Muqadas Awan
01-134181-050
Ahmed Hassan
01-134181-101

Entity Association Mining

Bachelor of Science in Computer Science

Supervisor:
Dr. Muhammad Muzammal

Department of Computer Science
Bahria University,Islamabad

January, 2022

Acknowledgement

"Starting with the Name of Allah who is the most merciful and the most beneficial."
First, we would like to thank Allah for giving us the opportunity to study in the Bahria University, Islamabad Campus and come to the stage to complete our Final Year Project. We have put all our efforts and knowledge gained during the whole degree in our project. We would like to thank our parents who have always been the pillars of the strength and support as well to taught us how hard work, devotion and working with the honest intentions can make us fly high. We would like to dedicate all our efforts, struggles and hardworking in our academic life to our dear parents and respected teachers, without them we are nothing. We would like to express our gratitude towards our supervisor **Dr. Muhammad Muzammal** for his support and guidance. We are truly thankful and appreciate his efforts. We would like to extend our sincere thanks to our class who has been with us from the start to the end of the degree who have been very supportive during the whole time and to the people who have helped us with their abilities in the project.

Muqadas Awan & Ahmed Hassan
Islamabad, Pakistan

January, 2022

Sponsor

The Final Year Project "Entity Association Mining" is sponsored by **NESCOM, Pakistan**. We would like to thank NESCOM organization for sponsoring our project. Also, we want to express our gratitude to their team for their support and guidance. Our project will be integrated with the NLP module of NESCOM organization. This was a one year paid project by NESCOM. All the functionalities incorporated in the project from requirements specification to methodology adopted are instructed by their team. In this one year with their team, we have learnt a lot of things from both academic and industrial point of view.

Muqadas Awan & Ahmed Hassan
Islamabad, Pakistan

January, 2022

"Man will not get anything unless he works hard."
(Surah al-Najm, 53:39)

Abstract

Newspaper is an important part of human life since the old times. First thing we do in the morning is to go through the newspaper. Everyone has their own way of keeping their self-informed about what is happening in the world. But all those efforts and struggles to keep in touch with what is happening around world are in vain until they do not know that about those people/events/ organization/country they are reading, are they associated with each other? In this busy world, no one has time to go through the news articles with the association perspective. But there is no platform available to analyze news articles and associate entities with respect to information stated in news articles.
EAM is a web platform for journalism community and those users who are interested in finding out association. The NEWS articles are extracted from five different international NEWS resources through web scraping, then it is passed through NLP pipeline to extract information from extracted NEWS articles. Then, entities are extracted from the extracted entities that are then is then normalized as per the occurrence in NEWS article which is then passed to FP-Growth algorithm that gives frequent item sets as output that is passed to association rules function to extract rules as association between entities, the output is decision matrix. Then the graph is generated from those extracted rules.

Contents

List of Figures

List of Tables

Acronyms and Abbreviations

AI Artificial Intelligence.

ASP Application Service Provider.

BFS Breath First Search.

CSS Cascading Style Sheet.

EAM Entity Association Mining.

FP Frequent Patterns.

HTML Hypertext Markup Language.

HTTP Hypertext Transfer Protocol.

KDD Knowledge Discovery in Database.

NER Named Entity Recognition.

NLP Natural Language Processing.

PHP Hypertext Preprocessor.

POS Parts of Speech.

SQL Structured Query Language.

TV Television.

WWW World Wide Web.

XML Extensible Markup Language.

Chapter 1

Introduction

Newspaper is an important part of human life since the old times. First thing we do in the morning is to go through the newspaper. Everyone has their own way of keeping their self-informed about what is happening in the world. Our young generation uses social media to keep them informed that what is happening around them. Elders watch news on TV and read newspapers of different companies all day. But main source of information is news resources.
But all those efforts and struggles to keep in touch with what is happening around world are in vain until they do not know that about those people/events/ organization/country they are reading, are they associated with each other? How important their association is? and how important are they (people/events/ organization/country) as individual? In a news resource.
In this busy world, no one has time to go through the news articles with the association perspective. People want results in hand, they are not interested in the process. But they are interested in digging out the association between entities. Not just two or three entities but to analyze how those entities are associated across different news articles of all different categories.
As every organization is shifting their data to online platform. News resources also publish their news articles in online platforms. There is a huge data of a news resource because there are thousands of news articles published by news resources every day and each news article of similar topic is telling a different perspective on an event or about association of two different people. Also, not only information across news articles is different but information on similar topics across news resources is different. Let us take example of Indian and Pakistani news resources news articles about Kashmir both countries news articles perspective is opposite to each other. We know that the more the data is, the better results in hand.

1.1 Problem Statement

In real world, there are many association rule mining systems are available but they are designed to analyze reviews of customers about different product[2].But there is no platform available to analyze news articles and associate entities with respect to information stated in news articles. The news analyzers must go through all news articles and draw maps to associate different people/events/organization/country and this waste a lot of their time. Also, while doing this they do not know that the association they are establishing is important or not. Or the method they are using to associate entities is helpful and valid or not. Due to this many of their efforts wasted when they get to know at the end of the

process that they have found wrong or invalid association which do not exist in real.

1.2 Objective

To create a web application which serves news analyzers and those people who are interested in finding out association between entities of news articles by providing a platform for observing association between entities features in a user-friendly manner by utilizing the best user experience practices in the industry.

- Provide an integrated platform which should cater all problems stated in the problem statement:
  - Integrated platform for all activities
  - Directed Graph system for users to analyze association between entities.
  - Also, a setting component in graph system to play with the graph like changing size of graph, 2D motion of graph etc.
  - Details system which contains information about each association between entities along with their support, confidence, and lift value.
- Extract information from NEWS articles.
- Design a model to detect entities.
- Design a model to associate entities.
- Design a model to create a graph of entities and their association.
- Platform where user can analyze association between different entities based on information extracted from news articles.
- Provide a user-friendly web application which focuses on UI and UX to maintain quality.
- Design and develop an application for journalism community to analyze association between entities.

1.3 Project Scope

The project is to provide people with a platform where they can analyze association between entities based on information extracted from news articles in the form of a web application. The idea behind this application is to give our journalism community a platform to analyze news articles with the perspective of association between entities.

- The main purpose of this application is to introduce a system that can classify entities and how these entities are associating with each other on websites, NEWS articles.
- This model can help to track that how entities are associated with each other on the basis of information provided by NEWS articles.
- This model is designed to create directed graph of entities and their association.

The project aims is to deliver a multi-platform web application with a single code-base able to run on both desktop and mobile. An online platform for journalism community to analyze association between entities.

1.4 Solution Application Areas

Entity Association Mining is a web platform for journalism community and those users who are interested in finding out association, they can analyze association between different entities in different news resources. In EAM, entities are associated based on NEWS articles data extracted from different news resources. This system is specially designed for news analyzers, journalist and a big community of world who is closely related to NEWS.

1.5 Feasibility Study

1.5.1 Product Feasibility Analysis

In this modern world, where every problem is solved through technology. There are still some problems that needs attention. One of them is association of entities based on the information extracted from NEWS articles. When this problem is solved, user does not need to serve a lot of time to analyze NEWS articles to extract entities and their association.

1.5.2 Target Market Feasibility Analysis

The target market of our application is journalism community that needs to analyze NEWS articles in more critical manner to extract entities and their association and create directed graph based on association found.

Chapter 2

Literature Review

In this chapter, we will discuss the concepts used to develop this system. We will also discuss the algorithm that will be used to develop this system. A brief discussion on the existing system and proposed system. At the end we will also draw a conclusion about the overall discussion.

2.1 Web Scraping

Web scraping is a method to extract data from websites. The information extracted from the websites is then exported into a format that is more useful to the user for other purposes. Web scraping of a web page involves it and extracting information from it. Fetching is the process of downloading a web page. So, web crawling is the most important component of the web scraping. Once the web pages are fetched, the next step is to extract data from those fetched web pages. The information of the web pages can be parsed, searched, reformatted, it data copied to spreadsheet to create csv or can be saved into notepad files. Web scraper basically take out information from the web pages to further use it for another useful purpose.

2.2 Natural Language Processing

NLP is a branch of AI that deals with text. It strives for human-like performance. It is mainly used in the pre-processing of raw data(text) to further use that data to extract implicit information from it. The main goal of NLP is to accomplish human-like language processing [4]. It is used in many areas like for information retrieval, sentiment analysis, information extraction, machine translation, question answering etc.

2.3 Data Mining

Data mining is a process of extracting interesting implicit information from raw data. Also, Data Mining is known as core process of KDD [9]. In the process of data mining, first raw data is cleaned and went through the pre-processing process. Then some algorithms are used to train the model using that preprocessed data. And then that model is used to predict against input. Also, model is evaluated using test data.

2.4 Association Rule Mining

Association rule mining is one of the types of data mining. It aims to extract interesting correlations, frequent patterns, associations, or casual structures among sets of items in the transaction databases or other data repositories [9]. It is an if-then statement which helps to find hidden association between different items from the given data. It is used to uncover relationship between item sets that are frequently appear together in data [3]. Rules are not formed based on how frequent two or more entities appear in the same instance together. But some measurements are used to confirm that this association exist in real or not. Those measurements are as follows:

2.4.1 Support

Support is defined as a measurement used to know how frequently an item set has appeared in data. Let us suppose that there is an association rule A→B. So, its support will be calculated as:

$$Support(A \rightarrow B) = \frac{Transactions\ containing\ both\ A\ and\ B}{Total\ number\ of\ transactions}$$

Value of support help us to find out associations that will further help us for analysis. If a rule has lower support value than it means that we have or there exist less information about that relationship hence no conclusion can be drawn. Value of Support lies between 0 and 1.

2.4.2 Confidence

Confidence is defined as a measurement used to find that how frequent consequent appears when antecedent is already in the data. Let us suppose that there is an association rule A→B. So, its confidence will be calculated as:

$$Confidence(A \rightarrow B) = \frac{Transactions\ containing\ both\ A\ and\ B}{Transaction\ containing\ A}$$

This is not important that what you have in the antecedent for such a frequent consequent. The confidence for an association rule having a most frequent consequent will always be high [1]. Considering just the value of confidence limits our capability to make any business inference. Value of confidence lies between 0 and 1.

2.4.3 Lift

Lift is defined as a measurement used to control value of support of consequent while calculating conditional probability of occurrence of consequent given antecedent. In literal meaning it is lift that provide antecedent confidence for having consequent in the data. Let us suppose that there is a rule A→B. So, its lift will be calculated as:

$$Lift(A \rightarrow B) = \frac{(Transactions\ containing\ both\ A\ and\ B)/(Transaction\ containing\ A)}{Fraction\ of\ Transaction\ containing\ B}$$

More the value of lift, there is more chances that where there is A, there is also B. Value of lift lies between 0 and infinity.

- If value of lift is greater than 1 than rule and its antecedent appear more often than expected. This means that rule has positive effect on occurrence its antecedent.

- If value of lift is less than 1 than rule and its antecedent appear less often than expected. This means that rule has negative effect on occurrence its antecedent.
- If value of lift is near to 1 than rule and its antecedent appear as often as expected. This means that rule has no effect on occurrence its antecedent.

2.5 Association Rule Mining Algorithms

2.5.1 Apriori Algorithm

Apriori algorithm is used to extract frequent item sets and association rule learning. The algorithm works in a level-wise search where k-item sets are used to explore (k+1) item sets to mine frequent item sets from transactional dataset. This algorithm uses BFS method and hash tree structure. There are two drawbacks of this algorithm that are as follows:

- Complex candidate generation process which uses a lot of time, space, and memory.
- It requires multiple scans of datasets [3].

2.5.2 FP-Growth Algorithm

In FP growth algorithm, an FP tree is constructed. To construct that tree, it needs two passes. It uses divide and conquer strategy. It extracts rules by scanning dataset two times. In first scan, it extracts all frequent item sets sorted in descending order. In second scan, it compresses dataset into FP tree. This algorithm performs mining on FP tree recursively. This algorithm extracts rules faster as there is no candidate generation in it. There are two subprocess of FP generation process those are:

- Construction of FP tree.
- Generation of frequent patterns from the FP tree.

FP tree is constructed in two passes. That are as follows:

Pass 1:

1. Scan data and find support of each item
2. Discard infrequent items
3. Sort frequent items in descending order based on their support value

Using the above steps FP tree can be built

Pass 2:

1. Nodes corresponds to item and it has a counter.
2. This algorithm reads one transaction at a time and then map it on the path.
3. Fix order will be used, so paths can overlap when transactions share same items/entities [3].

2.6 Comparison between Apriori and FP-Growth Algorithm

In Table 2.1, there is comparison between two most popular algorithm of association rule mining: Apriori and FP-growth algorithm.

