

INT2X: EXPLANATION FOR THE CAUSES OF A PREDICTION

2023-142

Project Proposal Report

Lakshani N.V.M.

**B.Sc. (Hons) Degree in Information Technology Specialized in Software
Engineering**


Department of Information Technology

**Sri Lanka Institute of Information Technology
Sri Lanka**

March 2023

DECLARATION

I declare that this is my 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 my 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.

Name	Student ID	Signature
Lakshani N.V.M.	IT20013950	

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

.....
Signature of the Supervisor

.....
Date

.....
Signature of the Co-Supervisor

.....
Date

ABSTRACT

With the development of the technologies, artificial intelligence has become more useful and powerful due to ability of automation systems with it. Even though AI made systems automated and easier to use, people started to recognize downside in it. Users find that it is not trustworthy when it comes to critical systems such as medical field and defense. These areas are very sensitive and needs more control over the decisions taken in these fields. As most of the machine learn models are black box, the output given by the AI systems cannot be trusted within these fields.

With the European Union's General Data Protection Regulation (GDPR), engineers tend to find solutions to turn these black box modals to white box modals. With this, Explainable AI were developed in 2016. XAI gave transparency for the black box modals. XAI is a way of explaining how the system predicts the outcome by combining the input and output.

K-Nearest Neighbor is also one of the modals which becomes a black box when there are many dimensions also known as curse of dimension in KNN. How KNN will be explainable for text classification by developing a novel counterfactual rule generation mechanism is addressed by this research.

Key words: KNN, Artificial Intelligence, Machine Learn, XAI

Table of Contents

DECLARATION	2
ABSTRACT.....	3
LIST OF ABBREVIATIONS	7
1. Introduction	8
1.1. What is XAI?	8
1.2. Research Area	9
2. Literature Review	10
2.1. Background	10
2.1.1. K-Nearest Neighbor	11
2.1.2. Counterfactual Explanation	11
2.2 Literature Survey	12
2.2.1. “Why should I trust you?” Explaining the predictions of any classifier	12
2.2.2. A unified approach to interpreting model predictions	13
2.2.3. “Why Should You Trust My Explanation?” Understanding Uncertainty in LIME Explanations	13
2.2.4. Using XAI Techniques to Persuade Text Classifier Results: A Case Study of Covid-19 Tweets 13	
3. Research Gap	14
4. Research Problem	16
5. Objectives.....	17
5.1 Main Objective	17
5.2 Specific Objectives	18
6. Methodology.....	19
6.1. System Architecture Diagram	19
6.2. Tools and Technologies	19
7. Requirements.....	20
7.1. User Requirements	20
7.2. Functional Requirements	20
7.3. Non-Functional Requirements	20
8. Gantt Chart.....	21
References	22
APPENDICES	23

Table of Figures

Figure 1: Work Breakdown Chart	18
Figure 2: System Architecture Diagram.....	19
Figure 3: Gantt Chart	21
Figure 4: Plagiarism Report.....	23

List Of Tables

Table 1: List of Abbreviations	7
Table 2: Existing tools	14
Table 3: Technologies	19

LIST OF ABBREVIATIONS

Table 1: List of Abbreviations

Abbreviation	Description
XAI	Explainable Artificial Intelligence
AI	Artificial Intelligence
GDPR	General Data Protection Regulation
KNN	K-Nearest Neighbor
SHAP	Shapley Additive explanations
LIME	Local Interpretable Model-Agnostic Explanations

1. Introduction

Explainable Artificial Intelligence (XAI) is becoming one of the most important technologies in coming future. This is because lack of trustworthiness of the predictions or the decisions given by the artificial intelligence (AI). When we consider about crucial and sensitive systems like cancer detections, defense systems, self-driving cars, finance systems, it is very important to take the most correct decision when they operate. But if these systems cannot be trusted or giving decisions with biasness and without considering important facts, these systems cannot no longer exists withing the commercial world. This is when XAI comes to the picture to solve this problem.

The requirement of the XAI comes to the image with the European Union's General Data Protection Regulation (GDPR) in 2016. GDPR stipulates a right to obtain "meaningful information about the logic involved" also known as the "right to an explanation" for consumers regarding automatic decisions [4]. They want an explanation of the decisions taken by machine learning models. This law affected the whole AI development, because all the models are black box type and even the developer are not aware of the internal functionalities of the model.

On 21 January 2019, the French Data Protection Authority (*Commission Nationale Informatique et Liberté* – "CNIL") imposed a fine of € 50 million on Google for infringing the General Data Protection Regulation 2016/679 (the "GDPR").

