Social Media Mine

By

Raja Aneeq Ashraf

Qanita Hassan

Asad Hussain

Asiya Mushtaq

Submitted to the Faculty of Computer Science

National University of Sciences and Technology, Rawalpindi in partial fulfillment for the requirements of a B.E Degree in Computer Software Engineering

May 2014

CERTIFICATE

Certified that the work contained in this thesis entitled **“Social Media Mine”** carried out by Raja Aneeq, Qanita Hassan, Asad Hussain and Asiya Mushtaq under the supervision of Dr. Hammad Afzal for partial fulfillment of Degree of Bachelor of Software Engineering is correct and approved.

Supervisor: _____

Dr. Hammad Afzal

_____ Department

MCS

Dated: _____

ABSTRACT

SOCIAL MEDIA MINE (REAL TIME ANALYZER OF SOCIAL MEDIA)

The Social Media Mine is designed to now-cast events in Pakistan. As per user choice, the relevant events would be plotted on map. This software system is a Real-time Analyzer for Social websites (Twitter) and News. This system provides the user, the opportunity to view the location of various events taking place in different cities of Pakistan. By mapping the user's required events on the Google map, the system shall meet the user's needs efficiently.

The system consists of a News Information Extractor that keeps track of news information from news websites. It searches and extracts customized information (related to news) based on user preferences. A Status (Tweets) Information Extractor keeps track of status (tweets) information from twitter website. It searches and extracts customized information (related to tweets) based on user preferences. An Event Information Storage that stores information extracted about news/tweets. Which is used further in pattern recognition and event mapping. A Pattern Extractor that identifies pre-defined events from the news/tweets and the location and time of the event. An Event Mapper module that plots the event on the Google maps.

Social Media Mine is a new product. The conception of its idea was originated with the aim of providing combine real time information from different types of media (news/social). Real time monitoring of social conditions is an important part of developing early warning systems for social and economic crunches for a particular region. Information from different media is available but there is no product which integrates this information. It is possible to gather valuable information by monitoring the social communication of people. A rich source of getting this info is social media, such as Twitter and News websites like Nation and Dawn news. By the analysis of this data one can get the opinion of people. Social media mine offers a new experience to users which will be time saver and different at the same time.

DECLARATION

We very solemnly declare that the work presented herewith is the result of sole effort of our group, comprising of Raja Aneeq, Qanita Hassan, Asad Hussain and Asiya Mushtaq, and is free of any kind of plagiarism in part or whole. We also declare that the dissertation has never been submitted previously in part or whole in support of another award or qualification either at this institution or elsewhere.

DEDICATION

To our respected teachers whose kind guidance and unfailing support made this mammoth task easy for us and to our very dear parents whose unceasing prayers and support gave us strength and courage to complete the work of this magnitude.

And

To Google, which made it all possible.

ACKNOWLEDGMENTS

We are, very humbly, grateful to Almighty Allah for bestowing us with the strength and resolve to undertake and complete the project. We owe a special debt of gratitude to our supervisor, Dr. Hammad Afzal for the continuous supervision, motivation and support provided to us and for their continuous and valuable suggestions, guidance, and instructions from time to time right through the project. We would also like to acknowledge and thank the faculty members of the department of Computer Software Engineering for their continuous support. It was their help and guidance which helped us complete the project in due time. We are also thankful to our class mates and our family members especially our parents for their support throughout the project in every possible way. Their faith kept us going.

Table of Contents

List of Figures

List of Tables

CHAPTER 1:
INTRODUCTION

1. Introduction

1.1 Purpose

The purpose of this document is to present a detailed description of the Social Media Mine. It will explain the purpose and features of the system, the interfaces of the system, what the system will do, the constraints under which it must operate and how the system will react to external stimuli. This document is intended for both the stakeholders and the developers of the system and will be proposed to the Evaluation Panel for its approval.

1.1.1 Document Conventions

- Words in bold in any paragraph refer to a specific term defined earlier or later in the document.
- The following pair of words have been used interchangeably in the document.
    - "Software" and "System".
    - "Device" and "Android Device"
- The word "System" refers to "Social Media Mine on Web".

1.2 Motivation

Real time monitoring of social conditions is an important part of developing early warning systems for social and economic crunches for a particular region. It is possible to gather valuable information by monitoring the social communication of people. By the analysis of large number of messages(Tweets) from a famous social networking site Twitter and extracted news from leading news websites i.e. The Nation and Dawn News we can deduce the information regarding different events taking place in the country. System purpose is to provide the users with the details of the on-going events in Pakistan

with their location mapped on the Google maps. We are not talking of forecasting or predicting, but we are using the term “nowcasting” to indicate the fact that we are performing real time measurements on the present state of the data.

1.3 Project Scope

The Social Media Mine is designed to now-cast events in Pakistan. As per user choice, the relevant events would be plotted on map. This software system will be a Real-time Analyzer for Social websites (Twitter) and News. This system will provide the user, the opportunity to view the location of various events taking place in different cities of Pakistan. By mapping the user’s required events on the Google map, the system will meet the user’s needs while remaining easy to understand and use. The Goal of the system is to provide our user the opportunity to view the location of various events taking place in different cities of Pakistan.

1.3.1 Project Vision

For	The software is intended for anyone who is interested in finding the location of an event
What	For a well-integrated information of on-going events
Name	Social Media Mine
Is	Real Time Social Media Analyzer
That	Facilitate the users in finding the details and the place of occurrence of a particular event.
Unlike	Tuning on the Television and listening to news
Our product	Will be the only available Real time social media analyzer

Table 1 Project Vision

1.4 Project Objective

The objective is to develop a web and android application that shall be installed on a web server and the android device. It shall allow the administrator to manage the extracted data from Twitter and news and shall allow the user to view the details and location of ongoing events in the country.

