Measuring the quality of explainable artificial intelligence algorithms applied to two classification tasks

Preliminary Project Report

George R.E Bradley

April 2021

1 Introduction

1.1 The increased adoption of artificial intelligence

A McKinsey & Company survey [1] carried out in February, 2018, which gathered responses from 2,135 participants provides insight into the adoption of artificial intelligence systems in industry. The survey suggested that 50% of participants have adopted and embedded artificial intelligence (AI) based systems into their business operations and circa 30 percent reported piloting the use of AI. Despite the relatively high rates of adoption the majority of respondents, circa 58% of participants disclosed that less than one-tenth of their companies' digital budgets goes toward AI. Over 70% of respondents however expect investments into AI to increase year on year.

These findings suggest that artificial intelligence systems are being deployed and used in industry today and that the integration of these systems will likely increase as their use value is realised and investment increases.

1.2 Artificial intelligence challenges

1.2.1 Interpretability

The interpretability challenges of AI techniques are commonly referred to as the *Black Box Problem.* An AI model is deemed a "black box" if it takes actions without being able to communicate the reasoning behind its actions. These actions usually include making predictions and decisions.

A major cause of the black box problem are machine-learning algorithms which find geometric patterns among multiple variables simultaneously which cannot be understood nor visualised by humans. The complexity of algorithms is another negating factor to AI explainability. For example large neural networks can consist of thousands of artificial neurons which make the reasoning behind actions of an AI system opaque [2].

"Contemporary models are more complex and less interpretable than ever; used for a wider array of tasks, and are more pervasive in everyday life than in the past; Justifying these decisions will only become more crucial, and there is little doubt that this field will continue to rise in prominence and produce exciting and much needed work in the future" [3].

1.2.2 Additional challenges

The Alan Turing institute's "Guide for the responsible design and implementation of AI systems in the public sector" identifies six potential "harms" caused by AI systems including the aforementioned interpretability challenges [4].

Four of the additional possible harms are:

1. **Bias and Discrimination**

    - **Societal bias reinforcement**
      Existing structures and dynamics of societies build the foundation from which AI systems learn. As a consequence of this, data driven approaches can "reproduce, reinforce, and amplify" any existing patterns of marginalisation and discrimination that exist in these societies.

    - **Designer bias**
      The implemented systems can potentially perpetuate the preconceptions and bias of the designers of the system. This is due to the designers making decisions concerning the underlying structure of the models.

    - **Data Bias**
      The dataset used to build a model can often be an incomplete representation of the population from which it draws inferences. Due to the data being unrepresentative the model can mirror any bias and or discrimination found in the initial dataset.

2. **Denial of Individual Autonomy, Recourse, and Rights**
   End users could become unable to designate responsibility for outcomes caused by algorithmically generated decisions. In cases where users are negatively affected by the automated decision making gaps in accountability may encroach on the autonomy and violate the rights of the affected party.

3. **Invasions of Privacy**
   Systems which use personal data of the end users could then be used to influence them. This use of personalised data could be interpreted as infringing upon the end user's ability to intentionally manage the effects the technology has on influencing them.

4. **Isolation and Disintegration of Social Connection**
   AI systems can lead to excessive curation of individual experiences. The curation of services might reduce the need for human-to-human interaction. Additionally this targeted curation could limit our exposure to worldviews different from our own which could lead to social polarisation.

AI systems have real impact, use cases and concerns. As the development and deployment of AI systems have continued to increase, ethical and societal pressures for these systems to provide explanations for their predictions has also increased.

1.3 Regulatory impact

The pressures for more transparent AI systems have culminated in new regulations being passed. Below are two examples which are indicative of these new regulations.

The European Union's General Data Protection Regulation restricts algorithms that make decisions based on user-level predictors which "significantly affect" users. In addition to this the law creates a "right to explanation" where a user can ask for an explanation of an algorithmic decision which was made about them [5].

In May 2019 the Illinois Legislature passed the Artificial Intelligence Video Interview Act [6], which addresses how employers use artificial intelligence to analyse job applicant's video interviews. Employers are required to notify applicants based in the state of Illinois that it plans to have their video interview analysed algorithmically, explain how the AI system analyses the interview and what characteristics are used for evaluation.

