

A Comparative Analysis of Local Explainability of Models for Sentiment Detection

Hooria Hajiyan, Heidar Davoudi, and Mehran Ebrahimi(✉)

Faculty of Science, Ontario Tech University, Oshawa, ON, Canada
{Hooria.Hajiyan,Heidar.Davoudi,Mehran.Ebrahimi}@ontariotechu.ca

Abstract. Sentiment analysis is one of the crucial tasks in Natural Language Processing (NLP) which refers to classifying natural language sentences by their positive or negative sentiments. In many existing deep learning-based models, providing an explanation of a sentiment might be as necessary as the prediction itself. In this study, we use four different classification models applied to the sentiment analysis of the Internet Movie Database (IMDB) reviews, and investigate the explainability of results using Local Interpretable Model-agnostic Explanation (LIME). Our results reveal how the attention-based models, such as Bidirectional LSTM (BiLSTM) and fine-tuned Bidirectional Encoder Representations from Transformers (BERT) would focus on the most relevant keywords.

Keywords: Explainable Artificial Intelligence · LIME · Sentiment Analysis · Attention-based model

1 Introduction

Machine learning has been widely used in various fields such as video captioning [1], big data analysis [2], Natural Language Processing (NLP) [3], text classification [4], and sentiment analysis [5] which has led to remarkable growth in Artificial Intelligence (AI) research [6]. Sentiment Analysis is a sub-field of NLP that combines tools and techniques from linguistics and computer science to systematically identify, extract, and study emotional states and personal opinions in natural languages. However, the main question is how an algorithm can detect if a text is expressing positive or negative sentiment. Despite the high accuracy and satisfactory performance of machine learning models in sentiment detection, still, the models can be complicated that provide no information about how the sentiment classification task is performed [7]. However, as these models are frequently used to predict people's preferences in recommendation engines, it may be useful to inspect how they learned this prediction knowledge. In social networks, negative sentiments can be shared quickly, which may become a problem if recommendation systems cannot explain the reasons behind their recommendations. Explanations supporting the output of a black-box model are crucial

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
K. Arai (Ed.): FTC 2022, LNNS 561, pp. 593–606, 2023.
https://doi.org/10.1007/978-3-031-18344-7_42

where experts require more information about the decision than a simple prediction. Recent research shows that most of the machine learning models used in NLP, text classification, and sentiment analysis work as black-box models, and the non-transparent structure of these models have led to different types of research in this field using XAI [7–10].

Explainable AI (XAI) emphasizes on understanding the cause and effect within the AI system by examining the sensitivities of the output to changes in the parameter inputs without needing to understand the complex computation of the model [6]. Furthermore, the explanations are helpful in supporting collaboration between AI agents and human experts in many applications [11–13].

Although the transformer models, which rely on attention-based mechanism have become very popular in the development of NLP tasks [14,15], other well-known approaches, like random forest, Support Vector Machine (SVM), Naive Bayes, and K-nearest neighbors can produce a quite good performance in sentiment analysis, their results are not explainable [16]. In this study, we investigate the high performance of attention models in sentiment analysis compared to the random forest and multinomial Naive Bayes, and sentiment explanation results at features level show how each relevant or irrelevant keywords contribute to true positive and false positive sentiment analysis.

To do so, we chose LIME as an explanation tool to evaluate the performance of four different classifiers. As LIME is a model agnostic Post-hoc explanation tool, it can be easily applied to any classification model, no matter if the decision process is interpretable by itself or not. Therefore, we can investigate and compare the sensitivity of attention-based models, and any other simple classifier in sentiment analysis, to locally check the importance of each token/word in the decision process.

In Sect. 2, explanation tools are categorized into local or global approaches and determining whether generating the explanation requires post-processing or not. Related work and the most XAI tools in NLP tasks are provided in Sect. 3. Section 4 discusses four different classification algorithms, besides the state-of-the-art XAI tool used in this study. Quantitative results of each model and sentiment explanations for some instances are provided in Sect. 5. Finally, we conclude the explanation results in Sect. 6.