1.5 Deliverables

- 1st Progress Report: including SRS Document
- 2nd Progress Report: including System Design Document
- 3rd Progress Report: including Demonstration
- Final Report: including complete documentation

CHAPTER 2:

LITERATURE REVIEW

2. Literature Review

2.1 Sentiment Analysis and Opinion Mining [1]

Sentiment analysis and opinion mining aim to automatically extract opinions expressed in the user-generated content.

2.1.1 Description

Sentiment analysis and opinion mining tools allow businesses to understand product sentiments, brand perception, new product perception, and reputation management. These tools help users to perceive product opinions or sentiments on a global scale. There are many social media sites reporting user opinions of products in many different formats. Monitoring these opinions related to a particular company or product on social media sites is a new challenge. Sentiment analysis is hard because languages used to create contents are ambiguous.

Major steps of sentiment analysis are

- Finding relevant documents
- Finding relevant sections
- Finding the overall sentiment
- Quantifying the sentiment
- Aggregating all sentiments to form an overview

Where basic components of an opinion are:

- an object on which opinion is expressed
- an opinion expressed on a object
- the opinion holder.

Objects are generally represented as a finite set of features, where each feature represents a finite set of synonymous words or phrases. Opinion mining tasks can be performed at

- The document level
- Sentence level
- feature level

2.2 Tweet Tracker [2]

Tweet Tracker is a Twitter-based analytic and visualization tool.

2.2.1 Description

The focus of the tool is to help HADR relief organizations to acquire situational awareness during disasters and emergencies to aid disaster relief efforts (Kumar et al. [3]). New social media platforms, such as Twitter microblogs, demonstrate their value and ability to provide information that is not easily attainable from traditional media. For example, during the Mumbai blasts of 2011 [4], firsthand information from the affected region was available on Twitter moments after the blast. TweetTracker is designed to help track, analyze, review, and monitor tweets. This is achieved through near-real-time tracking of tweets with specific keywords/hashtags and tweets generated from the region affected by the crisis. The tool supports monitoring and analysis of the collected tweets via real-time trending, data reduction, historical review, and integrated data mining techniques.

2.2.2 Components

TweetTracker consists of three main components:

- A Twitter stream reader
- A data storage module
- A data mining and visualization module.

The ***Twitter stream reader*** is a data collection module that continually crawls tweets through the Twitter streaming API (Application Programming Interface). Tweets are filtered based on user-specified keywords, hashtags, and geo-locations.

The ***data storage module*** is responsible for storing and indexing the collected tweets into a relational database for use by the visualization module.

The ***data mining and visualization module*** is a Web-based user interface to the collected tweets and a means to analyze the collected tweets. It provides geospatial visualization of tweets related to a particular event on a map, summarizes the tweets, and visualizes the trending keywords in the form of a word cloud, and it can identify popular resources (URLs) and users mentioned in the tweets. The tool also includes built-in language translation support for monitoring of multilingual tweets.

TweetTracker has been used in tracking, visualizing, and analyzing activities including the Arab Spring movement, the Occupy Wall Street movement, and various natural disasters such as earthquakes and cholera outbreaks.

2.3 Open Domain Event Extraction from Twitter [5]

2.3.1 Description

TwiCal extracts a 4-tuple representation of events which includes a named entity, event phrase, calendar date, and event type (see Table 1). This representation was chosen to closely match the way important events are typically mentioned in Twitter.

Given a raw stream of tweets, TwiCal system[6] extracts named entities in association with event phrases and unambiguous dates which are involved in significant events. First the tweets are POS (Parts of Speech) tagged, then named entities and event phrases are extracted, temporal expressions resolved, and the extracted events are categorized into types. Finally, the strength of association between each named entity is measured and date based on the number of tweets they co-occur in, in order to determine whether an event is significant.

2.4 Informing the Curious Negotiator [7]

The curious negotiator is an agent systems for negotiations described by **Debbie Zhang and Simeon J. Simoff**. The overall goal of its design was to exploit the interplay between contextual information and the development of offers in negotiation conducted in an electronic environment.

2.4.1 Description The negotiated items are pieces of information, specifically news articles. As a part of the system articles must be fetched, classified and stored for later use by negotiation agents.

A data extraction agent added to the system performs the following three stages:

- Data extraction
- Text filtering using a dynamic filter
- Keyword validation

The key ideas behind the extraction is that websites are constructed from hidden or visible nested tables in combinations with Cascading Style Sheet, Level 1 (CSS1) and that a ***news article*** is the largest block of text in a web page.

Extracting news from a page becomes ***the task of identifying the largest portion of text*** in a table. This is done by inserting all html tags of a page into an array, removing all tags which are not a table <table>.....</table> or within table tags. Iterate the array, for each text item, append it to a container which holds all text at this nesting depth. Once the array has been iterated, the container with the largest amount of text is returned.

A second page, preferably similar to the one first is then fetched, and extraction done in the same manner. The result is compared to the extracted body of text from the first page. Any identical sentences are considered static parts of the page and are then removed, filtered, from the end result.

To ensure that the UR L used during extraction was a valid one, or that nothing else went wrong, the result is validated. Validation is performed using keywords that should reasonably occur in the text. The keywords that are used are words appearing in the title, (except for stop words). If they are found to a satisfiable degree the text is accepted as an article.

2.5 ECON: An Approach to Extract Content from Web News Page(8)

A simple but effective approach, named ECON, to fully-automatically extract content from Web news page.

2.5.1 Description

ECON uses a DOM tree to represent the Web news page. ECON finds a snippet-node by which a part of the content of news is wrapped firstly, then backtracks from the snippet-node until a summary-node is found, and the entire content of news is wrapped by the summary-node.

Experimental results showed that ECON can achieve high accuracy and fully satisfy the requirements for scalable extraction. There have been many approaches existed to extract content from Web based on the techniques they use, the approaches can be divided into three classes:

- A ***wrapper*** can be generated by wrapper induction system for content extraction, since there are so many heterogeneous news sources, it is not practical to build wrappers for each news source.
- ***Techniques of Web mining***, such as classification and clustering. These approaches can improve the accuracy of extraction. However most of them need human interventions and the complexity of the underlying algorithms is not low, so this class of approaches has limited ability for scalable extraction.
- ***Extract content from Web page*** based on statistics. These approaches can usually perform the extraction in an unsupervised fashion, which is crucial for our task. However most of them rely on some weights or thresholds that are usually determined by some empirical experiments.

