

Use of explainable AI to interpret the results of NLP models for sentimental analysis

Vijaykumar Bidve¹, Pathan Mohd Shafi², Pakiriswamy Sarasu³, Aruna Pavate¹, Ashfaq Shaikh⁴,
Santosh Borde⁵, Veer Bhadra Pratap Singh¹, Rahul Raut¹

¹School of Computer Science and Information Technology, Symbiosis Skills and Professional University, Pune, India

²MIT Art, Design and Technology University, Loni Kalbhor, Pune, India

³Department of Computer Science and Engineering, Kalasalingam Academy of Research and Education, Tamil Nadu, India

⁴Department of Information Technology, M. H. Saboo Siddik College of Engineering, Mumbai, India

⁵Student Progression and Industry Relations Office, JSPM TSSM Group of Institutes, Pune, India

Article Info

Article history:

Received Dec 13, 2023

Revised Jan 16, 2024

Accepted Mar 23, 2024

Keywords:

Artificial intelligence

Explainability

Machine learning

Natural language processing

Sentimental analysis

ABSTRACT

The use of artificial intelligence (AI) systems is significantly increased in the past few years. AI system is expected to provide accurate predictions and it is also crucial that the decisions made by the AI systems are humanly interpretable i.e. anyone must be able to understand and comprehend the results produced by the AI system. AI systems are being implemented even for simple decision support and are easily accessible to the common man on the tip of their fingers. The increase in usage of AI has come with its own limitation, i.e. its interpretability. This work contributes towards the use of explainability methods such as local interpretable model-agnostic explanations (LIME) to interpret the results of various black box models. The conclusion is that, the bidirectional long short-term memory (LSTM) model is superior for sentiment analysis. The operations of a random forest classifier, a black box model, using explainable artificial intelligence (XAI) techniques like LIME is used in this work. The features used by the random forest model for classification are not entirely correct. The use of LIME made this possible. The proposed model can be used to enhance performance, which raises the trustworthiness and legitimacy of AI systems.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Vijaykumamr Bidve

School of Computer Science and Information Technology, Symbiosis Skills and Professional University

Pune, India

Email: vijay.bidve@gmail.com

1. INTRODUCTION

The use and need for machine learning (ML) models especially explainable artificial intelligence (XAI) deals with development and implementation of artificial intelligence systems and machine learning models that can provide understandable and interpretable justification for predictions and decisions [1]-[3]. The goal of XAI is to fill the gap between the black-box nature of many AI algorithms and the need for transparency, accountability, and trust in AI systems, particularly in critical domains like healthcare, finance, and autonomous vehicles [4], [5].

Some key features of explainable AI: interpretability: XAI models are more interpretable by humans. The inner workings of the AI model are understandable and transparent, allowing users to comprehend why a particular decision or prediction was made. Transparency: Transparent AI models provide clear and comprehensible information about how they arrive at their conclusions. This involves exposing the model's architecture, data used, and the weights or features that had the most influence on a decision.

Contextual Explanation: XAI systems consider the context of a specific decision or prediction. They explain not only what the model decided but also why it made that decision within the given context. User-Friendly Explanations: XAI provide explanations in a manner that is understandable to the end-users, who may be new to the AI or ML systems. This often involves using natural language explanations, visualizations, or other intuitive methods. Local vs. Global Explanations: XAI can provide explanations on both a local and global level. Local explanations deal with model's decision for a single data point, where global explanations deal with model behaves across the entire dataset. Model Agnostic Approaches: XAI is not tied to specific machine learning algorithms. They are applicable to a wide range of models, including deep neural networks, decision trees, random forests, and more. Regulatory Compliance: In many industries such as healthcare and finance AI systems used to provide explanations for their predictions. XAI helps companies to comply with these regulations while maintaining model performance. Bias and Fairness: XAI is also used for detecting and removing bias in AI models by explaining the factors that contribute to biased decisions. This helps in making AI systems fairer and more equitable. Debugging and Improvement: Explainable AI aid in debugging and improving AI models. By understanding how models work and where they make errors, developers can refine their models more effectively. Human-AI Collaboration: XAI is essential in facilitating collaboration between humans and AI systems. When humans can understand and trust AI, they are more likely to work alongside it in various tasks [6]-[8].

