A Literature Review on XAI and NLP Models of Sentimental Analysis

Shravan Nilesh Shah
Department of Computer Engineering, MIT ADT University, Loni Kalbhor, Pune

***Abstract*—The usage of artificial intelligence (AI) systems has increased significantly over the past few years. We have seen AI systems being implemented even for simple decision support systems. But the usage of black box models has created a major distrust among the regular users of the system. An AI system is expected to give accurate predictions still, it is also crucial that the decisions made by the AI systems are humanly interpretable i.e. anyone using the system should be able to understand and comprehend the results produced by the AI system. In this paper, we discuss the use of explainability methods such as LIME and SHAP to interpret the results of various black box models.**

***Index Terms*—** LIME, SHAP, Bag of words, TF-IDF, Logistic Regression.

I. INTRODUCTION

Today the use and need for machine learning models especially AI and deep learning models is growing faster than ever before. As the technology grows so does the power of these models due to the combination of extremely powerful machines and very easily accessible data.AI systems are easily accessible to the common man on the tip of their fingers.AI has been able to successfully help the everyday user in almost every day to day tasks as has become reliable than ever before. But the increase in usage of AI systems comes with its own limitation, i.e. its interpretability.

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, Much like the machine learning models that are used on regular basis.AI systems provide a solution without providing the reason for the solution. This black-box nature of the solution causes major distrust among the everyday user and even developers of that system. It is extremely important to know why a particular model makes a particular 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.

This paper discusses the use of various XAI and NLP methods to perform a sentimental analysis on a text dataset.

II. LITERATURE SURVEY

Explainable Artificial Intelligence (XAI):

The terms most encountered during the survey process were Interpretability and Explanability. But these terms are sometimes interchangeable or at least are used that way. A more clear definition for these words can be found in[1]. Model interpretability: The degree to which a human can understand the cause of a decision. Whereas, Model explainability is related to internal logic and mechanics inside a machine learning algorithm.

1. Objectives of XAI:
   a. Justification: To justify the model's decision in order to increase its credibility.
   b. Improvement: Some common problem that occurs during the training process is the data being imbalanced which leads to overfitting or underfitting.XAI helps in these situations.
   c. Fairness: XAI can help to prevent biases and discriminations in a machine learning process.
   d. Transparency: This means that when a model makes a decision, its whole process can be comprehended by a human.
   e. Understanding: It is a generalized term that teams up with other objectives.
2. Modality Of Explanation:
   a. Intrinsically Interpretable: Also known as Self-explaining models that provide some level of transparency eg. Decision tree.
   b. Post-hoc Interpretability: Several Post-hoc methods such as LIME and SHAP can be used to provide an explanation regarding the features.

3. Scope of Explanation: The methods used to explain the models can be classified into two types, i.e. mode specific and model agnostic.
   These two differ in the use of a specific structure of a machine learning model or completely independent of one.

a. Model Specific: Model-specific approach works based on the details of the machine learning models. Prior knowledge of the structure of the machine learning model in use is required. Hence it is also called as white box approach.
b. Model Agnostic: Model-agnostic approach works regardless of the knowledge of the machine learning algorithm in use and hence does not rely on any specific structure. Hence it is also known as the black-box approach.

III. METHODOLOGY

Several methods can be used to perform sentimental analysis on a dataset. However, This paper discusses the use of NLP methods such as Bag of Words, TF-IDF, Logistic regression, Random forest classifier and XAI methods such as LIME and SHAP.

A. Bag of Words: Any machine learning algorithm works on numbers i.e. the input to any algorithm is in numerical form. But the data available for sentimental analysis is in textual format. Hence we need to format the available text data such that it can be provided as input to the model. Bag of words or BOW is one such text modeling technique used in NLP.BOW processes the available textual data in a bag of words i.e. a vector that keeps count of the most frequently occurring words in the dataset.

B. Term Frequency-Inverse Document Frequency: Term Frequency-Inverse Document Frequency or TF-IDF is used to reflect how important a word is to a document or a collection of documents i.e. a corpus. It is calculated as follows:
   a. term frequency tf(t,d) = count of t in d / number of words in d
   b. document frequency df(t) = occurrence of t in documents
   c. inverse document frequency idf(t) = N/ df(t) = N/N(t)
   d. tf-idf(t, d) = tf(t, d) * idf(t)

where N(t) = Number of documents containing the term t

C. Logistic regression: Logistic regression is a supervised machine learning algorithm used for classification. The classification could be binomial or multinomial. Logistic regression is used to predict the probability that a particular instance belongs to a class or not.

D. Random Forest Classifier: Random forest classifier is a set of decision trees from a randomly selected subset of the training set and then it collects the votes from different decision trees to decide the final prediction.