2.5.2 Algorithm of Joint-para:

In a DOM tree of a Web news page, it can be observed that sometimes the entire text of news is broken into many short pieces by some nodes such as <p> and
. The input of Joint-para is a big-node. Joint-para checks its brother nodes to find a text-node-set.

And then get the text-para of the text-node-set and compute the punctuation-num of the text-para. If it is 0, the text-para will be regarded as noise and will not be output. Meanwhile, all the nodes that together wrap the noise piece are pruned. If the punctuation-num is not 0, the text-para will be output.

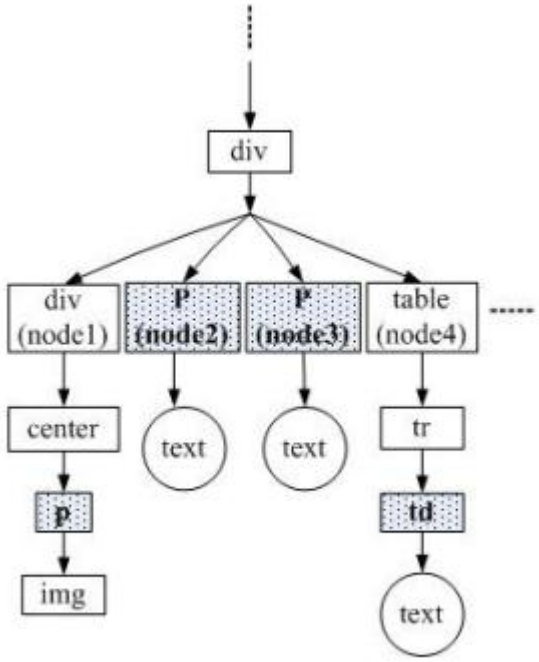


Figure 2-1 Algorithm of Joint-para

2.5.3 Algorithm of Extract-news:

The heuristics to detect when to stop backtracking is from such observation:

When backtracking from node1 to node2, if the content of news wrapped by node2 is more than node1, node1 must not be the summary-node, and the node-punc-num of node2 must be more than node1. Firstly, Extract-news transverse the DOM tree and perform the algorithm of Joint-para f or each big-node to get all text-paras. Then select randomly one node from the text-node-set that wraps the longest text-para on the way of backtracking, a sequence of distance can be obtained. The process of backtracking stops at the following condition: The distance appears 0 for the first time. For the distance 0, the child-node is regarded as the summary-node.

Two experiments were performed to check the efficiency of ECON and to compare ECON and CoreEx. ECON extracted correctly an average of 90% of the news pages, while 6% were wrongly extracted and 4% were missed. CoreEx extracted correctly an average of

92% of the news pages, while 2% were wrongly extracted and 6% were missed. It was concluded that ECON can deal with the testing pages nearly as well as CoreEx.

2.6 An Effective and Efficient Web News Extraction Technique for an Operational NewsIR System [9]

An automated approach to news recognition and extraction based on a set of heuristics about the articles structure, which is currently applied in an operational system.

2.6.1. Description

It proposed a method for news recognition and extraction from a set of web documents, based on domain specific heuristics that takes advantage of the common characteristics of the articles structure, resulting in an efficient and effective algorithm that solves the news data extraction problem.

It proposed a set of heuristics to identify and extract the article, or reject a not - news web page:

- News are composed of paragraphs that are next one each other, although it could appear some not desired content between them.
- Paragraphs have a minimum size. News also have a minimum size.
- Paragraphs are mostly text. Only styling markup and hyperlinks are allowed in paragraphs. Markup abounds in not desired content.
- A low number of hyperlinks are allowed in paragraphs. A high number indicates not desired content, probably section pages.

The evaluation routine had two phases. The first phase the proposed algorithm was run that implemented the heuristics against the data set. Result of this step records the set of new articles obtained. The extracted article set is then compared with the set of news extracted by URL pattern criteria.

After this stage the false positives (news extracted by the algorithm with an URL that does not match with the pattern criteria) and false negatives (that have not been extracted using our heuristics but which URL match with the pattern criteria) are manually inspected to determine whether they are news articles or not.

The implementation of these heuristics resulted in a linear complexity algorithm on the web page length. The algorithm followed the content of the HTML documents looking for paragraphs that match with the criteria and joins them to build the news body. They also designed an original evaluation strategy and built a data set that could be used by others. The evaluation assessed that the approach obtains very high values of precision and recall. This method is currently working in an operative news system

This method is used because it is easy to tune the parameters implied in the heuristics. A challenge that it considers adapts their algorithm to other fields and tasks where the content to recognize would be much more variable. Studies about the stability of the parameters and the use of statistical learning approaches to estimate the best parameter values would be necessary in these domains.

2.7 Nowcasting Events from the Social Web with Statistical Learning [10]

2.7.1. Description

The term “nowcasting”, commonly used in finance, expresses the fact that we are making inferences regarding the current magnitude $M(\varepsilon)$ of an event ε. For a time interval $u = [t - \Delta t, t]$, where t denotes the current time instance, consider $M(\varepsilon(u))$ as a latent variable. The web content $W(u)$ for this time interval is a partially observed variable; in particular, data from a social network, denoted as $S(u) \subseteq W(u)$ are being observed. In this work, $S(u)$ is used to directly infer $M(\varepsilon(u))$. For short time intervals u, we are inferring the present value of the latent variable, i.e. we are nowcasting the magnitude of an event.

The researchers presented a general framework for exploiting user input published in social media. It used the concept of Sparse learning that enables us to select a

consistent set of features (e.g. unigrams or bigrams) and then use it to perform inference via regression

The performance of the proposed methodology in the paper was evaluated by investigating two case studies.

