

INT₂X: EXPLANATION FOR THE CAUSES OF A PREDICTION

23-142

Project Proposal Report

Warnasooriya.S. D

B.Sc. (Hons) Degree in Information Technology

(Specialized in Data Science)

Department of Information Technology

Sri Lanka Institute of Information Technology


Sri Lanka

May 2023

DECLARATION

We declare that this is our own work, and this proposal does not incorporate without acknowledgement any material previously submitted for a degree or diploma in any other university or institute of higher learning, and to the best of our knowledge and belief it does not contain any material previously published or written by another person except where the acknowledgement is made in the text.

DECLARATION

| Name | Student ID | Signature |
|-------------------|------------|---|
| Warnasooriya S. D | IT20097660 |  |

The above candidate is carrying out research for the undergraduate Dissertation under my supervision.

.....

Signature of the Supervisor:

.....

Date

ABSTRACT

With the advancement of machine learning technologies, most of the decision-making processes are being automated and AI-based in recent years. When the topic of AI-based processes becomes more popular communities began to question the explainability and interpretability of those decision-making processes. The explainability problem of AI-based decisions was more pronounced after the European Union imposed General Data Protection Regulations (GDPR) in 2016. GDPR mentions that there is a right for consumers to know about the internal processes that use to make predictions by the black box models. As a solution for the explainability problem, the concept of Model Interpretability (XAI) is being raised as a hot topic and it is the next step of AI. Even though the topic has become popular, the number of studies that have been done related to that area is less. In this research, the main objective is to provide a novel XAI solution to enhance the model interpretability of black box models by developing a novel counterfactual rule generation mechanism related to the text classification domain. The proposed method will be specifically targeted the function-based classification models focusing on Support Vector Machine (SVM). SVM is a very popular classifier that performs well with both linear and non-linear data. However, the support vector machine is a black-box model because it is very difficult to explain the model predictions and challenging to understand the role of each input feature in the decision-making process. The black box behavior of SVM becomes more pronounced when SVM use to classify non-linear separable data by mapping input data to higher-dimensional space. Throughout the research, the drawbacks of the existing methods will be identified and addressed those by developing a novel XAI solution.

LIST OF CONTENTS

| | |
|---|-----------|
| DECLARATION..... | 2 |
| ABSTRACT..... | 3 |
| LIST OF CONTENTS..... | 4 |
| LIST OF FIGURES..... | 5 |
| LIST OF TABLES..... | 6 |
| LIST OF ABBREVIATIONS..... | 7 |
| LIST OF APPENDICES..... | 8 |
| 1.INTRODUCTION..... | 9 |
| 2. LITERATURE REVIEW..... | 13 |
| 2.1 Background..... | 13 |
| 2.2 Literature Survey..... | 16 |
| 2.2.1 White-box Induction from SVM Models: Explainable AI with Logic Programming [6]..... | 18 |
| 2.2.2 Anchors: High-Precision Model-Agnostic Explanations [7]..... | 19 |
| 2.2.3 Local rule-based explanations of black box decision systems [8]..... | 20 |
| 3. RESEARCH GAP..... | 22 |
| 4. RESEARCH PROBLEM..... | 23 |
| 5. OBJECTIVES..... | 25 |
| 5.1 Main Objective..... | 25 |
| 5.2 Specific Objectives..... | 26 |
| 5.3 Work Breakdown Structure..... | 27 |
| 6. METHODOLOGY..... | 28 |
| 6.1 System Architecture Diagram..... | 28 |
| 6.2 Tools and Technologies..... | 29 |
| 7. PROJECT REQUIREMENTS..... | 30 |
| 8. GANTT CHART..... | 31 |
| 9. REFERENCES..... | 32 |
| APPENDICES..... | 33 |

LIST OF FIGURES

| | |
|--|----|
| Figure 1-1: Husky Wolf Explanation..... | 10 |
| Figure2.1-1: Linearly Separable Data..... | 15 |
| Figure 2.1-2: Nonlinearly Separable Data..... | 15 |
| Figure2.2-1: Anchors Sample Explanation..... | 19 |
| Figure 2.2-2: Explanation rule and the counterfactual for compass dataset..... | 20 |
| Figure 5.3-1: Work Breakdown Structure..... | 27 |
| Figure 6.1-1: System Architecture Diagram..... | 28 |
| Figure 8-1: Tentative Gantt Chart..... | 31 |

LIST OF TABLES

| | |
|---|----|
| Table 3-1: Comparison between existing methods and proposed method..... | 22 |
|---|----|

LIST OF ABBREVIATIONS

| Abbreviations | Description |
|---------------|---|
| AI | Artificial Intelligence |
| XAI | Explainable AI |
| SVM | Support Vector Machine |
| GDPR | General Data Protection Regulation |
| VCS | Version Control System |
| SHAP | SHapley Additive exPlanations |
| LIME | Local Interpretable Model-agnostic Explanations |
| NLP | Natural Language Processing |
| TCAV | Testing with Concept Activation Vectors |

LIST OF APPENDICES

| | |
|------------------------------------|----|
| Appendix A: Plagiarism Report..... | 33 |
|------------------------------------|----|

1.INTRODUCTION

The concept of Interpretable AI also known as Explainable AI has come into the discussion with the evolution of Artificial Intelligence (AI) under the Machine Learning domain. Nowadays predictions made by artificial intelligence are used for critical and sensitive decision-making processes. But this traditional AI is a black box that can answer only "yes" and "no" type questions without elaborating how that answer is obtained. As a solution for that, there was a requirement of explainability to ensure the trustworthiness and transparency of AI making decisions.

At the dawn of the fourth industrial revolution, different domains are witnessing a fast and widespread adoption of artificial intelligence (AI) in the decision-making process under domain experts. Healthcare, Legal, military, transportation, and finance are some domains where AI models use for the decision-making process. Decisions made by the AI model are extremely critical in a situation like a medical diagnosis of a disease. Here AI model should be able to provide an explanation of the decision because patients are interested in the treatment process and want to know why these treatments are required. There are some situations in which the models can be biased and provide inaccurate results. In such situations model explainability is well needed to identify and avoid errors in decision making.

