A Survey of Explainable Artificial Intelligence Approaches for Sentiment Analysis

Bernadetta Maleszka(✉)

Faculty of Information and Communication Technology, Department of Applied Informatics, Wroclaw University of Science and Technology, Wybrzeze Wyspianskiego 27, 50-370 Wroclaw, Poland
Bernadetta.Maleszka@pwr.edu.pl

Abstract. Nowadays, the problems of sentiment analysis, opinion mining and fake news detection are very important. Artificial intelligence methods are widely used to analyze opinions in social media and to obtain the results in an efficient manner and with high accuracy. The most common approaches are ML methods using nonlinear models and complex structures, e.g. deep neural networks, SVM or random forest. These methods have only one disadvantage: they work as black-boxes so it is hard to understand how they predict the results and lowers trust to such methods. In this paper we present a survey of explainable artificial intelligence methods that are used in sentiment analysis area, analyze the differences between XAI in SA and feature selection methods and indicate trends and challenges in this area.

Keywords: Sentiment Analysis · Opinion Mining · XAI · Feature Selection

1 Introduction

Nowadays, sentiment analysis is a domain that develops rapidly. There are more and more models, methods and algorithms that help the user to form an opinion about particular topic, person, issue, service, etc. [1,5]. Developed artificial intelligence methods provide us with better and better results in this area.

To obtain better accuracy, more complicated structures of the model and more sophisticated methods are used. The problem arises when user asks how a particular result was achieved or how a particular sample has influenced the final model [34]. It is a problem of reliability of the system [35].

In this paper we present a survey of explainable artificial intelligence (XAI) methods that are used to increase user's trust to the system of sentiment analysis or opinion mining. We have provided the following sections: In Sect. 2 we present the basic concepts correlated to sentiment analysis and XAI. In Sect. 3 we describe related works for XAI methods in sentiment analysis. Some trends and challenges are provided in Sect. 4. Final remarks and summary are in Sect. 5.

N. T. Nguyen et al. (Eds.): ACIIDS 2023, LNAI 13996, pp. 52–62, 2023.
https://doi.org/10.1007/978-981-99-5837-5_5

2 Background

In this section we present the basic idea of sentiment analysis, explainable artificial intelligence and feature selection. We provide some definitions and description of the problem.

2.1 Sentiment Analysis

Nowadays, sentiment analysis is a very important issue as it can influence many aspects of everyday life. Before a user decides to buy or order a product or service, he or she tries to find the best offer but more and more often he or she looks for opinion from other users about the product or service.

In the last few years, the development of e-commerce systems and social networks has allowed the user to share his or her opinion easily [31]. On the other hand, the user can find a huge amount of e.g. product reviews, so that it is impossible to manage out all information. Many systems offer recommendations or decision support algorithms to improve user experience. Using sentiment analysis techniques allows to additionally enrich the accuracy of recommendations as they reflect users opinions.

The most popular tasks based on sentiment analysis are as follows: opinion mining [31], fake news detection [28,29] and stance detection [10]. The main contribution of sentiment analysis is to extract opinions from different modalities, e.g. text, image, video, etc. and usually combine them to obtain a final polarity. There arises a problem of opinion veracity and credibility which lead us to the fake news detection issues. It is possible to use sentiment analysis approaches to judge if a news is true or fake. The stance detection problem is correlated to users attitude toward a situation or an event. The user can agree or disagree with statements of other users.

According to Phan et al. [29] "Sentiment is the feeling, attitude, evaluation, or emotion of users toward specific aspects of topics or for the topics". The set of possible values of sentiment can be defined in many ways, e.g. [15,17,29]:

- $s = \{positive,\ negative\}$.
- $s = \{positive,\ neutral,\ negative\}$.
- $s = \{positive,\ neutral,\ negative,\ mixed\}$.
- $s = \{strong/very\ positive,\ positive,\ neutral,\ negative,\ strong/very\ negative\}$.
- $s = \{very\ very\ negative,\ very\ negative,\ negative,\ somewhat\ negative,\ neutral,\ somewhat\ positive,\ positive,\ very\ positive,\ and\ very\ very\ positive\}$.