People found that these black box testing models are difficult to trust. After this GDPR, engineers started developing XAI tools and techniques to make more trustworthy systems.

1.1. What is XAI?

Explainable artificial intelligence (XAI) is a set of processes and methods that allows users to trust the output, or the result created by machine learning algorithms. XAI is used to describe an AI model, its expected output and what are the reasons affected in taking that decision. Simply, XAI enables transparency throughout the machine learning models by giving analytical and logical reasonings which has affected taking the final decision for the mentioned problem. With XAI, black-box-type models have transformed into white-box models. So, with the development of XAI, using machine learning for systems like healthcare, defense, manufacturing, and autonomous vehicles have vastly increased.

1.2. Research Area

K-Nearest Neighbor (KNN) is one of the machine learning models which exist as a black box when the curse of dimension problem occurs. This research is intended to make this black box instance of KNN model to white box. The KNN algorithm decides a number k which is the nearest neighbor to that data point that is to be classified. The new system which is proposed to develop will use KNN algorithm for text classification to make it explainable by developing a novel counterfactual rule generation mechanism [3].

2. Literature Review

2.1. Background

With the European Union's General Data Protection Regulation (GDPR), the attention and focus of the AI development have moved to the concept of right to obtain "meaningful information about the logic involved". After this, engineers were more focused on developing systems to achieve this goal. That's when XAI were developed in 2016 with the intension of achieving the goal of explaining ability to the consumer. The concept of XAI is to explain what's happens within the black box model and give a brief understanding to the consumer why the decision was made.

These XAI models can be,

- I. model agnostic or model-specific,
 - a. Model agnostic - More general and can be applied to any machine learning model.
 - b. Model specific - Designed for a particular machine learning model or algorithm.
- II. post-hoc explainable or intrinsically explainable and,
 - a. Post hoc-Explain a trained model's decisions after it has been trained.
 - b. Intrinsic-Explain a model's decisions by analyzing its internal structure and parameters.
- III. global explainable or local explainable.
 - a. Local- Explanations that are specific to a single instance or prediction made by the model.
 - b. Global-Provide insights into the overall behavior and performance of the model across the entire dataset.

There is not any defined way of selecting what method to be selected when developing an XAI model [3]. Therefore, the developers can decide how and what they need to develop by considering their requirements.

2.1.1. K-Nearest Neighbor

As AI was used widely within advance applications, more attention has been taken by the Data Mining Technologies than ever before, due to the growth of the social networks. As one of the most popular and widely used ML Models, KNN can be considered as a Black-Box model under some conditions.

When there are fewer attributes, the KNN algorithm behaves as a White Box Algorithm since it votes for the most common label in classification. It does not have a modular interpretability since it is a non-parametric algorithm, and its interpretability is based on the single instances of the data set used for the algorithm. Transparent methods like KNN provides justification with local weights of features. So, KNN satisfies properties names as algorithmic transparency, decomposability, and simulatability. Simple KNN supports transparency, algorithmic transparency, and human centric simulation. KNN's transparency depends on the features, parameter N and distance function used to measure similarity. Higher value of K impacts simulation of model by human users.

The Black-Box situation of KNN algorithm occurs when it has large number of dimensions which is referred to 'Curse of Dimensionality' problem in KNN. The dimensionality curse problem affects all indexes when a KNN query searches for all points that are "close" to a given query point. According to this problem, searching for response points becomes inefficient above a certain dimensionality since it is no more expensive than a simple sequential scan of the entire dataset. This has become one of the most important problems regarding the interpretability of k-NN algorithm since addressing the interpretability initially high dimensionality problem has to solve. The proposed new model will overcome this KNN black box issue for text classification.

2.1.2. Counterfactual Explanation

Counterfactual explanations are a type of explanation that can be provided by explainable artificial intelligence (XAI) systems. Counterfactual explanations are designed to help users understand why a particular decision or prediction was made by an AI model by considering what would have happened if the input data were different [6].

In other words, a counterfactual explanation describes how the output of an AI model might have been different if the input data had been different in some way. By providing this type of explanation, XAI systems can help users better understand the decision-making process of AI models and identify potential biases or errors.

Counterfactual explanations can be particularly useful in situations where an AI model is making decisions that have significant real-world consequences, such as in healthcare or finance. By providing users with counterfactual explanations, XAI systems can help build trust in AI models and ensure that they are making fair and accurate decisions.