When machine learning models are used for the decision-making process, it is not enough to have high accuracy. The model should be able to explain how each feature maps with the final result. In this situation, two types of models come into the picture called "white box models" and "black box models".

White box models are machine learning models where the internal workings of the model are transparent and understandable to the user. In other words, the user has access to the model's underlying structure, parameters, and decision-making processes. White box models provide a better explanation because those models are not too complex. The accuracy provided by these models may not be enough for some predictions. Examples of white box models are,

- Regression models
- Decision tree models
- Naive Bayes model

When we come to black-box models, the internal workings of the model are not clear and not easily understandable. Black-box models can handle complex, high-dimensional data and provide more accuracy than white-box models. However, black box models can be difficult to interpret or explain, which can make it challenging to identify and correct any errors or biases in the model. Examples of black box models are

- Deep neural networks
- Random forest
- Boosted trees
- Support Vector Machine
- K nearest neighbor

There is a popular example called the "Husky" mistake that properly demonstrates the need of model explainability. [1]

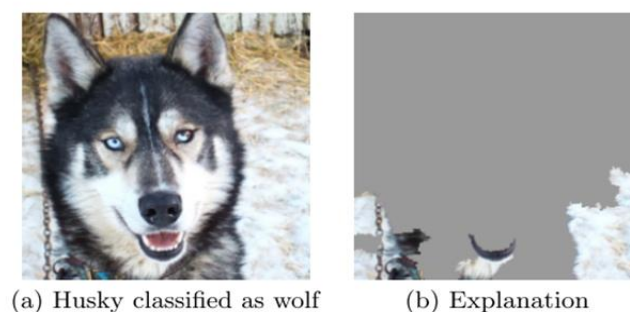


Figure 3-1: Husky Wolf Explanation

In that scenario, researchers trained a logistic regression neural network model by feeding the images of wolves and huskies as the input features. The images of wolves had a white colour background(snow) and the images of huskies had not that white background. After training the model they provided an image of a husky that had a white colour (snow) background. The model predicted that image as a wolf. In that situation, the model did the prediction based on the background colour without considering the animal's colour, position, size, and other important features. That means the model is trained as a bad classifier. If this happens in a critical decision-making process the situation must be dangerous. Using XAI we can avoid these situations by explaining the models and identifying if there is any bias.

There is no exact definition for Model Explainability/Interpretability and different authors prefer to use the most sensible definitions. The most popular and widely used definitions are **“Interpretability is the degree to which a human can understand the cause of a decision”** [2] and **“Interpretability is the degree to which a human can consistently predict the model's result”** [3] The main objective of XAI is to answer the "wh" questions related to AI-based decision-making. For example, XAI should be able to answer "why a particular answer was obtained?", "how a particular answer was obtained" and "when a particular AI-based system can fail?" [4] In [4] they proposed four reasons why explainability is needed. Those are,

➤ **Explain to justify.**

There were some situations AI/ML-based systems provide biased or discriminatory results. Therefore, the explanation of AI-based results is essential. The XAI systems provide the required information to justify the result when unexpected decisions are made.

➤ **Explain to control.**

Through the explainability process, users can get an understanding about the system's behavior. It provides greater visibility of unknown vulnerabilities and flaws and helps to rapidly identify and correct errors.

➤ **Explain to improve.**

A model that can be explained is one that can be improved easily. Because users know the relationship between inputs and output, and how to make the output smarter.

➤ **Explain to discover**

Explanation about the process is helpful to learn new facts, gather information, and gain knowledge. XAI models are useful to find new and hidden laws in different scenarios.

Nowadays Explainable AI (XAI) has numerous applications across various industries and domains. XAI can be used in medical diagnosis, treatment planning, and drug discovery to provide explanations for the recommendations made by AI models. In finance sectors, XAI is used in fraud detection, credit scoring, and investment analysis processes. This can help financial institutions comply with regulations, increase transparency, and improve customer trust. Autonomous vehicles also use XAI tools to provide explanations for the decisions made by the AI models that control these vehicles. Through that explanations, those systems can ensure safety and increase public trust in these emerging technologies. In sentimental analysis, XAI tools assist to detect biases, identify errors and evaluate fairness issues in text data. Furthermore, XAI tools use to identify patterns and anomalies in manufacturing processes, allowing for predictive maintenance and quality control as well.

This research mainly focuses to enhance the model interpretability of the Black-Box models related to a text classification task. Model explainability useful in text classification tasks because of several reasons. Through explainability, the system can identify whether the model is biased toward certain words or phrases in the text. This help to improve the model's accuracy and avoid making unfair predictions based on certain demographic factors, such as race or gender. Model explainability provides insights into the model's decision-making process, making it easier for users to understand how the model arrived at its predictions. This can assist to build trust in the model and make it more accessible to non-experts. When text classification models are not performing as expected, model explainability can help identify specific areas where the model has errors and the reasons for those errors. Further, by understanding which features are most important for the model to make accurate predictions, model explainability can help improve the model's performance by allowing developers to optimize the feature selection or weighting.

2. LITERATURE REVIEW

2.1 Background

The research area of XAI has been taken to the attention of the researchers and the communities with the fast growth of AI models and their applications. The need and the importance of the XAI concept have been raised when people consider the trustworthiness and transparency of AI-based decision-making processes. In the last few years, many researchers have been proposing many model-explainable methods for the research community. Because the knowledge in this area is not well matured most of the techniques are still in the research phase.

With the advancement in the XAI domain researchers have grouped the XAI explainability methods into different categories. They are the Local-Global method, model specific-model agnostic method, and Post hoc-Intrinsic method. When we consider the Local-Global explanation approach, Local explanation methods, aim to explain the model's predictions for individual instances or data points while Global explanation methods aim to provide an overall understanding of how a machine learning model works, and what factors or features are most important for the model's predictions. LIME (Local Interpretable Model-agnostic Explanations) and SHAP (SHapley Additive exPlanations) are local explanation methods and feature importance ranking, Partial dependence plots and decision trees are global explanation methods.