Sentiment analysis is a "process used to determine the sentiment orientation in opinions" [29]. The process can be treated as a classification problem: classify a given opinion o toward specific aspect or topic into one sentiment polarity from set s [6]. Sentiment analysis can be divided into three levels: document level (when we judge the polarity of the final conclusions of some report), sentence level (polarity of each sentence) and aspect level (polarity towards particular aspect).

Phan et al. [29] defines the problem in a wider way: sentiment is an attitude of a particular user u in a timestamp t towards a given topic p. The user u delivers an opinion about the topic p and the task is to judge whether the opinion is positive or negative.

The most popular methods for sentiment analysis are those based on machine learning approaches or those based on lexicon approaches (e.g. corpus or dictionary based approaches) [22,29,39]. In the first group one can use supervised methods (e.g. probabilistic classifiers: Naive Bayesian, Bayesian network, maximum entropy, linear classifier: SVM, neural network, decision tree, rule-based methods, etc.), semi-supervised methods (e.g. self-training, graph-based, generative models), unsupervised methods (k-means, fuzzy c-means, agglomerative and divisive algorithms) or deep learning methods (RNN, CNN, LSTM, GNN, GCN, etc.) [37]. We can also find many hybrid approaches that combine machine learning with lexicon-based approaches, especially deep neural networks and lexicon-based methods. Usually, methods from the last group obtain the best results.

To judge the efficiency of the method we can use typical efficient metrics, such as precision, recall, F-measure and accuracy. Usually more complicated methods obtain better results than linear or simple methods. On the other hand, these methods are hard to explain. It is not obvious how single opinion or statement affects the final result. This is the reason for the popularity of developing explainable methods.

2.2 Explainable Artificial Intelligence

Artificial intelligence has appeared in many aspects of our life, e.g. medicine, transport, e-commerce, intelligent houses, etc. The systems can help the doctor to analyze X-ray or magnetic resonance images [13], support car drivers [16], recommend us some personalized products or services [25], allows us to "talk" with ChatGPT [9], etc.

The main contribution of XAI is to increase user's trust in AI systems. User confidence is crucial in many situation, especially when the results of these systems affect our health or even life.

The main idea of the XAI is to explain why the system obtained the particular result. It can be illustrated with the Albert Einstein's quote: *"If you can't explain it simply, you don't understand it well enough"*. It is an important aspect of many deep algorithms were it is not obvious what information does the network contain or why does this particular input lead to that particular output [14].

The most frequent division of XAI approaches is into two groups: visualization methods and post-hoc analysis. In the first group, there exists a few algorithms that do not need any explanation as they are transparent enough. They are: linear or logistic regression, decision trees, kNN, rule based learners, general additive model, or Bayesian models [3]. The category of post-hoc analysis contains more sophisticated methods that do not allow to easily explain why a particular case was classified into particular class. They are e.g. tree ensembles, SVM, deep neural networks: multi-layer, convolutional or recurrent neural networks. Usually, the following techniques are used for explaining how they work:

model simplification, feature relevance, local explanations or visualization in the post-hoc step.

Athira et al. [4] differentiate two concepts: interpretability and explainability. In the first case, we have a simple structure and it can be used to interpret or explain how the method works (e.g. linear model, decision trees, association rules). It assumes that the used algorithms or methods are transparent and does not need any explanation. It can be also called model-based explainability, or explainability by design [24]. The category of post-hoc explanations tries to explain how a black box (an algorithm or a method) works based on the final results [38]. It is crucial for such models that are non-linear: ensemble methods or neural networks (e.g. CNN, RNN [2]).