2 Background

2.1 Transparency of a Black-Box Model

Transparency often refers to interpretability, which means a model is understandable by a human, such as regression, or decision tree [17]. However, explainability is associated with the notion of explanation as an interface between humans and a decision-maker [18]. There is an interchangeable misuse of interpretability and explainability in the literature. The notable difference among these concepts is interpretability refers to a passive characteristic of a model referring to the level at which a given model makes sense for a human, which is also expressed as transparency. In contrast to transparency, explainability refers to an active

characteristic of a model with its internal functions [13]. In other words, a model can be explained, but the interpretability of the model is something that comes from the design of the model itself [13].

2.2 Categorization of Explanation Methods

Explanations are often categorized into two main aspects [12,18]. The first one distinguishes whether the explanation is for an individual prediction, called *local explanation*, or the model's prediction process as a whole, called *global explanation*. Local explanation provides information or justification for the model's prediction on a specific input. Global explanation provides a similar justification by revealing how the model's predictive process works. In other words, global explanation describes the whole decision process as a human term independent of any particular input [8].

Whether the explanation is local or global, explanations differ on whether they arise as part of the prediction process or whether their generation requires post-processing following the model making a prediction [8].

Ante-Hoc. An Ante-hoc approach, which may also be referred to as directly interpretable [19], generates the explanation at the same time as the prediction is made, decision trees and rule-based models are examples of Ante-hoc explanation models.

Post-Hoc. In contrast, a Post-hoc approach requires an additional operation after the predictions are made. Local Interpretable Model-agnostic Explanations (LIME) [11] is an example of producing a local explanation using a surrogate model applied following the predictor's operation. LIME can be applied to any black-box model without the need to understand, change, or modify the decision process. In this research, we choose LIME as a model agnostic Post-hoc explanation method to evaluate four different classification models in sentiment analysis.

3 Related Work

To interpret a prediction of a model in NLP tasks, prior research has focused on developing post-hoc analysis algorithms to select or modify particular instances and explain the behavior of the model [20,21]. One of the post-hoc explanation tools which estimates the input contribution towards the output by computing the partial derivative is called first derivative saliency [22]. As suggested by its name and definition, first derivative saliency can enable feature importance explainability, especially on word/token-level features [8,23]. Gradient-weighted Class Activation Mapping (Grad-CAM) is an attribution method to calculate each input's contribution to the output in text classification problems and NLP tasks [24,25]. The other approach reveals attribute relevance to features computed in an intermediate layer of a neural network is Layer-wise Relevance Propagation (LRP). LRP has been used to enable feature importance explainability in fully

connected, convolution, and recurrent layers [26,27]. However, the main issue is needing access to the inner structure of the model. For this reason, model agnostic explanation methods, specifically perturbation-based, such as Local Interpretable Model-Agnostic Explanations (LIME), have been used among for explaining text classification problems such as sentiment analysis [7,11,28], as they are easy to understand and do not require access to the inner structure of the model [21]. In other words, model-agnostic explanation probe the black-box model by observing the probability change on the predicted class when erasing a certain word [11,29].

Although several studies have been conducted to provide an explanation for the results of sentiment analysis by LIME, they have been mostly focused on explaining a single model, specifically a fined-tuned attention-based model [30], using LIME to improve the explainability of a sentiment classifier with augmented data [30], or a comparison of sentiment analysis methods, then choosing the most accurate one to explain the results of the right predictions, with a high accuracy score [31]. More specifically, there is no work investigating how local explanations can reflect the accuracy of sentiment analysis methods in detecting positive or negative sentiments at the features level.

4 Methodology

4.1 Classification Methods

For this study, we considered four classification methods trained on IMDB review dataset to compare the results of a sentiment analysis, which could either be positive or negative.

Random Forest: Random forest is a classification algorithm consisting of many decisions trees. It uses bagging and feature randomness when building each tree to create an uncorrelated forest of trees whose prediction is more accurate than any individual tree. There have been many studies conducted using random forest and produced quite good performance in sentiment analysis on a large number of sentiments from online purchasing, movie reviews, YouTube, and Twitter social media [32–35]. Nevertheless, trusting the predictions and explaining the results are remain as the main issues.

