A Comparative Study of Explainable Artificial Intelligence Techniques for Sentiment Analysis Systems

***Abstract*—In our recent research, we have developed a cutting edge framework that combines machine learning with advanced explanatory artificial intelligence (XAI) methodologies. This innovative fusion aims to improve the dependability and transparency of sentiment analysis predictions. To validate our proposed model, we conducted experiments using the IMDB dataset and obtained compelling results that demonstrate its practicality and effectiveness. One particularly notable advantage of this approach is its ability to reveal how input data influences prediction outcomes, thereby providing valuable insights from different angles. Given the wide range of uses of sentiment analysis, including referral systems, sentiment recognition through behavioral cues, and public thought analytics, our method instills users with greater confidence in the system by delivering clear and evidence-supported analyses. Consequently, users can grasp and rely on the provided insights more effectively.**

***Index Terms*—Explainable Artificial Intelligence (XAI), Sentimental Analysis, LIME, Model Interpretability**

I. Introduction

The paper "explainable decision-making systems" by Buchanan and Shortliffe first came up with the term Explainable Artificial Intelligence (XAI) [1]. Researchers stated in 2004 that despite the increasing complexity of computer systems and AI, their ability to self-explain had not advanced [2]. This research presented a model to clarify the fundamental reasons behind the behaviors of Non-player Characters (NPCs) within a military simulation combat program. This was accomplished through the modification of their AI. The approach included dividing Command AI from Control AI, with Control AI acting as a vector to describe the in-game scenario. Command AI then evaluated the results of Behavioral AI and imposed relevant directives based on the analysis. After simulated battles, officers who interacted with NPCs analyzed the AI's actions using insights from Command AI's situational assessments and Control AI's instructions. As a result, they gained a deeper understanding of the AI's motivations [2].

In the year 2016, DARPA, a well-known institution leading advancements in science, launched a project named Explainable Artificial Intelligence (XAI) [3]. XAI is an approach that introduces the capability to provide explanations, allowing us to understand the logic behind the choices made by artificial intelligence models prior to reaching specific outcomes. This is particularly beneficial for individuals involved in machine learning, as it fosters confidence in the system. With the growing complexity of contemporary artificial intelligence, machines are processing data at a scale beyond human interpretive capabilities. Consequently, there is a necessity for deliberate endeavors and research to equip artificial intelligence with the capability to provide explanations through XAI. This guarantees rational interactions between AI systems and humans [4] [5]. Convolutional Neural Networks (CNN) and the rise of deep learning have also triggered an increased investigation into deciphering image data and modeling outcomes [6] [7]. An interpretable approach and a representative sentiment analysis model are used together to assess the model's capabilities to forecast and provide explanations. The possibility to improve the models' reliability and predictability becomes attainable through analyzing submodules, using real-world instances, and concentrating on efficient word vectorizers and models for classification.

The following is a brief overview of the installation and outline of our writing. In Chapter II, an introduction to the sentiment analysis system is presented, accompanied by a conversation about technologies linked to explainable artificial intelligence, and an exploration of the attributes of these interconnected technologies. In Chapter III, the strategy and methodology for incorporating explanatory capabilities are showcased through instances of how explainable AI is integrated into the proposed model. Ultimately, in Chapter IV, the paper wraps up by providing summaries and proposing potential avenues for future research endeavors.

II. Literature Review

In current years, the field of Explainable Artificial Intelligence has experienced a notable increase in research efforts, exploring its theories, tools, and methodologies. The study of XAI has gained significant prominence, attracting consistent scholarly attention. Among the earliest comprehensive investigations into this subject were conducted by Lacave and Diéz, who extensively examined methods of explanation specifically tailored for Bayesian networks [8]. They explored various probabilistic techniques in great detail. Building on these foundations, Ribeiro and his colleagues assessed multiple solutions aimed at improving the interpretability of AI/ML models [9]. These included additive models, decision trees, and attention-based networks, among others. Furthermore, they introduced a versatile approach that combines complex model predictions with perturbations in input data to uncover behavior patterns within these intricate systems [10].