- The daily amount of rainfall in five UK locations by using tweets
- The level of Influenza-like Illness (ILI) in the population of three UK regions based again on geo located Twitter content.

General claim was that statistical learning techniques can be deployed for the selection of features and, at the same time, for the inference of a useful statistical estimator. It used an algorithms i.e. Baseline Method: Feature Selection via Correlation Analysis.

The proposed methodology in the paper used tweets tagged with the location (longitude and latitude coordinates) of their author only. They used UK's 54 most populated urban centers and collected tweets geo-located within a 10km range from each one of them. The crawler exploited Atom feeds and periodically retrieved the 100 most recent tweets per urban centre.

All collected tweets were stored and indexed in a MySQL database. Text preprocessing such as stemming by applying Porter's Algorithm for English language [Porter 1980], stop word and punctuation removal as well as the computation of Vector Space representation (VSR) were performed by the software libraries.

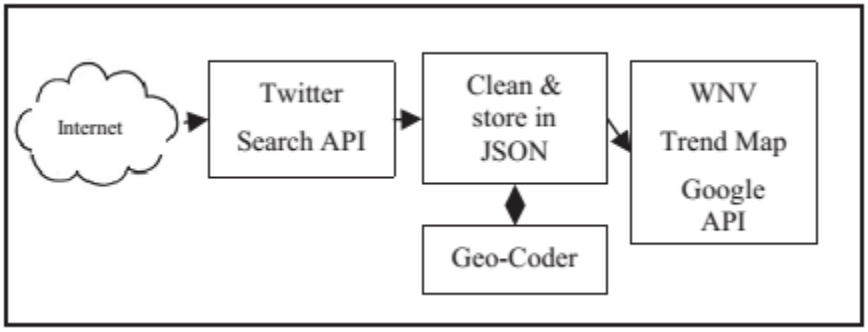
The results were drawn from two case studies, i.e. the benchmark problem of inferring rainfall rates and the real-life task of detecting the diffusion of Influenza-like Illness from tweets. In both case studies, the majority of selected features were directly related with the target topic and inference performance had been significant.

2.8 Real-time Spatio-temporal Analysis of West Nile Virus Using Twitter Data [11]

2.8.1. Description

West Nile virus (WNV) is one of the most geographically widespread parvoviruses in the world with cases occurring on all continents except Antarctica. The goal of study was to understand a real-time spatial temporal WNV activity using Twitter data. In this study, tweets for the entire world were collected using Twitter Search API with tags #WestNileVirus, and #WNV from August 31, 2011. Collected tweets were stored, cleaned, and geocoded. The Google API was used to display information on the web. The changes per week showed that the numbers were relatively high from August through October then gradually slowed down from

December through March. Research also found a very large increase in tweet numbers from March and April. This may be due to unusual higher temperature and mosquito activities in March and April.



CHAPTER 3:

Overview of Social Media Mine

3. Overview of Social Media Mine

3.1 Product Perspective

Social Media Mine is a new product. The conception of its idea was originated with the aim of providing combine real time information from different types of media (news/social).Information from different media is available but there is no product which integrates this information. Social media mine offers a new experience to users which will be time saver and different at the same time. Following diagram gives an overview of the Social Media Mine.

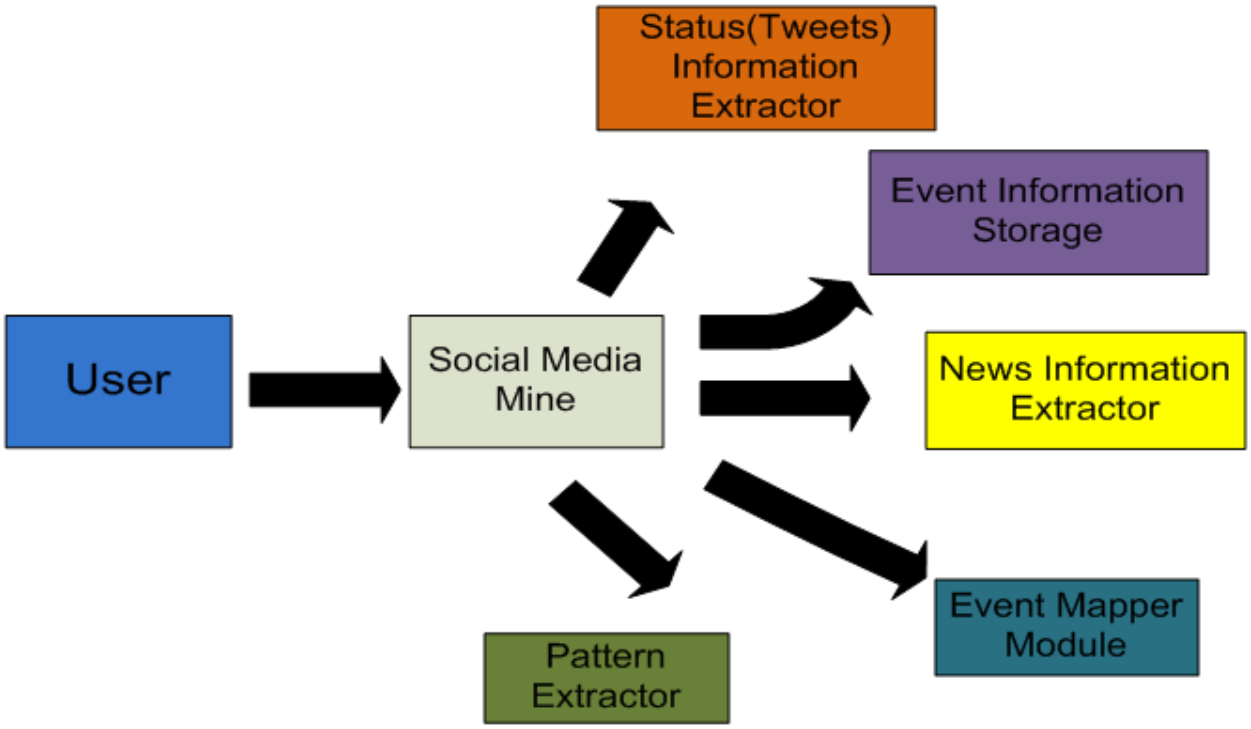


Figure 3-1 Product Perspective

3.2 Modules of the Project

3.2.1 News Information Extractor

The module helps keeping track of news information from news websites. This module will search and extract customized information (related to news) based on user preferences.

