	86.7	72.0	53.4	*57.8*	50.0	80.5	86.7	72.1
RoBERTa-base + IG	43.0	47.6	42.6	55.0	88.9	*80.0*	51.7	52.8	23.6	54.7	87.4	79.1	52.8	52.6	25.4	54.6	87.5	79.5
RoBERTa-large + IG	33.2	53.1	28.3	30.6	87.5	79.2	41.1	51.7	*19.3*	41.9	86.9	77.7	41.6	53.2	*21.6*	33.5	87.1	77.8
LSTM + ATT	*47.7*	49.8	49.2	87.1	87.7	70.1	55.3	49.5	56.7	*82.1*	87.6	70.9	*55.9*	48.4	58.2	*82.2*	87.7	70.8
RoBERTa-base + ATT	35.9	53.3	27.3	63.6	87.9	77.7	50.2	50.6	23.5	58.8	87.6	78.6	50.3	49.9	26.1	59.9	87.8	78.5
RoBERTa-large + ATT	40.6	49.8	*20.6*	24.1	87.5	*80.0*	54.4	52.0	19.5	30.5	86.6	78.6	54.5	50.7	22.2	25.2	87.4	*79.9*
LSTM + LIME	41.9	54.4	58.7	*92.8*	*90.4*	77.9	50.0	51.9	71.2	74.1	*88.5*	75.6	50.6	53.3	68.8	76.5	*88.7*	76.0
RoBERTa-base + LIME	45.9	48.7	54.6	69.6	*90.4*	*80.0*	*55.4*	51.6	29.5	55.7	88.0	*79.6*	55.4	52.5	30.1	55.3	88.1	79.5
RoBERTa-large + LIME	45.2	45.8	25.7	32.9	88.2	78.5	49.7	41.0	28.2	28.8	87.7	78.0	50.7	42.5	24.2	32.0	87.3	78.3

The results unequivocally demonstrate that the explanations generated by our framework based on ASW aggregation have significantly improved the rationality compared to the original saliency method. The loyalty metrics MAP, RBO comprehensiveness and RBO sufficiency have also seen a slight improvement, albeit not as much as the rationality indicator Token-F1. These results demonstrate that our method is capable of generating more plausible explanations while maintaining the fidelity of the original method to the model, with a slight improvement on this basis. Subsequently, the results obtained from the experiment will be scrutinised in detail from the three perspectives of evaluation metrics, interpretation methods, and models. We also deliberate upon the merits and demerits of this approach in the end.

5.1 Comparison between evaluation metrics

It can be observed from the experimental results that there is a certain degree of negative correlation between the Token-F1 and MAP, i.e. the higher the Token-F1, the lower the corresponding MAP. The intrinsic reason behind this phenomenon can be attributed to the fact that the calculation of MAP involves perturbing the evidence words through synonymous or antonymous substitution, which results in a change in the evidence words. As a consequence, the more reasonable the evidence extracted, the lower the consistency between the evidence words before and after the perturbation, leading to a decrease of MAP. However, this problem does not exist for RBO comprehensiveness and RBO sufficiency. These metrics we proposed measure the model's trust in the token importance score rankings obtained by the interpretation method, so there is no conflict with Token-F1.

The resemblance in trends between sufficiency and our proposed metrics is evident across all three models and saliency methods. Our proposed metrics surpass those of DeYoung et al. (2020) in precision by comprehensively considering all rationales, rather than a singular explanation. Calculating the confidence score based on the ranking of token importance scores is a more cogent approach. As such, this method affords a comprehensive evaluation of results obtained through saliency methods, in lieu of dependence on a solitary explanation.

5.2 Evaluation of models

Our framework generates superior explanations with maintaining the fidelity for all three models, across all explanation methods. Meanwhile, our experimental outcomes reveal that our proposed framework more effectively enhances the performance of transformer models than that of LSTM, across all saliency methods. This outcome may be attributed to the superior impact of ASW aggregation on transformer models.