In model specific-model agnostic approach, Model specific methods are designed to explain a specific machine-learning model, while model-agnostic methods can be used to explain any machine-learning model, regardless of its type or complexity. Furthermore, the explanation methods can be post hoc or intrinsic. Intrinsic methods aim to build explainability into the machine learning model itself by considering the internal structure and parameters of the model, while post-hoc methods analyze the model after it has been trained. Intrinsic methods are typically used for simpler models that can be designed to be inherently interpretable and post-hoc methods are used for more complex models.

As discussed in the introduction the concept of XAI has been focusing on black box models. Support Vector Machine is one of the most popular ML models because of its excellent predictions and generalization capabilities. It becomes a black box model due to its model complexity and low interpretability. SVM is one of the most popular supervised learning techniques, which is used for classification, regression tasks, and outlier detections. But it mainly focuses on the classification task. The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes with an optimal hyperplane. This best decision boundary is called a hyperplane. When we consider the visualization of SVM the model plots the data points in a dimensional space and the number of dimensions are equal to the number of features in the dataset. Also, SVM constructs a hyper-plane or set of hyper-planes in a high or infinite-dimensional space. There are two types of SVMs, linear SVM and nonlinear SVM. Linear SVM is used to handle linearly separable data, which means a dataset can be classified into two classes by using a single straight line while non-linear SVM is used to handle non-linear separable data.

When the input data is linearly separable, SVM uses a straight line to separate the data into two classes. Here SVM can be trained to find the optimal decision boundary that maximizes the margin between the two classes. To find the optimal hyperplane, the SVM solves an optimization problem that involves minimizing the norm of the weight vector subject to the constraint that all data points are classified correctly. This optimization problem can be solved efficiently using various optimization techniques such as quadratic programming. Once the optimal hyperplane is found, the SVM can make predictions on new input data by simply evaluating which side of the hyperplane the input data falls on.

When SVM is used to classify non-linearly separable data, it uses a non-linear kernel function to map the input data into a higher-dimensional space where a linear decision boundary can be found. The choice of the kernel function and its associated parameters can greatly impact the performance of the SVM. So, it may require significant experimentation and tuning to arrive at the best model. Once the optimal hyperplane is found, the SVM can make predictions on new input data by first mapping the input data into the higher-dimensional space using the kernel function and then evaluating which side of the hyperplane the transformed data falls on.

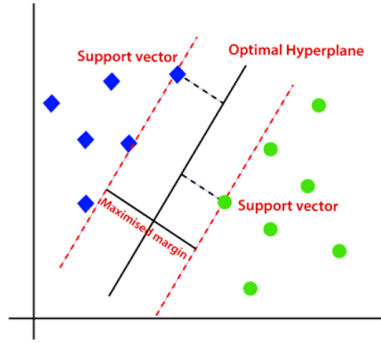


Figure 2.1-1: Linearly Separable Data

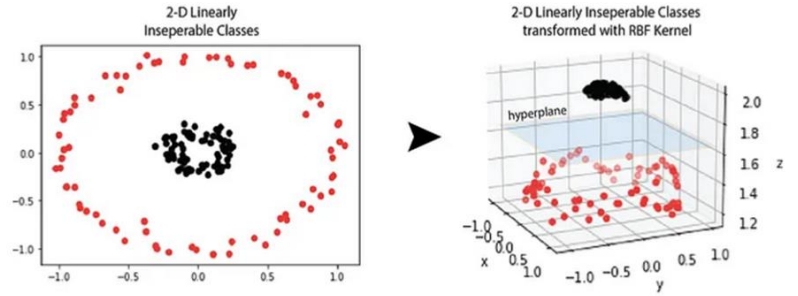


Figure 4.1-2: Nonlinearly Separable Data

The black box behavior of the SVM becomes more pronounced, **When SVM is used to classify non-linearly separable data.** When the input data is transformed into a higher-dimensional space using a non-linear kernel, the decision boundary can become highly complex and difficult to visualize. It may not be clear how the SVM is making its predictions, and it can be challenging to understand the role of each input feature in the decision-making process. Additionally, the kernel function used by the SVM may not have a direct interpretation in terms of the original input features, making it hard to understand which features are driving the SVM's predictions. That means challenging to understand the role of each input feature in the decision-making process.

When we consider the scope of SVM model usage natural language processing (NLP) is at the top. SVM is one of the most popular choices for text classification domain because SVM is a non-parametric algorithm and it can efficiently handle high-dimensional data by finding the optimal hyperplane that separates the different classes in the feature space. Also, SVMs are less prone to overfitting than other models, which is important in text classification because the number of features can be very large. This is because SVMs seek to maximize the margin between the different classes in the feature space, which helps to prevent overfitting. SVMs offer flexibility in the choice of kernel functions, which allows them to be used with different types of data and to capture different types of patterns. This makes SVMs a versatile algorithm for text classification.

The ultimate goal of this research is to overcome model explainability issues in the SVM model by providing a better XAI solution related to the text classification task.

2.2 Literature Survey.

XAI is dedicated to demystifying the black boxes by properly explaining the internal process of AI/ML models. It improves the trustworthiness, transparency, and responsibility of AI-based decision-making processes. Since the XAI research area is still in the starting era most of the techniques are in the research phase and researchers had built communities for do contributions to the XAI domain.

The author of [2] mentioned there are several explanation methods and strategies had proposed by researchers to make AI systems explainable. Among those techniques "Scoop Related Methods" and "Model related Methods" are commonly used for the explainable process. Scoop related methods are classified as global interpretability and local interpretability. Global interpretability facilitates the understanding of the whole mechanism of the model and the entire reasonings cause for the final output. Local interpretability focuses on explaining the reasons for a specific decision or a single prediction. Model-related methods are also classified into two categories as model specific interpretability (methods are limited to a specific model) and model-agnostic interpretability (methods are not tied to a specific ML model).