3.2.2 Status(Tweets) Information Extractor

The module helps keeping track of status (tweets) information from twitter website. This module will search and extract customized information (related to tweets) based on user preferences.

3.2.3 Event Information Storage

This module will store information extracted about news/tweets. This information will be used further in pattern recognition and event mapping.

3.2.4 Pattern Extractor

This module will identify pre-defined events from the news/tweets. It will also identify the location and time of the event.

3.2.5 Event Mapper

This module will plot the event on the maps.

3.3 Product Functions

- News Search: It allows user to search information from news websites based on user preferences.
- Manage tweets and News: It allows the administrator to manage the extracted news and tweets.
- Status (Tweets) Search: It allows user to search information from twitter based on user preferences.

- Selecting Location of Events: It allows user to select a particular city from the list in order to restrict searching of events to a particular city.
- Events Keyword Selection: It allows user to select an event keyword to search events related to that keyword.
- View Events List: It allows user to see list of events based on user previous preferences.
- View Events Description: It allows user to see details of a particular event that user has already selected.
- View Location on Map: It allows user to see location of a particular event that user has already selected on map.

Following diagram demonstrates the product functions of the Social Media Mine.

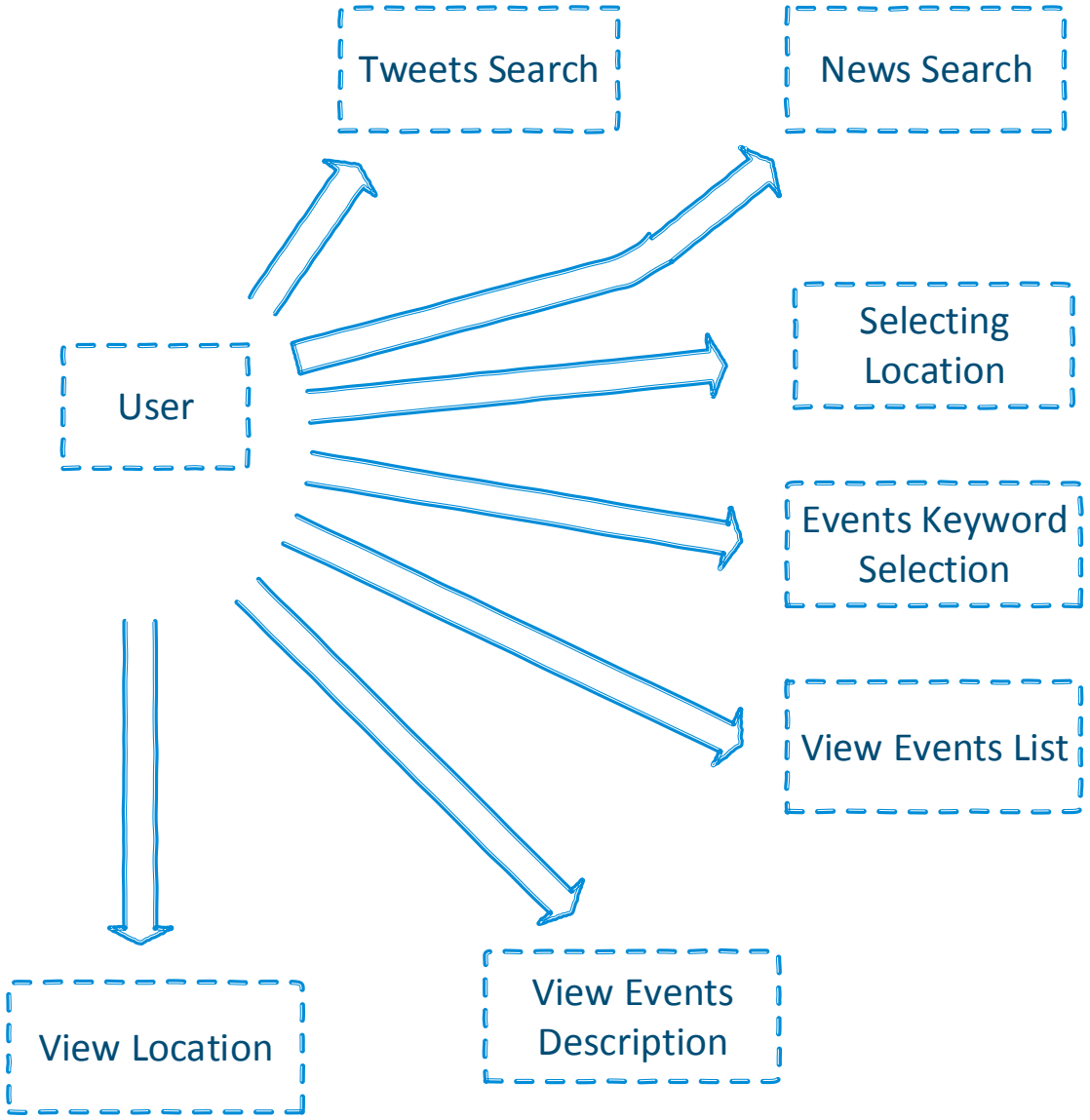


Figure 2: Product Functions

3.4 Operating Environment

Operating System	Windows 7/8 and android based mobile operating system
Client	Any Web Browser

Web Server	IIS server
Database	SQL server
Hardware Platform	Desktop/Laptop and android enabled device

Table 2: Operating Environment

3.5 Assumptions and Dependencies

Assumptions	
AS – 1	Assume that the project started on September 12, 2013.
AS – 2	Assume that the faculty members of the Department of CSE are the client party for this project.
AS – 3	Assume that the individuals in MCS form the user population, which for the moment, does not exceed 20 users.
Dependencies	
D-1	There will be a permanent dependency on Google Maps and reliability of system will depend on continuous / uninterrupted data from Google-Maps.
D-2	System is dependent on Twitter API to extract tweets for a good mix of information.
D-3	System is dependent on SQL server to store our data.