Table 2.1: Comparison Between Apriori and FP-Growth Algorithm

Characteristics	Apriori Algorithm	FP-Growth Algorithm
Data Support	Limited	Very Large
Speed in initial phase	High	High
Speed in later Phase	Slow	High
Accuracy	Less	More Accurate

2.7 Proposed System

As there is a lot of work has been done on association rule mining and some desktop applications have been developed, but that are used to perform association rule mining on customer reviews about product not NEWS articles [5]. Also, it uses apriori algorithm that has drawbacks. Also, this system only provides Decision matrix, no visualization. Figure 2.1 show GUI of existing system. Our proposed system focuses on the NEWS articles ex-

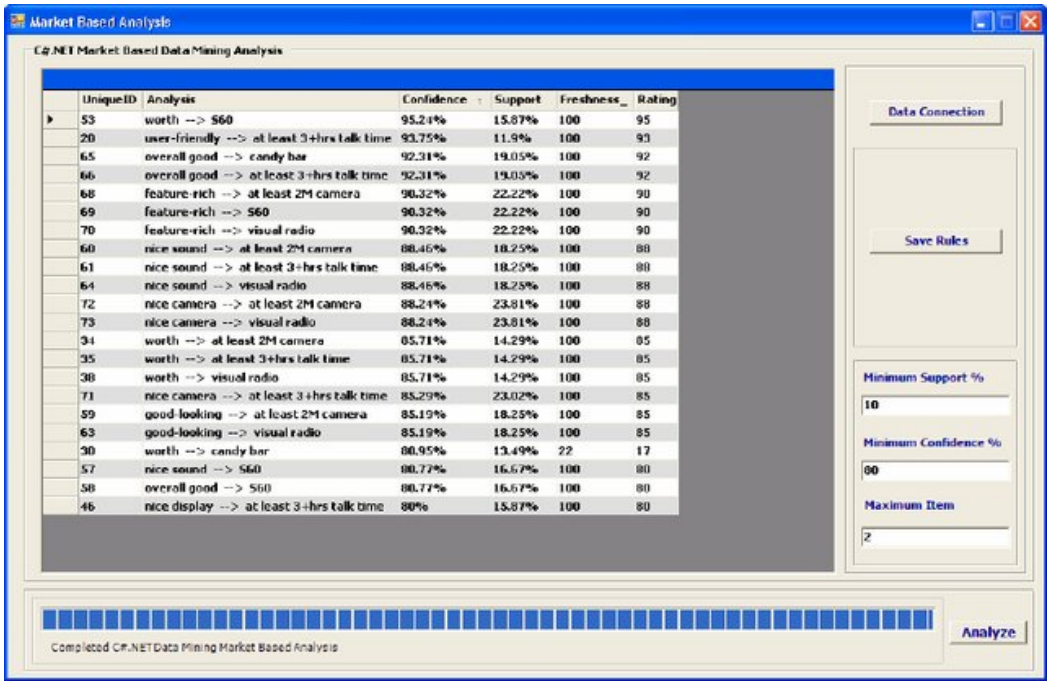


Figure 2.1: Existing System

tracted from different NEWS resources. The user will select news resource. Then a directed graph of huge number of entities extracted and their association will be generated along with decision matrix. Real time association will be provided to user. Existing systems do not provide facility to journalism community. Besides that, proposed system provide user

with visualizing association between entities, to better understand the relationship between entities. In Table2.2, there is comparison between proposed system and existing system.

Table 2.2: Comparison between Proposed System and Existing System

Proposed System	**Existing System**
Collects complete data from different NEWS resources which in result gives a complete visual association graph entities extracted from each NEWS resource.	Collects data about reviews of customers about mobile phones of specific company.
Provide complete data from different NEWS resources.	Provide data about a product of specific company.
Filtered NEWS articles with no repetitions.	Possibility of similar review from customers.
Web-based application	Desktop-based application
Entities and their associations can be visualized along with decision matrix.	Decision matrix is provided.

Our proposed system also focuses on GUI and make it user-friendly for Association rule mining. Our system will be a standalone system to not engage user in other applications. This will help user to minimize distraction and will increase productivity of knowledge.

2.8 Conclusion

After all the above discussion, we have come to conclude that technology is becoming an important part of life without that our survival is difficult in this world and everyone is busy in their lives. So, our proposed system will facilitate our journalism community by reducing their workload. Our proposed system will help user to analyze the entities and their association and make use of that information for their own purpose.

Chapter 3

Requirements Specifications

3.1 Existing System

For requirements specification, we used the existing system of "Market Basket Analysis" [5]. It is designed for association rule mining of reviews of customers about a mobile phone.

3.2 Proposed System

Our Proposed system is a web-based user-friendly system with modern and interactive interface that will allow user to analyze entities extracted from NEWS articles and their association in a directed graph. The user will select the NEWS resource and select the value of support and confidence, then a directed graph will be generated of extracted entities from selected NEWS resource and their association. A user can also drag the nodes of graph anywhere on the screen. Also, by changing the setting of graph, there will be change in the nature of the graph. User can also view support, confidence, and lift of the rule in decision matrix.

3.3 Intended Audience

This system is created for web users. The targeted audience of our system is an individual, organization, journalism community, politicians and others who want to analyze the NEWS articles with the perspective of finding association between entities.

3.4 Functional Requirements

3.4.1 Sign up

- User that is visiting our website first time will must sign up.
- For Sign up, user will fill sign up form and submit it.
- Admin will approve each user on sign up request.

3.4.2 Log in

- To get full access on website, user will need to log in.

- User will log in with the same email and password, he/she entered at the time of sign up.
- If user has entered correct email and password, then he/she will be logged in.
- Otherwise, an error message will be sent back to user.
- If log in is successful, it will take user to home page.

3.4.3 Generate graph

- User will select the NEWS resource and minimum value of support and confidence of the rule.
- Then Graph page will be loaded.
- In graph page, there is a directed graph of extracted entities from NEWS articles of selected NEWS resource and their association.

3.4.4 Setting of Graph

- User can change setting of graph.
- Then, there will be changes in graph physics, not in the entities and their association.

3.4.5 View Details of Graph

- User can view Decision matrix in Details page.
- In Detail page, all rules of graph along with their antecedent, consequent, antecedent's support, consequent's support, rule's support, confidence, and lift.

3.4.6 Search Entity

- After selecting NEWS resource and value of support and confidence, User can search entity with in that selected NEWS resource
- Then, a directed graph of the entities associated with that searched entity will be generated.

3.4.7 Log out

- User can log out at any time, he/she wants.

3.5 Non-Functional Requirements

3.5.1 Performance

The performance of a system is evaluated by its response time for a task.

- Our System allow user to work without any delays
- If a delay occurs, a processing bar will appear on screen to inform user that how much time will it take.

3.5.2 Responsiveness

A good system should always be responsive to the actions of the user. Our system is responsive towards user request. For example, user initiate the request to create new account or wants to log in

3.5.3 Security

The data used in our system is authentic and secure.

- Also, it keeps the data of user safe (Users data entered at the time of sign up).

3.5.4 Quality

Quality of the system is measured by the full filling the needs of the user and deliver a system with acceptable level of errors.
Usability: GUI and overall design of the system is kept simple and user friendly. Less complexity ensures a quick learning of user towards system. The usability can be defined in terms of:

- **Effectiveness:** Our system can give satisfactory and correct result against the task being performed.
- **Efficiency:** Our system can give effective result against the task performed in minimum time by utilizing minimum resources.
- **Easy to learn:** The system design is kept simple to make for ease for the user to learn it faster.

3.5.5 Availability

User can access website anytime he/she wants.

3.5.6 Dependability

As the system is web-based so user should have a good internet connection.

3.5.7 Scalability

The system is fully capable of handling huge amount of data.

3.6 Constraints, Assumptions and Dependencies

Every system has some constraints, assumptions, and dependencies against requirements. The constraints, assumptions, and dependencies of proposed system are as follows:

3.6.1 Constraints

- User should have a good internet connection.
- To get better results, there should be large amount of data available.
- The application should be used by those people who have at least a little knowledge of association rule mining.
- The data should be binomial as an input for Association rule mining algorithm.
- Development team should be familiar with the languages used to design the system. For Example, Python, Java script, HTML, CSS, PHP etc.
- Development should have knowledge about web scraping, NLP, and association rule mining.

3.6.2 Assumptions

- Our system will be able to generate 80
- The system will extract nouns from the articles.
- It should ignore date and place given at the start of every NEWS articles.

3.6.3 Dependencies

- The data set used for the system should be correct.
- Good internet connection is required to operate the website.

3.7 Use Cases

Use case is said to be a scenario written about how user will interact with the system and how system will perform a task. Every use case is described in a sequence of steps, that begin with user input and end with the results provided by system. The use cases that have been identified are as follows:

1. User sign up
2. User Log in
3. Generate Graph
4. Setting of graph
5. View Details of graph
6. Search entity
7. User Log out

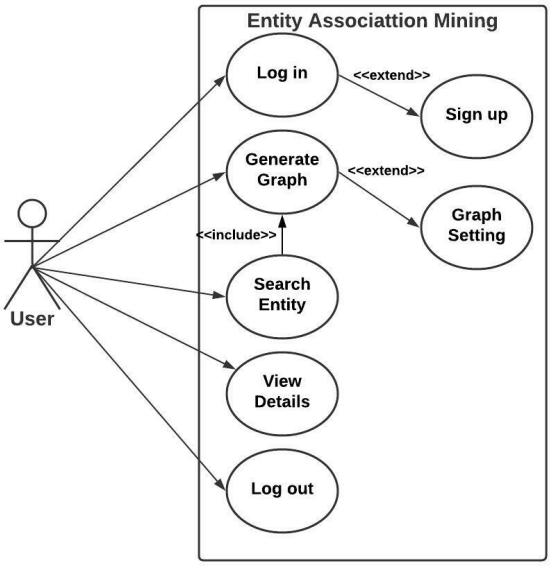


Figure 3.1: Main Use Case Diagram

Figure 3.1 shows the use case diagram of the whole system where user can log in, if user does not have account then he needs to first create the account. User can generate graph, search entity, view details and log out.

3.7.1 User Sign up

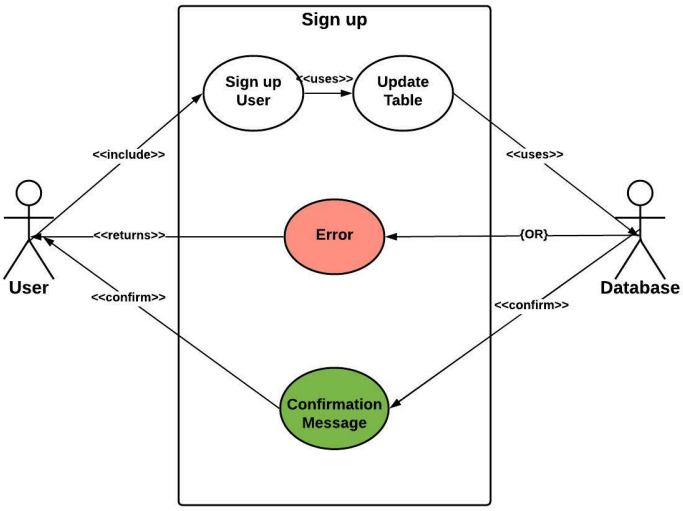


Figure 3.2: Sign up Use Case Diagram

In Figure 3.2, there are two actors User and database, first user will sign up then the table in database will be updated. After that a confirmation message will be sent back to user.

Table 3.1: Sign up Use Case Description

UC Name	Sign up
UC ID	01
Priority	High
Primary Actor	User
Description	This use case describes how user can sign up to the website. If user is visiting the website for the first time.
Basic Flow	User will go to sign up page and entered the required information to create an account.
Alternate Flow	Invalid credentials entered User will be prompted with error message
Pre-condition	Sign up page should be opened in user screen. Device should have a good internet connection
Steps	**Actor Actions:** Step 1: User will enter information required on the sign-up page. Step 2: User will make sure that information entered is correct. Step 3: User will click on the sign-up button **System Response:** Step 4: System will check if the information entered by user is correct Step 5: User will be directed to home screen
Post-condition	Home screen will be displayed

In Table 3.1, use case of sign up is described in detail. It includes all steps and constraints to be followed during the user will sign up.

3.7.2 User Log in

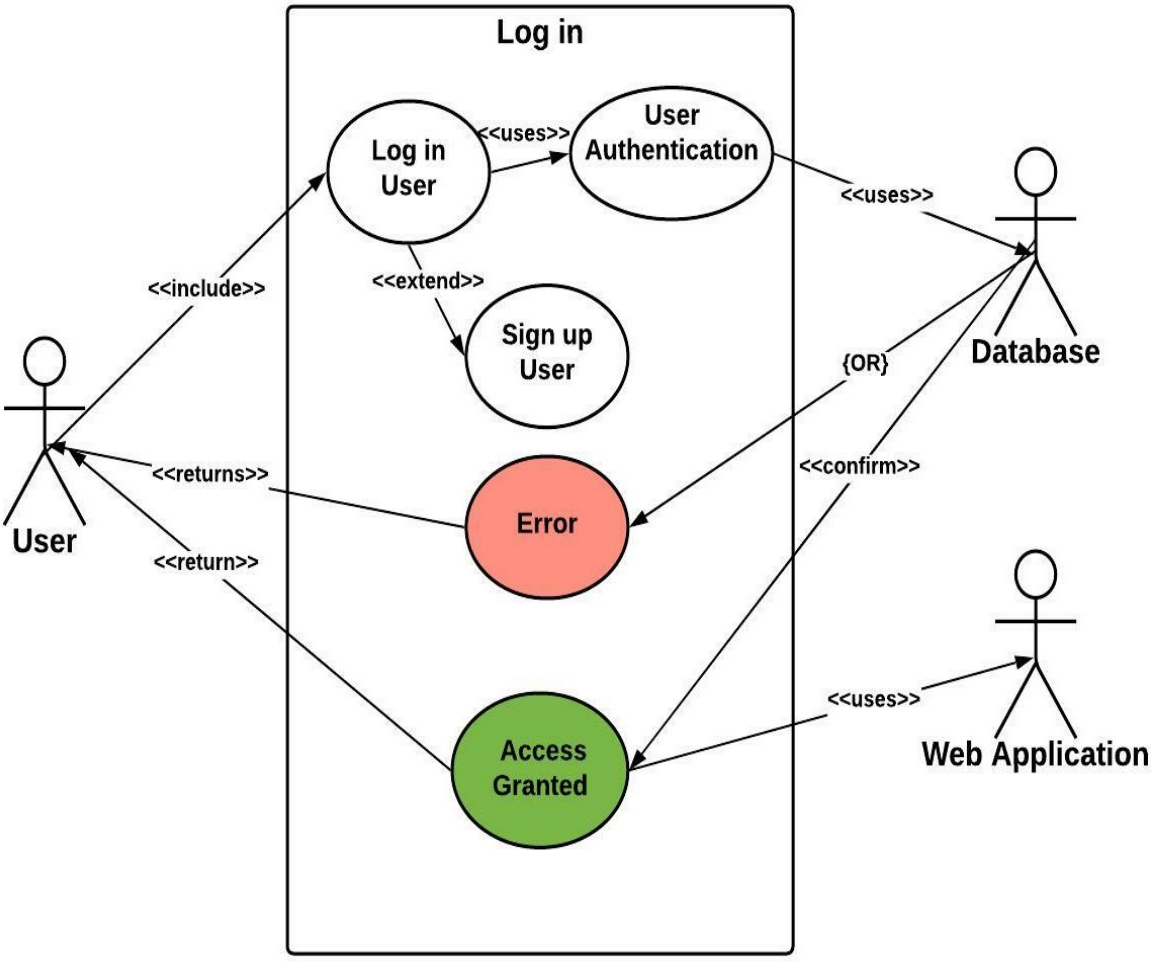


Figure 3.3: Log in Use Case Diagram

In Figure 3.3, there are three actors: User, Database, and Web application, user will first fill up log in form, then database will verify the user. If the user entered right credentials, user will be directed to web application. Else an error message will be sent back.