A counterfactual explanation for a KNN prediction would involve finding a set of input data points that are similar to the original input data point but lead to a different prediction. This can be done by modifying the features of the input data point and rerunning the KNN algorithm to see how the prediction changes.

For example, suppose a KNN model is used to predict whether a loan application should be approved based on factors such as income, credit score, and employment status. A counterfactual explanation could be provided by identifying a set of similar loan applications that were denied, and then modifying the features of the input data point to see how the prediction would have changed if the applicant had a higher income or a better credit score [6].

Counterfactual explanations can be useful in helping users understand how a KNN model makes predictions and can also be used to identify potential biases or errors in the model. By providing counterfactual explanations, XAI systems can help build trust in KNN models and ensure that they are making fair and accurate decisions.

2.2 Literature Survey

Development of XAI has started recently. Even though XAI development is at its early stage, there are few tools and techniques developed to support the concept of Explainable Artificial Intelligence. Shapley Additive explanations (SHAP), Local Interpretable Model-Agnostic Explanations (LIME) are the most famous XAI tools developed within the past few years [5].

2.2.1. “Why should I trust you?” Explaining the predictions of any classifier

LIME is a post-hoc model-agnostic explanation technique which aims to approximate any black box machine learning model with a local, interpretable model to explain each prediction. The authors suggest the model can be used for explaining any classifier, irrespective of the algorithm used for predictions as LIME is independent from the original classifier. Ultimately, LIME works locally which means that it's observation specific and, just like SHAP, it will provide explanations for the prediction relative to each observation. What LIME does is trying to fit a local model using sample data points that are like the observation being explained. The local model can be from the class of interpretable models such as linear models, decision trees, etc. [7]

2.2.2. A unified approach to interpreting model predictions

The SHAP framework, proposed by adapting a concept coming from game theory, has many attractive properties. In this framework, the variability of the predictions is divided among the available covariates; this way, the contribution of each explanatory variable to each point prediction can be assessed regardless of the underlying model. From a computational perspective, SHAP returns Shapley values expressing model predictions as linear combinations of binary variables that describe whether each covariate is present in the model or not.

Likewise, there are a few other tools such as ELI5, What-if tool, Skater, AIX360 etc. But all these tools are model agnostic while most of them are globally interpretable. Also, these tools give explanations for many kinds of data such as numbers, texts, images, and audios. Which made it difficult to figure out how exactly these models work specifically for an algorithm and text classification. The following are some researches regarding XAI tools, concerns regarding explainability of the data sets and usage of counterfactual analysis in XAI.

2.2.3. “Why Should You Trust My Explanation?” Understanding Uncertainty in LIME Explanations

As mentioned before, LIME explains the prediction of a desired input by sampling its neighboring inputs and learning a sparse linear model based on the predictions of these neighbors; features with large coefficients in the linear model are then considered to be important for that input’s prediction. As mentioned in this document, training LIME explanations has uncertainty due to sampling around the inputs to generate the explanation. So, sampling can lead to statistical uncertainty in interpretation. The concept of XAI is developed to help users to have trust over the predictions given by the black-box models. But if there is any uncertainty within the explanation, users not only tend to doubt the tool, also the model it-self too.[1]

2.2.4. Using XAI Techniques to Persuade Text Classifier Results: A Case Study of Covid-19 Tweets

This research depicts how they developed a new framework for sentimental data analysis using machine learn models such as Naive Bayes, random forest, logistic regression, and support vector machine (SVM), as well as four deep learning RNN, LSTM, GRU, and Bi-directional RNN. They used SHAP and LIME for explaining the predictions. According to this framework, support vector machine has the best performance than the other models. SVM has an average of 86% accuracy while RNN is 78.4%, LSTM 78.8%, 78.6% is GRU and 79% for Bi-directional RNN. So, this framework is also using various models, explanations even though they specifically designed for text classification.[2]

3. Research Gap

There are a few XAI tools and frameworks that have been developed since 2016. But all these tools are developed in a way where all of them use many machine learn models for their explanations. The following table shows how the existing tools developed and work.

Table 2: Existing tools

	Model specific method	Provide a counterfactual rule generation based explainable method	Locally Explainable	Text classification	Provide a user-friendly visualization
LIME	✗	✗	✓	✓	✗
SHAP	✗	✗	✗	✓	✗
Diverse Counterfactual Explanations	✓	✓	✗	✗	✗
XAI360	✗	✓	✓	✓	✗
GoogleXAI	✗	✓	✓	✓	✗
Explaining and Improving Model Behaviour with K Nearest Neighbour Representations	✗	✗	✓	✗	✗
Proposed KNN Explainable Method	✓	✓	✓	✓	✓

As shown in the above table (Table 1), the existing XAI tools are supporting many machines learn models. Also, they are developed to support many classification types. So, the existing tools are not specific to a model or a classification.