Arrieta et al. [3] and Ding et al. [8] have defined more aspects of explainability: understandability – user can understand how the algorithm works without any additional explanation about the internal structure; comprehensibility – the result of the learning algorithm should be understandable for human, it is also connected with the model complexity; transparency – the model by itself is understandable.

Another division of XAI models is into global and local explanation [38]. The global one aims to explain how the input variables influence the model. The local explanation focuses on how each feature influences the result (e.g. SHAP algorithm [20]).

Dazeley et al. [7] claim that full XAI system should implement two processes: social and cognitive. The first process should take into account interactions with other actors like people, animals, other agents, etc. The cognitive process should identify general causes and counterfactuals [11].

The authors have proposed the following levels of explanations according to the factors of user beliefs and motivations [7]:

- Reactive: it is an explanation of an agent's reaction to immediately perceived inputs – like instinctive behaviour of animals in dangerous situation.
- Disposition: it is an explanation of an agent's underlying internal disposition towards the environment and other actors that motivated a particular decision – the agent's decision is based on its beliefs or desires.
- Social: it is an explanation of a decision based on an awareness or belief of its own or other actors' mental states.
- Cultural: it is an explanation of a decision made by the agent based on what it has determined is expected of it culturally, separate from its primary objective, by other actors.
- Reflective: it is an explanation detailing the process and factors that were used to generate, infer or select an explanation.

The first four levels are object-level explanations based on decisions or arguments and the last meta-explanation is based on the scenario structure or historical decisions or justifications.

In the literature one can find many methods for XAI but the majority of them can be classified to the lower levels: reactive, disposition or social.

In the next part of this section we present the most popular approaches to XAI. The methods in the group of visualization are based on visual form of explanation, like highlighted text in natural language processing [23] or explicit visualization of the results according some subsets of features [33]. The post-hoc explanations' aim is to find feature relevance, model simplification, text explanation or explanation by example [3]. In many cases the post-hoc methods also use visualization approaches.

Visualization. Nowadays, there is more and more methods to train the model but it is hard to explain why we obtained any specific final results, what was an impact of particular set of features or cases during the training process and how they have influenced the final prediction mechanism [33].

Visualization approach allows us to take a look inside the data in a simpler way than using analytical methods. It can provide us with some intuitions about data distribution or differences between some subsets of cases.

So et al. [33] claim that the basics of explanation is the set of features that can be visualized. They differentiate the following aspects:

- feature importance – it calculates how the feature of all observations impacts the prediction. The most popular method is SHAP (SHapley Additive exPlanations) [20] or counterfactual explanations [11];
- additive variable attributions – it estimates which instances of the dataset are outliers;
- what-if analysis – one can use ceteris-paribus plot to analyze a relationship between features and response.

One of the most effective algorithms for sentiment analysis uses CNN architecture. Souza et al. [36] proposed five different PIV (particle image velocimetry) techniques to visualize the flow of the method. They are as follows: guided backpropagation (GBP), saliency (SAL), integrated gradients (IGR), input $\times$ gradients (IXG) and DeepLIFT (DLF).

Post-hoc XAI Methods. An input for a post-hoc XAI methods is a trained model. An expected output of the method is an approximate model that explains how the original model works [24]. It can also reflect decision logic or generate some representation of the model that is understandable, e.g. set of rules, feature importance score or heatmaps.

Most of the XAI methods dedicated for the text processing are model-specific approaches [3].

Some exemplary methods are described below [24]:

- LIME (Local Interpretable Model-agnostic Explanations) – the algorithm introduces some perturbations to real samples and provides observations about the output of the model;
- If-then rules – they should reflect the dependencies between the features. The generated rules should represent the original black-box model; determining the optimal set of rules is an optimization task.

The results obtained from post-hoc XAI methods that have found some dependencies between features, can be used for the feature selection methods. The main aim of feature selection methods is to reduce the dimensionality of the dataset and the complexity of the solution. It is possible because a lot of data is redundant [21,32]. The task is to delete (or omit) some data as it does not significantly change the result of the algorithm.