The GDPR has had a significant impact on research contributions in recent years. Early investigations examined explainability from various perspectives. One study conducted

a bibliometric analysis on XAI to uncover research trends, identify key contributors and regions, and explore new avenues of study [11] . Another group combined traditional concepts with modern ideas like deep learning. They compared the advantages and disadvantages of transparent models versus opaque, black-box models [12]. Several review articles emerged advocating for the use of transparent models, particularly in contexts requiring critical decisions [8]. Research reviews also analyzed the methodologies behind explainability across different sectors and stakeholders [13]. Some studies focused on terminology, practical applications, and challenges in achieving accountable AI [14]. Separate investigations categorized different techniques for explainability including text-based explanations, visualizations, and numerical interpretations [15]. A substantial body of literature exists that explores specific strategies for elucidating AI/ML models. For example, Robnik-Sikonja's work scrutinized explanations based on perturbation as predictive frameworks, [16] while Zhang's team shed light on visual elucidation methods for advanced deep learning systems [17]. Additionally, Daglarli provided insights into XAI strategies tailored specifically for deep meta-learning designs [18].

Vilone and Longo assumed the responsibility of conducting comprehensive review studies, offering a detailed overview on the latest trends in XAI [19] . Their research diligently classified various XAI techniques and assessment metrics, presenting a more thorough perspective compared to other literature reviews. However, they did not explore specific sectors or tasks that benefit from advancements in XAI. While experts in domain-specific fields have examined potential opportunities and challenges, most of the existing literature predominantly revolves around the medical and healthcare sectors [20]. Nevertheless, there are also reviews available covering other domains such as industrial sectors, software development, and automotive industries [21].

In the aforementioned investigations, researchers delved into the complexities and frameworks of XAI. They explored challenges and potential remedies in various sectors, without focusing on specific application domains or tasks. Building upon this foundation, our latest exploration introduces a groundbreaking framework that seamlessly merges machine learning with state-of-the-art XAI techniques. The primary objective behind this innovation is to enhance the reliability and clarity of sentimental analysis outcomes. To validate our initiative's credibility, we conducted experiments using the IMDB dataset, vividly showcasing its practicality and effectiveness. One remarkable feature of our method is its ability to elucidate how input variables influence predictions, offering insights from multiple perspectives. Given the broad implementation of analyzing sentiment in guidance engines, emotion discernment through facial indicators, and tapping into public sentiment, our methodology empowers users with enhanced trustworthiness supported by solid evidence. This enables users to comprehend and rely on the provided insights. Furthermore, while assessing emotional analytics data and leveraging the explanatory power of our model using prominent XAI paradigms such as LIME, we acknowledge that these complexities may pose challenges for average users. Consequently, there is an emerging need for an intuitive interface that demystifies model interpretations.vAs our research progresses along this trajectory, it becomes imperative to authenticate the interpretative potency of our system through quantifiable performance benchmarks juxtaposed with real-world user satisfaction indices.

III. Methodology

LIME stands for "Local Interpretable Model-agnostic Explanations." It's an algorithm designed to provide explanations for the predictions made by classifiers or regression models. It achieves this by creating an approximation of the data within a localized context centered around the specific data point under examination. LIME is a technique within the realm of Explainable Artificial Intelligence (XAI) that reveals the specific aspects that a model focuses on when analyzing individual instances and the factors it utilizes as the foundation for making predictions. Notably, LIME is a method that is independent of the underlying model's learning approach, making it "model-agnostic." As a result, regardless of the particular learning model being employed, LIME may therefore be used to offer every AI model the capacity to provide explanations [22].