The proposed model was designed as a model specific and local explainable. This model uses K-Nearest Neighbor as the model for text classification. Within this model, it is planned to do sentimental analysis by using the Twitter Sentimental Analysis dataset. So, in order to explain this text, within this research, it had planned to develop a novel post-hoc, model-specific, locally explainable XAI solution by developing counterfactual rule generation mechanism. So, this model will only focus on text classification through KNN model which makes it easier for the users to understand how the predictions were explained by the proposed system and how the results will be changed according to the changes in the inputs. This will reduce the uncertainty of the explanations given by the existing tools and provides more model specific way to explain the classification of text.

4. Research Problem

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

The proposed method is a model specific, locally explainable model, which follows a very different approach from existing tools such as LIME, SHAP etc. The most important difference between the proposed and existing tools is that the proposed model is only focused on text classification by using KNN.

We can use KNN for text classification tasks to make the model more interpretable. KNN is a simple and effective classification algorithm that can be used in text classification tasks. KNN can be used to identify the k nearest neighbors of a new data point and classify it based on the majority class of its neighbors.

To make the KNN model explainable, we can use counterfactual rule generation to generate a set of alternative reviews that would have been classified differently 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 classification for the original review.

Similarly, we can use counterfactual analysis to detect and mitigate bias in KNN 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 KNN text classification model is making certain errors.

To overcome the problems discussed in the research gap section, it is important to provide a user-friendly and understandable XAI solution to make the KNN model more explainable related to the text classification tasks.

5. Objectives

5.1 Main Objective

When the number of dimensions is low, KNN is considered as white box model. In such cases the result is clear and do not need an explanation about the result or the prediction. But when the number of the dimensions are high, the KNN model becomes a black box model. Which make it unable to understand how the predictions were made. Simply, when the number of dimensions is high, the transparency of the model disappears. So, in such cases having an explainable approach for the KNN is very important.

Even though there are many XAI tools which are supporting KNN model for text classification, it is hard to vary how they explain text classification through those models. As mentioned in the paper “Why Should You Trust My Explanation? Understanding Uncertainty in LIME Explanations” there are uncertainness within the explanations given by these tools [2]. These uncertainties occurred due to the randomness in sampling procedure, variation with sampling proximity, and variation in explained model credibility for different data points.

Therefore, the focus of this research is to minimize those uncertainties, biasness within the explanations given for the predictions of KNN model by developing a **novel post-hoc, model-specific, local XAI solution to enhance the model interpretability of function-based classification models focus on k-NN by developing a novel counterfactual rule generation mechanism related to the text classification domain**. Also, 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.