The methods and techniques of Explainable Artificial Intelligence presented above focus on the feature – how a particular feature influences the result. They take care about the form of explanation, use a subset of features to obtain the result and they are separated from the model [11].

These methods also have disadvantages: they cannot show us, e.g. what is a minimal set of samples or instances that guarantees the obtained results [41] and using these methods, is not clear which input instance has determined the final result.

3 Explainable AI in Sentiment Analysis

Sentiment analysis is an area where transparency is a crucial feature of the user's trust in the system [2]. Before a user makes a purchase decision or decides to use the service, he or she may decide to check the opinion about the topic, product or delivered service, etc.

Explainable artificial intelligence techniques allow us to better understand prediction of the model [12]. More and more methods and models in this area are predictive – to increase user's confidence in the system, it should provide transparent and trustworthy results. As the authors claim, more effective algorithms mean less transparency.

The main objective of the XAI methods in sentiment analysis area is to answer the query: "How can XAI methods reveal potential bias in trained machine learning models for the prediction of product ratings?" [34].

In this section we present a classification of the existing solution for XAI in sentiment analysis domain. Most commonly used methods focus on the following aspects [34]:

- Feature importance – it approximates the global relevance of the feature in the model. It depends on the model, e.g. for models based on trees it can split the tree and for linear models it is correlated with regression coefficient.
- Local attributions – this approach allows to visualize the impact of a single feature's variance as it can be missed by the analysis of global feature importance.
- Partial dependency plot – it presents how each feature or several features can impact the final result.

Above mentioned methods are based on the visualization of the results. They can be used both to explain how the model works and to improve it: a feature can be not used in the model when it is not important, it has too high variance or it has weak relationship with other attributes.

The improvement of the interpretability or explainability can be achieved mostly by high transparency of the model that can be developed from structure of the network, feature importance, local gradient information, redistribution the function's value on the input variables, specific propagation rules for neural networks [2].

In Table 1 we summarize existing papers focused on XAI in SA. Each paper is analyzed according to the main problem, feature and techniques used for sentiment analysis and type of explainability.

All these papers developed models for predicting sentence polarity. Most of them work on text reviews of documents or movies or simply tweets using a wide range of possible models of the data and methods for sentiment analysis: naive Bayes, decision trees, random forests, LSTM, softmax attention, neural networks (CNN, RNN, etc.).

The most popular approach to provide explanations of the results is a visualization method: SHAP, BertViz [12], LIME [12,35], feature importance [19,26,40], local feature attributions and partial dependency plots [34], contextual importance and utility [30].

4 Trends and Challenges

The area of XAI methods is more and more developed to ensure more transparent and confident results that user can trust them. There are still many aspects that should be taken into account.

The challenges of XAI methods in sentiment analysis are correlated with development of new methods for SA, especially deep neural network approaches. As they become more and more popular and are used by wider and wider group of people (sometimes they even use them without thought or awareness how they work), it is important to take care about the responsibility of the results. Arrieta et al. [3] highlighted the need of preparing and using a set of principles that should be satisfied. They called this trend as responsible AI – it should include the following issues: fairness, privacy, accountability, ethics, transparency, security and safety.

XAI algorithms used in presented papers focus on visualization approaches. It allows us to see the impact of a feature or set of features on the final result. It increases the transparency and it can help to reduce the dimensionality of the problem.

There appears more and more sophisticated algorithms that take into account more information and obtain more accurate results. Unfortunately, they do not focus on the interpretability.

Most of responsible AI aspects are still not introduced to SA methods. The users would like to have trustworthy methods for analyzing opinion mining so explainable sentiment analysis is a promising investigation area. Due to the wide variety of the SA methods, better explainable algorithms should be also created.