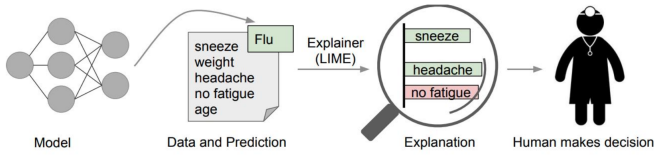

Fig. 1. Illustrative description of individual predictions

Using the LIME explainer, Figure 1 illustrates the idea of distinct cases [22]. In this example, the model provided a prediction of the patient contracting the flu. The LIME explainer presents a visual representation of the patient's medical background, highlighting the elements contributing to the flu prediction. Specifically, symptoms such as sneezing and headache were influential in predicting "flu," while there was contradictory evidence for "fatigue." This empowers medical professionals to form educated judgments regarding the reliability of the model's forecasts [22]. The equation for the model-agnostic LIME is provided below.

$$\xi\left(x_i\right)=\underset{g \in G}{\arg \min } L\left(f, g, \pi_x\right)+\Omega(g) \tag{1}$$

The model denoted as "g" operates as an interpretable model tailored for a specific instance labeled as "x". This model "g" strives to minimize a loss function "L", which is commonly quantified using Mean Squared Error (MSE). The assessment is based on how closely the forecasts generated by the interpretable model match the forecasts produced by the original model "f". This alignment is accomplished while

also ensuring that the simplicity of the model, represented as $\Omega(g)$, remains minimal. For instance, this simplicity could involve the utilization of fewer variables. You can draw a parallel between $\Omega(g)$ and the idea of regularization techniques (like Lasso, Ridge, Elastic Net) applied in linear regression models. In this analogy, heightened regularization corresponds to reducing variables to a minimum, effectively reducing the number of variables employed. On the contrary, reduced regularization corresponds to allowing a greater number of variables to have a substantial role in the model.

SHAP, an acronym for SHapley Additive exPlanations, is an explanatory technique designed to enhance the comprehensibility of a model. When applied to individual instances, it reveals the internal mechanisms and the degree to which the SHAP values, gathered for a comprehensive view, impact the final prediction in a positive or negative manner. SHAP offers various analysis techniques:

1) Individual Instance Analysis: It illustrates SHAP's functioning in specific cases.
2) Feature Relationship Analysis: This method examines how specific features relate to each other.
3) Comprehensive Instance Analysis: SHAP allows a thorough evaluation of all instances.
4) Model Influence Analysis: It analyzes the impact of all features on the model's predictions.

When incorporated within an AI application system, SHAP offers a notable benefit by facilitating in-depth explanations and evaluations of prediction results [23].

IV. Data Collection and Analysis

In the context of classifying subjects in general natural language texts and predicting the polarity of reviews, a predictive model has been employed, and its outcomes are assessed using accuracy metrics. However, these metrics alone do not elucidate the underlying pathways that led to these predictions. The fusion of the model for sentiment analysis system and Explainable Artificial Intelligence (XAI) presents a configuration that facilitates the comprehension of model predictions. This utilization of XAI technology is versatile and can be used regardless of the model's nature. Shown in Figure 2, it presents a system setup that integrates an explanatory structure adaptable to different model forms. The central objective is to pinpoint the most efficient model, delivering its predictions to analysts or decision-makers. During this process, it's possible to provide both visual and textual representations of the model's behavior, which aids in understanding how it operates.

A model is created using newsgroup data obtained from the sci-kit-learn library. Its performance is then assessed using test data. The dataset contains 20 distinct news categories. Furthermore, this narrative will meticulously outline the process of integrating a descriptive module into a sentiment analysis system. The IMDB Movie Review dataset obtained from Kaggle will be utilized as a basis for demonstration [23]. The IMDB dataset contains 25,000 film reviews categorized as either 'positive' or 'negative'.

A machine learning model employs a method of vectorization to transform text strings into a format suitable for advanced processing. This process of vectorization applies the TF-IDF (Term Frequency-Inverse Document Frequency) technique. After completing the preprocessing phase, a suitable machine learning model is utilized in the training step to classify new datasets. The naive Bayes algorithm for polynomial distribution precisely calculates the possibility of a document belonging to a particular class by analyzing the vectorized input. Subsequently, the naive Bayes model contrasts the results probabilistically, evaluating the similarity between the input news and word usage frequency obtained from the previously established model. It's noteworthy to mention that the alpha parameter is influential in this procedure. Setting a very small alpha value could result in the model overfitting the training data. Conversely, if the alpha value is chosen to be too large, the model may underfit the data. The appropriate setting of the alpha parameter is crucial for achieving balanced and accurate model performance.