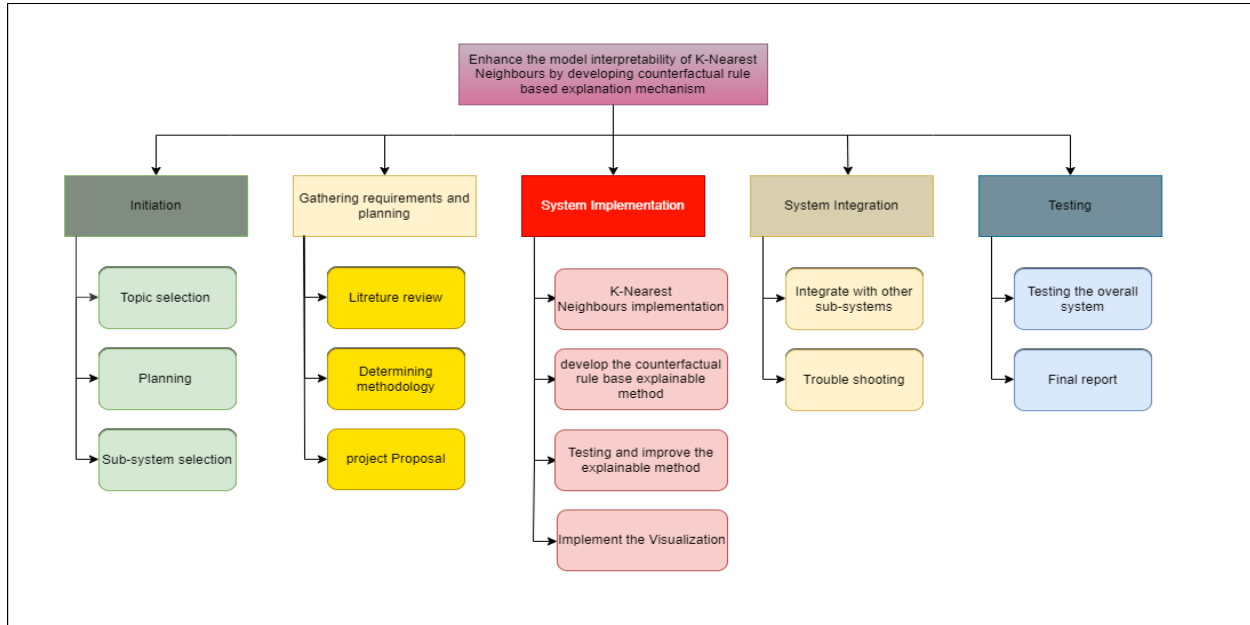
5.2 Specific Objectives

The sub-objectives mentioned below should be targeted to achieve the main goal.

- I. Exploring Prepare the dataset and implement the k-NN classifier.
- II. Develop the novel counterfactual rule generation mechanism related to the text classification task.
- III. Test the output with existing explainable methods.
- IV. Do experiment to improve the XAI solution.

5.3 Work Breakdown Structure

Figure 1: Work Breakdown Chart

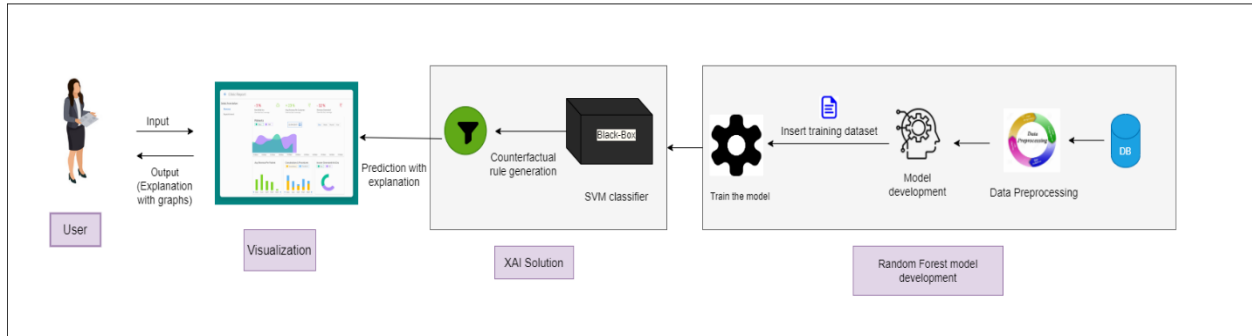


6. Methodology

There are few milestones that needs to be complete to develop the proposed system. These goals that needs to be achieved to complete this research is mentioned in this section.

6.1. System Architecture Diagram

Figure 2: System Architecture Diagram



6.2. Tools and Technologies

Table 3: Technologies

Frontend	<ul style="list-style-type: none">• ReactJS• Flask• Bootstrap
Backend	<ul style="list-style-type: none">• Python
Version Control	<ul style="list-style-type: none">• GitHub
Tool	<ul style="list-style-type: none">• VS Code• Google Colab