Figure 3-2 Assumptions and Dependencies

CHAPTER 4:

SYSTEM REQUIREMENTS & SPECIFICATION

4. System Requirements & Specification

4.1 System Features

4.1.1. Search Tweets for Location

4.1.1.1. Description and Priority:

This use case describes how the system will search the extracted tweets based on city. The system shall take the keyword (city) from the user and then search the extracted tweets on the basis of the keyword.

Priority = High

4.1.1.2. Stimulus and Response Sequence:

Stimulus: The user requests to view the location and details of some event.
Response: The system queries user for the city.

Stimulus: The user selects a city.
Response: The system searches the extracted tweets (on the basis of location), marks events and displays the link to the details of event(s).

Stimulus: The user clicks on a link.
Response: The system displays the details of event(s) and a google map with the marked location.

4.1.1.3 Functional Requirements (FR):

FR #1: The system shall get the location (city) from user.
FR #2: The system shall extract location-based tweets from Twitter.
FR #3: The system shall store the extracted tweets in database.

FR #4: The system shall analyze the tweets to extract event information.
FR #5: The system shall display the detail of event on web-page.
FR #6: The system shall mark the location of event on the map.

4.1.1.4. Use Case Description

USE CASE NAME	Search Tweets for Location
ACTOR	User, Tweets Information extractor
NORMAL COURSE	1) The process is initiated when the user selects a location from a list of cities. 2) System shall search the extracted geo-located tweets as per the location chosen by the user. 3) System shall analyze all the tweets to get information about on-going events. 4) System shall mark all categories of event going-on in the selected location on Google map. 5) System shall display a link to the detail of the event(s).
ALTERNATE COURSE	5) (a) If no event is found in the location, System shall display a message of "No event Going" on the screen.
PRE CONDITION	✓ A list of cities is predefined in the system. ✓ News have been extracted from Nation and Dawn news websites.
POST CONDITION	✓ A marker (pointer) is set on the places where the events are taking place on the Google map. ✓ Link(s) to the event(s) is/are displayed.
ASSUMPTIONS	✓ Events are taking place in various parts of the country. ✓ Geo-located tweets exists on twitter and have been extracted.

Table 2 Search Tweets for Location

4.1.1.5. Use Case Diagram

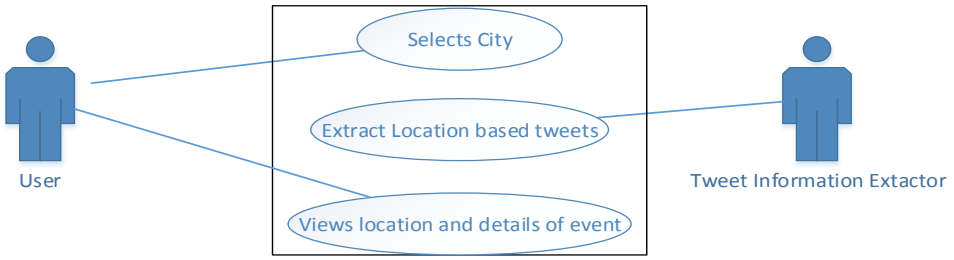


Figure 4-1 Search Tweets for Location

4.1.2 Search Tweets for Keyword

4.1.2.1 Description and Priority:

This use case describes how the system will search the extracted tweets, based on keyword. The system shall take the keyword from the user and then search the extracted tweets on the basis of the keyword.

Priority = High

4.1.2.2 Stimulus and Response Sequence:

Stimulus: The user requests to view the details and location of event(s).
Response: The system queries user for the keyword.

Stimulus: The user enters a keyword.
Response: The system searches the extracted tweets (on the basis of keyword), marks events and displays the link to the details of event(s).

Stimulus: The user clicks on a link.
Response: The system displays the details of event(s) and a google map with the marked location.

4.1.2.3 Functional Requirements (FR):

FR #1: The system shall get a keyword from user
FR #2: The system shall search the extracted tweets on the basis of the keyword.
FR #3: The system shall analyze the tweets to extract event information.
FR #4: After analysis, the system shall mark the location of event on the map.
FR #5: The system shall display the detail of event on web-page.

4.1.2.4. Use Case Description

USE CASE NAME	Search Tweets for Keyword
ACTOR	User, Tweets Information extractor
NORMAL COURSE	1) The process is initiated when the user enters a keword for search. 2) System shall analyze all the tweets to get information about on-going event related to the keyword. 3) System shall mark all events related to keyword, on Google map. 4) System shall display links to detail of event(s).
ALTERNATE COURSE	5) (a) If no event is found, system shall display a message of "No event Going" on the screen.
PRE CONDITION	✓ Geo located tweets have been extracted already.
POST CONDITION	✓ A marker (pointer) is set on the places where the events are taking place on the Google map. ✓ Links to the detail of event(s) are displayed.
ASSUMPTIONS	✓ Events are taking place in various parts of the country. ✓ Geo-located tweets exists on twitter.