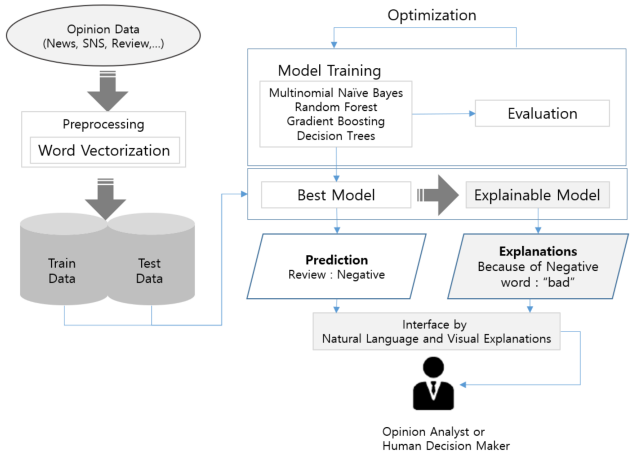

Fig. 2. Sentiment Analysis System Model based on XAI

Once the model training is completed, the performance is evaluated using testing data. The effectiveness is measured by the f1-score. The model forecasts results on the test dataset and subsequently contrasts these predictions with the real outcomes, yielding the f1-score as the output. Table 1 illustrates the effectiveness of various classifiers, excluding the naive Bayesian classifier. When employing both the TF-IDF vectorizer and the naive Bayes classifier together, an impressive f1-score of 83.5% is achieved. This outcome signifies that this particular model configuration yields the most favorable performance, outperforming other classifiers in the comparison.

V. Results and Analysis

Integrating LIME involves utilizing Python's module designed for providing explanations in text form. The LIME explainer allows for the customization of parameters such as feature selection, the bag-of-words (BOW) approach, and

Classifier	Parameter	F1-score
Multinomial Naïve Bayes	Alpha = 0.01	0.835
Random forest	n_estimators = 500	0.7687
Bagging	Base classifier = KNeighborsClassifier() n_estimators = 500 max_samples = 0.5 max_features = 0.5	0.64426
Gradient boosting	n_estimators = 500	0.79093
Decision tree		0.72314
Extra trees	n_estimators = 500	0.81830

Fig. 3. Performance Evaluation of Several Classifier

the size of the kernel. This application focuses solely on fundamental categories.

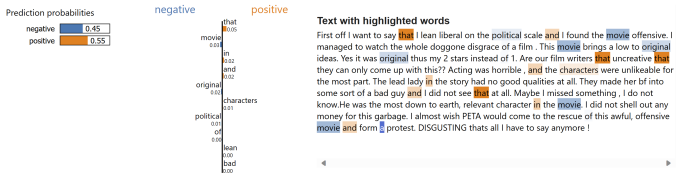

Fig. 4. Result Analysis of LIME illustration enactment for instance 40000

In Figure 4, the outcome of the LIME descriptor's sub-module output for the initial case of the test data is depicted. In the illustration, sub-modules are emphasized through textual highlighting. In this particular movie review dataset, the categorization has been allocated to the 'sentiment' class, with a probability of 50.67%.Even individuals without movie expertise can easily recognize the correlation between terms like "brilliant" and "cinematography" and their association with the "sentiment" category.

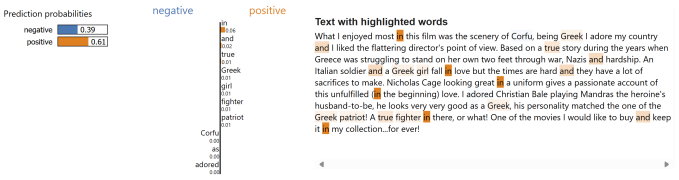

Fig. 5. Result Analysis of LIME illustration enactment for instance 45000

The successful application of LIME to textual data has been verified. LIME effectively validated the rationale behind the model's decision-making process, using submodule highlighting to emphasize the relevant components. LIME's text explainer serves the purpose of uncovering the rationale behind AI-driven decisions and providing insights for enhancing the model's performance.