When evaluating model interpretability, our focus is specifically directed towards the IG and ATT, given that the LIME remains agnostic to the model. In comparison to LSTM, based on the IG method, transformer models exhibit superior performance on plausibility and faithfulness for English. However, for Chinese, LSTM is competitive in terms of faithfulness and performs better on plausibility. Conversely, when using ATT, LSTM surpasses transformer models on plausibility for both English and Chinese. In our comparison of RoBERTa-base and RoBERTa-large, we discover that the former outperforms the latter on plausibility for both English and Chinese, utilising these two saliency methods. However, RoBERTa-large outperforms RoBERTa-base on plausibility with ATT for Chinese. Interestingly, in terms of faithfulness evaluation, RoBERTa-base outperforms RoBERTa-large with both the IG and ATT methods, for both English and Chinese.

5.3 Evaluation of saliency methods

The findings indicate that our framework yields the most significant improvement in LIME, followed by ATT, with IG exhibiting the lowest improvement, among the three saliency methods. Based on our assessment, it is evident that LIME outperforms other methods in terms of Token-F1, sufficiency, R-Suf, and R-com metrics. This is attributed to the fact that LIME's rationales more precisely emulate the decision-making mechanism of deep learning models. In comparing IG and ATT, it is noticeable that ATT demonstrates better performance in terms of plausibility, whereas IG exhibits superior performance with respect to faithfulness.

5.4 Advantages and limitations

Our proposed framework significantly enhances the plausibility of explanations generated by attribution methods across all three models, while preserving their original fidelity. This improvement can be attributed to our framework's foundation of utilising ASW aggregation to construct explanations, which we augment to improve their effectiveness. Furthermore, we observe that in contrast to the framework that relies on dependency analysis, the one that is founded upon deep learning models exhibits a competitive level of faithfulness and yields superior results in plausibility. This can be attributed, in significant part, to the heightened precision of deep learning models in extracting ASW.

Last but not least, our framework still has some limitations. The most significant issue is how to extract the final interpretations based on token scores, given that the current ratio-based method is not particularly effective. Additionally, there is scope for improving the extraction of ASW, with better methods that could enhance the rationality of subsequent explanations.

6 Conclusions

This paper proposed an augmented interpretive framework based on ASW aggregation to bolster the coherence and precision of interpretations yielded by saliency methods used in text sentiment classification models. In addition, we also proposed more reasonable loyalty metrics, namely RBO comprehensiveness and RBO sufficiency. These metrics evaluated the faithfulness of interpretations extracted by saliency methods through calculating RBO between two rankings. They eliminated the negative correlation with Token-F1, unlike MAP, and adhere to the rules proposed by Jacovi and Goldberg (2020). Compared to DeYoung et al. (2020), they are more reasonable to calculate the confidence score based on all rationales selected from the ranking of token importance scores, rather than a singular explanation. The experimental results demonstrated that our proposed framework significantly enhances the rationality of explanations extracted by saliency methods on three typical models, while preserving the faithfulness to the models.

Acknowledgements

This work was supported by the National Key R&D Program of China (Grant Number 2021YFC3300204).

Declarations

Funding: National Key R&D Program of China (2021YFC3300204).

Data availability: All datasets are open-source, and the sources are cited.

Authors' contributions: The presented ideas were conceived by all authors. The methodology, code implementation, experiment, and data analysis were performed by Chao Li. Yingsi Zhao provided guidance on the implementation methods and steps of the experiment, while Qing-An Zeng and Bo Shen offered guidance on the final analysis of the experimental results. Chao Li wrote the original draft, while Bo Shen and Qing-An Zeng provided critical review, commentary, and manuscript revision. All authors reviewed the manuscript before submission.

References

Abnar, S. and Zuidema, W. (2020) 'Quantifying attention flow in transformers', *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, pp.4190–4197.