Table 3.2: Log in Use Case Description

UC Name	Log in
UC ID	02
Priority	High
Primary Actor	User
Description	This use case describes how user can log in to the website. If user already have an account affiliated with the website.
Basic Flow	User will go to log in page and entered username and password.
Alternate Flow	Invalid username or password entered User will be prompted with error message
Pre-condition	Log in page should be opened in user screen. Device should have a good internet connection
Steps	**Actor Actions:** Step 1: Log in page should be opened in user screen. Step 2: User will make sure that username and password entered are correct. Step 3: User will click on the log in button **System Response:** Step 4: System will check if the username and password entered by user is correct Step 5: User will be directed to home screen
Post-condition	Home screen will be displayed

In Table 3.2, use case of log in is described in detail. It includes all steps and constraints to be followed during the user will log in.

3.7.3 Generate Graph

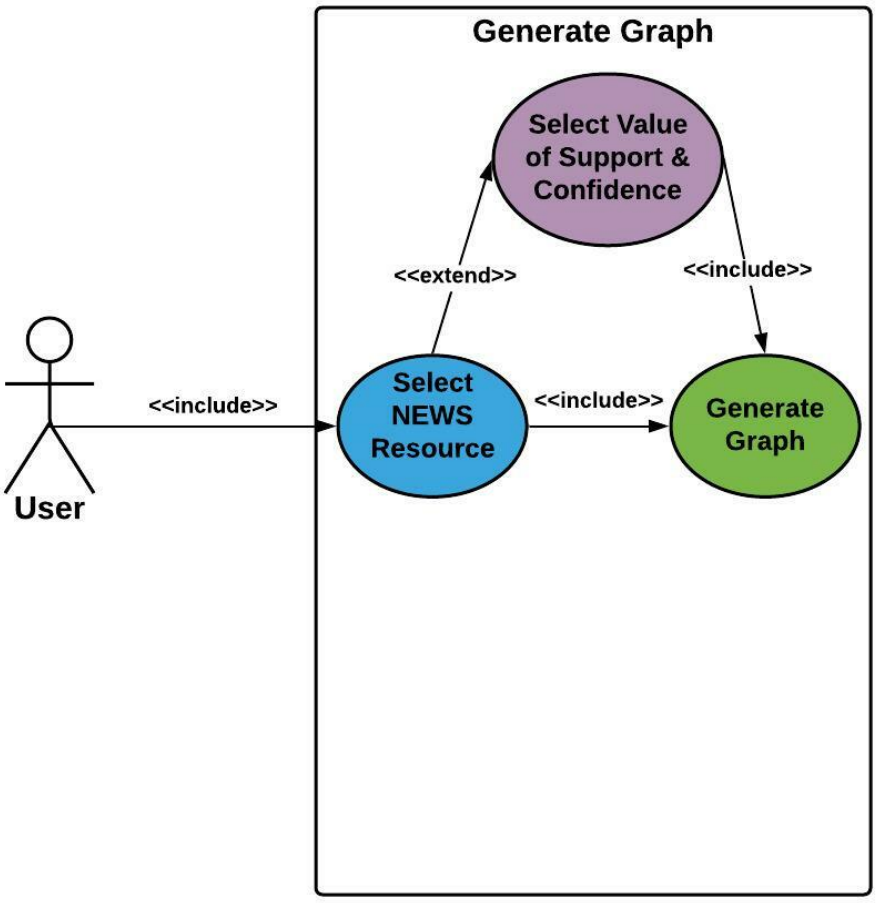


Figure 3.4: Generate Graph Use Case Diagram

In Figure 3.4, there is one actor: User. User will first select NEWS resource, then user can also select minimum value of support and confidence and then graph will be generated.

Table 3.3: Generate Graph Use Case Description

UC Name	Generate Graph
UC ID	03
Priority	High
Primary Actor	User
Description	This use case describes how graph will be generate against user input to the website. User must log in.
Basic Flow	User will go to home page and select NEWS resource and minimum value of support and confidence.
Alternate Flow	NEWS resource is not selected. User will still be on home page.
Pre-condition	Home page should be opened in user screen. Device should have a good internet connection
Steps	**Actor Actions:** Step 1: User will select the NEWS resource and minimum value of support and confidence. **System Response:** Step 2: System will check whether the NEWS resource is selected or not. Step 3: If selected than which NEWS resource is selected and what is the value set by user for minimum support and confidence Step 4: Then system will generate graph in graph page
Post-condition	Graph page will be displayed

In Table 3.3, it is explained that how user will give input to the system and how system will generate graph against user's input.

3.7.4 Setting of Graph

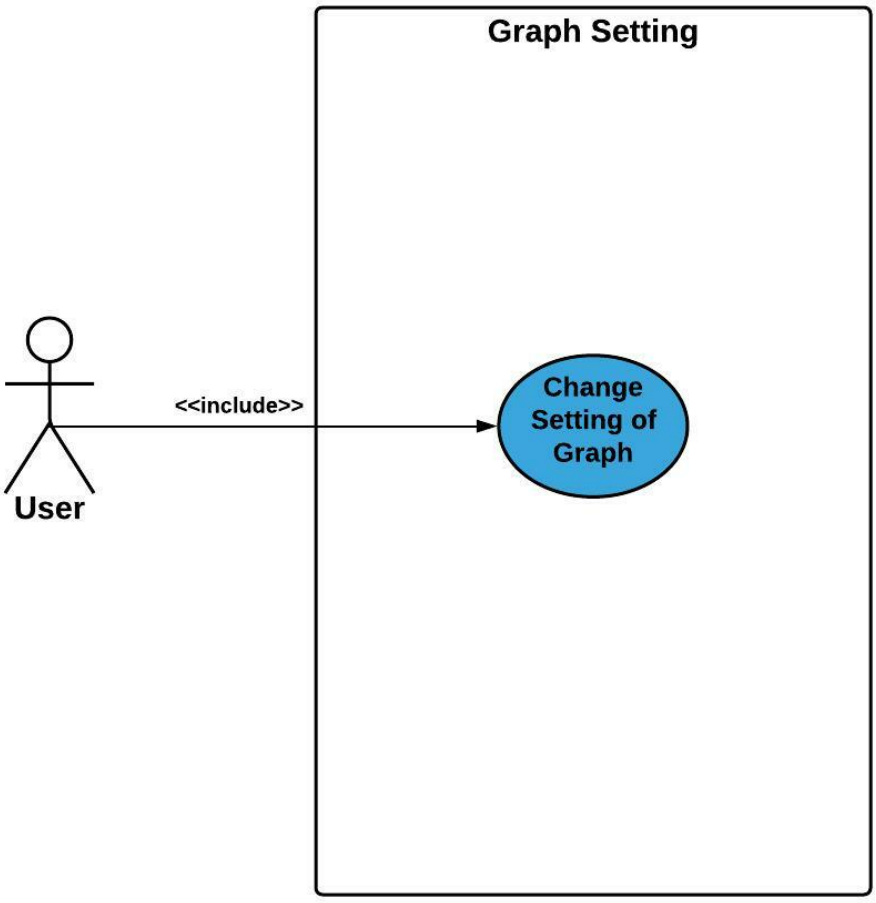


Figure 3.5: Graph Setting Use Case Diagram

In Figure 3.5, User is allowed to change setting of the graph. Changing setting of graph will change its physics.

Table 3.4: Setting of Graph Use Case Description

UC Name	Setting of Graph
UC ID	04
Priority	Low
Primary Actor	User
Description	This use case describes how user can change setting of the graph to the website. If the graph is already generated.
Basic Flow	User will go to graph page and change setting of graph.
Alternate Flow	Invalid input entered User will be prompted with error message.
Pre-condition	Graph page should be opened in user screen. Device should have a good internet connection
Steps	**Actor Actions:** Step 1: User will go to setting of graph. Step 2: User will make changes in the setting of graph. **System Response:** Step 3: System will check the changes in the setting done by user. Step 4: if there is an invalid input in setting than there will be no change and an error message will be sent back to user. Step 5: Otherwise, system will generate the graph with changes in its physics.
Post-condition	Graph page will be displayed

In Table 3.4, it has been explained that how user can change the settings of graph.

3.7.5 View Details of Graph

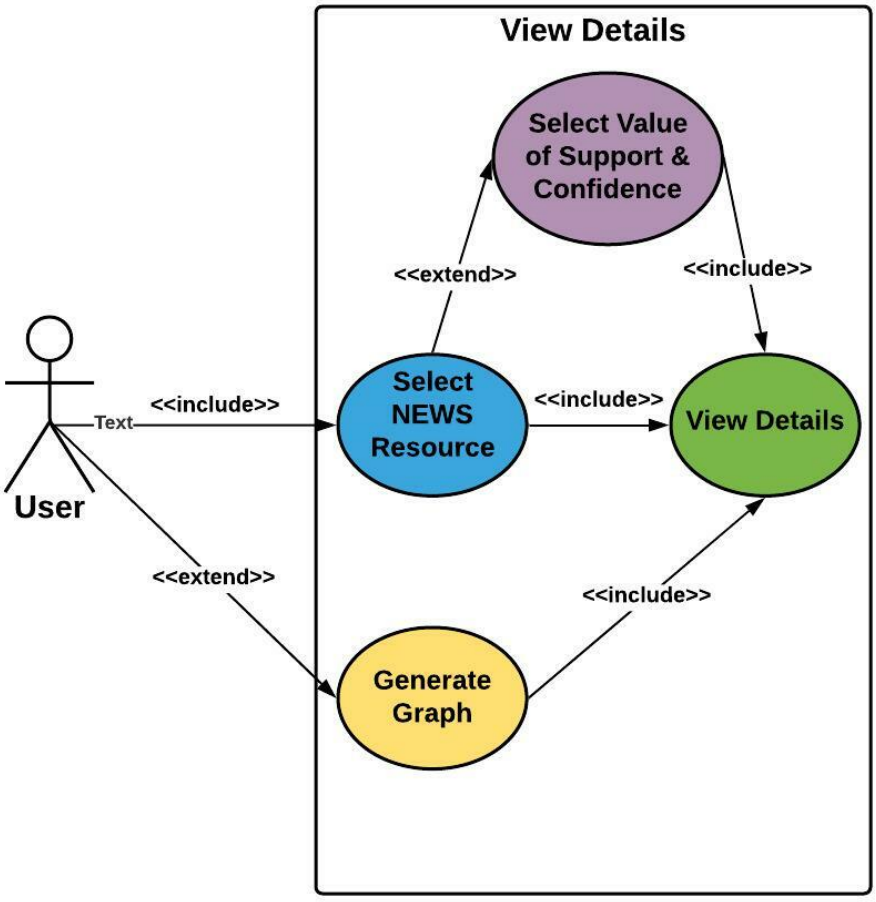


Figure 3.6: View Details Use Case Diagram

In Figure 3.6, there is one actor: User. User can view decision matrix but first he/she needs to either select NEWS resource or generate graph.

Table 3.5: View Details of Graph Use Case Description

UC Name	View Details of Graph
UC ID	05
Priority	Medium
Primary Actor	User
Description	This use case describes how user can view details of the generated graph to the website. If graph is already generated or NEWS resource is selected.
Basic Flow	User will go to detail page and see the decision matrix of generated graph
Alternate Flow	Graph is not generated and user click on detail page. User will be prompted with error message.
Pre-condition	Home / Graph page should be opened in user screen. Device should have a good internet connection
Steps	**Actor Actions:** Step 1: If User is on Home Page, he/she has to select NEWS resource and then click on Detail link. Step 2: If user is on graph page, then directly click Detail link. **System Response:** Step 3: User will be directed to Detail page.
Post-condition	Detail page will be displayed

In Table 3.5, it is described that how user can view Decision Matrix.

3.7.6 Search Entity

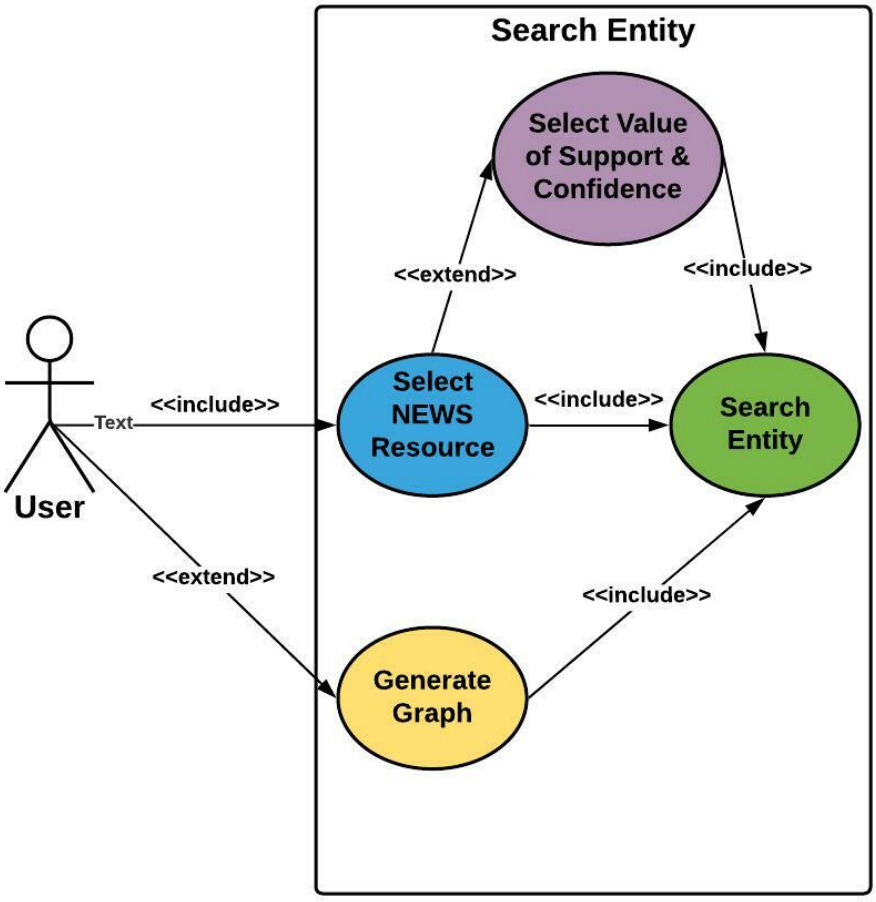


Figure 3.7: Search Entity Use Case Diagram

In Figure 3.7, there is one actor: User. User can search entity but first he/she needs to either select NEWS resource or generate graph.