Multinomial Naive Bayes: Naive Bayes is one of the most popular algorithms used in a variety of classification problems because of its fast processing time and high level of effectiveness [36]. This algorithm uses statistical methods to calculate the probability of a class based on its attributes, then find the highest probability value to classify the data to the most appropriate category. Due to the basic concept of the Naive Bayes algorithm, it is more often used in classifying texts and sentiments analysis as it combines the probability of words and categories documents [37,38]. The Multinomial Naive Bayes (Multinomial NB) method is the probability-based algorithm suitable for classification with discrete features such as word counts for text classification. This approach considers the term frequency and calculates the probability of each label given the input text [39,40].

Bidirectional LSTM: Bidirectional Recurrent Neural Network (BRNN) is a kind of recurrent neural networks that can be trained using all available input information in the past and future of a specific time frame. The idea is to split the state neurons of a regular RNN in a part that is responsible for the positive time direction, forward states, and apart for the negative time direction, backward states [41]. These kinds of networks have been the most suitable method in text classification, NLP, text mining, and sentiment analysis in recent years [42,42–47]. Because in RNNs the gradient becomes vanished in long sequences, a Bidirectional LSTM Network (BiLSTM) has been proposed as a solution to this problem [48]. BiLSTM has the ability to extract the contextual information from the feature sequences by dealing with both forward and backward dependencies. In this study, we used BiLSTM with a forward LSTM, which processes the sequence in chronological order, and a backward LSTM, which handles the sequence in reverse order. The output is then the concatenation of the states of the forward and backward LSTM, followed by a classifier layer to provide the probability of each positive or negative sentiment.

Bidirectional Encoder Representations from Transformers-BERT: BERT is designed to pre-train deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers. As a result, the pre-trained BERT model can be fine-tuned with just one additional output layer to create state-of-the-art models for a wide range of tasks [49]. BERT improves the fine-tuning based approaches by using a Masked Language Model (MLM) pre-training objective inspired by Cloze test [50]. The MLM randomly masks some of the tokens from the input, and the objective is to predict the original vocabulary of the masked word based only on its context. The MLM objective also enables the representation to fuse the left and the right context, which allows us to pre-train a deep bidirectional transformer. In other words, BERT models are usually pre-trained on a large corpus of text, then fine-tuned for specific tasks. We chose BERT because it has been wildly successful on a variety of tasks in NLP, text classification, and sentiment analysis [4,39,51–54]. We used this pre-trained model and modify the output layer, which acts as a classifier to predict the probability of each class, then train the model on the sentiment analysis dataset.

4.2 LIME

Since the post-hoc explanation methods have received increasing attention in recent years, and model agnostic explanation has been widely used in sentiment analysis and NLP tasks [7,11,28,29], we chose LIME as a local explanation approach in this research. LIME is a post-hoc explanation tool that is easy to understand and it can be applied to any classification method when the prediction process is done, and it is a model agnostic explanation as it does not need to get access to the inner structure of a complex model [11]. LIME can faithfully explain the predictions of any black-box model, which takes an instance and creates some neighbours around that called perturbations, then builds a

locally linear model around the predictions of an opaque model to explain it. The number of perturbations around each instance and the total number of features represented in the output are two main parameters of LIME that should be fine-tuned based on the given problem. According to the literature, the number of perturbations will guarantee the stability of the resulting explanation [55], and it should be 10 times larger than the number of words [56]. As the instances are of different lengths, we adjusted this parameter based on the size of the given instance. Therefore, the linear model fits on these perturbations and returns the n important features. We decided to highlight 10% of the length of each instance as the most important words detected by LIME.

5 Experiments

This study was conducted on the IMDB review dataset for binary sentiment analysis containing 25,000 reviews for training and 25,000 for testing. In addition, there is 50,000 additional unlabeled data for unsupervised use as well, Table 1.

Table 1. IMDB Movie Dataset.

IMDB review (negative - positive)	
Train ('text', 'label')	25,000
Test ('text', 'label')	25,000
Unsupervised	50,000

We trained four classification methods on this dataset, and the classification reports for training and test are provided in Table 2. In order to avoid the overfitting problem on a large dataset, we trained random forest using Grid Search Cross-Validation (GridSearchCV) with a set of possible values for each parameter. GridSearchCV loops through all possible values and combinations of the hyperparameter and fits the model on the training data. Finally, it will return a suitable combination of parameter values. The chosen parameters are as follows: n-estimators = 800, max-depth = 10, and max-features = 100. Although we expected to see the attention models such as BRRN and BERT have higher accuracy than random forest and Multinomial NB, the main purpose of this study is to interpret the decision made by each classifier using LIME. In other words, the results of LIME will represent which words/tokens the model was looking at to make a prediction.