Table 1. Summary of the XAI methods in SA problems.

No. Paper	Main problem	Feature used	Techniques used	Type of explainability
[18]	generating opinions summary (aspect-based SA)	review content – new dataset of opinions about one entity in the restaurant domain	aspect extraction, sentences grouping, rules of interest extraction	discovering subgroup of features using statistical methods; generating the rules of classification to subgroups; developing quality measures that are easy to understand for humans
[12]	comparing effectiveness of selected NLP models	tweets	explainable FEs (EFEs); pre-trained DL FEs that do not require training on task-specific data; and trainable DL FEs that require training on task-specific data	local interpretable model-agnostic explanations (LIME), the variant called submodular pick LIME (SP-LIME); Shapley additive explanations (SHAP); BertViz, designed specifically for transformer LMs
[34]	sentiment analysis of online reviews; extracting features for product rating prediction	online reviews	knn, support vector machines, random, forests, gradient boosting machines, XGBoost	local feature attributions and partial dependency plots
[30]	model for sentence polarity for the Italian language	sentence content	BERT model, Long-Short Term Memory (LSTM) and the WMAL-based text representation module	Lexicon-driven classification explanation; contextual importance and utility; explanatory and WMAL attention
[26]	investigation of the capability of an attention mechanism to "attend to" semantically meaningful words	text from videos of Stanford Emotional Narratives Dataset (SEND) - dataset consisting of videos of people narrating emotional events in their lives	Window-Based Attention (WBA) consisting of a hierarchical, two-level long short-term memory (LSTM) with softmax attention	attention based explanation: word deletion experiments and visualizations of results
[40]	leveraging a sentiment knowledge graph to better capture the sentiment relations between aspects and sentiment terms	online learner review dataset	knowledge-enabled language representation model BERT for aspect-based sentiment analysis	knowledge-enabled BERT model delivers explainable information to boost performance
[27]	examination of interpretable HMMs methods performance under various architectures, parameters, orders and ensembles	annotated datasets of documents or movies reviews	interpretable Hidden Markov Models (HMM)-based methods for recognizing sentiments in text	visual interpretation of the HMM
[19]	attention-based multi-feature fusion method for intention recognition	movie comments	fusing features extracted from frequency-inverse document frequency (TF-IDF), convolutional neural networks (CNNs), long short term memory (LSTM)	attention mechanism for measuring feature importance
[35]	securing the reliability of machine learning-based sentiment analysis and prediction	movie reviews	multinomial naïve Bayes, random forest, random boosting, decision trees	LIME - visualization

5 Summary

In this paper we have presented the explainable artificial intelligence methods that are used in sentiment analysis. We have described definitions and an

overview of the existing methodologies used in SA. The second part focuses on explainable methods that are more and more popular in general area of artificial intelligence. And finally, we presented exemplary research articles that use XAI methods in the opinion mining.

Most of presented paper uses only visualization methods to help the user to interpret the result so it is still a potential research domain.