Table 3.6: Search Entity Use Case Description

UC Name	Search Entity
UC ID	06
Priority	High
Primary Actor	User
Description	This use case describes how user can view details of the generated graph to the website. If graph is already generated or NEWS resource is selected.
Basic Flow	User can be on any page on the website and entered the entity in the text box.
Alternate Flow	User has not entered any entity and click on search button. User will be prompted with error message.
Pre-condition	Home/Graph/Detail page should be opened in user screen. Device should have a good internet connection
Steps	**Actor Actions:** Step 1: If User is on Home Page, he/she has to select NEWS resource Step 2: Else, enter entity and click on search button **System Response:** Step 3: Graph and Decision matrix both will be generated.
Post-condition	Graph/Detail page will be displayed

In Table 3.6, there is a step-by-step explanation of how user can search entity to get its association as a result.

3.7.7 User Log out

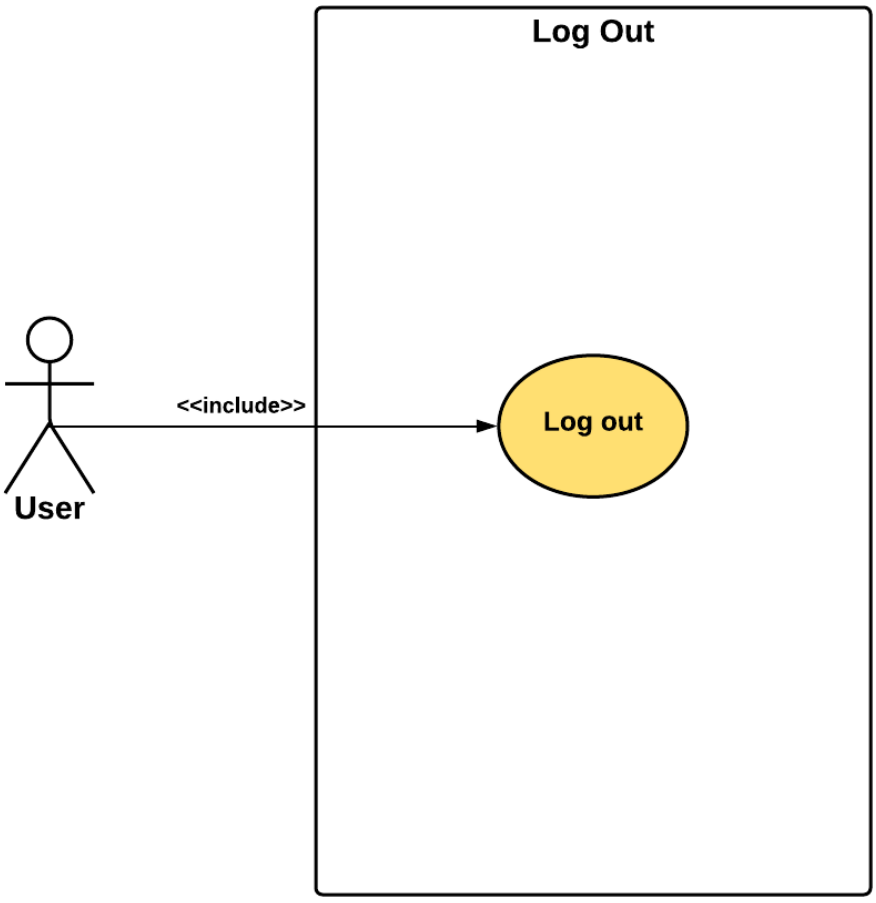


Figure 3.8: Log out Use Case Diagram

In Figure 3.8, there is one actor: User. User can log out his/her account at any time.

Table 3.7: Log out Use Case Description

UC Name	Log out
UC ID	07
Priority	High
Primary Actor	User
Description	This use case describes how user can log out to the website. If user has already logged in.
Basic Flow	User will click on log out button and alert will appear.
Alternate Flow	User has a bad internet connection.
Pre-condition	User has already logged in. Device should have a good internet connection
Steps	**Actor Actions:** Step 1: User will click on log out button. **System Response:** Step 2: System will check if the user has already logged in. Step 3: System will send a message to user that user has been logged out.
Post-condition	Log in page will be displayed

In Table 3.7, use case about how user can log out from the website is explained.

Chapter 4

System Design

This chapter will cover the design of proposed system. It states all information required for the development of the proposed system. To achieve the best results from the project, the developers need to have the right design in hand about the system. When developers have all the relevant information to understand the requirements of the proposed system, then they can easily develop the proposed system that will be according to requirements specified in Chapter 3. A good design is the one which reflects all requirements and users needs. This chapter consist of a detailed diagrams of each module of the system. This chapter represent the system architecture, design methodology and high-level design.

4.1 System Architecture

A system architecture is the conceptual model that defines the structure and behavior of a system. The architecture description of a system is the formal description and representation of a system that is organized in a way that supports reasoning about the structures and behaviors of the system [7].Figure 4.1 is explaining the system architecture of the proposed system.

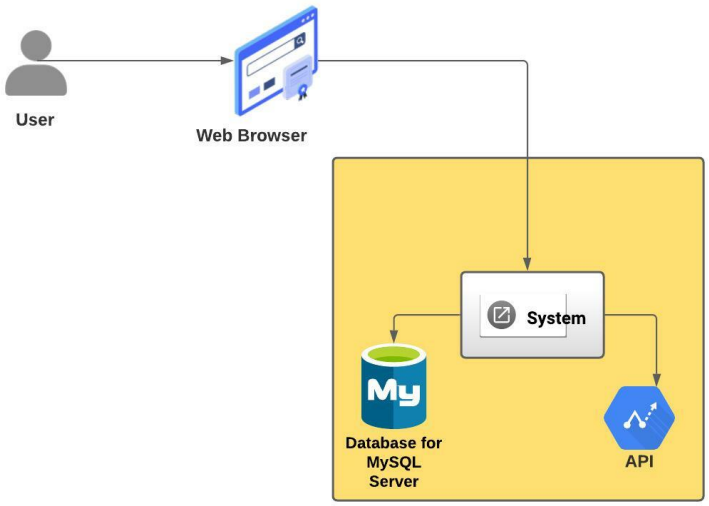


Figure 4.1: System Architecture

4.2 Design Methodology

Design methodology will describe any constraints in the system design and include any assumptions made during development of system design. Figure 4.2 has explained the design methodology of the proposed system.

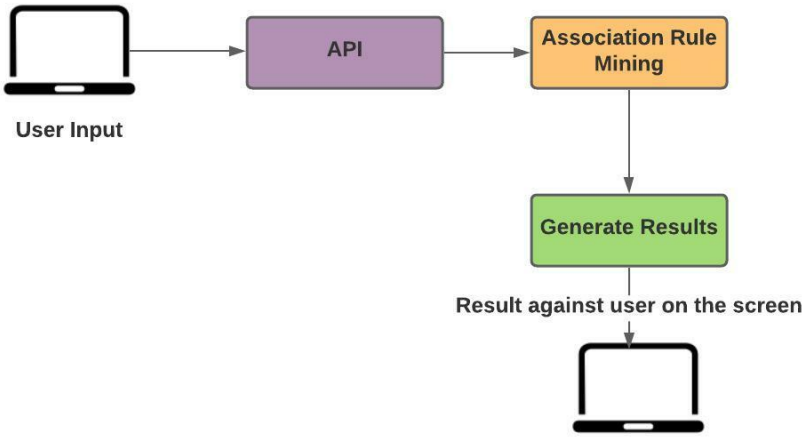


Figure 4.2: Design Methodology

4.3 High Level Design

4.3.1 Data Flow Diagram Level 0

DFD level 0 is also known as context diagram. When user interacts with the system, he/she will basically interact with the presentation layer. Presentation layer will send user input to the Logic layer. Then this layer will respond to user input and sent back results to presentation layer. Figure 4.3 shows context diagram of the proposed system.

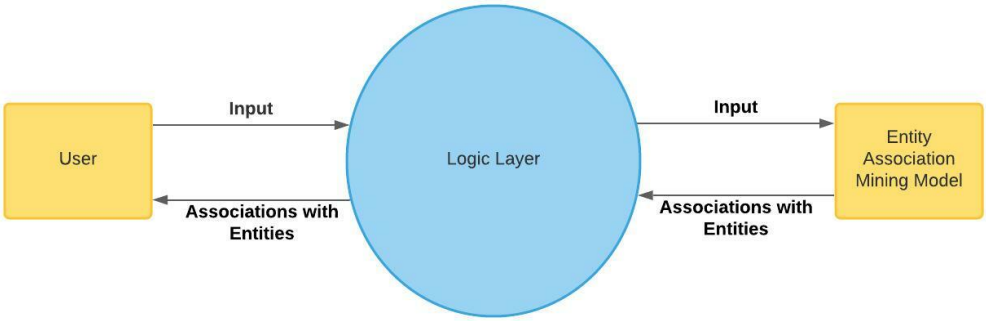


Figure 4.3: Context Diagram

4.3.2 Sequence Diagram

Sequence diagram is a diagram that tells the developer that how user will interact with the system and how system will respond the user.

Sign up

If user has entered information in correct pattern, then an account will be created else system will send an error message to user. Figure 4.4 shows sequence diagram of the sign up operation.

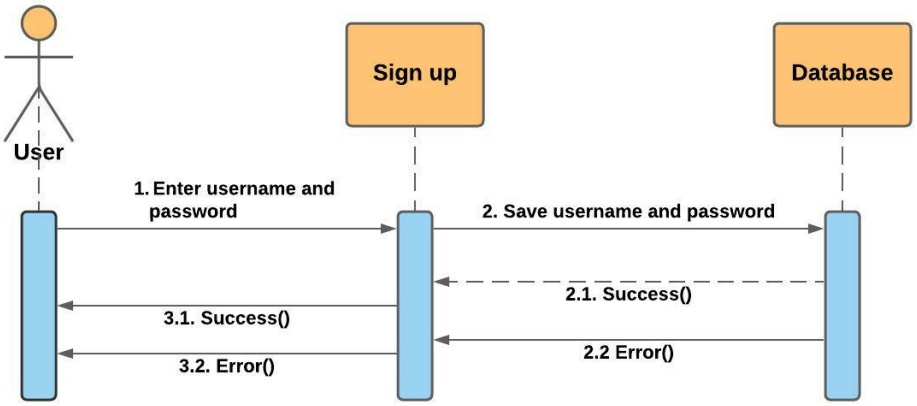


Figure 4.4: Sign up Sequence Diagram

Log in

If user has entered correct username and password, then system will let user to use the website else system will send an error message to user. Figure 4.5 shows sequence diagram of the log in operation.

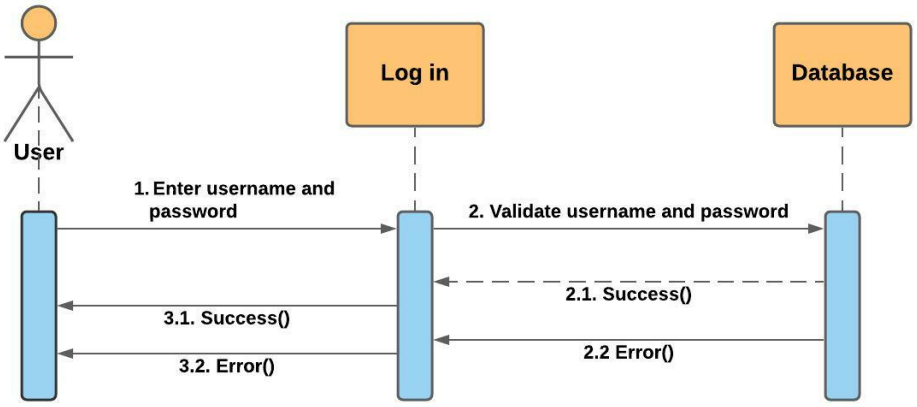


Figure 4.5: Log in Sequence Diagram

Generate Graph

If user has selected NEWS resource and select the minimum value of support and confidence then system will let user to view graph. Figure 4.6 shows sequence diagram of how graph will be generated after user input.

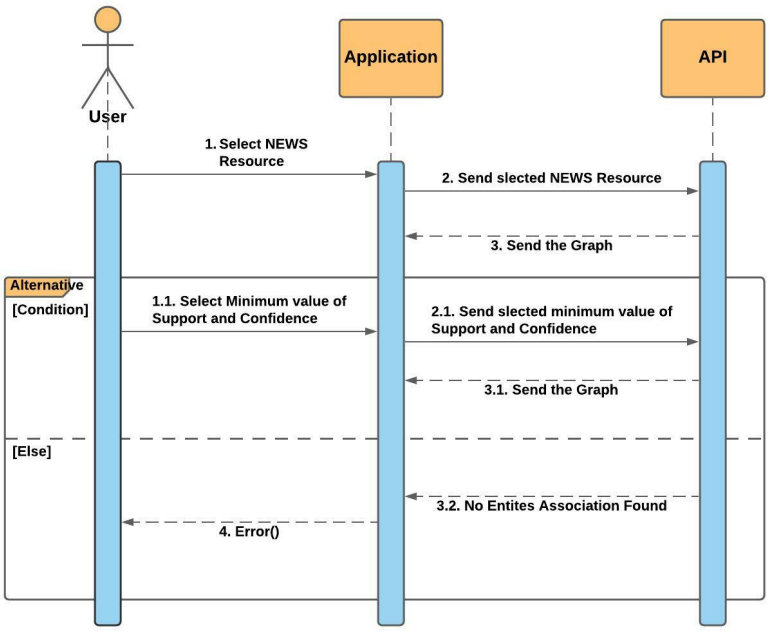


Figure 4.6: Generate Graph Sequence Diagram

Setting of Graph

Once the graph is generated, user can change the physics of graph changing the setting. Figure 4.7 shows sequence diagram of how physics of graph can be changed by user input.