There are various techniques and methods for achieving explainability in AI, including feature importance analysis, model-agnostic interpretability tools (e.g., LIME), rule-based systems, attention mechanisms, and more. The selection of technique depends on the application, and the level of explainability required [9], [10]. Explainable AI is a rapidly evolving field, and ongoing research continues to advance our understanding of making AI systems more transparent, interpretable, and accountable for their decisions [11].

The term Black Box describes a system that can be defined in terms of its input and output without knowing the internal workings of the system, most like the machine learning models that are used on regular basis [12]. AI systems provide a solution without providing the reason for the solution. This black-box nature of the solution causes significant distrust among the everyday user and even developers of that system [13].

It is crucial to know why a particular model makes a specific decision in order to retain its reliability and to further increase the strength of the model. Here explainability comes into the picture. Model Explainability facilitates the debugging process, bias detection, and increasing trust toward the results of the AI system [14].

With the increasing significance of artificial intelligence in the day-to-day world, it is crucial that the trust between the system and its user also increases. One way this is possible is if the user understands the output produced by the AI system i.e. the user can interpret how and why a particular decision was made by the AI system [15]. This can be done by the use of explainable AI systems. The significance of this research is to understand the working of the explainable AI methods and to understand the research gaps that are present in this field of study [16]. The rest paper is organized in three main sections including method, result and discussion, and conclusion as given in subsequent part of this paper.

2. METHOD

2.1. Literature review

This section provides a detailed literature review of the recent work available in the domain of the explainable artificial intelligence. The purpose of the literature review is to know the depth of work form the literature in the domain of the explainable artificial intelligence. The literature review focused on methodologies used and limitations of existing work.

Schwalbe and Finzel [17] described that many terminologies have been developed in the field of explainable artificial intelligence (XAI). As the XAI methods are growing, taxonomy methods are also needed by the practitioners to select right method for any specific context. Multiple varying taxonomies are found in the literature which have different focus and overlapping also on some points. Authors claimed to present complete taxonomy of methods related to XAI. The survey about foundation for context sensitive research is provided by the authors which is useful for researchers and practitioners.

Amjad *et al.* [18] stated that importance of AI systems for decision support is increasing day by day. Use of black box approach in AI has been raised, so it is important to understand how AI systems are involved in decision making. In clinical decision support systems, natural language processing has been used to extract information from textual data. The concept of 'Explainability' is more important in the decision making as compared to black box approach. The authors mainly focused the area of medical databases to find use of Explainability components. Authors concluded that, the attention mechanism is the most dominant approach for Explainability. The attention mechanism used by different researchers is different. The graphing and hybrid techniques are emerging trends of Explainability.

Buijsman [19] explained that explainable artificial intelligence (XAI) is useful to understand black box algorithms. Explanation includes characteristics of counterfactual cases. In artificial intelligence the generalization with more features leads to wide scope and better accuracy. This definition of XAI helps to identify characteristics of a good explainable AI. Authors provided a clear definition of explanation and explanatory depth in the context of explainable artificial intelligence (XAI). The manipulation definition of explanation from the philosophy of science, which holds that an explanation consists of a generalization that reveals what happens in counterfactual instances, provides good solutions to these problems. The account maintains that a generalization with more abstract variables, a bigger scope, and/or greater accuracy is preferable in terms of explanatory depth. The author aims to help define what a decent explanation for AI is by applying these concepts and contrasting them with alternative definitions in the XAI field.

Danilevsky *et al.* [20] described that there are advances happening in explainable models but those are leading to less interpretable models. The authors discussed the main categorizations of explanations and ways to visualize it. The operations and Explainability techniques to generate explanations for natural language processing (NLP) model predictions are also discussed by the authors. White box models, such as rule-based models and decision trees, are still used but are less frequently presented as interpretable or explicable, and as a result, they are not the main force behind the field's current trajectory. These techniques are used as resources for community model development. The gaps in the current system are identified to fix direction of the future work.

Saranya and Subhashini [21] discussed that the AI simulates human intelligence to solve real life problems. Machine learning and deep learning algorithms can be used to predict the outcome more accurately without human intervention. Explainable artificial intelligence models provide explanation for the decisions and predictions. XAI strive to increase transparency, reliability, and accountability of various public systems.