From the proposed techniques some have become more popular among the research community and new contributions are also done on top of these techniques. LIME (Local Interpretable Model-agnostic Explanation) [1], SHAP (SHapely Additive exPlanations) [5], XAI360 and Google XAI are examples for that. When we consider LIME, it interprets the model and explains the classification of the model in a faithful manner. It provides local optimum explanations which compute the important features by generating samples of the feature vector. Those samples are following a normal distribution. After getting the predictions from the samples it assigns weights to each of the rows to get an idea of how close they are from the original sample. Then LIME uses a feature selection technique to identify the most significant features.

SHAP is a unified approach that has been developed based on coalitional game theory. In that theory, they assign a reward to the game players according to their contributions to the game. SHAP assigns feature importance value for each feature that affects a particular output result. It maps the input features with the output results based on that Sharpley value. The key difference between LIME and SHAP is in the way that assigns weights to the input features. LIME uses a cosine measurement while SHAP uses Shapley formula. In this review, rule-based explanation methods will be discussed.

XAI360 is a commercial explainable AI (XAI) tool that provides a model-agnostic explanation. Also, it provides both local and global explanations to enhance the model interpretability of black box models. XAI360 uses various techniques to do the model explainability task. Feature importance ranking, decision tree method, counterfactual explanation, and visualizations are some of those techniques. Through feature importance ranking, XAI360 can identify the features or variables that had the most influence on the model's output and understand which factors were most important in driving the model's predictions or recommendations. XAI360 can generate decision trees that visualize the decision-making process of the black box model. These trees show how the model arrived at its output by breaking down the decision process into a series of smaller, more understandable steps.

Google XAI (Explainable AI) is a research initiative focused on developing techniques and tools for making AI more transparent and interpretable. This is also a model-agnostic method that provides both local and global explanations. Testing with Concept Activation Vectors (TCAV), integrated gradients, and attention mechanisms are some of the explainable techniques developed by Google XAI. TCAV provides a technique for understanding how models make decisions by analyzing the activations of different neurons in the model's deep learning architecture. Integrated Gradients calculate the importance of input features by integrating the gradients of the model's output with respect to the input features. Furthermore, attention mechanisms are used in deep learning architectures to identify which parts of an input are most important for the model's output.

2.2.1 White-box Induction from SVM Models: Explainable AI with Logic Programming [6]

The author of [6] had come up with a model-specific explainable approach to SVM that focuses on inducing logic programs. Inductive Logic Programming (ILP) is a subfield of machine learning and here the models do the learning process in the form of logic programming rules (Horn Clauses) that are comprehensible to humans. They get used SHAP to calculate the feature importance value of the input features. This paper makes a novel contribution to introducing a novel ILP algorithm called SHAP-FOIL that iteratively learns a single clause for the most influential support vector. According to the paper, FOIL is a top-down sequential covering inductive logic programming algorithm and there were some issues with those types of algorithms. To overcome those problems, they proposed this new SHAP-FOIL algorithm.

This SHAP-FOIL method has been introduced as a statistical learning-based ILP method. Also, the SHAP-FOIL algorithm has the capability of learning the logic programs based on the global behavior of the SVM model. Explain the model as a logic program leads to greater comprehensibility for the users because of its well-defined declarative and operational semantics. The intuition behind the SHAP-FOIL algorithm is: If a subset of feature values explains the decision on a particular support vector, it explains the decision on data points that are “similar” to that support vector too. Similarity is measured using an equation.

2.2.2 Anchors: High-Precision Model-Agnostic Explanations [7]

In this research, they have introduced a novel model-agnostic methodology that explains the behavior of complex models with high-precision if-then rules called "anchors". Since this method is model agnostic it can be applied to any model that exists. This study shows how a model would behave on unseen data instances with less effort and higher precision, as compared to existing explanations. In that approach, once the model identifies the features that hold Anchors, the changes that happen for the rest of the features will not be considered. Therefore, the prediction will be almost the same all the time according to the anchors.

`{"not", "bad"} → Positive` `{"not", "good"} → Negative`

Figure2.2-1: Anchors Sample Explanation

Above example was given in the paper [7]. In this example, the method considers the words “not bad” and “not good” to virtually predict the sentiment of the statement (“not bad” as positive and “not good as negative”). This method has been applied to various machine learning scenarios and domains including Text Classification, Structured Prediction, Tabular Classification, Image Classification, and Visual Question Answering to prove the usefulness of the method. Further, the study shows that Anchors can effectively identify the most relevant parts of the input that contribute to the classifier's prediction.

2.2.3 Local rule-based explanations of black box decision systems [8]

In this study, the authors have proposed a method called "LORE" which is a local agnostic approach that provides interpretable explanations on black box models based on logic rules. First, it learns a local interpretable predictor on a synthetic neighbourhood generated by a genetic algorithm. Then it derives the logic from the local interpretable predictor to provide a meaningful explanation that caused the prediction. Apart from providing an explanation for decision making LORE provides a set of counterfactual rules, which propose the changes in the features that lead to a different outcome.

According to the study [8], it supports for relational, tabular data to generate explanation rules. The high-level idea of the process has been explained like this. There is a given black box predictor which used to perform binary predictions and a specific instance 'x' labeled with outcome 'y' by the binary predictor. First, it will develop an interpretable predictor using a balanced dataset of neighbourhood of the given instance 'x' through a genetic algorithm. Then the local explanation for the given instance 'x' is extracted using the obtained decision tree classifier. Furthermore, it generated a set of counterfactual rules, that explains which features and conditions should be changed in order to invert the given prediction 'y'. As the authors of [8] say, the compass dataset, it may come up with the below explanations for both decision and counterfactuals.

$\{age \leq 39, race = African-American, recidivist = True\} \rightarrow High Risk$ and
the counterfactuals $\{age > 40\}, \{race = Native-American\}$.