1.4 Importance of explainable AI

Explainable AI (XAI) represents a set of techniques and algorithms which enable the better understanding and greater transparency of AI systems. XAI is a field which facilitates the building of confidence and trust with end users whilst ensuring that the developed models and systems work as intended.

XAI plays a critical role in the tackling of the main challenges of artificial intelligence (AI) models, the nature of XAI primarily aims to addresses the interpretability challenges of AI models and consequentially tackles the possible "harms" of AI outlined by the Alan Turing institute. With further transparency and explanations of the workings of AI models we are able to identify the possible perpetuation of biases and discrimination of models whilst being able to help shift control to the end users who with further understanding of the functioning of AI models can take actions to prevent being influenced negatively by them.

The increased importance of XAI is correlated with the increased use of AI systems. This importance being mirrored in the ethical and societal pressures which have led to recent regulatory impact and research output.

1.5 Project Aim

This project aims to apply contrastive and game theoretic XAI approaches to two classification tasks applied to text and image data respectively. A sample group is then constructed and surveyed in order to measure the quality of the algorithmically developed explanations.

1.6 Objectives

- Select the implementation software/toolkit.
- Select and extract suitable datasets.
- Apply algorithms to the datasets to produce relevant explanations.
- Create sample group for surveying.
- Distribute the explanations to the same group, collect and aggregate results.

2 Dataset

The project aims to apply two XAI approaches to two data types, image and text data. The datasets were chosen using criteria based selection. Each of the selection criteria for both data types are outlined below.

2.1 Image data

The selection of the image dataset is based on three criteria, data type, data size and data structure.

- **Data Type**
  In order to fulfill the data type criterion, the dataset must consist of image data in the format of JPG or PNG. These two file formats were selected as they are two of the most widely used image formats [7] with readily available and free programs for accessing and editing the data.
- **Data Size**
  When developing classification models there are challenges associated with the amount of data needed to create a viable model. There are no standard guidelines suggesting the minimum amount of data needed for classification tasks. In general larger datasets produce better outcomes. In the selection of the dataset a minimum size of 10,000 records per class was deemed appropriate. Therefore in order to satisfy the criterion the dataset must have more than or equal to 10,000 records per class.
- **Data Structure**
  The CEMExplianer algorithm mentioned in section 4.2 is deployed to algorithmically develop contrastive explanations. The CEMExplainer algorithm requires labelled data, therefore in order to fulfill the data structure criterion the dataset is required to have the following characteristics:
  - Labelled.
  - Structured in the form train, test and validation.

2.1.1 Selected image dataset

The selected image dataset [8] consists of 140,000 records, 70,000 images of real faces and 70,000 imaged of fake faces (generated by StyleGAN).

Table 1 below depicts the selection criteria for the image dataset showing that the dataset fulfills the defined criteria.

Selection Criteria	**Y/N**
Data in format JPG or PNG	Y
>= 10,000 records per class	Y
Labelled data	Y
Structured in form train, test and validation	Y

Table 1: Image dataset selection criteria matrix

2.2 Text data

The selection of the text dataset is based on the same three criteria as the image dataset, data type, data size and data structure using different fulfilment requirements.

- **Data Type**
  In order to fulfill the data type criterion, the dataset must consist of data in csv format. The csv data format was preferred as it allows for easier manipulation and preparation of the data for the creation of a classification model.

- **Data Size**
  As mentioned in 1.2 there are challenges associated with the amount of data needed to generate a viable classification model. In the selection of the text dataset a minimum size of 10,000 records per class was deemed appropriate. Therefore in order to satisfy the criterion the dataset must have more than or equal to 10,000 records per class.

- **Data Structure**
  In order to fulfill the data structure criterion the dataset is required to be in the structured in the form train, test and validation.

2.2.1 Selected text dataset

The selected text dataset [9] consists of circa 50,000 records, where each class has a circa 25,000 records.

Table 2 below depicts the selection criteria for the text dataset showing that the dataset fulfills the defined criteria.

Selection Criteria	**Y/N**
Data in format CSV	Y
>= 10,000 records per class	Y
Structured in form train, test and validation	Y

Table 2: Text dataset selection criteria matrix

3 Technical specification

The classification models and the XAI algorithms used in the project are implemented using the Python programming language. Python was selected as the chosen implementation language due to two reasons.