Balkir *et al.* [22] stated explainable artificial intelligence (XAI) methods are often used to detect, measure, and mitigate bias in machine learning models. The exact methods to reduce biases are not specified in details. The authors focused to identify current practices in which explainability methods are applied to find biases. The trend of explainability and fairness in natural language processing (NLP) research also reviewed by the authors. There are many challenges while applying explainability to increase the fairness of NLP models.

Neely *et al.* [23] discussed the concept of "attention as explanation" in natural language processing. The attention-based explanation is not much corresponding to the techniques of feature attribution. The transformer-based model is not having significant correlation with any of the theoretical method. The authors contend that the evaluation of attention-based explanations should no longer be based on rank correlation. Testing different explanation techniques and having human intervention to ascertain whether the explanations are in line with human insight for the specific use case at hand is more important.

Saeed and Omlin [24] stated that the artificial intelligence (AI) has advanced significantly in the last decade. As a result, algorithms were used to solve a wide range of issues. This achievement has come with the cost of using opaque, black-box AI models and increased complexity of the models. Explainable AI (XAI) has emerged as a solution to this demand, with the goal of increasing AI transparency and accelerating its adoption in crucial sectors. Authors divided methodical meta-survey into two themes regarding XAI's problems and potential future research areas. It is mainly based on the broad XAI research challenges and directions in the form of design, development, and deployment phases of the machine learning life cycle.

Arrieta *et al.* [25] provided a comprehensive overview of the field of explainable AI (XAI). The authors reviewed the XAI literature and discussed taxonomy of recent contributions pertaining to the explainability of various Machine Learning models. The models those aiming at Deep Learning techniques for which a second taxonomy is established. This existing study is used to provide context for a number of difficulties that XAI faces. The impasse between data fusion and explainability is also discussed. The authors predictions point to the idea of "Responsible Artificial Intelligence" which is a paradigm for the extensive application of AI methods in actual businesses that has impartiality, model explainability, and accountability.

Chi and Liao [26] proposed a system which uses quantitative argumentation to detect fake news on social media. The system is designed to be automated and explainable. It means that it can provide a clear explanation of how it arrived at its decision. The authors used a dataset of fake news articles and real news articles to train the system. The system uses a set of features to represent each article, such as the number of words, the number of sentences, and the number of named entities. The system exhibits superior interpretability and transparency when compared to machine learning techniques, and it may leverage data knowledge more effectively than other argumentation-based approaches. The technique of machine learning algorithms is based on quantitative reasoning can produce competitive or greater performance compared to another pure machine learning. The explanation model offers a means of refining the algorithms when certain issues are found. The authors constructed an argumentation graph, which is used to evaluate the credibility of each article.

Islam *et al.* [27] described a systematic literature review (SLR) on the recent developments of explainable artificial intelligence (XAI) methods and evaluation. The review found that visual explanations are more appealing to end users. Strong evaluation metrics are being developed to critic the quality of explanations, and XAI methods are mainly established for safety-critical domains. Care need to be taken to produce explanations for common users from delicate fields like banking and the legal system.

Table 1 summarises the reviewed literature, the title of the paper, author, methodology, and remarks are the attributes of the table. The methodology focuses on the key techniques and/or algorithms used in each of the paper. The remarks describe the advantage/disadvantages of the work presented in each of the paper. The summary of literature review is used to find the gap in the existing work.

Table 1. Summary of literature review

Sr. No.	Title	Author	Methodology	Limitation
1	A comprehensive taxonomy for explainable artificial intelligence: a systematic survey of surveys on methods and concepts	Schwalbe and Finzel [17]	Context sensitive research.	Main focus on literature review.
2	Attention-based explainability approaches in healthcare natural language processing	Amjad <i>et al.</i> [18]	Attention mechanism, natural language processing (NLP) to extract information.	Attention mechanism is not discussed in details.
3	Defining explanation and explanatory depth in XAI	Buijsman [19]	Artificial intelligence with generalization.	Implementation of generalization is not specified.
4	A survey of the state of explainable AI for natural language processing	Danilevsky <i>et al.</i> [20]	Explanations with NLP models.	NLP model description not explained in details.
5	A systematic review of explainable artificial intelligence models and applications: recent developments and future trends	Saranya and Subhashini [21]	Machine learning and deep learning.	algorithm description is missing.
6	Challenges in applying explainability methods to improve the fairness of NLP models	Balkir <i>et al.</i> [22]	Explainability methods like LIME and SHAP.	No implementation-based work.
7	A song of (Dis)agreement: evaluating the evaluation of explainable artificial intelligence in natural language processing	Neely <i>et al.</i> [23]	DeepSHAP and GradSHAP, LIME.	Uncertainty against the assumption
8	Explainable AI (XAI): a systematic meta-survey of current challenges and future opportunities	Saeed and Omlin [24]	Black-box AI model.	Main focus on literature review.
9	Explainable artificial Intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI	Arrieta <i>et al.</i> [25]	Machine learning and deep learning.	Algorithm description is missing.
10	A quantitative argumentation-based Automated eXplainable decision system for fake news detection on social media	Chi and Liao [26]	Machine learning.	Algorithm description is missing.
11	A systematic review of explainable artificial intelligence in terms of different application domains and tasks	Islam <i>et al.</i> [27]	Visual explanations, evaluation metrics.	Main focus on literature review.