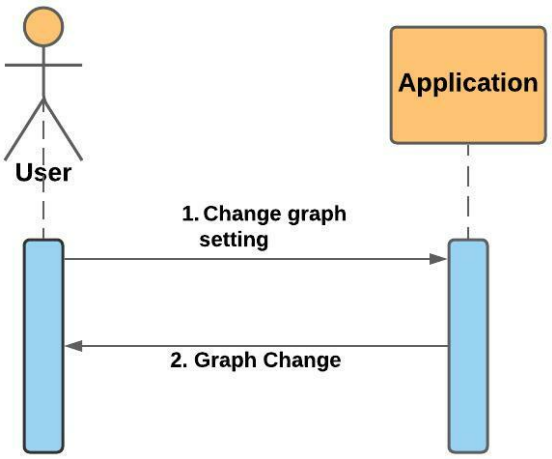


Figure 4.7: Setting of Graph Sequence Diagram

Search Entity

If user has selected the NEWS resource and value of support and confidence than user can search entity within that NEWS resource. Figure 4.8 shows sequence diagram of how user can search an entity to get its association.

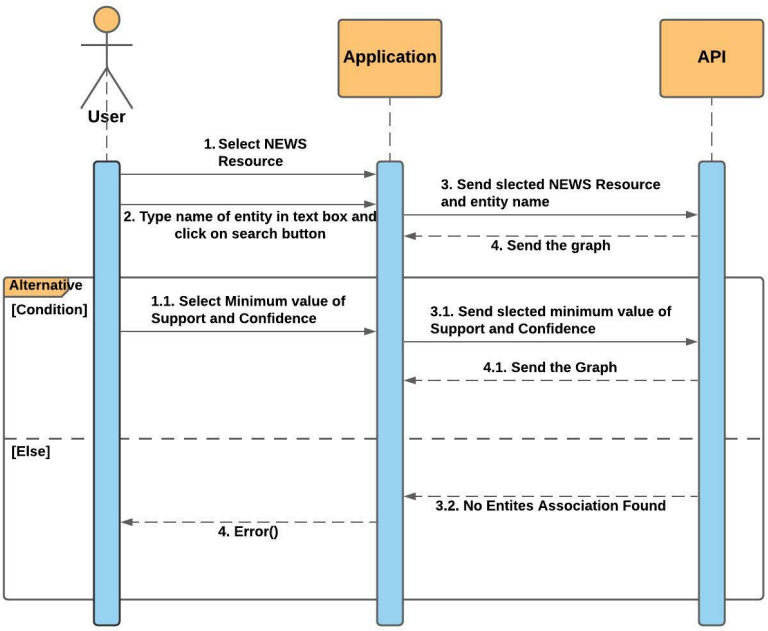


Figure 4.8: Search Entity Sequence Diagram

Log out

If user is already logged in than he/she can log out. Figure 4.9 shows sequence diagram of the log out process.

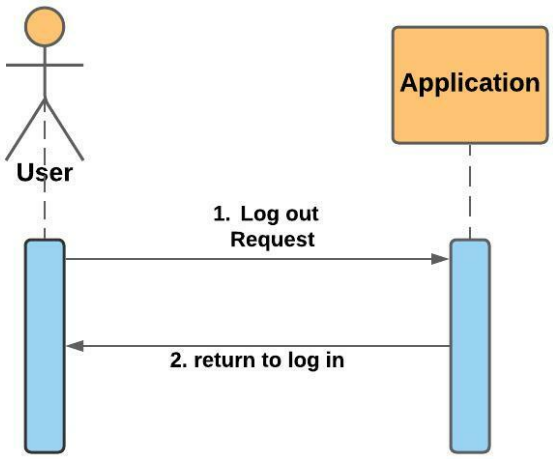


Figure 4.9: Log out Sequence Diagram

4.3.3 Deployment Diagram

Deployment diagram is a physical diagram that show how system will be deployed in the real environment. The system is deployed on windows server. The web server consist of python execution which enable system to run web framework and python scripts. The website is connected to an API and a database. User can interact with system as well as perform operations through web browser. Figure 4.10 shows deployment diagram of the proposed system.

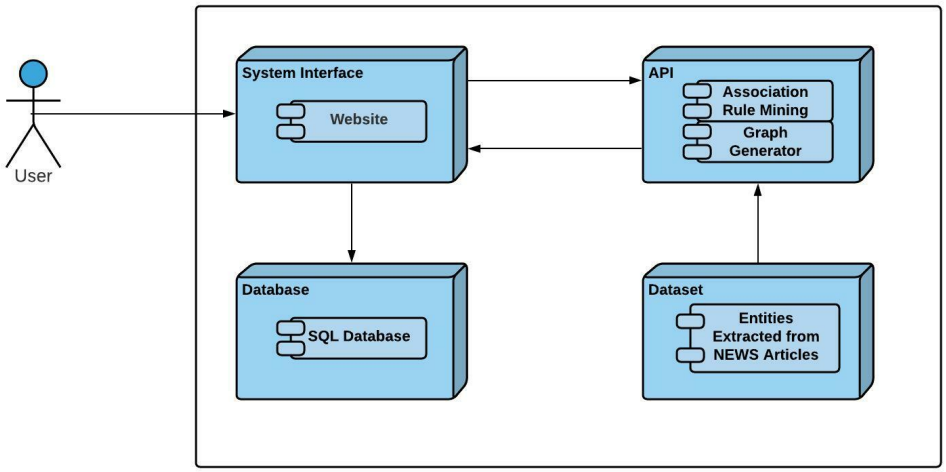


Figure 4.10: Deployment Diagram

4.3.4 Activity Diagram

Activity diagram is a graphical representation of workflow of step-by-step activities and actions performed for choice, iteration, and concurrency.

Activity Diagram I: User Input

Figure 4.11 activity diagram shows that how user can input data. First user will open the website on browser. An interactive interface will appear containing a drop down box having different NEWS resources, two sliders to set value of support and confidence and a search text box to search entity.

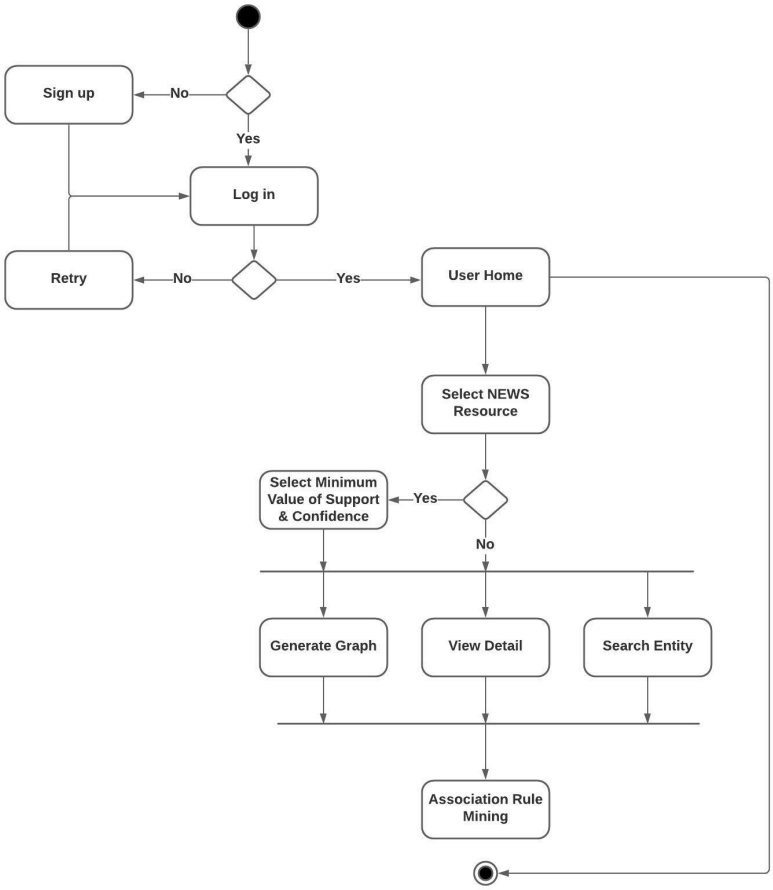


Figure 4.11: User Input Activity Diagram

4.4 GUI design

The user interface of a system is designed to provide a medium for user to interact with the system. The interface of a system should be kept system to facilitate user to understand and learn it quickly. Our system is designed to achieve effective and efficient result. Screenshots of the proposed system are given below:

4.4.1 Icon

Figure 4.12 shows the icon of the web application. The icon of web application is designed in such a way that it should fully implicate the name of web application. It is designed in Adobe Illustrator.

Figure 4.12: Icon

4.4.2 Sign up Page

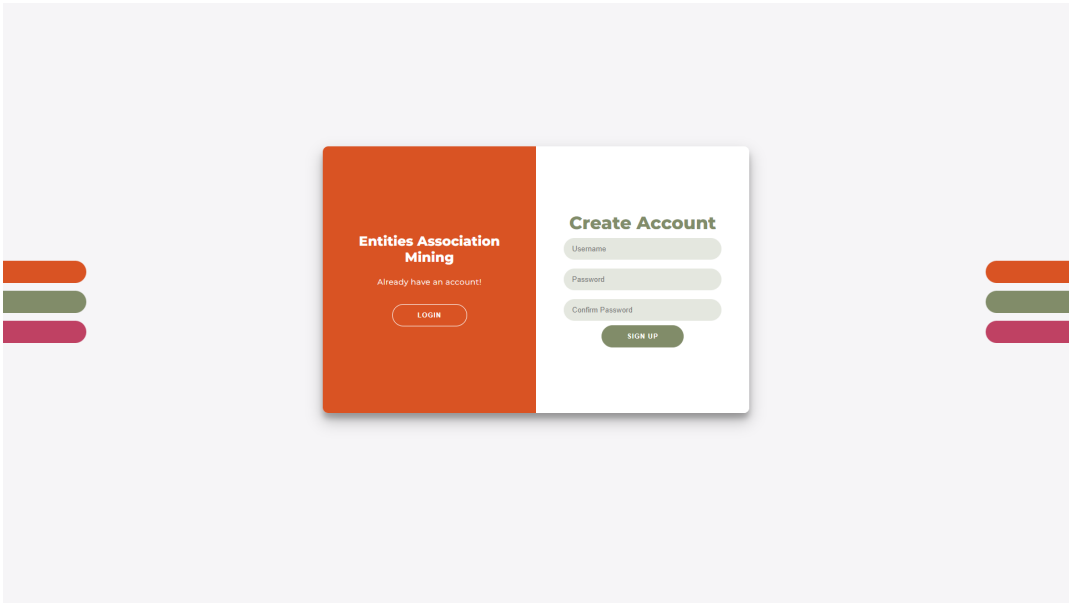


Figure 4.13: Sign up Page

Figure 4.13 shows sign up page. If user has visited the website for the first time than user needs to create account. User will enter username and password and then click on sign up button.

4.4.3 Log in Page

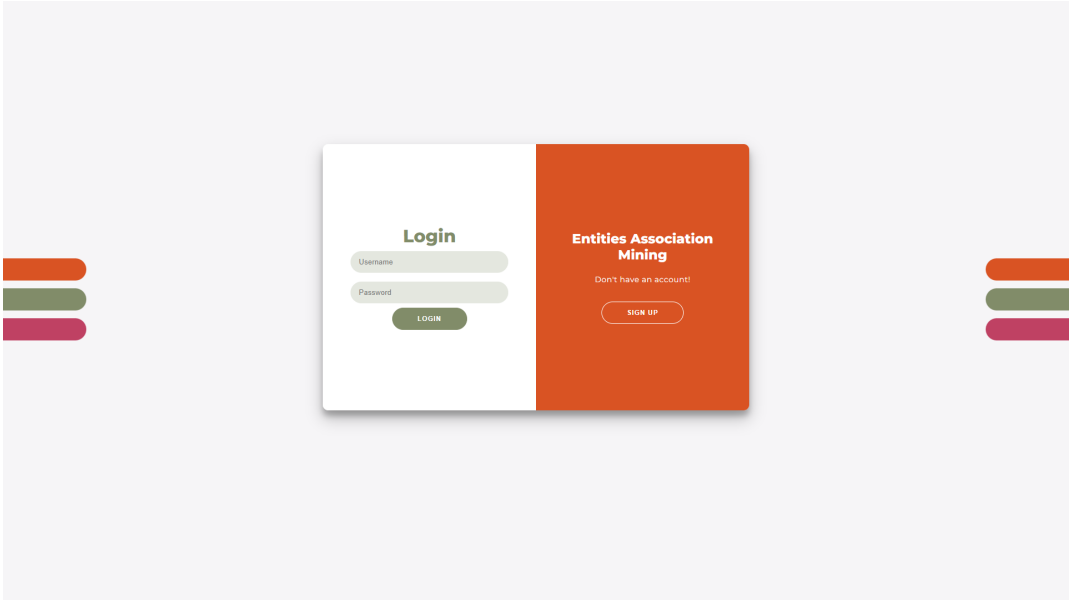


Figure 4.14: Log in Page

Figure 4.14 shows log in page. If user want to access website, he/she has to log in first. User will enter username and password and click on log in button.

4.4.4 Home Page

Figure 4.15: Home Page

Figure 4.15 shows home page. After log in user will be directed to home page. In home page, user will select NEWS resource but before that user can change value of support and confidence. Once NEWS resource is selected, graph page will be loaded.