1. Python's extensive packages for the development of classification models.
2. IBM's AI Explainability 360 toolkit is developed for use with Python [10].

IBM's AI Explainability 360 toolkit is used in the implementation of the project as the toolkit offers the largest and most comprehensive toolkit of XAI algorithms and performance metrics.

4 Methodology

The project's methodology is comprised of four steps, outlined below.

4.1 Creating classification models

Classification models for both the image and text dataset are created. For the image classification model Keras is used in the creation of the model. For the semantic analysis of the text dataset, a model is produced using a Python sentiment analysis pipeline.

4.2 Applying XAI algorithms

The CEMExplainer algorithm [11] is used to compute a contrastive explanation for the image dataset. The explanation is created by finding what is minimally sufficient, pertinent positive and what should be necessarily absent, pertinent negative to maintain the original classification.

The SHAP (SHapley Additive exPlanations) algorithm [12] which is a game theoretic approach is used to explain the output of the semantic analysis of the text dataset. It connects optimal credit allocation with local explanations using the classic Shapley values from game theory and their related extensions.

4.3 Creating sample group

The peer group was used for the creation of the sample group as a familiarity with Computer Science principles is deemed preferable. A possible risk of the sample group is that there will not be enough participants to be statistically significant however it should be indicative of a process which can be applied to a larger sample size.

Special consideration has to be given due to COVID-19. The survey will be distributed using google forms and the answers asked will be quantifiable using a likert scale to determine the extent to which the participant agrees about statements concerning the explanations.

4.4 Collecting and aggregating survey findings

A series of explanations from both classification tasks with questions are formulated and distributed with the sample group. A 10 day period is used to collect completed surveys. The results are aggregated and analysed.

5 Literature Review

Within the literature there is a distinction between models which by design have a level of interpretability and others which require the use of XAI techniques in order to build explanations.

When AI models are non-transparent by design, post-hoc explainability techniques are used in order to generate information about how a developed model produces its predicted values. Post-hoc explainability techniques leverage common techniques used by humans for explanations. Techniques which can be applied to models in general are feature relevance estimation, visualisation techniques and model simplification.

Feature relevance explanation

Feature relevance explanation describes the process of explaining the function of models by ranking the influence that each feature of the model has on the prediction made.

SHAP (SHapley Additive exPlanations) [13] is a XAI technique which uses game theory Shapley values. Shapley value based explanation "uses fair allocation results from cooperative game theory to allocate credit for a model's output among its input features". In game theory a player can either join or not join a game. To replicate this for the means of explaining a opaque model a feature either has "joined a model" when the value of that feature is known or it has not joined when the value is not known. The SHAP value for a specific feature is the difference between the expected model output and the partial dependence plot at the feature's value.

Visual explanation techniques

[14] presents a series of sensitivity analysis (SA) visual explanation approaches which can be deployed in the explanation of black-box models. Sensitivity analysis works by querying fitted AI models with sensitivity samples and recording the responses. As a consequence no information from training is used allowing for SA methods to be universally applicable to any supervised learning method.

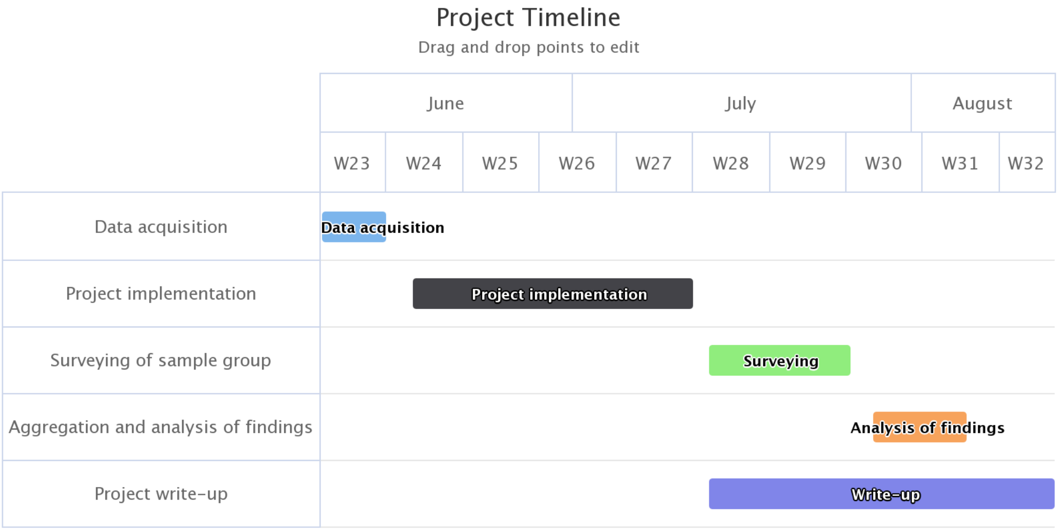
The CEMExplainer [11] algorithm can be used to compute contrastive explanations for image data by finding what is necessarily, minimally and sufficiently present and analogously absent in in a model's predicition in order to maintain and justify the original classification.