5.1 Quantitative Results

This Section provides classification reports shown in Table 2. The results show that BiLSTM and BERT make the most accurate predictions in text classification. In the next section, we explain the predictions made by classifiers for an individual instance and reveal the performance of each method at features level.

Table 2. Classification accuracy.

	Train	Test
Random forest	0.80	0.79
Multinomial NB	0.85	0.82
Bidirectional RNN	0.87	0.85
BERT	0.93	0.86

5.2 Sentiment Explanation Results

To explain the sentiment predictions, we have evaluated the performance of each model using LIME on two instances of positive and negative sentiment. The instances showing true positive and false positive to discuss the accuracy of each method at the features level, and the explanations are shown visually. More precisely, we chose two instances to reveal how false positive prediction may arise by looking at irrelevant words, and on the other hand, a correct prediction can stem from an inaccurate model, where the prediction can not be explained in a meaningful way. Note that first the input instance is fed into the model, and then LIME creates some neighbours or perturbations by masking some tokens or words. These perturbations are new input instances, and the model predicts the labels. Finally, a linear model has been fitted on these neighbours, and LIME returns the coefficients, called weights, showing the importance of tokens toward each label. LIME represents the weight of tokens that affect the prediction positively and negatively.

The first sentiment is: *I really tried, but movie just didn't work for me. The action scenes were dull, acting was surprisingly poor, and some of these characters were TOO stereotypical to even be funny. Pam Grier tries, but when you have nothing to work with, even her considerable talent can not prevent a disaster. Even by standards of weak genre, film is pretty bad.*

The actual label of this review is negative. The predicted label made by random forest and Multinomial NB is positive, which is wrong. However, BiLSTM and BERT made a correct prediction by considering the whole context and looking at the words which affected the prediction toward the negative label. The prediction probabilities and local explanations are provided in Fig. 1, 2, 3, and 4, respectively. Figure 1 and 2 show that random forest and Multinomial NB have looked at some irrelevant words and considered each word/token as a feature instead of the whole context. This is an example of a false-positive error that happened by these two methods and the predicted probabilities are shown for each sentiment analysis. Although it is not a long sentence, this explanation shows why the models made a wrong prediction by looking at some stop words, or irrelevant tokens, as these classification models look at each token/word individually without considering the contextual meaning. We removed those stop words and then passed sentences to the same models, but there was no specific improvement in the results, and again, some irrelevant words/tokens were

highlighted by LIME. So as we keep the sentences as is to have more consistency in our comparison.

From another point of view, the predicted probabilities in Fig. 1, 2, and the weight of each word toward the positive and negative sentiments are slightly different, however, the highlighted words remain the same. Therefore, these words seem to be the most frequent ones in the input instance, and the behavior of both of these methods would be alike in this case. Figure 3, and 4 show the attention models made the correct prediction for the same instance by looking at the most relevant words.

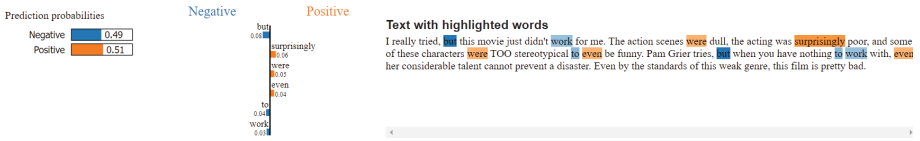


Fig. 1. Actual: negative, predicted label by random forest: positive.

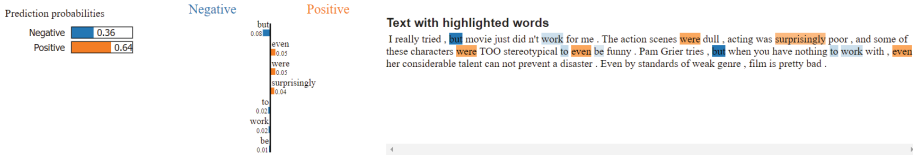


Fig. 2. Actual: negative, predicted label by multinomial NB: positive.