7. Requirements

7.1. User Requirements

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

7.2. Functional Requirements

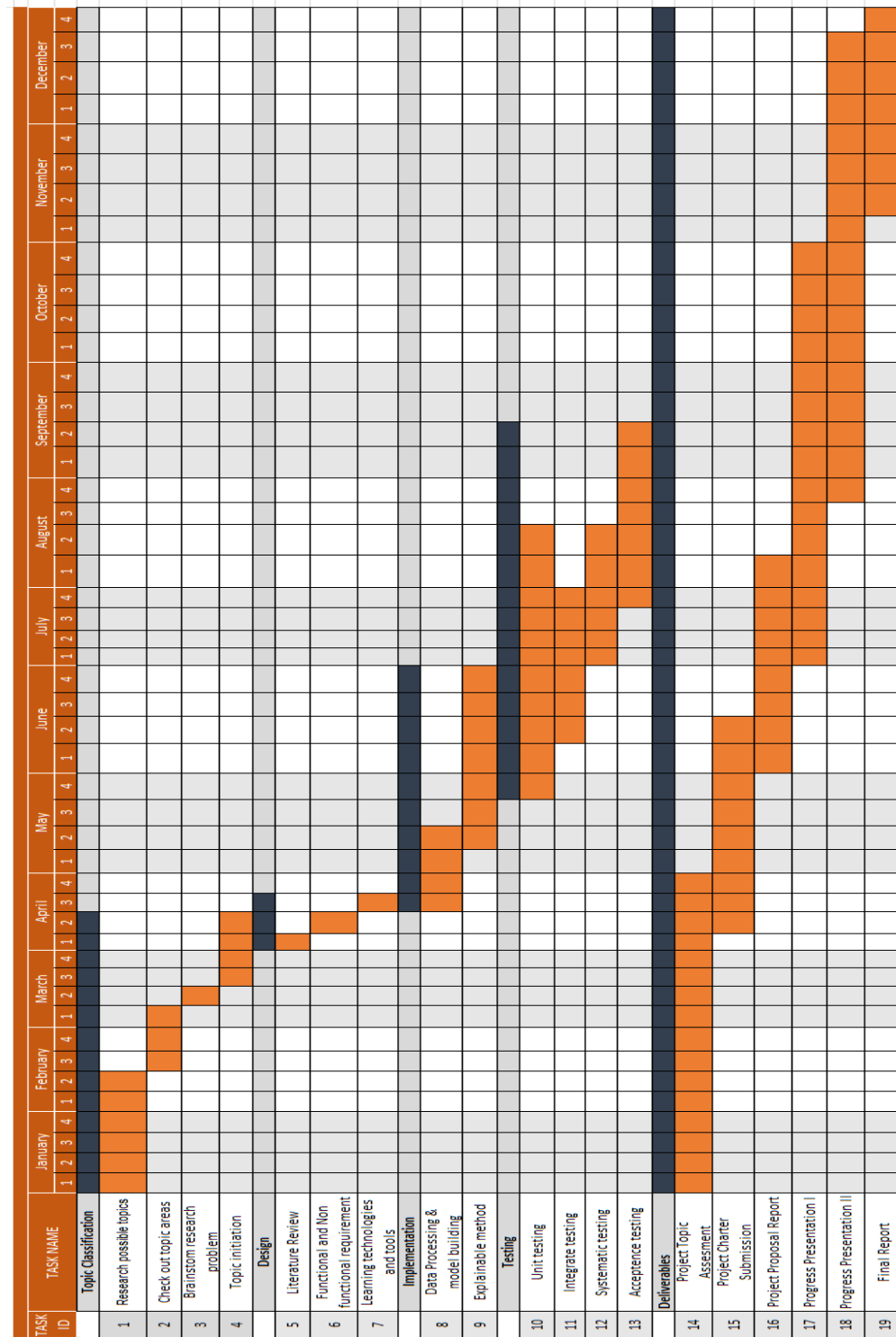
- I. Provide the counterfactual rules.
- II. System should be able to provide appropriate visualizations when needed.
- III. Model accuracies should be provided by the system.

7.3. Non-Functional Requirements

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

8. Gantt Chart

Figure 3: Gantt Chart



References

- [1] H. Hemdan, H. Elbakry, H. Elghareeb, S.S. Elhishi, "Using XAI Techniques to Persuade Text Classifier Results: A Case Study of Covid-19 Tweets." In Indian Journal of Science and Technology, August 2022.
- [2] Y. Zhang, K. Song, Y. Sun, S. Tan, M. Udell, "Why Should You Trust My Explanation? Understanding Uncertainty in LIME Explanations" in Cornell University, V2, June 2019.
- [3] P. Gohel, P. Singh, M. Mohanty, " Explainable AI: current status and future directions", July 2021.
- [4] R. Confalonieri, L. Coba, B. Wagner, T. R. Besold, "Introduction" in A historical perspective of explainable Artificial Intelligence. John Wiley & Sons, Ltd, Jan 2021.
- [5] S. M. Lundberg and S. I. Lee, "A unified approach to interpreting model predictions," in Proc. Adv. Neural Inf. Process. Syst., 2017, pp. 4768–4777.
- [6] Grath RM, Costabello L, Van CL, Sweeney P, Kamiab F, Shen Z, Lecue F. Interpretable credit application predictions with counterfactual explanations. arXiv preprint arXiv:1811.05245. 2018 Nov 13.
- [7] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 1135-1144). ACM.
- [8] Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. In Advances in Neural Information Processing Systems (pp. 4765-4774).

APPENDICES

Figure 4: Plagiarism Report