Figure 2.2-2: Explanation rule and the counterfactual for compass dataset

The logic of this method is common as the methods discussed early in this review such as LIME [1], and Anchors [7]. The novelty of this method is it uses genetic algorithm to capture the decision boundary in the neighbourhood. Therefore, it produces high quality training dataset that will be used for learning the decision tree classifier in order to explore the decision rules.

2.2.4 Explaining Machine Learning Classifiers through Diverse Counterfactual Explanations [9]

Concept of Counterfactuals requires imaginations of hypothetical realities that can be happened other than the existing situation. In this research, to understand the concept of Counterfactuals it has given a real-time scenario that can be happened in day-to-day life. It will be discussed below. In a situation where a loan has been rejected. Using interpretable models, the organization may give an explanation for that action. As an example, it can be because of “poor credit” history. When we consider the person’s point of view who has been applied for the loan, the explanation does not help him to decide “what should he/she do next?” If the system is able to suggest some solutions to improve the chance to get the loan in future, that would be more effective for both sides.

However, there might be some situations where the most importance feature or features like gender, race may not be sufficient to flip or change the prediction. Therefore, it is pretty much important to provide alternative rules which are actionable. As the counterfactual explanation for the above loan example, the paper has provided below information [9].

“You would have received the load if your income was higher by \$10000”.

When the person gets that kind of explanation, he/she would be able to identify which features that need to be improved to be eligible for the load in future.

3. RESEARCH GAP

The main target of this research is to make the SVM model more interpretable related to the text classification domain. we can use a counterfactual rules generation mechanism, which proposes the changes that need to be done in the input features to flip the final predictions to achieve the SVM model explainable task. The proposed local rule-based explanation methods are mainly focusing on binary predictors. Because of that LORE [8] and SHAP-FOIL algorithm [6] can't be used for text classification tasks. Also, those two approaches are not used a counterfactual explanation method for the explanation process. When we consider Anchors[7], it can be used for text classification tasks but it does not follow a counterfactual explanation approach. In LIME[1], SHAP[5] XAI360[4], and Google XAI[10] tools, they provide text classification explanation mechanisms. But only XAI360 provides a counterfactual explanation for explaining the internal behavior of black box models.

Further, model explainability visualizations provide by most of the XAI tools are very complex and those visualizations can understand only by the subjects' experts. So, it is very important to provide a user-friendly more understandable visualization of the internal process of the black box models to the end users.

| | Anchors [7] | LORE [8] | SHAP- FOIL [6] | SHAP [5] | LIME [1] | Google XAI [10] | XAI360 [4] | Diverse Counterfactual Explanations [9] | Proposed SVM Explainable Method |
|--|----------------|-------------|----------------------|-------------|-------------|-----------------------|---------------|--|--|
| Model Specific Approach. | X | X | √ | X | X | X | X | X | √ |
| Text classification explainable task | √ | X | X | √ | √ | √ | √ | √ | √ |
| Provide a Counterfactual Rule Generation based Explainable Method. | X | X | X | X | X | X | √ | √ | √ |
| Provide a visualization of Explainable mechanism for end users. | X | X | X | √ | √ | √ | √ | X | √ |

Table 3-1: Comparison between existing methods and proposed method

4. RESEARCH PROBLEM

How to get a counterfactual rule generation-based explanation for the Support Vector Machine classifier, when it handles non-linear separable data in text classification?

When it comes to the applications of Machine Learning models, most of the time those models are used in decision classifier systems. Therefore, the importance of the XAI is highlighted in those situations. As discussed in the previous sections, multiple approaches have been taken by the researchers to provide a proper explainable mechanism to explain those models. Most of these proposed solutions are still in the research phase because of the immaturity of the XAI domain.

This research mainly focuses to enhance the model interpretability of the SVM model related to a text classification task. As discussed in the background SVMs are popular in the text classification domain because of its ability to efficiently handle high-dimensional data, robustness to overfitting, and flexibility in kernel choice. When we consider about make the SVM model explainable, a counterfactual rule generation mechanism can use for it. Counterfactual explanations assist to make SVM models more interpretable by providing concrete examples of how a change in input variables would affect the output. Also, through these counterfactual rules, we can identify the key features that are driving those predictions and properly map the input features with the output result.

For Example, suppose there is a text classification problem that needs to classify customer reviews of a product as positive or negative. The dataset contains the customer reviews with their corresponding labels, and after training the SVM model it predicted that a particular customer review was positive. Here we can use counterfactual rule generation to generate a set of alternative reviews that would have been classified as negative by the model. These alternative reviews would differ from the original review in one or more key features, such as the presence or absence of certain words or phrases. By these counterfactual explanations, we can identify which words or phrases were most important in driving the model's positive classification for the original review. That information can use to provide an XAI solution for the SVM model that makes the model more interpretable.

When we consider the issues in the text classification domain being biased for a particular topic is major. It may cause for happen a fairness issue as well. There were some situations in the model trained on news articles that disproportionately includes articles from certain news sources or about certain topics. For instance, if the training data contains mostly articles from conservative news sources or about conservative politicians, the resulting model may perform poorly on articles about liberal politicians or liberal-leaning news sources. Similarly, if the training data includes mostly articles about crime, the model may be biased towards classifying all text related to crime as negative, even if some articles are discussing positive developments in crime prevention or criminal justice reform. Such biases can lead to inaccurate and unfair text classification results. So, it is very important to identify the biased classifiers and avoid inaccurate results while improving the fairness of decisions. For that purpose, we can apply counterfactual rule-generation techniques.

Counterfactual analysis assists to detect and mitigate bias in SVM text classification models by generating hypothetical scenarios where a particular protected attribute (such as race or gender) is changed. Here we can observe whether the model's output is affected by that attribute. If the model's output changes significantly when the protected attribute is changed, it may be a sign that the model is biased. Also, this Counterfactual analysis can be used to understand why a particular SVM text classification model is making certain errors. As the previous bias detection solution, it generates hypothetical scenarios to identify where the input is slightly changed to cause model misclassification. By observing what aspects of the inputs are causing the error we can identify areas for improvement in the model. Furthermore, this counterfactual analysis can use to evaluate the fairness of SVM text classification models. The rules can identify where certain groups are overrepresented or underrepresented and if the model is affected by the group differences.