As can be seen, the attention-based models are trying to capture the contextual meaning of a text. Figure 3, and 4 show the true negative sentiment analysis made by BiLSTM and BERT, and the probabilities are about 1. Though the highlighted words and their weights detected by LIME are slightly different in BiLSTM and BERT, the whole context conveys the same sentiment and it is what we hope to see in attention-based methods.



Fig. 3. Actual: negative, predicted label by BiLSTM: negative.



Fig. 4. Actual: negative, predicted label by BERT: negative.

The second sentiment is: *"I first saw this movie on IFC. Which is a great network by the way to see underground films. I watched this movie and was thinking it was going to be pure drama and a story line that doesn't hold water. But it really was a worth while watch. The main character is in such rough shape, and you hate to see him deny help, but no matter what you just can't hate him. His devotion to The Beatles and John Lennon is a great metaphor for his life and the helplessness he feels. The atmosphere of the film is also great. At times, you feel like you can see what he sees, feel what he feels in some situations. This movie does not leave you wanting to know more, or disliking a loophole in the plot. There are NO loopholes (in my opinion). I have always been a fan of foreign films, especially now with movies being made so poorly in America. I really enjoy the foreign settings because I feel it can take you on a trip, and sometimes understand a different culture. This movie did all those things to me and more. Please watch this movie and if you're new to foreign films, this is a great start."*

The actual label of this instance is positive. The prediction made by each classifier besides the local explanations are as follows.

According to Fig. 5 and 6, although the predictions of random forest and Multinomial NB is correct and the probability of the correct label is significantly higher than the opposite one, again like the first instance, most of the highlighted words detected by LIME seem to be irrelevant to the sentiment analysis task, like the most frequent stop words, *a, is, and,* etc. The point is that a model can make a correct prediction, but it does not seem to be meaningful if we explain the reason behind that. That is why we cannot easily trust the model's prediction, besides it might raise the question is this correct prediction happened accidentally.

The same instance has been fed into BiLSTM and BERT. The prediction probabilities and explanation results are shown in Fig. 7, and 8. The prediction probabilities of the actual label are close to 1, and the highlighted parts of the sentence are more precise. In Fig. 7, and 8 the word *great* got more weight, besides other relevant words such as *worth, enjoy* detected by LIME. Although these attention-based models consider the contextual meaning of the given instance, these highlighted words are the most affecting ones towards the positive label.

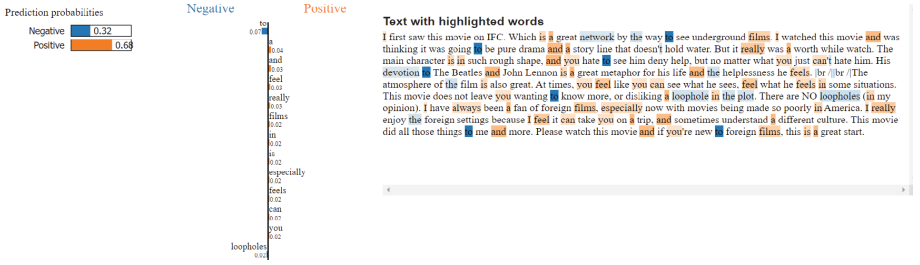


Fig. 5. Actual: positive, predicted label by random forest: positive.



Fig. 6. Actual: positive, predicted label by multinomial NB: positive.

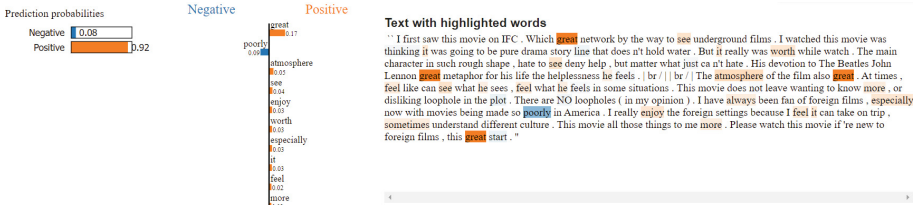


Fig. 7. Actual: positive, predicted label by BiLSTM: positive.