Table 3 Search Tweets for keyword

4.1.2.5. Use Case Diagram

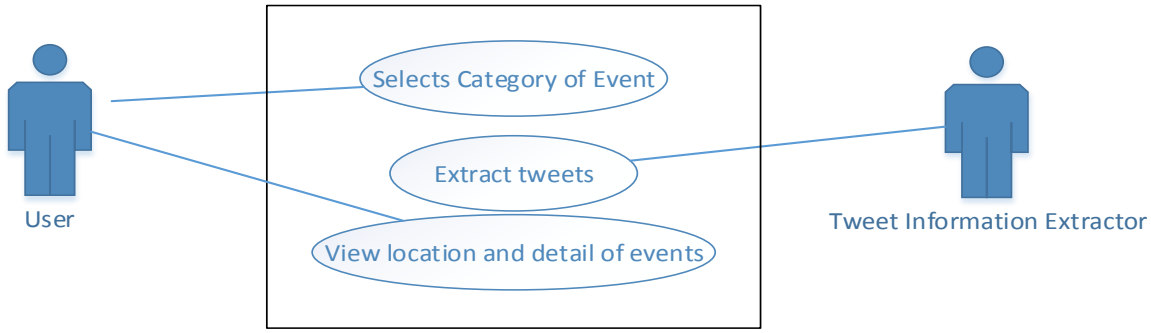


Figure 4-2 Search Tweets for keyword

4.1.3 Mark Location on Google Map:

4.1.3.1 Description and Priority:

This use case describes the process with which the events are marked on the Google maps (by the system). The user enters a keyword, the system searches for events related to the keyword and marks their location on the Google map.

Priority = High

4.1.3.2 Stimulus and Response Sequence:

Stimulus: The user requests to search for events related to a keyword.
Response: The system displays a list of event's link related to the keyword.

Stimulus: User selects a link by clicking on it.
Response: System displays the details of the event with its location marked on the Google map.

4.1.3.3 Functional Requirements (FR):

FR #1: The system shall get the analyzed data from database (i.e. event location and description).
FR #2: The system shall mark the event on Google Map.

4.1.3.4. Use Case Description

USE CASE NAME	Mark Location on Google Map
ACTOR	User, Event mapper
NORMAL COURSE	1) The process is initiated when the user enters a keyword for search. 2) The system searches the extracted tweets and news for the keyword. 3) A list of links to events is displayed by the system. 4) The user clicks on a link. 5) A Google map is displayed on the screen, which has event marked on it along with the details of the event.
ALTERNATE COURSE	2) (a) If no event is found in the location System shall display a message of "No event Going" on the screen.
PRE CONDITION	✓ Tweets and news have been extracted and saved in the database.
POST CONDITION	✓ The system displays the extracted event's information from tweets and news. ✓ A marker (pointer) is set on the place where the event is taking place on the Google map.
ASSUMPTIONS	✓ Events are taking place in various parts of the country. ✓ Geo-located tweets exists on twitter.

Table 4 Mark Location on Google Map

4.1.3.5. Use Case Diagram

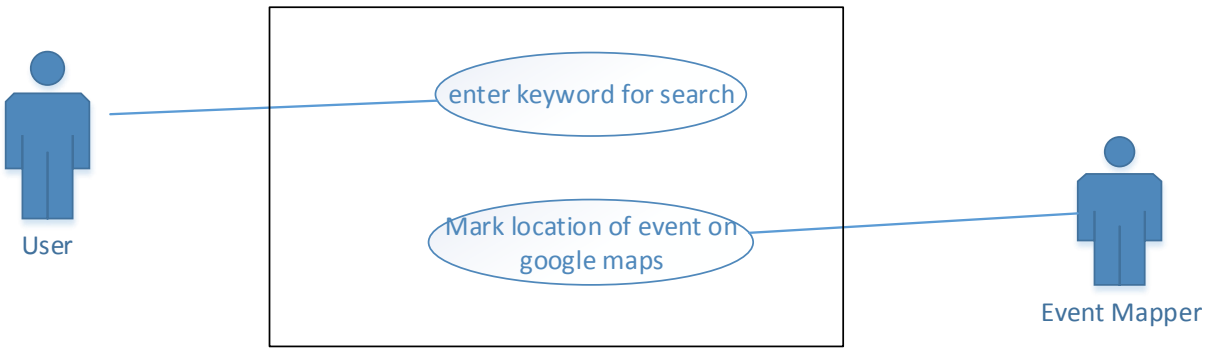


Figure 4-3 Location on Google map

4.14 Search Extracted News for Keyword:

4.1.4.1 Description and Priority:

This use case describes the process through which the system searches the extracted news (from the Dawn and Nation news websites) for the keyword. The system shall take the keyword from the user and then perform the search on the extracted news.

Priority = High

4.1.4.2 Stimulus and Response Sequence:

Stimulus: The user requests to view the details and location of event(s).
Response: The system queries user for the keyword.

Stimulus: The user enters a keyword.
Response: The system searches the extracted news (on the basis of keyword), marks events and displays a list of links, of event(s).

Stimulus: The user clicks on a link.
Response: The system displays the details of event(s) and a google map with the marked location.

4.1.4.3 Functional Requirements (FR):

FR #1: The system shall get a keyword from user.
FR #2: The system shall extract news based on keyword provided by user.
FR #3: The system shall analyze the news to extract event information.
FR #4: After analysis, the system shall mark the location of event on the map.
FR #5: The system shall display the detail of event on web-page.

4.1.4.4. Use Case Description

USE CASE NAME	Search Extracted News for Keyword
ACTOR	User, Pattern extractor
NORMAL COURSE	1) The process is initiated when the user enters a keyword for search. 2) System shall analyze all the extracted news to get information about on-going event related to the keyword. 3) System shall mark all events related to keyword, on Google map. 4) System shall display links to detail of event(s).
ALTERNATE COURSE	5) (a) If no event is found in the location System shall display a message of "No event Going" on the screen.
PRE CONDITION	✓ News have been extracted from Nation news and Dawn news already.
POST CONDITION	✓ A marker (pointer) is set on the places where the events are taking place on the Google map. ✓ Links to the detail of event(s) are displayed.
ASSUMPTIONS	✓ Events are taking place in various parts of the country. ✓ News exists on the Nation news and Dawn news websites.