E. LIME: LIME or Local Interpretable Model-agnostic Explanations is used as the name states, to explain what actually the machine learning models are doing. In short, LIME is used to explain the prediction of a machine learning model i.e. it performs the role of an explainer. The output of LIME is a set of explanations representing the contribution of each feature to a prediction for a single sample, which is a form of local interpretability[13].
   One major advantage of using LIME is that it works well with images, text, and tabular data and can easily be implemented using Python[14].

F. SHAP: SHAP or SHapley Additive exPlanations is the most powerful Python package for understanding and debugging your models. It can tell us how each model feature has contributed to an individual prediction. By aggregating SHAP values, we can also understand trends across multiple predictions.SHAP allows us to use multiple plot diagrams to easily understand the prediction of the model such as
   a. Waterfall plot
   b. Force plot
   c. Mean SHAP plot
   d. Beeswarm plot
   e. Dependence plots

IV. CONCLUSION AND FUTURE WORK

The survey showcases that the explainability of a system increases the trust and credibility of an AI

system. But a lot of work has to be done to bridge the gap between XAI and NLP. The above-discussed Natural Language Processing (NLP) Methods and Explainable Artificial Intelligence (XAI) Methods can be used to design a system that provides a sentimental analysis of text-based data and to further provide logical explanations for the output of the system.XAI methods do provide some transparency about the black-box models that are used for AI systems but this transparency is not absolute. There is an increasing need for further research and improvement in the field of XAI.

REFERENCES

[1] Ahmad Haji Mohammadkhania, Nitin Sai Bommib, Mariem Daboussic, Onkar Sabnisd, Chakkrit Tantithamthavorne, Hadi Hemmatif.- A Systematic Literature Review of Explainable AI for Software Engineering, arXiv:2302.06065v1 [cs.SE] 13 Feb 2023

[2] AKM Bahalul Haque, A.K.M. Najmul Islam, Patrick Mikalef- Explainable Artificial Intelligence (XAI) from a user perspective: A synthesis of prior literature and problematizing avenues for future research, Technological Forecasting & Social Change 186 (2023) 122120

[3] Jože M. Rožanec, Blaž Fortuna, Dunja Mladenić- Knowledge graph-based rich and confidentiality preserving Explainable Artificial Intelligence (XAI), Information Fusion 81 (2022) 91–102

[4] Nadhila Nurdin1, Dimas Adi - Explainable Artificial Intelligence (XAI) towards Model Personality in NLP task, IPTEK, The Journal of Engineering, Vol. 7, No. 1, 2021 (eISSN: 2337-8557)

[5] Erik Cambria, Lorenzo Malandri, Fabio Mercurio, Mario Mezzanzanica, Navid Nobani- A survey on XAI and natural language explanations, Information Processing and Management 60 (2023) 103111.

[6] Esma Balkır, Svetlana Kiritchenko, Isar Nejadgholi, and Kathleen C. Fraser- Challenges in Applying Explainability Methods to Improve the Fairness of NLP Models, arXiv:2206.03945v1 [cs.CL] 8 Jun 2022

[7] Michael NEELY, Stefan F. SCHOUTEN, Maurits BLEEKER, Ana LUCIC-A Song of (Dis)agreement: Evaluating the Evaluation of Explainable Artificial Intelligence in Natural Language Processing, arXiv:2205.04559v1 [cs.CL] 9 May 2022

[8] Waddah Saeed, Christian Omlin -Explainable AI (XAI): A systematic meta-survey of current challenges and future opportunities, Knowledge-Based Systems 263 (2023) 110273

[9] Alejandro Barredo Arrieta, Natalia D´ıaz-Rodr´ıguez, Javier Del Sera, Adrien Bennetotb, Siham Tabikg, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopeza, Daniel Molinag, Richard Benjamins, Raja Chatilaf, and Francisco Herrera -Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI, arXiv:2205.04559v1 [cs.CL] 9 May 2022, Information Fusion Volume 58, June 2020, Pages 82-115

[10] Haixiao Chi, Beishui Liao -A quantitative argumentation-based Automated eXplainable Decision System for fake news detection on social media, Knowledge-Based Systems 242 (2022) 108378

[11] Saranya A., Subhashini R. -A systematic review of Explainable Artificial Intelligence models and applications: Recent developments and future trends, Decision Analytics Journal 7 (2023) 100230

[12] Marina Danilevsky, Kun Qian, Ranit Aharonov, Yannis Katsis, Ban Kawas, Prithviraj Sen. -A Survey of the State of Explainable AI for Natural Language Processing, arXiv: 2010.00711v1 [cs.CL] 1 Oct 2020

[13] Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin -"Why Should I Trust You?": Explaining the Predictions of Any Classifier, arXiv:1602.04938v3 [cs.LG]

[14] Vignesh Mathivanan -Everything You Need to Know about LIME, [Online] Available:https://www.analyticsvidhya.com/blog/2022/07/everything-you-need-to-know-about-lime/