Ahern, I., Noack, A., Guzman-Nateras, L., Dou, D., Li, B. and Huan, J. (2020) *Normlime: A New Feature Importance Metric for Explaining Deep Neural Networks*, CoRR http://arxiv.org/abs/1909.04200.

Alvarez-Melis, D. and Jaakkola, T. (2017) 'A causal framework for explaining the predictions of black-box sequence-to-sequence models', *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Copenhagen, Denmark, pp.412–421.

De Cao, N., Schlichtkrull, M.S., Aziz, W. and Titov, I. (2020) 'How do decisions emerge across layers in neural models? interpretation with differentiable masking', *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Association for Computational Linguistics, pp.3243–3255.

DeYoung, J., Søgaard, A. and Kiritchenko, S. (2020) 'Eraser: a benchmark to evaluate rationalized nlp models', *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp.7632–7651.

Ding, S. and Koehn, P. (2021) 'Evaluating saliency methods for neural language models', *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Association for Computational Linguistics, pp.5034–5052.

Feng, S., Wallace, E., Grissom II, A., Iyyer, M., Rodriguez, P. and Boyd-Graber, J. (2018) 'Pathologies of neural models make interpretations difficult', *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Brussels, Belgium, pp.3719–3728.

Ho, N-H., Yang, H-J., Kim, S-H. and Lee, G. (2020) 'Multimodal approach of speech emotion recognition using multi-level multi-head fusion attention-based recurrent neural network', *IEEE Access*, Vol. 8, pp.61672–61686, DOI: 10.1109/ACCESS.2020.2984368.

Hochreiter, S. and Schmidhuber, J. (1997) 'Long short-term memory', *Neural Computation*, Vol. 9, No. 8, pp.1735–1780.

Jacovi, A. and Goldberg, Y. (2020) 'Towards faithfully interpretable NLP systems: how should we define and evaluate faithfulness?', *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, pp.4198–4205.

Jain, S. and Wallace, B.C. (2019) 'Attention is not explanation', *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, North American Chapter of the Association for Computational Linguistics, June, Minneapolis, Minnesota, Vol. 1, pp.3543–3556, DOI: 10.18653/v1/N19-1357.

Kenton, J.D.M-W.C. and Toutanova, L.K. (2019) 'Bert: pre-training of deep bidirectional transformers for language understanding', *Proceedings of NAACL-HLT*, Vol. 1, p.2.

Kessler, J. and Nicolov, N. (2009) 'Targeting sentiment expressions through supervised ranking of linguistic configurations', *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 3, pp.90–97.

Klinger, R. and Cimiano, P. (2013) 'Joint and pipeline probabilistic models for fine-grained sentiment analysis: extracting aspects, subjective phrases and their relations', *2013 IEEE 13th International Conference on Data Mining Workshops*, pp.937–944.

Kobayashi, N., Iida, R., Inui, K. and Matsumoto, Y. (2006) 'Opinion mining on the web by extracting subject-aspect-evaluation relations', *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, pp.86–91.

LeCun, Y., Bottou, L., Bengio, Y. and Haffner, P. (1998) 'Gradient-based learning applied to document recognition', *Proceedings of the IEEE*, Vol. 86, No. 11, pp.2278–2324.

Li, J., Monroe, W. and Jurafsky, D. (2016) *Understanding Neural Networks through Representation Erasure*, CoRR, abs/1612.08220.

Li, X., Bing, L., Li, P., Lam, W. and Yang, Z. (2018) 'Aspect term extraction with history attention and selective transformation', *Proceedings of the 27th International Joint Conference on Artificial Intelligence, IJCAI'18*, AAAI Press, pp.4194–4200.

Lundberg, S.M. and Lee, S-I. (2017) 'A unified approach to interpreting model predictions', *Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17*, Long Beach, California, USA, Vol. 30, No. 10, pp.4768–4777.