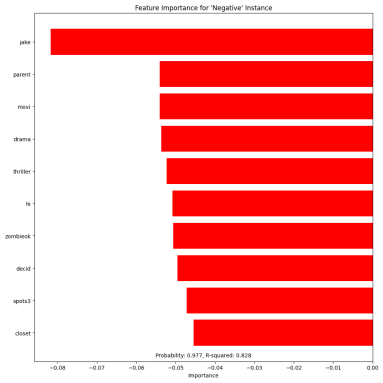

Fig. 6. "Negative" Labeled Instance's Feature significant plot

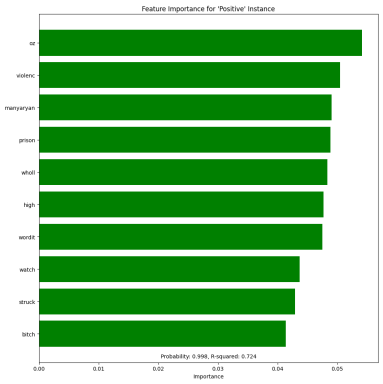

Fig. 7. "Positive" Labeled Instance's Feature significant plot

Figure 6 displays a feature importance visualization for individual instances that have been classified as 'negative'. In this plot of figure 7, the green color indicates the extent to which each feature contributes to the prediction outcome, while the red color signifies how much each feature contradicts the prediction. For instance, terms like "love" "good"," and "true" correspond to features supporting the 'positive' prediction. Conversely, the term "fall" opposes the prediction result. Additionally, Figure 7 showcases important plots for individual instances classified as 'positive'.

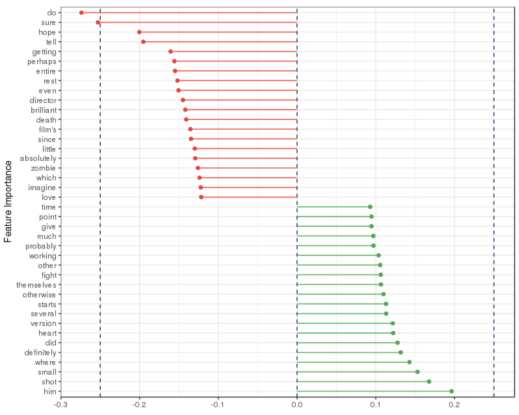

Fig. 8. Instance's Feature significant plot

Figure 8 aims to visually illustrate and explain the rank-based correlation. Its purpose is to identify attributes that show stronger associations with positive and negative reviews compared to others. Words like ”hope,” ”yes,” and ”getting” are often emphasized in comments suggesting enhancements in feedback. For example, one might find a review saying, ”I wish the movie adopts a nuanced style, elevating its overall caliber.” Conversely, critical film feedback tends to spotlight words like ”movie,” ”all,” ”rest,” and ”director,” hinting at concerns about the film’s excellence and potential areas for enhancement. A more detailed analysis is warranted for the words highlighted in green, which contribute to positive reviews. The analysis aims to uncover the reasons behind the prevalence of moderately ambiguous phrases in positive reviews, contrasting them with words that have explicit patterns or meanings. When praising a movie, the specific expressions used heavily depend on the film’s style, resulting in certain phrases dominating the compliments. Consequently, a description indicative of a positive review mirrors a transitional term. It can be inferred that expressions within reviews lauding a movie might manifest in highly specific terms connected to the movie itself.

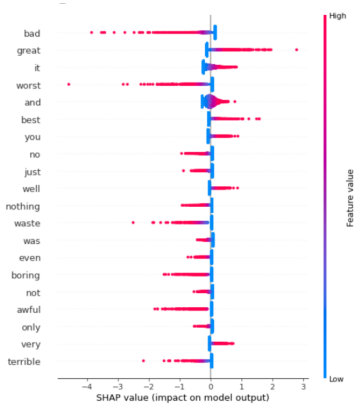

Fig. 9. (Scatter Plot) A graphical representation of the impact on the distribution of Shapley values across all attributes.