4.4.5 Graph Page

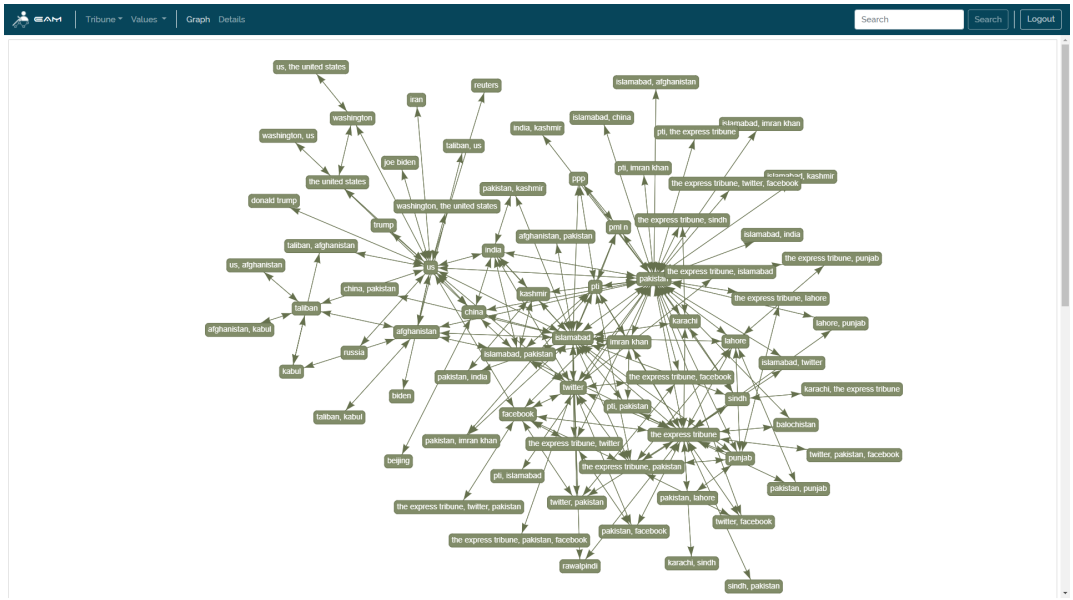


Figure 4.16: Graph Page

Figure 4.16 shows Graph page. It is loaded once user has selected NEWS resource. User can see entities and their association there. Also, user can search entity and view details of

graph by clicking on Details in navigational bar.

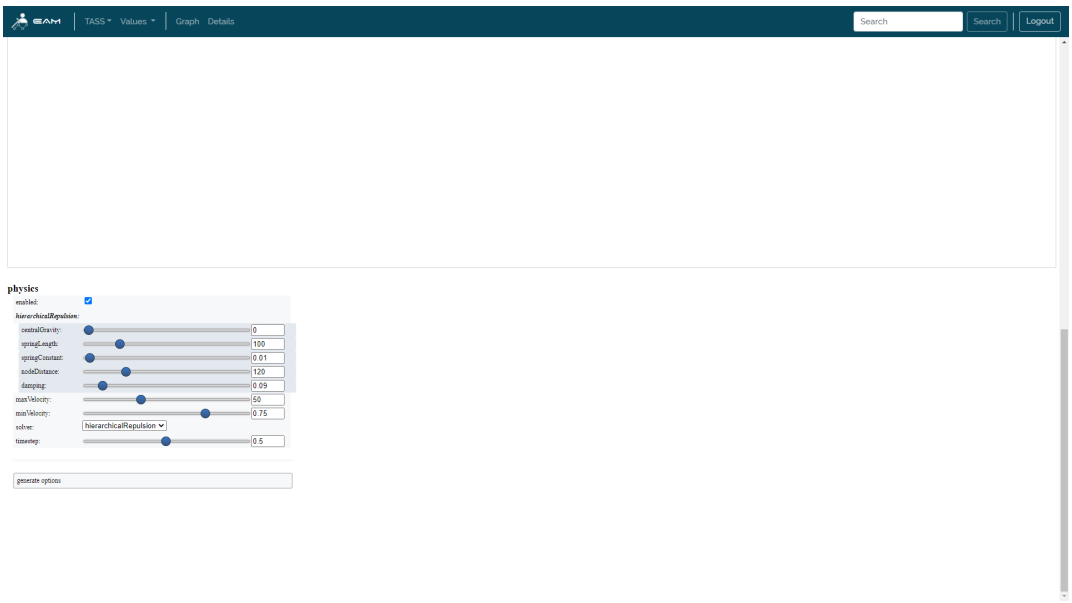
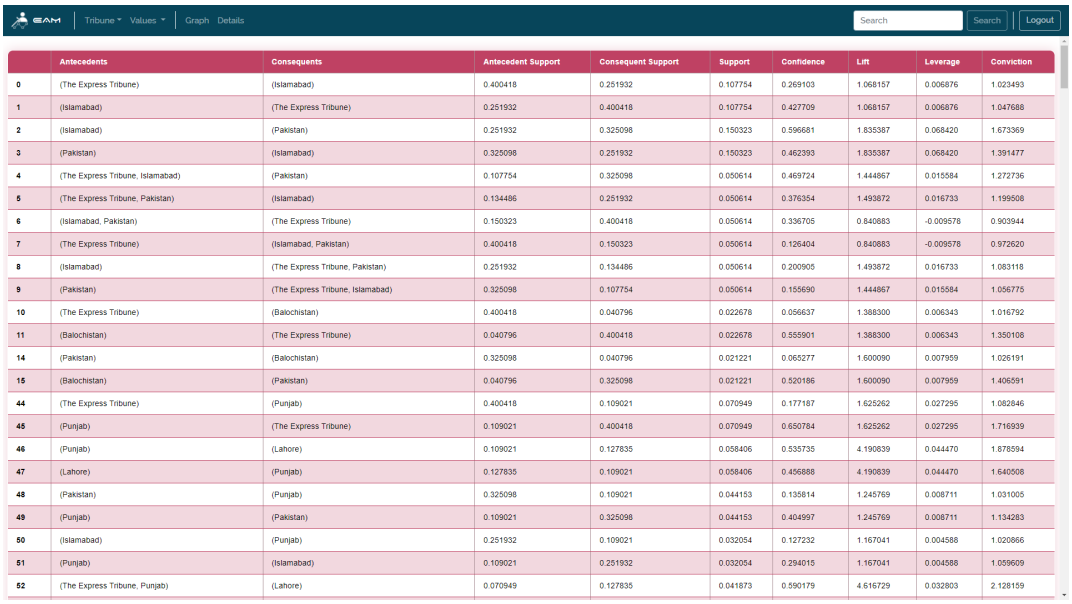


Figure 4.17: Graph Setting

Scrolling down on graph page graph physics setting appear Figure 4.17. User can change physics of graph by changing its setting.

4.4.6 Details Page



	Antecedents	Consequents	Antecedent Support	Consequent Support	Support	Confidence	Lift	Leverage	Conviction
0	(The Express Tribune)	(Islamabad)	0.400418	0.251932	0.107754	0.269103	1.068157	0.006876	1.023493
1	(Islamabad)	(The Express Tribune)	0.251932	0.400418	0.107754	0.427709	1.068157	0.006876	1.047688
2	(Islamabad)	(Pakistan)	0.251932	0.325098	0.150323	0.596681	1.835387	0.068420	1.673369
3	(Pakistan)	(Islamabad)	0.325098	0.251932	0.150323	0.462393	1.835387	0.068420	1.391477
4	(The Express Tribune, Islamabad)	(Pakistan)	0.107754	0.325098	0.050614	0.469724	1.444867	0.015584	1.272736
5	(The Express Tribune, Pakistan)	(Islamabad)	0.134486	0.251932	0.050614	0.376354	1.493872	0.016733	1.199508
6	(Islamabad, Pakistan)	(The Express Tribune)	0.150323	0.400418	0.050614	0.336705	0.840883	-0.009578	0.903944
7	(The Express Tribune)	(Islamabad, Pakistan)	0.400418	0.150323	0.050614	0.126404	0.840883	-0.009578	0.972620
8	(Islamabad)	(The Express Tribune, Pakistan)	0.251932	0.134486	0.050614	0.200905	1.493872	0.016733	1.083118
9	(Pakistan)	(The Express Tribune, Islamabad)	0.325098	0.107754	0.050614	0.155690	1.444867	0.015584	1.056775
10	(The Express Tribune)	(Balochistan)	0.400418	0.040796	0.022678	0.056637	1.388300	0.006343	1.016792
11	(Balochistan)	(The Express Tribune)	0.040796	0.400418	0.022678	0.555901	1.388300	0.006343	1.350108
14	(Pakistan)	(Balochistan)	0.325098	0.040796	0.021221	0.065277	1.600090	0.007959	1.026191
15	(Balochistan)	(Pakistan)	0.040796	0.325098	0.021221	0.520186	1.600090	0.007959	1.406591
44	(The Express Tribune)	(Punjab)	0.400418	0.109021	0.070949	0.177187	1.625262	0.027295	1.082846
45	(Punjab)	(The Express Tribune)	0.109021	0.400418	0.070949	0.650784	1.625262	0.027295	1.716939
46	(Punjab)	(Lahore)	0.109021	0.127835	0.058406	0.535735	4.190839	0.044470	1.878594
47	(Lahore)	(Punjab)	0.127835	0.109021	0.058406	0.456888	4.190839	0.044470	1.640508
48	(Pakistan)	(Punjab)	0.325098	0.109021	0.044153	0.135814	1.245769	0.008711	1.031005
49	(Punjab)	(Pakistan)	0.109021	0.325098	0.044153	0.404997	1.245769	0.008711	1.134283
50	(Islamabad)	(Punjab)	0.251932	0.109021	0.032054	0.127232	1.167041	0.004588	1.020866
51	(Punjab)	(Islamabad)	0.109021	0.251932	0.032054	0.294015	1.167041	0.004588	1.059609
52	(The Express Tribune, Punjab)	(Lahore)	0.070949	0.127835	0.041873	0.590179	4.616729	0.032803	2.128159

Figure 4.18: Details Page

Once graph is generated, user can see decision matrix of the graph by clicking on details.In Figure 4.18 user can see antecedent, consequent, support, confidence, lift, etc.

Chapter 5

System Implementation

In this chapter, we will discuss how the proposed system is built, which methodology we adopt to build the system and what tools and technologies we have used to build the system. This chapter covers up each step that has been involved to full fill both functional and non-functional requirements that are discussed in the previous two chapters (Chapter 3 and 4) of the system have been implemented.

5.1 System Architecture

Our proposed system is based on 3-tier architecture with:

1. Topmost and first layer is Presentation Layer.
2. Then the middle and second layer are Business Logic Layer.
3. The last and third layer is Data access layer.

5.1.1 Presentation Layer

The Presentation Layer is consisting of GUI that is an interactive interface for the users. Presentation layer is the main layer of the proposed system which will get input from user, process that and system will send feedback to user. This layer contains the interface design of the system that is provided in Figure 4.1.

5.1.2 Business Logic Layer

The Business Logic Layer is the main body of the web application which cover up all functionalities with which the system will work such as the taking input of user in presentation layer to association rule mining model, generating graph, and then sending the result back to presentation layer.

5.1.3 Data Access Layer

The Data Access Layer deals with the operations of databases. Data Access Layer save and verify the user information in the Database on the request of logic layer.

5.2 Methodology

Methodology of the proposed system explains the whole process from data collection to generating results.

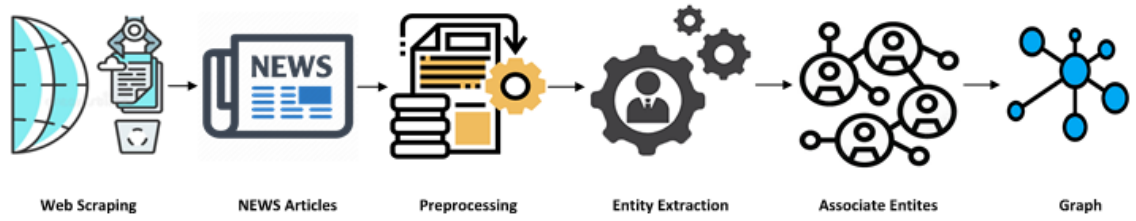


Figure 5.1: Methodology

5.2.1 Data Collection

For data collection, we have extracted information from five NEWS resources that are as follows:

1. Asian NEWS International (ANI, NEWS articles: 71,593)
2. Associated Press of Pakistan (APP, NEWS articles: 23,997)
3. Russian NEWS Agency TASS (NEWS articles: 17,512)
4. The Hills (NEWS articles: 64,803)
5. The Express Tribune (NEWS articles: 15,786)

The NEWS articles are extracted from NEWS resources through web scraping. In total, we have extracted 193,691 NEWS articles. Each NEWS article is saved in a .txt file. That file contains NEWS resource name, link of NEWS article, date of publication, date of extraction, Author name, title of NEWS article and content of the article. Articles extracted from each NEWS resource are saved in their separate folder (name of folder is NEWS resource name).

5.2.2 Data Preprocessing

Before preprocessing, NEWS articles extracted via web scraping are stored in csv file for each NEWS article separately. In data preprocessing, the csv file is converted into data frame using pandas library, all instances with any null value are dropped. The Attributes (NEWS resource name, link of NEWS article, date of publication, date of extraction, Author name and title of NEWS article) are not going to use in the project. So, they are also dropped from the data frame.

5.2.3 Entity Extraction

The contents of each NEWS article are passed through NLP pipeline built by spacy to normalize text and extract entities.

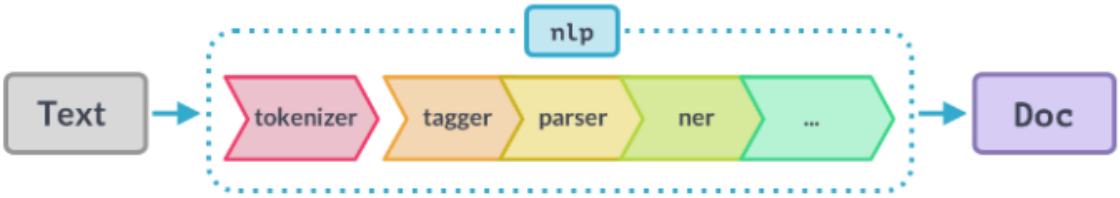


Figure 5.2: NLP Pipeline

In NER, every entity has a tag which classify that which type of entity is it. The type entities extracted are mentioned in Table 5.1 The extracted entities were saved in a file

Table 5.1: NER tags used

Type	Description	Example
PERSON	People including fictional	Imran Khan
NORP	Nationalities, religious, political groups	Pakistani
FAC	Buildings, airports, highways, bridges, etc.	Express Highway
ORG	Companies, agencies, institutions, etc.	Google, Envicrete
GPE	Countries, cities, states.	Pakistan, Turkey
LOC	Non GPE location, mountain ranges, bodies of water.	Europe, Nile River
PRODUCT	Objects, Vehicles, foods, etc. (Not services)	Formula 1
EVENT	Named hurricanes, battle, wars, sports events, etc.	PSL
WORK OF ART	Title of books, songs, etc.	The Mona Lisa
LAW	Named documents made into laws.	Roe V. Wade

created by pickle library in the form of lists within a list.