Ma, D., Li, S., Wu, F., Xie, X. and Wang, H. (2019) 'Exploring sequence-to-sequence learning in aspect term extraction', *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Florence, Italy, pp.3538–3547.

Madsen, A., Meade, N., Adlakha, V. and Reddy, S. (2022) 'Evaluating the faithfulness of importance measures in NLP by recursively masking allegedly important tokens and retraining', *Findings of the Association for Computational Linguistics: EMNLP 2022*, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, pp.1731–1751.

Mathew, B., Saha, P., Yimam, S.M., Biemann, C., Goyal, P. and Mukherjee, A. (2021) 'Hatexplain: a benchmark dataset for explainable hate speech detection', *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35, pp.14867–14875.

Murdoch, W.J., Liu, P.J. and Yu, B. (2018) 'Beyond word importance: contextual decomposition to extract interactions from LSTMs', *International Conference on Learning Representations*.

Nivre, J. (2003) 'An efficient algorithm for projective dependency parsing', *Proceedings of the Eighth International Conference on Parsing Technologies*, Nancy, France, pp.149–160.

Peng, H., Xu, L., Bing, L., Huang, F., Lu, W. and Si, L. (2020) 'Knowing what, how and why: a near complete solution for aspect-based sentiment analysis', *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34, pp.8600–8607.

Pontiki, M., Galanis, D., Pavlopoulos, J., Papageorgiou, H., Androutsopoulos, I. and Manandhar, S. (2014) 'SemEval-2014 task 4: aspect based sentiment analysis', *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, Association for Computational Linguistics, Dublin, Ireland, pp.27–35.

Qiu, G., Liu, B., Bu, J. and Chen, C. (2011) 'Opinion word expansion and target extraction through double propagation', *Computational Linguistics*, Vol. 37, No. 1, pp.9–27.

Ribeiro, M.T., Singh, S. and Guestrin, C. (2016) ''Why should i trust you?' Explaining the predictions of any classifier', *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp.1135–1144.

Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D. and Batra, D. (2017) 'Grad-CAM: visual explanations from deep networks via gradient-based localization', *2017 IEEE International Conference on Computer Vision (ICCV)*, pp.618–626.

Simonyan, K., Vedaldi, A. and Zisserman, A. (2014) 'Deep inside convolutional networks: visualising image classification models and saliency maps', in Bengio, Y. and LeCun, Y. (Eds.): *2nd International Conference on Learning Representations, ICLR 2014, Workshop Track Proceedings*, 14–16 April, Banff, AB, Canada.

Sixt, L., Granz, M. and Landgraf, T. (2020) 'When explanations lie: why many modified BP attributions fail', *Proceedings of the 37th International Conference on Machine Learning, ICML'20*, JMLR.org.

Smilkov, D., Thorat, N., Kim, B., Viégas, F. and Wattenberg, M. (2017) *SmoothGrad: Removing Noise by Adding Noise*, arXiv preprint arXiv:1706.03825.

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C.D., Ng, A. and Potts, C. (2013) 'Recursive deep models for semantic compositionality over a sentiment treebank', *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Seattle, Washington, USA, pp.1631–1642.

Somasundaran, S. and Wiebe, J. (2010) 'Recognizing stances in ideological on-line debates', *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pp.116–124.

Sundararajan, M., Taly, A. and Yan, Q. (2017) 'Axiomatic attribution for deep networks', *International Conference on Machine Learning*, PMLR, pp.3319–3328.

Tian, H., Gao, C., Xiao, X., Liu, H., He, B., Wu, H., Wang, H. and Wu, F. (2020) 'SKEP: sentiment knowledge enhanced pre-training for sentiment analysis', *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, pp.4067–4076.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. and Polosukhin, I. (2017) 'Attention is all you need', *Advances in Neural Information Processing Systems*, Vol. 30, pp.5998–6008.