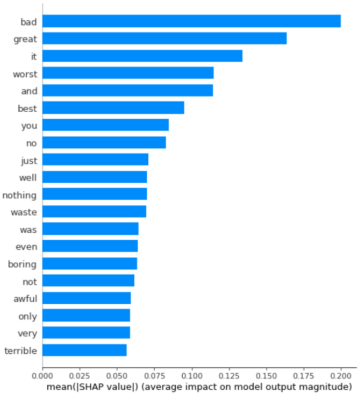

Fig. 10. (Bar Chart) A graphical representation of the impact on the distribution of Shapley values across all attributes.

The impact of variations in the top 20 token values on the Shapley value is illustrated in Figure 9 through a scatter plot graph. Observations reveal that features with higher values like ”bad”, ”great”, ”it”, and ”worst” exert a more significant influence on the model compared to those with lower values like ”only”, ”very” and ”terrible”. Additionally, the token ”bad” exhibits the highest variability, as supported by the accompanying bar graph depicted in Figure 10. The word ’bad’ emerges as the most vital determinant of review polarity (positive/negative), followed by ’great,’ ’it,’ and ’worst,’ among others, highlighting their influential role across the model.

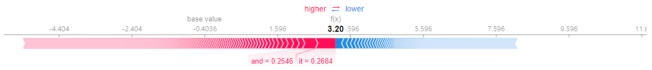

Fig. 11. Positive Reviews Instances Identified

Fig. 12. Negative Reviews Instances Identified

Figures 11 and 12, the color red plays a significant role in amplifying the impact of positive reviews while simultaneously diminishing the influence of negative comments. This feature effectively heightens the power of commendations and minimizes any detrimental effects from unfavorable feedback.

The SHAP (SHapley Additive exPlanations) Explainer provides insightful explanations for individual instances. In Figure 11, we observe a case that is predicted as a positive review. The outcome is determined by calculating the Shapley value based on the feature’s importance, as shown in Figure 10. The highest Shapley value recorded is 3.20, indicating a positive prediction for this review. Notably, the key influential features found in this text are ’and’ and ’it’, with ’and’ being particularly significant across the entire corpus as it represents a higher-level concept of text structure. A high frequency of ’and’ strongly suggests a more positive review. In contrast, Figure 12 displays negative prediction results with the most prominent Shapley value recorded at -2.80. This indicates that the corresponding review was predicted to be negative. The features that significantly influenced this prediction are ’no’ and ’why’, appearing in that sequential order.

VI. Conclusion

This research introduces a model that merges explainable artificial intelligence models. The goal is to enhance the dependability of sentiment analysis and predictions derived from machine learning. The model’s efficacy was evaluated and explained using the IMDB dataset. This approach has a significant benefit - it offers comprehensive insights into how data impacts the predictive results of the model. Across diverse domains employing sentiment analysis, like recommendation systems, emotion analysis via facial expression recognition, and opinion assessment, this method holds the potential to foster user trust. By delivering more detailed and substantiated

analytical outcomes to users, the system can establish a higher level of credibility.

This research focused on examining emotion analysis data and assessing the model's explanatory capabilities using outcomes obtained from two prominent model-agnostic xAI algorithms: LIME and SHAP. However, this methodology might not be readily comprehensible to everyday users. Consequently, there arises a necessity for further investigation aimed at creating a user-friendly explanatory interface module based on Human-Computer Interaction (HCI) principles. This module would aim to provide an interpretation of the model's results in a manner that is accessible and understandable for a wider audience.

Subsequent phases of this research will necessitate validating the interpretive efficacy by implementing the proposed sentiment analysis system. This entails the introduction of a quantifiable performance metric for the descriptor. Additionally, it's important to factor in the actual user satisfaction rating during this process.

REFERENCES

[1] E. H. Shortliffe and B. G. Buchanan, "A model of inexact reasoning in medicine," *Mathematical Biosciences*, vol. 23, no. 3, pp. 351–379, 1975. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0025556475900474

[2] M. van Lent, W. Fisher, and M. Mancuso, "An explainable artificial intelligence system for small-unit tactical behavior," in *AAAI Conference on Artificial Intelligence*, 2004. [Online]. Available: https://api.semanticscholar.org/CorpusID:7286175

[3] D. Gunning, M. Stefik, J. Choi, T. Miller, S. Stumpf, and G.-Z. Yang, "Xai—explainable artificial intelligence," *Science Robotics*, vol. 4, no. 37, p. eaay7120, 2019. [Online]. Available: https://www.science.org/doi/abs/10.1126/scirobotics.aay7120

[4] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal, "Explaining explanations: An overview of interpretability of machine learning," in *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*, 2018, pp. 80–89.