In the XAI research area, there are some proposed counterfactual rule generation-based SVM model explainable methods. But most of those counterfactual rule-based mechanisms cannot be used in text classification tasks because the text classification scenarios contain high dimensional data. Also as mentioned in the research gap section model explainability visualizations provided by proposed methods are not user-friendly and understandable for end users. To overcome above discussed problems, it is important to provide an XAI solution to make the SVM model more explainable related to the text classification tasks.

5. OBJECTIVES

5.1 Main Objective.

Provide a novel post-hoc, model-specific, local XAI solution to enhance the model interpretability of function-based classification models focus on SVM by developing a novel counterfactual rule generation mechanism related to the text classification domain.

For generating counterfactual rules, we can use the neighbourhoods of given input instances. The current related works have been done using genetic algorithms to generate neighbourhoods in order to deduce the local rules. Therefore, in this approach, there will be a comparison between the method based on the genetic algorithms and the other methods. End of the research the most efficient and appropriate method will be selected, and the selected method will be applied to generate the neighbourhood from training data.

The identified neighbourhood will be passed to a SVM classifier to elicit the counterfactual rules. Eventually, the proposed solution will be visualized using the most appropriate Graphical User Interface (GUI) techniques that provide a better user experience to the end users. Through the visualization, the user can get a better understanding about the relationship between input features and how input features are mapped to an output.

5.2 Specific Objectives.

To achieve the main objective, there are few milestones to reach.

1. Prepare the dataset and implement the SVM classifier.

To perform the experiments, there should be identified dataset that aligns with the requirements. Twitter Sentimental Analysis and 20 Newsgroups datasets are used for this research. Twitter Sentimental Analysis dataset contains 1,600,000 tweets extracted using the Twitter API. By using this dataset, we can classify tweets into positive, neutral, and negative emotions. 20 Newsgroups dataset contain a collection of newsgroup documents. There is file (list.csv) that contains a reference to the document-id number and the newsgroup it is associated with. There are 20 document files that are related to different newsgroups, as one document for one newsgroup.

2. Develop the novel counterfactual rule generation mechanism related to the text classification task.

To generate counterfactual explanation rules I am going to use neighbourhoods of instances. Here the neighbourhoods have to be generated from the training data by considering the instance that is going to be predicted from the model.