Wang, D., Zhao, K. and Wang, Y. (2020) 'Retracted article: based on deep learning in traffic remote sensing image processing to recognize target vehicle', *International Journal of Computers and Applications*, pp.1–7, Taylor & Francis.

Wang, H., Du, M., Yang, F. and Zhang, Z. (2019) *Score-CAM: Improved Visual Explanations via Score-Weighted Class Activation Mapping*, CoRR, abs/1910.01279.

Wang, L., Shen, Y., Peng, S., Zhang, S., Xiao, X., Liu, H., Tang, H., Chen, Y., Wu, H. and Wang, H. (2022) 'A fine-grained interpretability evaluation benchmark for neural NLP', *Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL)*, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Hybrid), pp.70–84.

Webber, W., Moffat, A. and Zobel, J. (2010) 'A similarity measure for indefinite rankings', *ACM Transactions on Information Systems (TOIS)*, Vol. 28, No. 4, pp.1–38.

Wiegreffe, S. and Pinter, Y. (2019) 'Attention is not not explanation', *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Association for Computational Linguistics, Hong Kong, China, pp.11–20.

Wu, Y., Zhang, Q., Huang, X-J. and Wu, L. (2009) 'Phrase dependency parsing for opinion mining', *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pp.1533–1541.

Xu, L., Chia, Y.K. and Bing, L. (2021) 'Learning span-level interactions for aspect sentiment triplet extraction', *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)*, Association for Computational Linguistics, pp.4755–4766.

Yang, B. and Cardie, C. (2012) 'Extracting opinion expressions with semi-Markov conditional random fields', *Conference on Empirical Methods in Natural Language Processing*.

Yang, B. and Cardie, C. (2013) 'Joint inference for fine-grained opinion extraction', *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Association for Computational Linguistics, Sofia, Bulgaria, pp.1640–1649.

Yin, F., Shi, Z., Hsieh, C-J. and Chang, K-W. (2022) 'On the sensitivity and stability of model interpretations in nlp', *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp.2631–2647.

Yin, Y., Wei, F., Dong, L., Xu, K., Zhang, M. and Zhou, M. (2016) 'Unsupervised word and dependency path embeddings for aspect term extraction', *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI'16*, AAAI Press, pp.2979–2985.

Yuan, T., Li, X., Xiong, H., Cao, H. and Dou, D. (2021) 'Explaining information flow inside vision transformers using Markov chain', *Explainable AI Approaches for Debugging and Diagnosis*.

Yuan, Y., Zhou, X., Pan, S., Zhu, Q., Song, Z. and Guo, L. (2020) 'A relation-specific attention network for joint entity and relation extraction', *IJCAI*, Vol. 2020, pp.4054–4060.

Zaidan, O. and Eisner, J. (2008) 'Modeling annotators: a generative approach to learning from annotator rationales', *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pp.31–40.

Zeiler, M.D. and Fergus, R. (2014) 'Visualizing and understanding convolutional networks', *Computer Vision-ECCV 2014: 13th European Conference, Proceedings, Part I*, 6–12 September, Zurich, Switzerland, Springer, pp.818–833.

Zhang, L. and Liu, B. (2012) 'Sentiment analysis and opinion mining', *Encyclopedia of Machine Learning and Data Mining*.

Zhang, W., Huang, Z., Zhu, Y., Ye, G., Cui, X. and Zhang, F. (2021) 'On sample based explanation methods for NLP: faithfulness, efficiency and semantic evaluation', *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)*, Association for Computational Linguistics, pp.5399–5411.

Zhuang, L., Jing, F. and Zhu, X-Y. (2006) 'Movie review mining and summarization', *Proceedings of the 15th ACM International Conference on Information and Knowledge Management*, pp.43–50.

Zhuang, L., Wayne, L., Ya, S. and Jun, Z. (2021) 'A robustly optimized BERT pre-training approach with post-training', *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, Chinese Information Processing Society of China, Huhhot, China, pp.1218–1227.