[5] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," *ArXiv*, vol. abs/1311.2901, 2013. [Online]. Available: https://api.semanticscholar.org/CorpusID:3960646

[6] C. Gan, N. Wang, Y. Yang, D.-Y. Yeung, and A. G. Hauptmann, "Devnet: A deep event network for multimedia event detection and evidence recounting," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 2568–2577.

[7] S. M. Mathews, "Explainable artificial intelligence applications in nlp, biomedical, and malware classification: A literature review," *Advances in Intelligent Systems and Computing*, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:199011996

[8] C. Lacave and F. J. Díez, "A review of explanation methods for bayesian networks," *The Knowledge Engineering Review*, vol. 17, no. 2, pp. 107–127, 2002.

[9] M. T. Ribeiro, S. Singh, and C. Guestrin, "Model-agnostic interpretability of machine learning," 2016.

[10] ——, "" 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*, 2016, pp. 1135–1144.

[11] J. M. Alonso, C. Castiello, and C. Mencar, "A bibliometric analysis of the explainable artificial intelligence research field," in *International conference on information processing and management of uncertainty in knowledge-based systems*. Springer, 2018, pp. 3–15.

[12] O. Loyola-Gonzalez, "Black-box vs. white-box: Understanding their advantages and weaknesses from a practical point of view," *IEEE access*, vol. 7, pp. 154 096–154 113, 2019.

[13] F. K. Došilović, M. Brčić, and N. Hlupić, "Explainable artificial intelligence: A survey," in *2018 41st International convention on information and communication technology, electronics and microelectronics (MIPRO)*. IEEE, 2018, pp. 0210–0215.

[14] F. Xu, H. Uszkoreit, Y. Du, W. Fan, D. Zhao, and J. Zhu, "Explainable ai: A brief survey on history, research areas, approaches and challenges," in *Natural Language Processing and Chinese Computing: 8th CCF International Conference, NLPCC 2019, Dunhuang, China, October 9–14, 2019, Proceedings, Part II 8*. Springer, 2019, pp. 563–574.

[15] A. Adadi and M. Berrada, "Peeking inside the black-box: a survey on explainable artificial intelligence (xai)," *IEEE access*, vol. 6, pp. 52 138–52 160, 2018.

[16] M. Robnik-Šikonja and M. Bohanec, "Perturbation-based explanations of prediction models," *Human and Machine Learning: Visible, Explainable, Trustworthy and Transparent*, pp. 159–175, 2018.

[17] Q. Zhang, Y. N. Wu, and S.-C. Zhu, "Interpretable convolutional neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8827–8836.

[18] E. Dağlarli, "Explainable artificial intelligence (xai) approaches and deep meta-learning models," *Advances and applications in deep learning*, vol. 79, 2020.

[19] G. Vilone and L. Longo, "Explainable artificial intelligence: a systematic review," *arXiv preprint arXiv:2006.00093*, 2020.

[20] A. Holzinger, G. Langs, H. Denk, K. Zatloukal, and H. Müller, "Causability and explainability of artificial intelligence in medicine," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 9, no. 4, p. e1312, 2019.

[21] K. Gade, S. C. Geyik, K. Kenthapadi, V. Mithal, and A. Taly, "Explainable ai in industry," in *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, 2019, pp. 3203–3204.

[22] D. Dave, H. Naik, S. Singhal, and P. Patel, "Explainable ai meets healthcare: A study on heart disease dataset," *ArXiv*, vol. abs/2011.03195, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:226277988

[23] A. Ghosh, "Sentiment analysis of imdb movie reviews : A comparative study on performance of hyperparameter-tuned classification algorithms," in *2022 8th International Conference on Advanced Computing and Communication Systems (ICACCS)*, vol. 1, 2022, pp. 289–294.