References

1. Alsaif, H.F., Aldosssari, H.D.: Review of stance detection for rumor verification in social media. Eng. Appl. Artif. Intell. **119**, 105801 (2023)
2. Arras, L., Montavon, G., Muller, K.R., Samek, W.: Explaining recurrent neural network predictions in sentiment analysis. In: Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, pp. 159–168 (2017)
3. Arrieta, A.B., et al.: Explainable Artificial Intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. Inf. Fusion **58**(2020), 82–115 (2020)
4. Athira, A.B., Kumar, S.D.M., Chacko, A.M.: A systematic survey on explainable AI applied to fake news detection. Eng. Appl. Artif. Intell. **122**, 106087 (2023)
5. Birjali, M., Kasri, M., Beni-Hssane, A.: A comprehensive survey on sentiment analysis: approaches, challenges and trends. Knowl.-Based Syst. **226**(2021), 107–134 (2021)
6. Chaturvedi, I., Satapathy, R., Cavallari, S., Cambria, E.: Fuzzy commonsense reasoning for multimodal sentiment analysis. Pattern Recogn. Lett. **125**(2019), 264–270 (2019)
7. Dazeley, R., Vamplew, P., Foale, C., Young, Ch., Aryal, S., Cruz, F.: Levels of explainable artificial intelligence for human-aligned conversational explanations. Artif. Intell. **299**, 103525 (2021)
8. Ding, W., Abdel-Basset, M., Hawash, H., Ali, A.M.: Explainability of artificial intelligence methods, applications and challenges: a comprehensive survey. Inf. Sci. **615**(2022), 238–292 (2022)
9. Dwivedi, Y.K., Kshetri, N., et al.: "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. Int. J. Inf. Manage. 71, 102642 (2023). https://doi.org/10.1016/j.ijinfomgt.2023.102642. ISSN 0268–4012
10. Esuli, A., Sebastiani, F.: SentiWordNet - a publicly available lexical resource for opinion mining. In: Proceedings of the 5th Conference on Language Resources and Evaluation (LREC 2006), pp. 417–422 (2006)
11. Fernandez, C., Provost, F., Han, X.: Explaining data-driven decisions made by AI systems: the counterfactual approach (2020). arXiv:2001.07417v1. Accessed 5 Mar 2023
12. Fiok, K., Karwowski, W., Gutierrez, E., Wilamowski, M.: Twitter account: comparison of model performance and explainability of predictions. Expert Syst. Appl. **186**, 115771 (2021)
13. Fuhrman, J.D., Gorre, N., Hu, Q., Li, H., El Naqa, I., Giger, M.L.: A review of explainable and interpretable AI with applications in COVID-19 imaging. Med. Phys. **49**(1), 1–14 (2022). https://aapm.onlinelibrary.wiley.com/doi/10.1002/mp.15359

14. Gilpin, L.H., Bau, D., Yuan, B.Z., Bajwa, A., Specter, M., Kagal, L.: Explaining explanations: an overview of interpretability of machine learning (2019). arXiv:1806.00069v3. Accessed 18 Mar 2023
15. Gutierrez-Batista, K., Vila, M.-A., Martin-Bautista, M.J.: Building a fuzzy sentiment dimension for multidimensional analysis in social networks. Appl. Soft Comput. **108**, 107390 (2021)
16. Hacohen, S., Medina, O., Shoval, S.: Autonomous driving: a survey of technological gaps using google scholar and web of science trend analysis. IEEE Trans. Intell. Transp. Syst. **23**(11), 21241–21258 (2022)
17. Hussein, D.M.E.D.M.: A survey on sentiment analysis challenges. J. King Saud Univ. Eng. Sci. **2018**(30), 330–338 (2018)
18. López, M., Martínez-Cámara, E., Luzón, V., Herrera, F.: ADOPS: Aspect Discovery OPinion Summarisation Methodology based on deep learning and subgroup discovery for generating explainable opinion summaries. Knowl.-Based Syst. **231**, 107455 (2021)
19. Liu, C., Xu, X.: AMFF: a new attention-based multi-feature fusion method for intention recognition. Knowl.-Based Syst. **233**, 107525 (2021)
20. Lundberg, S.M., Lee, S.-I.: A unified approach to interpreting model predictions. In: NIPS 2017: Proceedings of the 31st International Conference on Neural Information Processing Systems, pp. 4768–4777 (2017)
21. Lötsch, J., Ultsch, A.: Enhancing explainable machine learning by reconsidering initially unselected items in feature selection for classification. Biomedinformatics **2**, 701–714 (2022). https://doi.org/10.3390/biomedinformatics2040047
22. Medhat, W., Hassan, A., Korashy, H.: Sentiment analysis algorithms and applications: a survey. Ain Shams Eng. J. **5**, 1093–1113 (2014)
23. Montavon, G., Samek, W., Muller, K.R.: Methods for interpreting and understanding deep neural networks (2017). https://arxiv.org/pdf/1706.07979.pdf. Accessed 21 Mar 2023
24. Moradi, M., Samwald, M.: Post-hoc explanation of black-box classifiers using confident itemsets. Expert Syst. Appl. **165**, 113941 (2021)
25. Nabizadeh, A.H., Leal, J.P., Rafsanjani, H.N., Shah, R.R.: Learning path personalization and recommendation methods: a survey of the state-of-the-art. Expert Syst. Appl. **159**, 113596 (2020)
26. Nguyen, T.-S., Wu, Z., Ong, D.C.: Attention uncovers task-relevant semantics in emotional narrative understanding. Knowl.-Based Syst. **226**, 107162 (2021)
27. Perikos, I., Kardakis, S., Hatzilygeroudis, I.: Sentiment analysis using novel and interpretable architectures of Hidden Markov Models. Knowl.-Based Syst. **229**, 107332 (2021)
28. Phan, H.T., Nguyen, N.T., Hwang, D.: Fake news detection: a survey of graph neural network methods. Appl. Soft Comput. **139**, 110235 (2023)
29. Phan, H.T., Nguyen, N.T., Hwang, D.: Sentiment analysis for opinions on social media: a survey. J. Comput. Sci. Cybern. **37**(4), 403–428 (2021)
30. Polignano, M., Basile, V., Basile, P., Gabrieli, G., Vassallo, M., Bosco, C.: A hybrid lexicon-based and neural approach for explainable polarity detection. Inf. Process. Manage. **59**, 103058 (2022)
31. Serrano-Guerrero, J., Romero, F.P., Olivias, J.A.: Fuzzy logic applied to opinion mining: a review. Knowl.-Based Syst. **222**, 107018 (2021)
32. da Silva, M.P.: Feature Selection using SHAP: an Explainable AI approach. University of Brasilia. Doctoral thesis (2021)