The summary of literature survey described in the Table 1 leads to the conclusion that, most of the research performed in the domain of explainable artificial intelligence XAI is survey based and a smaller number of articles are with full proof implementation. The dataset used for research works is not significant and it is possible that the data used is noisy and biased. Sufficient work with complete implementation is not available in the domain of explainable artificial intelligence and hence it is not widely used even though it has been many years since its introduction.

2.2. Working

The Figure 1 shows the proposed architecture of the system developed by the authors. The architecture mainly includes following steps. The architecture is mainly divided in three parts. The first part includes data collection and preprocessing. Second part deals with NLP techniques and machine learning models. The third and last part deals with model accuracy, LIME and results.

2.2.1. Data collection

This step is responsible to choose appropriate dataset and searching for viable source to find the dataset. This work has used IMDB dataset. There are multiple steps one must consider before performing data collection such as, Type of data required: In this research project for sentimental analysis textual data is required. Size of data required: A suitable amount of data is required for the training and testing process i.e. somewhere between 40,000 and 70,000 records to perform sentimental analysis. The dataset needs 80% records for training and 20% records testing set i.e. around 40,000 records in training set and 10,000 records in the testing set. Reliability of data: Irrelevant data can hinder the analytical process of the research project hence it is required to understand the data before using it. This project is checking reliability of data before using it. The IMDB dataset is standard as well as reliable.

2.2.2. Data pre-processing

This step involves preparing the dataset to be passed on machine learning models for the analytical process. This step involves multiple data pre-processing techniques are viz Tokenization, removing stop-words, removing bold texts, removing special characters, stemming, normalizing training and testing set etc. Pre-processing phase generates output for the main processing phase.

2.2.3. NLP techniques

NLP techniques are used to modify textual data into algorithm-suitable data. The steps involved in this process are bag of words, term frequency-inverse document frequency (TF-IDF), Pad Sequencing. These techniques are used to process the input data to produce the results.

2.2.4. Machine learning models

Machine learning techniques are used to perform classification process over the dataset. The main techniques of machine learning used are Logistic regression, Random Forest classification, Bidirectional-Long Short-Term Memory. The machine learning models mainly deals with pre-processed data to produce the results in the form of predictions.

2.2.5. Comparing model accuracy, LIME and results

This step checks model accuracy of each machine learning model to determine which model performs best on the available dataset. The LIME procedures are used in this step to interpret and understand the predictions of a machine-learning models. This step displays the results in the form of findings of the LIME procedure.

The proposed work uses three NLP techniques including bag of words, Term Frequency-Inverse Document Frequency, and Pad Sequencing. The use and significance of these algorithms is explained in this section. Also, the proposed work uses three machine learning algorithms including random forest classifier, logistic regression model, bidirectional LSTM. Working of these machine learning algorithms is also explained in this section.

- Bag of words: The machine learning models cannot work directly on raw text data. The text data must be converted into machine readable format i.e. numbers. This technique is called feature extraction. bag of words (BOW) is one of the techniques of the feature extraction. Bag of words is a representation of occurrence of words in a document. The bag of words includes two things viz vocabulary of words, occurrence of words. The information of structure and order of the words is not maintained in the bag of words. This model is only concerned about known words in the document, not about the place of words in the documents.
- Term frequency-inverse document frequency: It is technique to quantify words in a set of documents. This technique intended to show importance of a word in a document. The term frequency measures occurrences of word in a document. The term frequency is different for each word and document, it can be calculated as given in equation 1. The terminologies used to calculate term frequency are, t - words, d - documents, N - count of corpus, Corpus - total document set.

$$TF(t, d) = \text{Frequency of } t \text{ in } d / \text{Total words in } d \quad (1)$$

Document frequency measures importance of a document in a set of corpuses. Term frequency counts t i.e. terms (words) from the document whereas document frequency counts the frequency of the term t in a corpus N . Both the terms are measured and count is maintained separately. These term frequencies are used to understand context of a document.