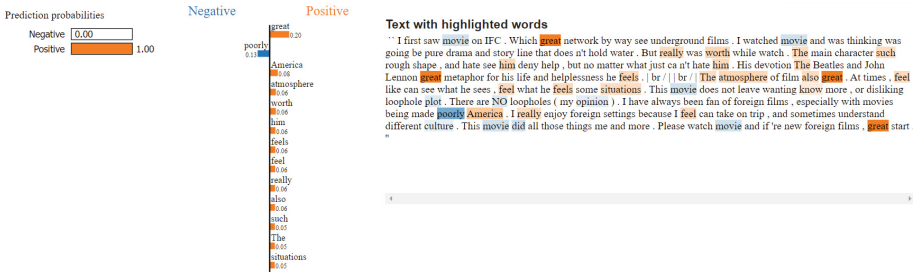


Fig. 8. Actual: positive, predicted label by BERT: positive.

6 Conclusion

Sentiment analysis has become very popular in both research and business due to the availability of vast amount of opinions currently produced by users on social media. Standard sentiment analysis deals with classifying the overall sentiment of a text by considering the importance of each word within a context. This study investigated the accuracy of four classification models for sentiment analysis using local explanations. Precisely, we performed four different classification methods, including random forest, Multinomial NB, BiLSTM, and BERT on the IMDB review, then used LIME to explain the predictions. We investigated two case studies from the given dataset chosen as true positive and false positive predictions, then revealed the importance of each keyword affecting the predicted label. The results proved that although random forest and Multinomial NB would predict the actual label, the prediction might not be precise in long sentences when some irrelevant words are highlighted in the explanations. In contrast, attention-based models like BiLSTM and BERT accurately predicted the correct label by focusing on the most relevant parts of the sentence. This study shows how a correct prediction, even with a high prediction probability, cannot be accepted blindly. In other words, a correct prediction may arise from an inaccurate model when we explain the model's behavior at the features level. Although many classification methods have been used in sentiment analysis, attention-based models represent the high performance through a decision process, when the results are explained. For further research, we can examine the sensitivity of attention-based models to different words through an explanation framework by replacing the most affecting tokens with their synonyms and then explaining the decisions again to see if the new words can change the contextual meaning of a sentence or not. Furthermore, since LIME is a model-agnostic explanation technique, we can combine it with other NLP tasks, e.g. summarizing, extraction, and question-answering, which we leave for our future work.

Acknowledgments. This work has been supported in part by the Natural Sciences and Engineering Research Council of Canada (NSERC).

References

1. Zhou, L., Zhou, Y., Corso, J.J., Socher, R., Xiong, C.: End-to-end dense video captioning with masked transformer. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 8739–8748 (2018)
2. Najafabadi, M.M., Villanustre, F., Khoshgoftaar, T.M., Seliya, N., Wald, R., Muharemagic, E.: J. Big Data, **2**(1), 1–21 (2015)
3. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in Neural Information Processing Systems, pp. 3111–3119 (2013)
4. Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M., Gao, J.: Deep learning-based text classification: a comprehensive review. ACM Comput. Surv. (CSUR) **54**, 1–40 (2021)