5.2.4 Associate Entities

For association rule mining data should be normalized. We have normalized extracted entities based on their occurrence in each NEWS article by transaction encoder. Then, that normalized data is passed to FP-growth algorithm with minimum support threshold of 0.01 to extract frequent item sets and minimum confidence threshold of 0.01 to extract rules from the extracted frequent item sets. The result of Association rule mining is the complete the decision matrix of each NEWS resource. That is saved in a file using pickle library.

5.2.5 Graph

Then, decision matrix gained from association rule mining is passed to a graph function to generate graph of the entities and their association. That is end output of the proposed system.

5.3 Tools and Technologies

5.3.1 Visual Studio Code

Visual Studio Code is a stand-alone code editor to facilitate developers for their code edit-build-debug cycle that runs on Windows, Linux and Mac OS. It includes support for debugging, syntax highlighting, intelligent code completion, snippets, code refactoring and embedded Git. Developers can change theme, keyboard shortcuts, preferences and install extensions that add functionality.

5.3.2 XAMPP Control Panel

It is a management tool that controls each component of text server. It is very easy to install Apache distribution for Windows, Mac OS, and Linux. The package includes the Apache web server, MySQL, PHP, Perl, an FTP server and phpMyAdmin.

5.4 Languages and Libraries used

5.4.1 Python

Python is an interpreted high-level general-purpose language. Its language constructs as well as its object-oriented approach aim to help programmers write clear, logical code for small and large-scale projects. Python is dynamically typed and garbage collected. It supports multiple programming paradigms, including structured (particularly, procedural), object-oriented and functional programming. It is often described as a "batteries included" language due to its comprehensive standard library [8].

mlxtend

mlxtend is a library is a tool and extension for data analysis and machine learning. It has many interesting functions. It contains functions to extract frequent patterns by implementing the Algorithms stated in Chapter 2.

Pandas

Pandas is a library of python used for data manipulation and analysis. Basically, it consists of data structures and operations for manipulating numerical tables and time series. The name is derived from the term "panel data", an econometric term for data sets that include observations over multiple time periods for the same individuals [6]

pyvis

pyvis is a python library for visualizing networks. It is built around visjs that is a JavaScript visualization library.

pickle

Pickle is module of python library that is used to serializing and deserializing a python object structure.

IPython

It is a python library for interactive computing that offers introspection, rich media, shell syntax, tab completion and history. It provides an interactive shell for data visualization and use GUI tool kits.

Flask

It is a micro web framework written in python. It is an integrated support for unit testing.

Beautiful Soup

It is a python library for parsing HTML and XML documents. It creates a parse tree for parsed pages that can be used to extract data from web pages that is useful for web scraping.

Requests

It is a HTTP library of python. The library makes HTTP requests simpler and more human-friendly. It abstracts the complexities of making requests behind a beautiful, simple API so that we can focus on interacting with services and consuming data in our application.

spacy

spacy is a python open-source library for NLP. It features NER, POS tagging, dependency parsing, word vectors etc.

5.4.2 HTML

HTML is a standard markup language for the documents designed to be displayed on web browser. HTML elements are building blocks of web pages. It uses tags written using angular brackets.

5.4.3 CSS

It is a style sheet language used for describing of a HTML document. It is a cornerstone technology of the WWW, alongside HTML and JavaScript.

5.4.4 Java Script

It is sometimes abbreviated as js. It is a high-level language that has dynamic typing, prototype-based object-orientation and first-class functions. It is only used in web browsers. Its syntax is like JAVA.

5.4.5 PHP

It is general purpose open-source scripting language towards web development. It is used for making dynamic and interactive web pages. It is a widely used, free, and efficient alternative to competitors such as Microsoft's ASP. PHP works well with HTML and database.

5.4.6 MySQL

It is a database service that is fully managed database to deploy cloud-native applications. It is a standard database system for websites with huge volumes of both data and end-users.

5.4.7 Bootstrap

It is an open-source CSS framework used to make the website responsive. It contains CSS and JavaScript based design templates for forms, buttons, navigation, and other components of interface. We have used to make the interface responsive.

Chapter 6

System Testing and Evaluation

As we have completed the design and development of the proposed system. Now, this is time to test our system that does this system full fill both functional and non-functional requirements specified in Chapter 3. To check the behavior of the system we have applied different techniques to get the best result. Testing plays an important yet crucial role to assure that the system will satisfy user needs and the execution of system is carried out without errors.

6.1 Graphical User Interface (GUI) Testing

The purpose of GUI testing was to evaluate the interaction between the user and the system. The GUI should be user-friendly to make him/her to understand and learn system. The text in the different components of the interface should be clear. The interface should be according to design guidelines.