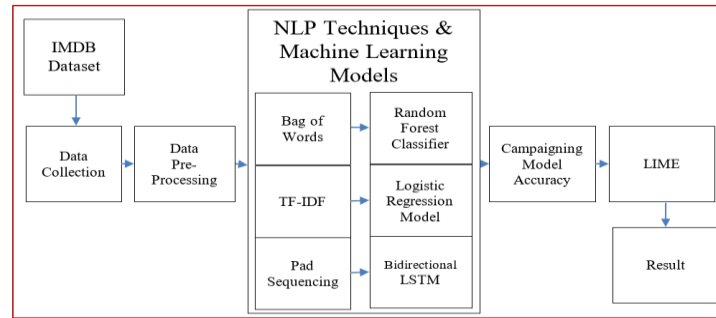


Figure 1. System architecture

- Pad Sequencing
Pad sequencing ensures that all the sequences are of the same length. To make all the sequences of the same length padding 0 is done before each sequence till all the sequences become of the same length. The sequence length should of the longest sequence after padding zeros to minimum length sequences.
- Random Forest Classifier
To create a decision tree subset of datapoint and subset of features is selected. In the set of k records n record and m features are randomly selected. The decision trees are for each data sample individually. Every individual decision tree will create an output with respect to sample data. The final output of the decision tree is based on majority voting or classification.
- If the features of the random forest algorithm are discussed, each individual tree is different with its own attributes and features. Every decision tree does not represent all features hence the feature space is reduced. Every tree is created independently with distinct data and attributes hence CPU parallelism can be used to build random forests. In random forest there is no need to separate data for training and testing. The stability of the result is ensured through majority voting and averaging.
- Logistic Regression
It predicts a probability that the object belongs to a certain class or not. The logistic regression involves following terminologies. Independent variable: Input variable used to find value of a dependent variable. Dependent variable: The output or target variable whose value is to be calculated. Logistic function: This is a relation between independent and dependent variable. Coefficient: A factors shows how independent and dependent variable relates each other. Maximum likelihood estimation: The method used to estimate coefficient of logistic regression.
- Bidirectional LSTM
It is based on the concept of recurrent neural network (RNN). RNN is used for natural language processing and speech Recognition. RNN keep record of the sequence of data and data pattern to make predictions. The feedback loop is the key component of RNN which makes it different from rest neural network systems. The feedback loops share data to nodes and predictions are made accordingly based on gathered information. In this work the Bidirectional LSTM model is trained over 12 epochs and it is observed that with each epoch being trained the accuracy of Bidirectional LSTM increases whereas the loss decreases as shown in Figure 2.

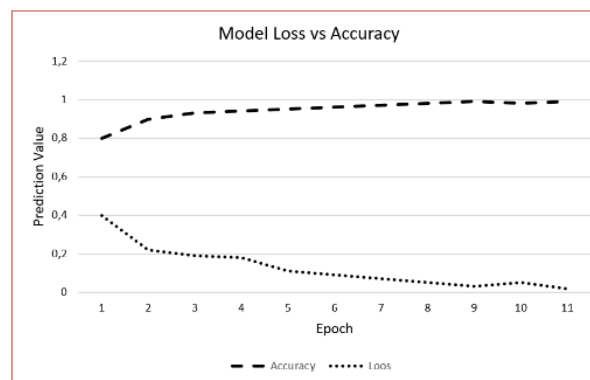


Figure 2. Bidirectional LSTM loss vs accuracy

3. RESULTS AND DISCUSSION

Local interpretable model-agnostic explanations (LIME): It approximates Blackbox model to interpretable model to explain the predictions. The data is tested with Blackbox model to observe the output. The output is used as a model for sample weights with certain variations. The original data is correlated with newly trained explanation model.