3. Test the output with existing explainable methods.

4. Do experiments to improve the XAI solution more.

5. Do the visualization using the most appropriate Graphical User Interface (GUI) technique.

5.3 Work Breakdown Structure

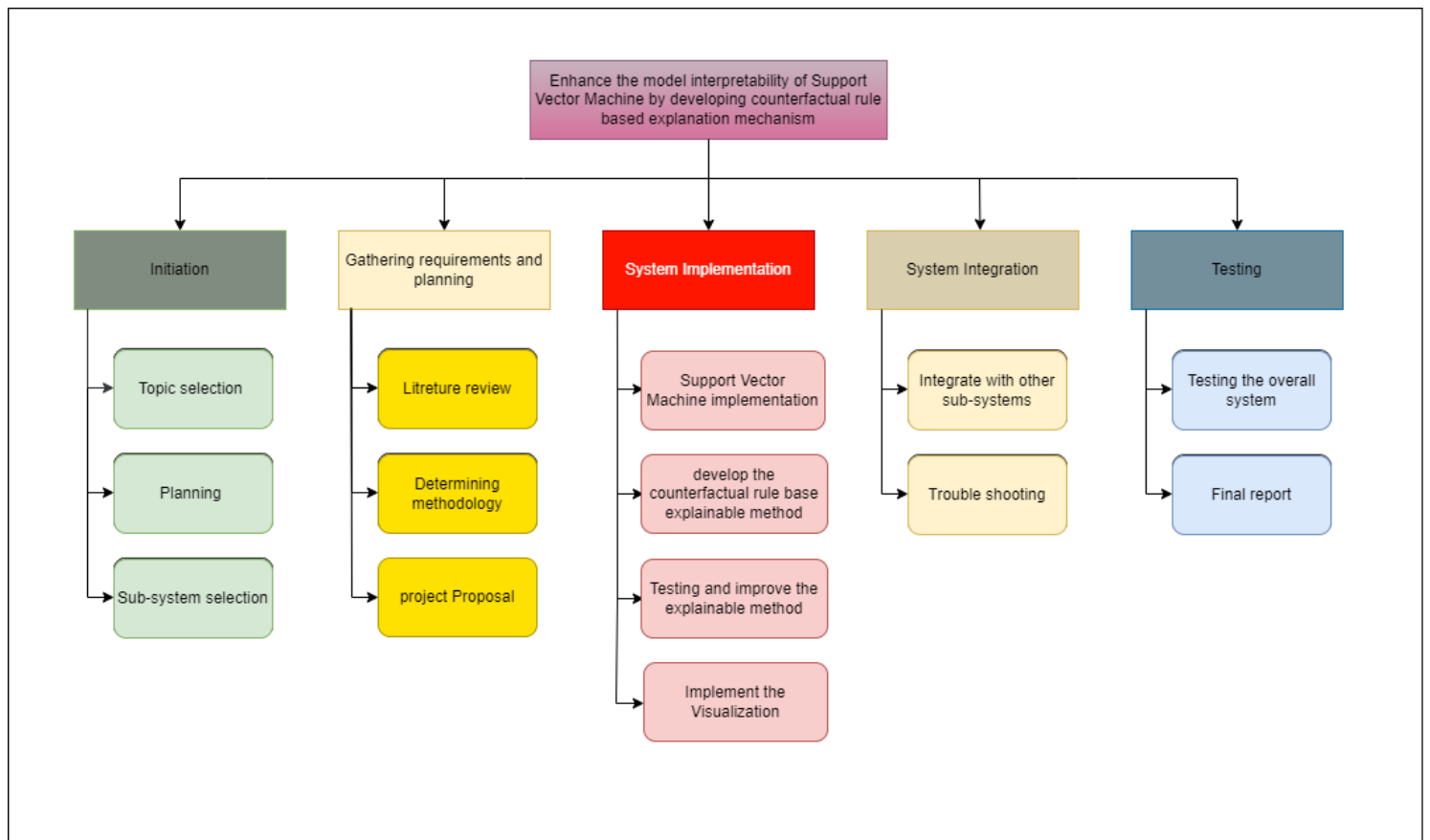


Figure 5.3-1: Work Breakdown Structure

6. METHODOLOGY

In order to implement the proposed solution, there are few milestones to accomplish. In this section, it will be discussed about the steps that need to be followed to carry out the study and the tools and technologies that are going to be required when accomplishing the sub-tasks.

6.1 System Architecture Diagram

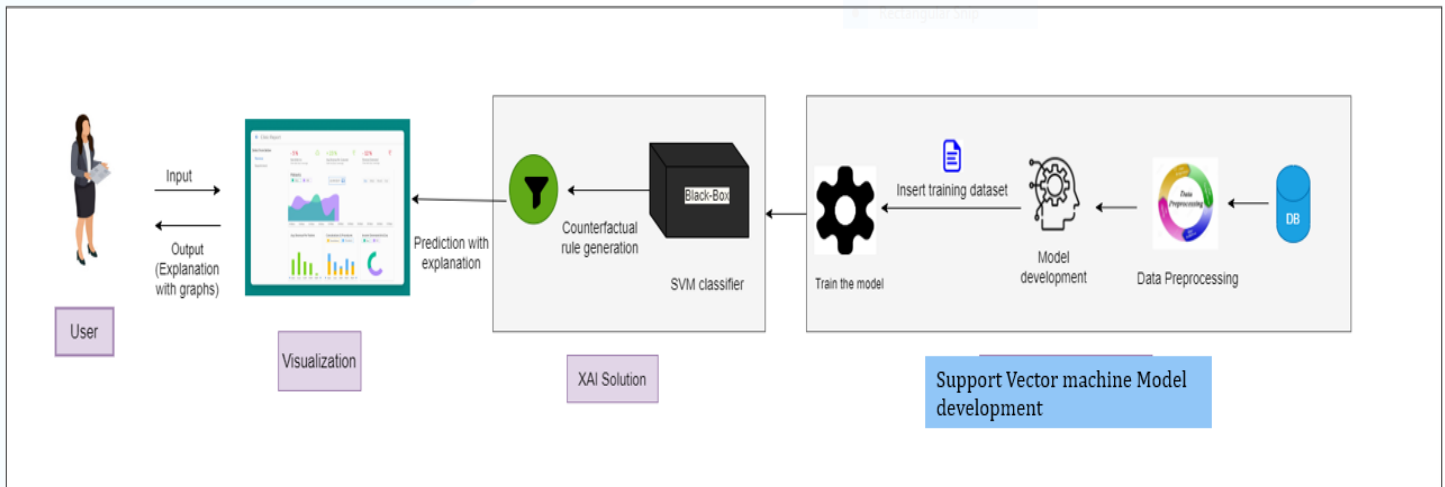


Figure 6.1-1: System Architecture Diagram

The above diagram illustrates the high-level system architecture of the proposed solution. As the initial step, the user has to provide the relevant training dataset and the instance that need to be predicted to the system through a GUI. The provided data will be applied to the support vector machine (SVM) model to get the prediction. To extract the explanation rules, a neighbourhood from the given training dataset will be generated and the extracted neighbourhood will be applied to a SVM classifier. Meanwhile, the counterfactual rules will be generated by the SVM classifier. After extracting the counterfactual rules, it is important to evaluate the novel explanation mechanism by testing with existing explainable tools and analyzing experts' feedback. Finally, the outcomes of the process (Counterfactual Rules) will be transferred to the GUI with appropriate visualizations to be more understandable for the users.

6.2 Tools and Technologies

➤ **Frontend:**

ReactJS and Bootstrap will be used as front-end technologies. Flask will be used to handle the communication between front-end and back-end.

➤ **Backend:**

Python will be used as the programming language to back-end development.

➤ **Version Control:**

Gitlab will be used as the VCS Platform.

➤ **Tools:**

VS Code and Google Colab will be used as editors for frontend backend developments.

7. PROJECT REQUIREMENTS

➤ User Requirements

- User should have a knowledge of decision-making systems based on machine learning.
- Sometimes the researchers will be the users.
- Dataset should be pre-processed, and appropriate data engineering techniques should be applied.
- Instance that needs to be predicted should be provided by the user.

➤ Functional Requirements

- Provide the counterfactual rules.
- System should be able to provide appropriate visualizations when needed.
- Model accuracies should be provided by the system.

➤ Non-Functional Requirements

- Output should be understandable.
- Visualization should be user-friendly, accurate and interactive.