6.1.1 Sign up Page

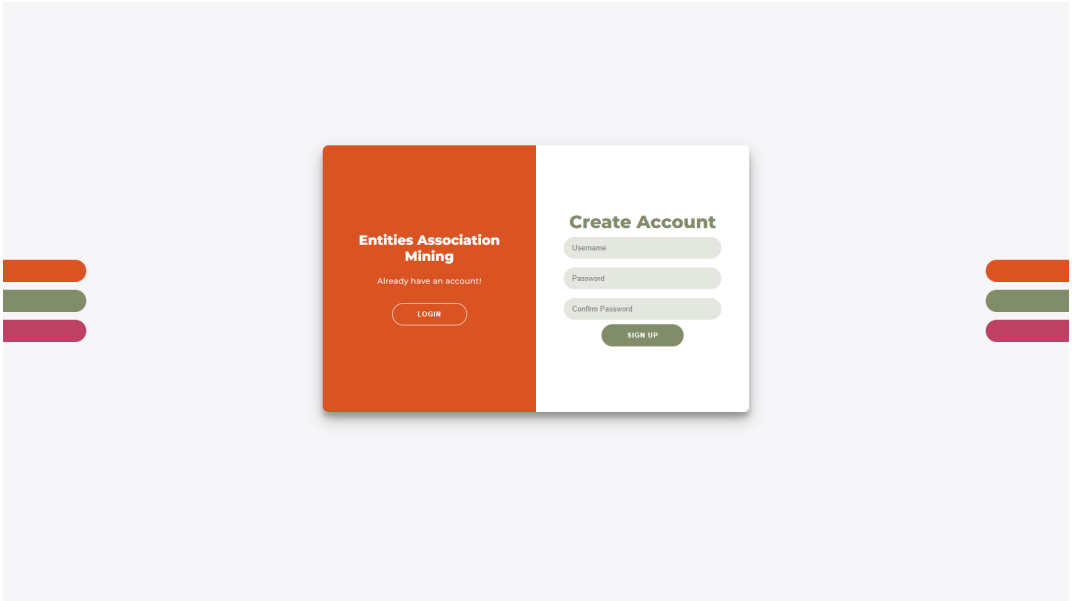


Figure 6.1: Sign up Page

6.1.2 Log in Page

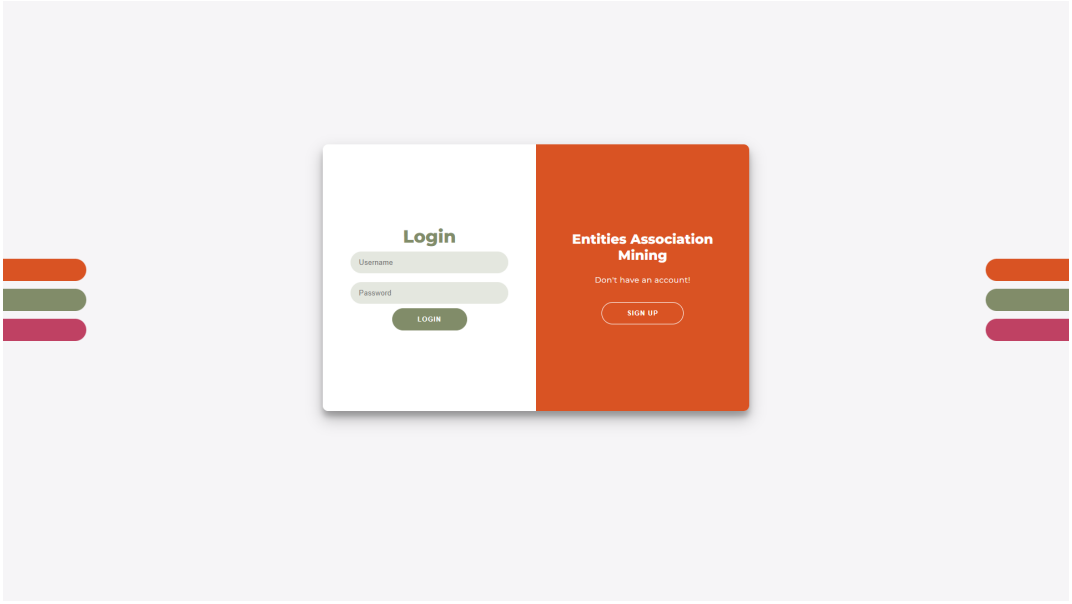


Figure 6.2: Log in Page

6.1.3 Home Page

Figure 6.3: Home Page

6.1.4 Graph Page

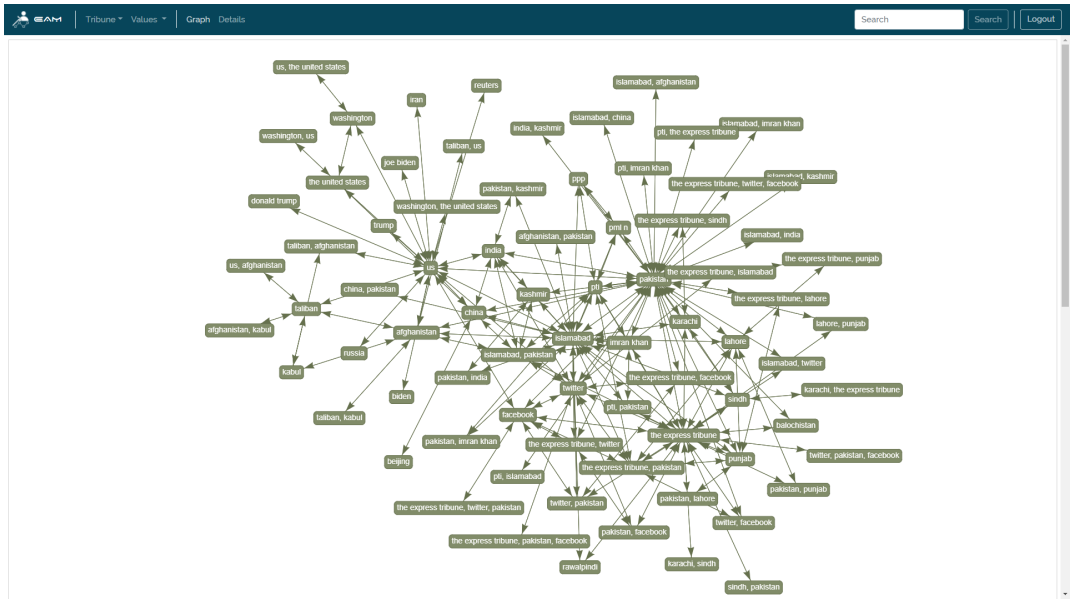


Figure 6.4: Graph Page

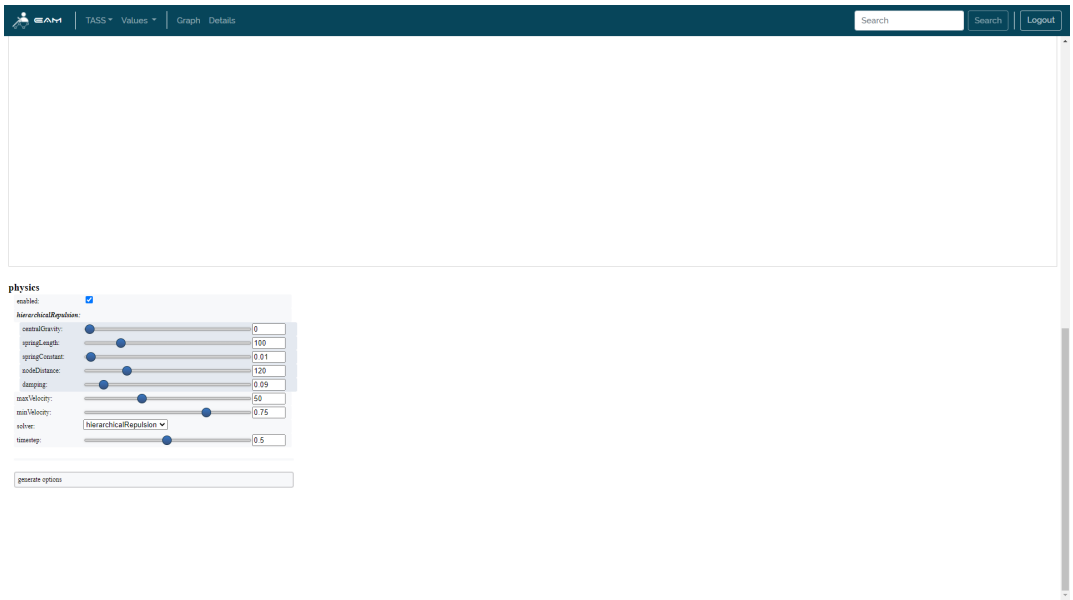


Figure 6.5: Graph Setting

6.1.5 Detail Page

	Antecedents	Consequents	Antecedent Support	Consequent Support	Support	Confidence	Lift	Leverage	Conviction
0	(The Express Tribune)	(Islamabad)	0.400418	0.251932	0.107754	0.269103	1.068157	0.006876	1.023493
1	(Islamabad)	(The Express Tribune)	0.251932	0.400418	0.107754	0.427709	1.068157	0.006876	1.047688
2	(Islamabad)	(Pakistan)	0.251932	0.325098	0.150323	0.596681	1.835387	0.068420	1.673369
3	(Pakistan)	(Islamabad)	0.325098	0.251932	0.150323	0.462393	1.835387	0.068420	1.391477
4	(The Express Tribune, Islamabad)	(Pakistan)	0.107754	0.325098	0.050614	0.469724	1.444867	0.015584	1.272736
5	(The Express Tribune, Pakistan)	(Islamabad)	0.134486	0.251932	0.050614	0.376354	1.493872	0.016733	1.199508
6	(Islamabad, Pakistan)	(The Express Tribune)	0.150323	0.400418	0.050614	0.336705	0.840883	-0.009578	0.903944
7	(The Express Tribune)	(Islamabad, Pakistan)	0.400418	0.150323	0.050614	0.126404	0.840883	-0.009578	0.972620
8	(Islamabad)	(The Express Tribune, Pakistan)	0.251932	0.134486	0.050614	0.200905	1.493872	0.016733	1.083118
9	(Pakistan)	(The Express Tribune, Islamabad)	0.325098	0.107754	0.050614	0.155690	1.444867	0.015584	1.056775
10	(The Express Tribune)	(Balochistan)	0.400418	0.040796	0.022678	0.056637	1.388300	0.006343	1.016792
11	(Balochistan)	(The Express Tribune)	0.040796	0.400418	0.022678	0.555901	1.388300	0.006343	1.350108
14	(Pakistan)	(Balochistan)	0.325098	0.040796	0.021221	0.065277	1.600090	0.007959	1.026191
15	(Balochistan)	(Pakistan)	0.040796	0.325098	0.021221	0.520186	1.600090	0.007959	1.406591
44	(The Express Tribune)	(Punjab)	0.400418	0.109021	0.070949	0.177187	1.625262	0.027295	1.082846
45	(Punjab)	(The Express Tribune)	0.109021	0.400418	0.070949	0.650784	1.625262	0.027295	1.716939
46	(Punjab)	(Lahore)	0.109021	0.127835	0.058406	0.535735	4.190839	0.044470	1.878594
47	(Lahore)	(Punjab)	0.127835	0.109021	0.058406	0.456888	4.190839	0.044470	1.640508
48	(Pakistan)	(Punjab)	0.325098	0.109021	0.044153	0.135814	1.245769	0.008711	1.031005
49	(Punjab)	(Pakistan)	0.109021	0.325098	0.044153	0.404997	1.245769	0.008711	1.134283
50	(Islamabad)	(Punjab)	0.251932	0.109021	0.032054	0.127232	1.167041	0.004588	1.020866
51	(Punjab)	(Islamabad)	0.109021	0.251932	0.032054	0.294015	1.167041	0.004588	1.059609
52	(The Express Tribune, Punjab)	(Lahore)	0.070949	0.127835	0.041873	0.590179	4.616729	0.032803	2.128159

Figure 6.6: Details Page

6.2 Usability Testing

The purpose of usability testing is to evaluate that how a user will use the system. It is conducted by allowing a non-technical user to use the system. Then, performance of the user is monitored and observed to review the relevant issues regarding usability and get them resolved by the developer.

6.3 Software Performance Testing

In this testing approach, the system is evaluated based on non-functional requirements. The system should be effective and efficient both. The system should pass this test because non-functional requirements play an important role to make a good application.

6.4 Compatibility Testing

In this testing approach, the system is tested on different browsers to check the compatibility of the system. Since our system is web-based.

6.5 Exception Handling

Exception Handling is an approach to make system that much capable that it can handle a failure. As the system has a front-end and back-end, the exception handling is applied to both to avoid exceptions. The exception handling is applied to the system to ensure the smooth execution of our system.

6.6 Load Testing

As our system is web-based, therefore it will run on a web browser. Also, in the backend of a website there are multiple process running. On average our system took 3 to 15 seconds to load full website.

6.7 Software Testing Technique

Software testing is the process by which each component of the system is evaluated according to the requirements specified in Chapter 3. Following are some software testing techniques which are used to test the website:

1. Functional Testing
2. Performance Testing
3. Acceptance Testing

6.7.1 Functional Testing

In this testing, instead of testing the whole system, each component or functionality of the system is tested. It can be defined as giving input to the function to evaluate the output. It has following types:

Unit Testing

Testing the program files individually if they are working fine. All resources and functions have been tested and working fine.

Integrating Testing

Testing all programming files together to verify if they have been integrated successfully. The integration of all resources and functions has been tested and produced fair results.

System Testing

Testing the behavior of the whole system to verify that the system full fill all non-functional requirements.

White Box Testing

Testing the system with the knowledge of system's internal structure, lines of code and methodology followed. System has produced fair results.

Black Box Testing

Testing the system with the results produced by the system. To do black box testing website is loaded to the web browser and given to user to test. After using the website, user was satisfied with the results.

6.7.2 Performance Testing

Performance testing can be defined as the process of testing the performance of the system. This testing approach is used to ensure that how reliable, accurate and effective the system. It has following types:

Stress Testing

Testing the limits of the system. It is done by overloading the system by setting the lower value of support and confidence. It generates result within a minute.

Configuration Testing

Testing the system on different hardware and software platforms. The application was tested on different systems and browsers. Results produced were acceptable.

6.7.3 Acceptance Testing

Testing the system that whether it is ready to be deployed in real environment or not. This type of testing reveals the errors and omissions in the system requirements. The system is testing in real environment and it is working fine.

6.8 Test Cases

In this phase we have tested the effectiveness of the developed system. The system and its modules are tested with different inputs and the results provided by the system is compared with expected results from the system.

6.8.1 Opening of web application

Table 6.1 shows test case of opening web application.

Table 6.1: Test Case no.1: Opening the Web Application

Test Case ID	1
Function to be tested	Opening the web application
Initial State	Web browser should be already opened
Input	Type the website address in browser
Expected Output	Web application launches successfully
Actual Result	Web application launches successfully.
State	Pass

6.8.2 Sign up

Table 6.2 shows test case of user sign up to web application.

Table 6.2: Test Case no.2: Sign up

Test Case ID	2
Function to be tested	Sign up
Initial State	Web application already opened
Input	User Data (Username and password)
Expected Output	User will successfully sign up.
Actual Result	User signed up successfully.
State	Pass

6.8.3 Log in

Table 6.3 shows test case of user log in with valid input.

Table 6.3: Test Case no.3: Log in with valid input

Test Case ID	3
Function to be tested	Log in using valid input
Initial State	Web application already opened and user has account
Input	Log in credentials (Username and password)
Expected Output	User will be navigated to Home Page.
Actual Result	User is navigated to Home Page.
State	Pass

Table 6.4 shows the test case of user log in with invalid input.

Table 6.4: Test Case no.4: Log in with invalid input

Test Case ID	4
Function to be tested	Log in using invalid input
Initial State	Web application already opened and account does not exist
Input	Log in credentials (Username and password)
Expected Output	System will send an error message
Actual Result	System has sent an error message
State	Pass

6.8.4 Generate Graph

Table 6.5 shows test case of generating graph through selecting NEWS resource.

Table 6.5: Test Case no.5: Generate Graph by selecting NEWS resource

Test Case ID	5
Function to be tested	Generate Graph by selecting NEWS resource
Initial State	Web application already opened and user has already logged in
Input	NEWS resource and minimum value of support and confidence.
Expected Output	Graph will be generated.
Actual Result	Graph is generated.
State	Pass

Table 6.6 shows test case of generating graph by not selecting NEWS resource.

Table 6.6: Test Case no.6: Generate Graph by not selecting NEWS resource

Test Case ID	6
Function to be tested	Generate Graph by not selecting NEWS resource
Initial State	Web application already opened and user has already logged in
Input	User clicked on Graph link.
Expected Output	No graph will be generated and error message will be sent.
Actual Result	Graph is not generated and error message is sent.
State	Pass

6.8.5 Setting of Graph

Table 6.7 shows test case of changing graph's physics setting.

Table 6.7: Test Case no.7: Change Graph's Physics Setting

Test Case ID	7
Function to be tested	Change Graph's Physics Setting
Initial State	Graph has been generated
Input	**U**ser done changes in the graph setting.
Expected Output	Visualization/physics of graph will change.
Actual Result	Visualization/physics of graph changed.
State	Pass

6.8.6 View Details of Graph

Table 6.8 shows test case of viewing details of entities and their association through selecting NEWS resource.

Table 6.8: Test Case no.8: View details with NEWS resource selected

Test Case ID	8
Function to be tested	View details with NEWS resource selected
Initial State	NEWS resource is selected
Input	**U**ser will click on Details link.
Expected Output	Decision matrix will be displayed
Actual Result	Decision matrix is displayed.
State	Pass

Table 6.9 shows test case of viewing details of entities and their association through not selecting NEWS resource.

Table 6.9: Test Case no.9: View details with NEWS resource not selected

Test Case ID	9
Function to be tested	View details with NEWS resource not selected
Initial State	NEWS resource not selected
Input	**U**ser will click on Details link.
Expected Output	No Decision matrix will be displayed and error message will be sent.
Actual Result	Decision matrix is not displayed and error message is sent.
State	Pass

6.8.7 Search Entity

Table 6.10 shows test case of searching entity's association after selecting NEWS resource.

Table 6.10: Test Case no.10: Search entity with NEWS resource selected

Test Case ID	10
Function to be tested	Search entity with NEWS resource selected
Initial State	NEWS resource is selected
Input	**Step 1:** User will type entity in text box. **Step 2:** User will click on search button.
Expected Output	Graph of association of entities with searched will generate.
Actual Result	Graph of association of entities with searched is generated.
State	Pass

Table 6.11 shows test case of searching entity's association by not selecting NEWS resource.

Table 6.11: Test Case no.11: Search entity with NEWS resource not selected

Test Case ID	11
Function to be tested	Search entity with NEWS resource not selected
Initial State	NEWS resource not selected
Input	**Step 1:** User will type entity in text box. **Step 2:** User will click on search button.
Expected Output	No graph will be generated.
Actual Result	No graph is generated.
State	Pass

Table 6.12 shows test case of user click search button while text box is empty.

Table 6.12: Test Case no.12: Click search button with empty text box when NEWS resource is selected

Test Case ID	12
Function to be tested	Click search button with empty text box when NEWS resource is selected
Initial State	NEWS resource is selected
Input	User will click on search button.
Expected Output	No graph will be generated.
Actual Result	No graph is generated.
State	Pass

6.8.8 Log out

Table 6.13 shows test case of user log out from web application.

Table 6.13: Test case no.13: User Log out

Test Case ID	13
Function to be tested	User Log out
Initial State	User has been logged in
Input	User will click on log out button.
Expected Output	User account will log out.
Actual Result	User account is log out.
State	Pass

6.9 Result

We have extracted NEWS articles from the NEWS resources mentioned in Chapter 5 and train our model to extract entities and the associate of them with each other. The results are accurate up to 80% on the extracted data.

Chapter 7

Conclusion and Future Work

The final year project **"Entity Association Mining"** has come to its end, with the web application that is fully functional and working. Association rule mining and NLP has been an exciting field to work on. The aim of this project was to develop a web application that allows people to analyze NEWS articles with the perspective of identifying the association between entities. After the hardworking of us and the guidance and support of our supervisor, the system has been developed successfully.This project was started with the aim of learning a lot in the field AI and data mining.
In this busy world, people always think that technology can make their life easy. This project is our first step towards making their life a bit more easy. This web application can make a huge impact on the working life of our journalism community. We have learnt a lot while developing this project with a well reputed government organization's team in severe pandemic environment across the country. Thanks to Allah Almighty for helping us to bring this project into life in these difficult times.
While developing this web application, we have learnt a lot about NLP and how information can be retrieved from web platforms. We have done research on association rule mining algorithms. For interface creation, we have learnt HTML, CSS, JavaScript and PHP. Also, we have learnt that how front end can be integrated with python scripts and we can make python scripts data secure using pickle library. This project helped us a lot to gain industrial experience and a chance to worked with an experienced team of well reputed organization.

7.1 Recommendations

There are more features that can added to this application. Like add more NEWS resources data to the application to check association of entities in them.

7.2 Learning Outcomes

From this final year project, the learning outcomes can be listed down as follows:

- Project planning needs to be as detailed and comprehensive as possible
- Deadlines should be set with a realistic approach
- Technology for development should be chosen while keeping cost-effectiveness and project scope in consideration
- Project deadlines should be met at any cost

The overall learning experience can be used to initiate and complete projects in IT firms.

7.3 Future Work

This is the first version of this application with an aesthetic GUI and limited functionalities. The initial goal of the project was to bring it into reality with limited resources. In future, we hope there is still room of improvement through the following aspects:

7.3.1 Run on multiple devices

The application is only limited to web browser. It will be released for desktop and mobile devices both as well in future.

7.3.2 Add more NEWS resources

There is a limited amount of NEWS resources in the web application. In future more NEWS resources will be added.

7.3.3 Expand dependency of associations

In the developed project, we have only incorporated entities. In future, title, content and NEWS resource of NEWS articles will be associated.

References

[1] Anisha Garg. Complete guide to association rules (1/2), 2018.

[2] Won Young Kim, Joon Suk Ryu, Kyu Il Kim, and Ung Mo Kim. A method for opinion mining of product reviews using association rules. In *Proceedings of the 2nd international conference on interaction sciences: Information technology, culture and human*, pages 270–274, 2009.

[3] Trupti A Kumbhare and Santosh V Chobe. An overview of association rule mining algorithms. *International Journal of Computer Science and Information Technologies*, 5(1):927–930, 2014.

[4] Elizabeth D Liddy. Natural language processing. 2001.

[5] Si Jie Phua, Wee Keong Ng, Haifeng Liu, Bin Song, Xiang Li, et al. A rule mining approach to emotional design in mass customization. In *DS 42: Proceedings of ICED 2007, the 16th International Conference on Engineering Design, Paris, France, 28.-31.07. 2007*, pages 813–814, 2007.

[6] Wikipedia contributors. Pandas (software) — Wikipedia, the free encyclopedia. https://en.wikipedia.org/w/index.php?title=Pandas$_(software)oldid = 1059759509, 2021. [$Online; accessed 16 - January - 2022$].

[7] Wikipedia contributors. Systems architecture — Wikipedia, the free encyclopedia. https://en.wikipedia.org/w/index.php?title=Systems$_architectureoldid = 1043291148, 2021. [$Online; accessed 16 - January - 2022$].

[8] Wikipedia contributors. Python (programming language) — Wikipedia, the free encyclopedia. https://en.wikipedia.org/w/index.php?title=Python$_(programming_language)oldid = 1065081237, 2022. [$Online; accessed 16 - January - 2022$].

[9] Qiankun Zhao and Sourav S Bhowmick. Association rule mining: A survey. *Nanyang Technological University, Singapore*, 135, 2003.