The proposed model is tested for different types of texts including positive and negative. The sentiments are analysed and percentage accuracy with different algorithms is calculated. This work is tested for the accuracy with different models, as per the results obtained the accuracy of random forest classifier is 86.97%, the accuracy for logistic regression with bag of words model is 75.12 whereas for TF-IDF model is 75.

4. CONCLUSION

This work shows that the bidirectional LSTM algorithm has more accuracy for sentiment analysis. The features used by random forest model for classification are not much accurate. The XAI technique such as LIME demonstrates the working of a black box model such as random forest classifier. The use of LIME increases accuracy of sentiment analysis. Thus, the explainability helps in better understanding a model and it can be used to improve the model performance which in turn increases the reliability and credibility of the AI systems. In future the explainability techniques such as SHAP can be implemented over various machine learning models.

ACKNOWLEDGEMENTS

The authors acknowledge to all who supported directly or indirectly in the accomplishment of this work. We are grateful to our parents for their blessing and family members, friends and colleagues for their best wishes for this work. We extend our special thanks to our employers for their huge support all the time in all of research work.





REFERENCES

- [1] G. Vilone and L. Longo, "Classification of explainable artificial intelligence methods through their output formats," *Machine Learning and Knowledge Extraction*, vol. 3, no. 3, pp. 615–661, Aug. 2021, doi: 10.3390/make3030032.
- [2] K. Fiok, F. V. Farahani, W. Karwowski, and T. Ahram, "Explainable artificial intelligence for education and training," *The Journal of Defense Modeling and Simulation: Applications, Methodology, Technology*, vol. 19, no. 2, pp. 133–144, Apr. 2022, doi: 10.1177/15485129211028651.
- [3] M. Neely, S. F. Schouten, M. J. R. Bleeker, and A. Lucic, "Order in the court: explainable AI methods prone to disagreement," arXiv preprint arXiv, 2021. doi: 10.48550/arXiv.2105.03287.
- [4] W. Samek and K.-R. Müller, "Towards explainable artificial intelligence," in *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, 2019, pp. 5–22. doi: 10.1007/978-3-030-28954-6_1.
- [5] D. Cirqueira *et al.*, "Explainable sentiment analysis application for social media crisis management in retail," in *Proceedings of the 4th International Conference on Computer-Human Interaction Research and Applications*, SCITEPRESS - Science and Technology Publications, 2020, pp. 319–328. doi: 10.5220/0010215303190328.
- [6] D. Minh, H. X. Wang, Y. F. Li, and T. N. Nguyen, "Explainable artificial intelligence: a comprehensive review," *Artificial Intelligence Review*, vol. 55, no. 5, pp. 3503–3568, Jun. 2022, doi: 10.1007/s10462-021-10088-y.
- [7] H. Liu, Q. Yin, and W. Y. Wang, "Towards explainable NLP: a generative explanation framework for text classification," in *ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, 2020, pp. 5570–5581.
- [8] S. Gite, H. Khatavkar, K. Kotecha, S. Srivastava, P. Maheshwari, and N. Pandey, "Explainable stock prices prediction from financial news articles using sentiment analysis," *PeerJ Computer Science*, vol. 7, p. E340, Jan. 2021, doi: 10.7717/peerj-cs.340.
- [9] M. Oussalah, "AI explainability. a bridge between machine vision and natural language processing," 2021, pp. 257–273. doi: 10.1007/978-3-030-68796-0_19.
- [10] J. El Zini and M. Awad, "On the explainability of natural language processing deep models," *ACM Computing Surveys*, vol. 55, no. 5, pp. 1–31, May 2023, doi: 10.1145/3529755.
- [11] M.-H. Song, "A study on explainable artificial intelligence-based sentimental analysis system model," *International Journal of Internet, Broadcasting and Communication*, vol. 14, no. 1, pp. 142–151, 2022, doi: 10.7236/IJIBC.2022.1.142.
- [12] S. M. Mathews, "Explainable artificial intelligence applications in NLP, biomedical, and malware classification: a literature review," in *Advances in Intelligent Systems and Computing*, 2019, pp. 1269–1292. doi: 10.1007/978-3-030-22868-2_90.
- [13] F. Bodria, A. Panisson, A. Perotti, and S. Piaggese, "Explainability methods for natural language processing: applications to sentiment analysis," in *CEUR Workshop Proceedings*, 2020, pp. 100–107.
- [14] L. Bacco, A. Cimino, F. Dell'Orletta, and M. Merone, "Extractive summarization for explainable sentiment analysis using transformers," in *CEUR Workshop Proceedings*, 2021, pp. 62–73.
- [15] A. K. Mohankumar, P. Nema, S. Narasimhan, M. M. Khapra, B. V. Srinivasan, and B. Ravindran, "Towards transparent and explainable attention models," in *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 4206–4216. doi: 10.18653/v1/2020.acl-main.387.
- [16] N. Burkart and M. F. Huber, "A survey on the explainability of supervised machine learning," *Journal of Artificial Intelligence Research*, vol. 70, pp. 245–317, Jan. 2021, doi: 10.1613/jair.1.12228.