33. So, Ch.: Understanding the prediction mechanism of sentiments by XAI visualization. In: 4th International Conference on Natural Language Processing and Information Retrieval, Sejong, South Korea, 18–20 December 2020. ACM (2020)
34. So, C.: What emotions make one or five stars? Understanding ratings of online product reviews by sentiment analysis and XAI. In: Degen, H., Reinerman-Jones, L. (eds.) HCII 2020. LNCS, vol. 12217, pp. 412–421. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-50334-5_28
35. Song, M.H.: A study on explainable artificial intelligence-based sentimental analysis system model. Int. J. Internet Broadcast. Commun. **14**(1), 142–151 (2022). https://doi.org/10.7236/IJIBC.2022.1.142
36. de Souza Jr., L.A., et al.: Convolutional Neural Networks for the evaluation of cancer in Barrett's esophagus: explainable AI to lighten up the black-box. Comput. Biol. Med. **135**, 104578 (2021)
37. Ventura, F., Greco, S., Apiletti, D., Cerquitelli, T.: Explaining the Deep Natural Language Processing by Mining Textual Interpretable Features (2021). https://arxiv.org/abs/2106.06697. Accessed 31 Mar 2023
38. Zacharias, J., von Zahn, M., Chen, J., Hinz, O.: Designing a feature selection method based on explainable artificial intelligence. Electron. Mark. **32**, 2159–2184 (2022). https://doi.org/10.1007/s12525-022-00608-1
39. Zhang, L., Wang, S., Liu, B.: Deep learning for sentiment analysis: a survey (2018). https://doi.org/10.1002/widm.1253. Accessed 11 Mar 2023
40. Zhao, A., Yu, Y.: Knowledge-enabled BERT for aspect-based sentiment analysis. Knowl.-Based Syst. **227**, 107220 (2021)
41. https://elula.ai/feature-importances-are-not-good-enough/. Accessed 10 Mar 2023