8. GANNT CHART

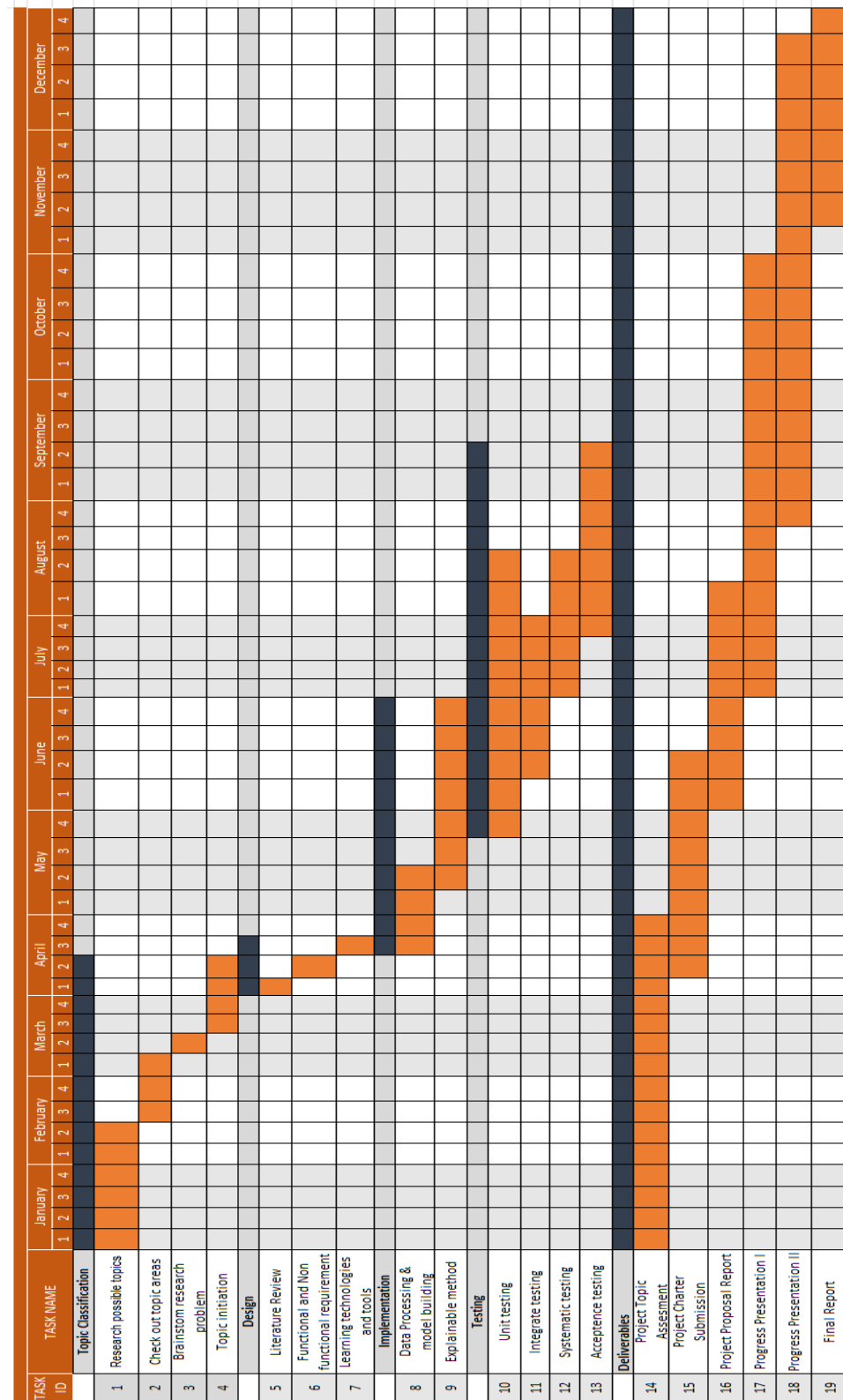


Figure 8-1: Tentative Gantt Chart

9. REFERENCES

- [1] M. T. Ribeiro, S. Singh, and C. Guestrin, ““Why should i trust you?” Explaining the predictions of any classifier,” in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Aug. 2016, vol. 13-17-Aug, pp. 1135–1144, doi: 10.1145/2939672.2939778.
- [2] Miller, Tim. "Explanation in artificial intelligence: Insights from the social sciences." arXiv Preprint arXiv:1706.07269. (2017).
- [3] Kim, Been, Rajiv Khanna, and Oluwasanmi O. Koyejo. "Examples are not enough, learn to criticize! Criticism for interpretability." *Advances in Neural Information Processing Systems* (2016).
- [4] A. Adadi and M. Berrada, “Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI),” *IEEE Access*, 2018, doi: 10.1109/ACCESS.2018.2870052.
- [5] S. M. Lundberg and S. I. Lee, “A unified approach to interpreting model predictions,” *Adv. Neural Inf. Process. Syst.*, vol. 2017-Decem, no. Section 2, pp. 4766–4775, 2017.
- [6] F. Shakerin and G. Gupta, “White-box Induction from SVM Models: Explainable AI with Logic Programming,” *Theory Pract. Log. Program.*, vol. 20, no. 5, pp. 656–670, 2020, doi: 10.1017/S1471068420000356.
- [7] M. T. Ribeiro and C. Guestrin, “Anchors : High-Precision Model-Agnostic Explanations,” pp. 1527–1535.
- [8] R. Guidotti, A. Monreale, S. Ruggieri, D. Pedreschi, F. Turini, and F. Giannotti, “Local rule-based explanations of black box decision systems,” *arXiv*, no. May, 2018.
- [9] R. K. Mothilal and C. Tan, “Explaining Machine Learning Classifiers through Diverse Counterfactual Explanations.”, 2019
- [10] Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., ... & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information fusion*, 58, 82-115

APPENDICES

| ORIGINALITY REPORT | | | |
|--------------------|--|--------------|----------------|
| 10% | 7% | 6% | 5% |
| SIMILARITY INDEX | INTERNET SOURCES | PUBLICATIONS | STUDENT PAPERS |
| PRIMARY SOURCES | | | |
| 1 | Submitted to Sri Lanka Institute of Information Technology Student Paper | 3% | |
| 2 | arxiv.org Internet Source | 1% | |
| 3 | hdl.handle.net Internet Source | 1% | |
| 4 | wintergreenresearch.com Internet Source | 1% | |
| 5 | research.library.mun.ca Internet Source | <1% | |
| 6 | Submitted to University of Wales Institute, Cardiff Student Paper | <1% | |
| 7 | Amina Adadi, Mohammed Berrada. "Peeking inside the black-box: A survey on Explainable Artificial Intelligence (XAI)", IEEE Access, 2018 Publication | <1% | |

Appendix A: Plagiarism Report