- [17] G. Schwalbe and B. Finzel, "A comprehensive taxonomy for explainable artificial intelligence: a systematic survey of surveys on methods and concepts," *Data Mining and Knowledge Discovery*, Jan. 2023, doi: 10.1007/s10618-022-00867-8.
- [18] H. Amjad *et al.*, "Attention-based explainability approaches in healthcare natural language processing," in *Proceedings of the 16th International Joint Conference on Biomedical Engineering Systems and Technologies*, SCITEPRESS - Science and Technology Publications, 2023, pp. 689–696. doi: 10.5220/0011927300003414.
- [19] S. Buijsman, "Defining explanation and explanatory depth in XAI," *Minds and Machines*, vol. 32, no. 3, pp. 563–584, Sep. 2022, doi: 10.1007/s11023-022-09607-9.
- [20] M. Danilevsky, K. Qian, R. Aharonov, Y. Katsis, B. Kawas, and P. Sen, "A Survey of the state of explainable AI for natural language processing," 2020. doi: 10.48550/arXiv.2010.00711.
- [21] A. Saranya and R. Subhashini, "A systematic review of explainable artificial intelligence models and applications: recent developments and future trends," *Decision Analytics Journal*, vol. 7, 2023, doi: 10.1016/j.dajour.2023.100230.
- [22] E. Balkır, S. Kiritchenko, I. Nejadgholi, and K. C. Fraser, "Challenges in applying explainability methods to improve the fairness of NLP models," in *TrustNLP 2022 - 2nd Workshop on Trustworthy Natural Language Processing, Proceedings of the Workshop*, 2022, pp. 80–92. doi: 10.18653/v1/2022.trustnlp-1.8.
- [23] M. Neely, S. F. Schouten, M. Bleeker, and A. Lucic, "A song of (Dis)agreement: evaluating the evaluation of explainable artificial intelligence in natural language processing," 2022. doi: 10.3233/FAIA220190.
- [24] W. Saeed and C. Omlin, "Explainable AI (XAI): a systematic meta-survey of current challenges and future opportunities," *Knowledge-Based Systems*, vol. 263, p. 110273, Mar. 2023, doi: 10.1016/j.knosys.2023.110273.
- [25] A. B. Arrieta *et al.*, "Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI," *Information Fusion*, vol. 58, pp. 82–115, Jun. 2020, doi: 10.1016/j.inffus.2019.12.012.
- [26] H. Chi and B. Liao, "A quantitative argumentation-based automated eXplainable decision system for fake news detection on social media," *Knowledge-Based Systems*, vol. 242, p. 108378, Apr. 2022, doi: 10.1016/j.knosys.2022.108378.
- [27] M. R. Islam, M. U. Ahmed, S. Barua, and S. Begum, "A systematic review of explainable artificial intelligence in terms of different application domains and tasks," *Applied Sciences*, vol. 12, no. 3, p. 1353, Jan. 2022, doi: 10.3390/app12031353.

BIOGRAPHIES OF AUTHORS







Dr. Vijaykumar Bidve     is Associate Professor at School of CSIT, Symbiosis Skills and Professional University, Kiwale, Pune, Maharashtra, India. He Holds a Ph.D. degree in Computer Science and Engineering with specialization in Software Engineering. His research areas are Software Engineering, Machine Learning, Cyber Security. Dr Vijaykumar has published six patents. He has published more than 40 research articles in national and international journals. He is a life member of ISTE. He is working as an expert for various subjects. Also, he has worked as a reviewer for various conferences and journals. He can be contacted at email: vijay.bidve@gmail.com.