One area of discussion which is underserved in the literature is that considerable onus is on the individual to interpret the explanation produced by visual explanation techniques.

Explanation by simplification

The majority of simplification based explanations use rule extraction techniques. Local Interpretable Model-Agnostic Explanations (LIME) [15]. LIME provides local explanations around the models predicted value. The algorithm by default produces 5000 samples of the feature vector following normal distribution. Once this is done, using the prediction model the target variable from the 5000 samples is obtained. Once a surrogate dataset is created each row of the dataset is weighted based on its proximity to the original sample. A feature selection technique is then applied to select the most significant features.

6 Project schedule



References

[1] “Ai adoption advances, but foundational barriers remain, mckinsey & company, nov. 13, 2018.” https://www.mckinsey.com/featured-insights/artificial-intelligence/ai-adoption-advances-but-foundational-barriers-remain.

[2] “Y. bathaee, “the artificial intelligence black box and the failure of intent and causation,” harvard journal of law technology, vol. 31, no. 2, 2018.” https://jolt.law.harvard.edu/assets/articlePDFs/v31/The-Artificial-Intelligence-Black-Box-and-the-Failure-of-Intent-and-Causation-Yavar-Bathaee.pdf.

[3] “Biran, o. and cotton, c., 2017, august. explanation and justification in machine learning: A survey. in ijcai-17 workshop on explainable ai (xai) (vol. 8, no. 1, pp. 8-13).”

[4] "Understanding artificial intelligence ethics and safety a guide for the responsible design and implementation of ai systems in the public sector dr david leslie public policy programme," doi: 10.5281/zenodo.3240529."

[5] "B.goodman and s.flaxman, eu regulations on algorithmic decision-making and a right to explanation, jun. 28, 2016."

[6] "D. b. pasternak, "illinois and city of chicago poised to implement new laws addressing changes in the workplace - signs of th...," lexology.com, jun. 05, 2019."

[7] "Research guides: All about images: Image file formats, umich.edu, 2021." https://guides.lib.umich.edu/c.php?g=282942p=1885348.

[8] "Image dataset." https://www.kaggle.com/xhlulu/140k-real-and-fake-faces?select=test.csv.

[9] "Text dataset." https://www.kaggle.com/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews.

[10] "Ai explainability 360," mybluemix.net, 2021." http://aix360.mybluemix.net/.

[11] "A. dhurandhar et al., "explanations based on the missing: Towards contrastive explanations with pertinent negatives," arxiv.org, 2018." ://arxiv.org/abs/1802.07623.

[12] "Local white box explainers — aix360 0.1 documentation, readthedocs.io, 2018." https://aix360.readthedocs.io/en/latest/lwbe.htmlshap-explainers.

[13] "Welcome to the shap documentation — shap latest documentation, readthedocs.io, 2021." https://shap.readthedocs.io/en/latest/.

[14] "P. cortez and m. j. embrechts, using sensitivity analysis and visualization techniques to open black box data mining models, information sciences, vol. 225, pp. 1–17, mar. 2013, doi: 10.1016/j.ins.2012.10.039."

[15] "Why should i trust you?' — proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, acm.org, 2016." https://dl.acm.org/doi/abs/10.1145/2939672.2939778.