5. Hasan, A., Moin, S., Karim, A., Shamshirband, S.: Machine learning-based sentiment analysis for twitter accounts. Math. Comput. Appl. **23**, 11 (2018)
6. Linkov, I., Galaitsi, S., Trump, B.D., Keisler, J.M., Kott, A.: Cybertrust: from explainable to actionable and interpretable artificial intelligence. IEEE (2020)
7. Bodria, F., Panisson, A., Perotti, A., Piaggesi, S.: Explainability methods for natural language processing: applications to sentiment analysis (Discussion Paper) (2020)
8. Danilevsky, M., Qian, K., Aharonov, R., Katsis, Y., Kawas, B., Sen, P.: A survey of the state of explainable AI for natural language processing arXiv preprint arXiv:2010.00711 (2020)
9. Liu, H., Yin, Q., Wang, W.Y.: Towards explainable NLP: a generative explanation framework for text classification arXiv preprint arXiv:1811.00196 (2018)
10. Wiegreffe, S., Marasović, A.: Teach me to explain: a review of datasets for explainable NLP. arXiv preprint arXiv:2102.12060 (2021)
11. Ribeiro, M.T., Singh, S., Guestrin, C.: 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, pp. 1135–1144 (2016)
12. Adadi, A., Berrada, M.: Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). IEEE Access **6**, 52138–52160 (2018)
13. Arrieta, A.B., et al.: Explainable Artificial Intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, vol. 58. Elsevier (2020)
14. Grimsley, C., Mayfield, E., Bursten, J.: Why attention is not explanation: surgical intervention and causal reasoning about neural models (2020)
15. Brunner, G., Liu, Y., Pascual, D., Richter, O., Ciaramita, M., Wattenhofer, R.: On identifiability in transformers. arXiv preprint arXiv:1908.04211 (2019)
16. Daeli, N.O.F., Adiwijaya, A.: Sentiment analysis on movie reviews using Information gain and K-nearest neighbor. J. Data Sci. Appl. **3**, 1–7 (2020)
17. Lipton, Z.C.: The mythos of model interpretability. In: Machine learning, the Concept of Interpretability is Both Important and Slippery. ACM New York, NY, USA,, vol. 16, no. 3, pp. 31–57. Queue Press (2018)
18. Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., Pedreschi, D.: A survey of methods for explaining black box models. In: ACM Computing Surveys (CSUR) (2018)
19. Arya, V., et al.: One explanation does not fit all: a toolkit and taxonomy of AI explainability techniques. arXiv preprint arXiv:1909.03012 (2019)
20. Kenny, E.M., Keane, M.T.: Twin-systems to explain artificial neural networks using case-based reasoning: comparative tests of feature-weighting methods in ANN-CBR twins for XAI. In: Twenty-Eighth International Joint Conferences on Artificial Intelligence (IJCAI), Macao (2019)
21. Keane, M.T., Smyth, B.: Good counterfactuals and where to find them: a case-based technique for generating counterfactuals for explainable AI (XAI). In: International Conference on Case-Based Reasoning (2020)
22. Saltelli, A., et al.: Global sensitivity analysis: the primer (2008)
23. Sundararajan, M., Taly, A., Yan, Q.: Axiomatic attribution for deep networks. In: International Conference on Machine Learning (2017)
24. Gorski, L., Ramakrishna, S., Nowosielski, J.M.: Towards grad-cam based explainability in a legal text processing pipeline. arXiv preprint arXiv:2012.09603 (2020)
25. Lertvittayakumjorn, P., Toni, F.: Human-grounded evaluations of explanation methods for text classification. arXiv preprint arXiv:1908.11355 (2019)

26. Poerner, N., Roth, B., Schütze, H.: Evaluating neural network explanation methods using hybrid documents and morphological agreement. arXiv preprint arXiv:1801.06422 (2018)
27. Croce, D., Rossini, D., Basili, R.: Explaining non-linear classifier decisions within kernel-based deep architectures. In: Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP (2018)
28. Alvarez-Melis, D. Jaakkola, T.S.: A causal framework for explaining the predictions of black-box sequence-to-sequence models. arXiv preprint arXiv:1707.01943 (2017)
29. Chen, H., Zheng, G., Ji, Y.: Generating hierarchical explanations on text classification via feature interaction detection. arXiv preprint arXiv:2004.02015 (2020)
30. Chen, H., Ji, Y.: Improving the explainability of neural sentiment classifiers via data augmentation. arXiv preprint arXiv:1909.04225 (2019)
31. Aljuhani, S.A., Alghamdi, N.S.: A comparison of sentiment analysis methods on Amazon reviews of Mobile Phones. Int. J. Adv. Comput. Sci. Appl. **10**, 608–617 (2019)
32. Karthika, P., Murugeswari, R., Manoranjithem, R.: Sentiment analysis of social media network using random forest algorithm. In: 2019 IEEE International Conference on Intelligent Techniques in Control, Optimization and Signal Processing (INCOS) (2019)
33. Singh, J., Tripathi, P.: Sentiment analysis of Twitter data by making use of SVM, Random Forest and Decision Tree algorithm. In: 2021 10th IEEE International Conference on Communication Systems and Network Technologies (CSNT) (2021)
34. Munshi, A., Arvindhan, M., Thirunavukkarasu, K.: Random forest application of twitter data sentiment analysis in online social network prediction. In: Emerging Technologies for Healthcare: Internet of Things and Deep Learning Models (2021)
35. Aufar, M., Andreswari, R., Pramesti, D.: Sentiment analysis on YouTube social media using decision tree and random forest algorithm: a case study. In: 2020 International Conference on Data Science and Its Applications (ICoDSA) (2020)
36. Novendri, R., Callista, A.S., Pratama, D.N., Puspita, C.E.: Sentiment analysis of YouTube movie trailer comments using Naïve Bayes. Bull. Comput. Sci. Electr. Eng. **1**, 26–32 (2020)
37. Dey, S., Wasif, S., Tonmoy, D.S., Sultana, S., Sarkar, J., Dey, M.: A comparative study of support vector machine and Naive Bayes classifier for sentiment analysis on Amazon product reviews. In: 2020 International Conference on Contemporary Computing and Applications (IC3A) (2020)
38. Li, Z., Li, R., Jin, G.: Sentiment analysis of danmaku videos based on Naïve Bayes and sentiment dictionary. IEEE Access (2020)
39. Dhola, K., Saradva, M.: A comparative evaluation of traditional machine learning and deep learning classification techniques for sentiment analysis. In: 2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence) (2021)
40. Rahman, R., Masud, M.A., Mimi, R.J., Dina, M.N.S.: Sentiment analysis on bengali movie reviews using multinomial Naïve Bayes. In: 2021 24th International Conference on Computer and Information Technology (ICCIT) (2021)
41. Schuster, M., Paliwal, K.K.: Bidirectional recurrent neural networks. IEEE Trans. Signal Process. **45**, 2673–2681 (1997)
42. Nistor, S.C., Moca, M., Moldovan, D., Oprean, D.B., Nistor, R.L.: Building a twitter sentiment analysis system with recurrent neural networks. Sensors **21**, 2266 (2021)