Dr. Pathan Mohd Shafi     is Professor at MIT Art, Design and Technology University, Loni Kalbhor, Pune, India. He Holds a Ph.D. degree in Computer Science & Engineering with specialization in public key cryptography for cross-realm authentication in Kerberos. He has worked as the resource person for workshops and seminars. He has published more than 55 research articles in national and international journals. He is a life member of ISTE and CSI. His research areas are cryptography and machine learning. Dr Shafi has published number of patents. He is working as an expert for various subjects. Also, he has worked as a reviewer for various conferences and journals. He can be contacted at email: shafipathan@gmail.com.






Dr. Pakiriswamy Sarasu     is Director International and Industry relations, Professor in Computer Science Engineering Dept. at Kalasalingam Academy of Research and Education, Tamilnadu, India. She holds a Ph.D. Degree in Computer Science and Engg. She did her masters in Embedded Systems Technology and Bachelors in Computer Science engineering. Her research areas include chaotic systems, cryptography and autonomous vehicle. She played a major role in creating innovation entrepreneurial ecosystem and more than 120 startups are created under her guidance and mentorship. She is continuously involved in mentoring students and faculty members for innovation and entrepreneurship. One patent is granted for her as one of the inventors. She can be contacted at email: sarasujivat@gmail.com.






Dr. Aruna Pavate    is working as Assistant Professor at School of CSIT, Symbiosis Skills and professional University, Pune, India. Her research interests include Machine learning and security, data mining, data science, and cyber security. She is a member of various professional bodies including ISTE, IAENG, AICTSD, Insc and IEEE and editor for ASMS, IIP. She has worked as program chair for many the conferences and for journals such as JEET, Journal of experimental and theoretical artificial intelligence, expert systems with applications, Applied artificial intelligence, Adhoc reviewer for international journal of ambient computing and intelligence. She has published more than 50 articles in various journals and conferences. She can be contacted at email: arunaapavate@gmail.com.






Dr. Ashfaq Shaikh    is Working as Assistant Professor in M. H. Saboo Siddik College of Engineering, Byculla, Mumbai, India. He is Ph.D. Computer Engineering with a specialization in big data analytics, machine learning, recommendation system, information and cyber security. Working as Assistant Professor in M. H. Saboo Siddik College of Engineering Byculla Mumbai India, over 23 year of teaching experience. His passion for teaching and innovation contribution resulted in winning several awards and recognition such as Mastek Deep Blue Winner in 2017, AICTE Best Team Award in Smart India Hackathon in 2018, Best Faculty Award in year 2021. He can be contacted at email: ashfaq.mhss@gmail.com.






Dr. Santosh Borde    Dr Santosh Borde is working with JSPM TSSM Group of Institutes as Asst Director for Student Progression and Industry Relations Office. He has 21 years of experience in the field of Education. He handles various responsibilities towards training and development, corporate relations and alumni relations. His area of interest is in the field of Human Computer Interaction. He worked on e learning model using human computer aspects of usability during his Ph.D. work. He can be contacted at email: spraborde@gmail.com.



Mr. Veer Bhadra Pratap Singh    is working as Assistant Professor in the School of CSIT at Symbiosis Skill and Professional University, Kiwale, Pune, Maharashtra. He had completed his B. Tech. from UPTU, Lucknow, India, Master of Science in Web Information Systems from University of Ulster, United Kingdom, and M.Tech. from Dr. A P J University, Indore, M.P., and is currently pursuing his Ph.D. from Maha Kaushal University, Jabalpur, M.P. He has published more than 12 papers in various Journals and Conferences of repute and also published 10 patents. Area of interest is machine learning, artificial intelligence, and algorithms. He can be contacted at email: er.veerpratap@gmail.com.



Mr. Rahul Raut    is working as Assistant Professor in the School of CSIT at Symbiosis Skill and Professional University, Kiwale, Pune, Maharashtra. He has completed his M.Tech. Degree from S.G.B. Amravati University, MS, India. He is Currently a Research fellow with S.G.B. Amravati University, Maharashtra, India. He has Published two Books with Reputed Publisher, one Book Chapter and over 15 Conference and Journal Papers. His Research interests broadly include vehicular ad-hoc network, mobile ad-hoc network, signal processing for communication, machine learning, and neural network. He can be contacted at email: mr.rahulraut@gmail.com.