43. Islam, M.S., Sultana, S., Roy, U.K., Al Mahmud, J., Jahidul, S.: HARC-new hybrid method with hierarchical attention based bidirectional recurrent neural network with dilated convolutional neural network to recognize multilabel emotions from text. Jurnal Ilmiah Teknik Elektro Komputer dan Informatika (JITEKI) (2021)
44. Abid, F., Li, C., Alam, M.: Multi-source social media data sentiment analysis using bidirectional recurrent convolutional neural networks. Comput. Commun. **157**, 102–115 (2020)
45. Cai, Y., Huang, Q., Lin, Z., Xu, J., Chen, Z., Li, Q.: Recurrent neural network with pooling operation and attention mechanism for sentiment analysis: a multi-task learning approach. Knowl.-Based Syst. **203**, 105856 (2020)
46. Turek, J., Jain, S., Vo, V., Capotă, M., Huth, A., Willke, T.: Approximating stacked and bidirectional recurrent architectures with the delayed recurrent neural network. In: International Conference on Machine Learning (2020)
47. Elfaik, H., et al.: Deep bidirectional LSTM network learning-based sentiment analysis for Arabic text. J. Intell. Syst. **30**, 395–412 (2021)
48. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**, 1735–1780 (1997)
49. Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: Bert: pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
50. Taylor, W.L.: "Cloze procedure": a new tool for measuring readability. J. Quart. **30**, 415–433 (1953)
51. Habimana, O., Li, Y., Li, R., Gu, X., Yu, G.: Sentiment analysis using deep learning approaches: an overview. Sci. China Inf. Sci. **63**, 1–36 (2020)
52. Chauhan, P., Sharma, N., Sikka, G.: The emergence of social media data and sentiment analysis in election prediction. J. Ambient. Intell. Humaniz. Comput. **12**, 2601–2627 (2021)
53. Karimi, A., Rossi, L., Prati, A.: Adversarial training for aspect-based sentiment analysis with Bert. In: 2020 25th International Conference on Pattern Recognition (ICPR) (2021)
54. Hoang, M., Bihorac, O.A., Rouces, J.: Aspect-based sentiment analysis using BERT. In: Proceedings of the 22nd NORDIC Conference on Computational Linguistics (2019)
55. Zhou, Z., Hooker, G., Wang, F.: S-lime: stabilized-lime for model explanation. In: Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining (2021)
56. Garreau, D., Mardaoui, D.: What does LIME really see in images? In: International Conference on Machine Learning (2021)