Table 5 Search Extracted News for Keyword

4.1.6.5. Use Case Diagram

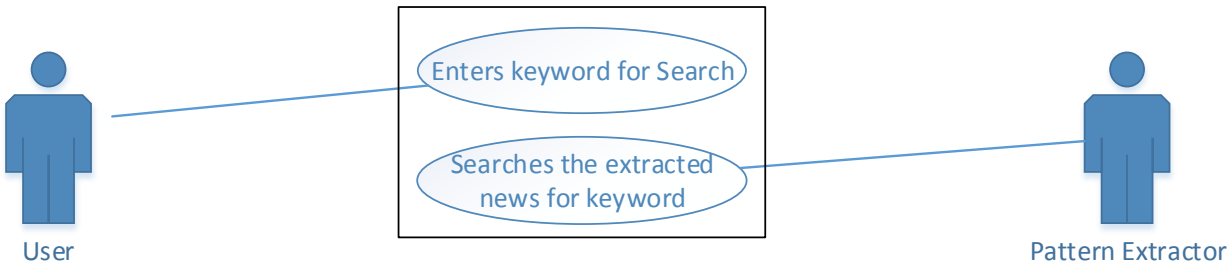


Figure 4-4 Search Extracted News for Keyword

4.1.5 Search News for Location:

4.1.5.1 Description and Priority:

This use case describes how the system will search the extracted news based on city selected by user. The system shall take the keyword (city) from the user and then search the extracted news on the basis of the keyword.

Priority = High

4.1.5.2 Stimulus and Response Sequence:

Stimulus: The user requests to view the location and details of some event.
Response: The system queries user for the city.

Stimulus: The user selects a city.
Response: The system searches the extracted news (on the basis of location), marks events and displays the link to the details of event(s).

Stimulus: The user clicks on a link.
Response: The system displays the details of event(s) and a google map with the marked location.

4.1.5.3 Functional Requirements (FR):

FR #1: The system shall get the city from user.
FR #2: The system shall analyze the news to extract event information.
FR #3: After analysis, the system shall mark the location of event on the map.
FR #4: The system shall display the detail of event on web-page.

4.1.5.4. Use Case Description

USE CASE NAME	Search News for Location
ACTOR	User, pattern extractor
NORMAL COURSE	1) The process is initiated when the user selects a location from a list of cities. 2) System shall search the extracted news as per the location chosen by the user. 3) System shall analyze all the news to get information about on-going events. 4) System shall mark all events going-on in the selected location on Google map. 5) System shall display a link to the detail of the event(s).
ALTERNATE COURSE	5) (a) If no event is found in the location System shall display a message of "No event Going" on the screen.
PRE CONDITION	✓ A list of cities is predefined in the system. ✓ News have been extracted from Nation and Dawn news websites.
POST CONDITION	✓ The system extracted event's information from news. ✓ All the events taking place in the selected city are marked on the Google map. ✓ Link(s) to the event(s) is/are displayed.
ASSUMPTIONS	✓ Events are taking place in various parts of the country. ✓ News are available on websites.

Table 6 Search News for Location

4.1.5.5. Use Case Diagram

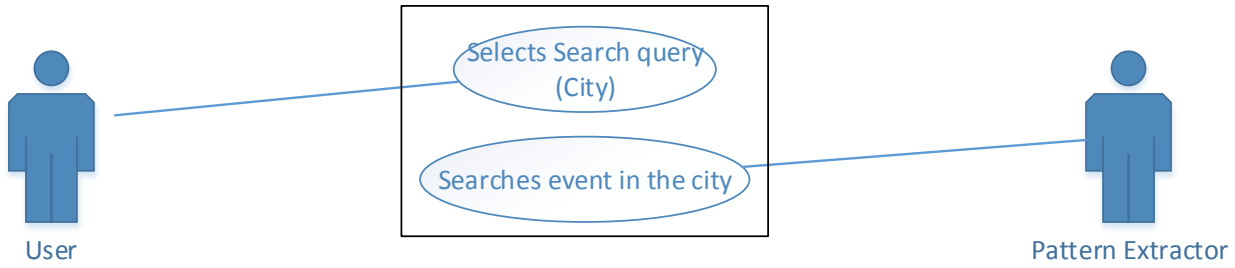


Figure 4-5 Search News for Location

4.1.6 Administrator Login:

4.1.6.1 Description and Priority:

This use case describes the process through which the Administrator logs in to the system.

Priority = High

4.1.6.2 Stimulus and Response Sequence:

Stimulus: The admin requests to login.
Response: The system queries admin for the username and password.

Stimulus: The admin enters username and password and presses the Submit button.
Response: The system matches the username and password with the username and password of the admin saved in the system. If the username and password is correct the system directs the admin to the admin homepage or else an error message is displayed: 'incorrect username or password'.

Stimulus: The admin requests to logout by clicking the logout button.
Response: The system directs the admin to the SMM's main page.

4.1.6.3 Functional Requirements (FR):

FR #1: The system shall get the username and password from the admin.
FR #2: System shall match the username and password with the admin's username and password.
FR #3: System shall take the admin to his homepage if the Username and password entered is correct.
FR #4: The system shall give an error if the username and password is incorrect.
FR#5: The system shall allow the admin to logout of the system.

4.1.6.4. Use Case Description

USE CASE NAME	Administrator Login
ACTOR	Administrator
NORMAL COURSE	1) The process is initiated when the clicks the Submit button. 2) System shall check the username and password is correct or not 3) If the username and password is correct the system shall direct admin to his homepage. 4) The system shall direct the admin to SMM's home page when the admin presses the logout button.
ALTERNATE COURSE	3) (a) If the username or password is incorrect the system shall display an error message: " Incorrect username or password ".
PRE CONDITION	✓ Administrator account had already been created.
POST CONDITION	✓ The admin is directed to his homepage when he logs in successfully.
ASSUMPTIONS	N/A

Table 7 Administrator Login

4.1.6.5. Use Case Diagram

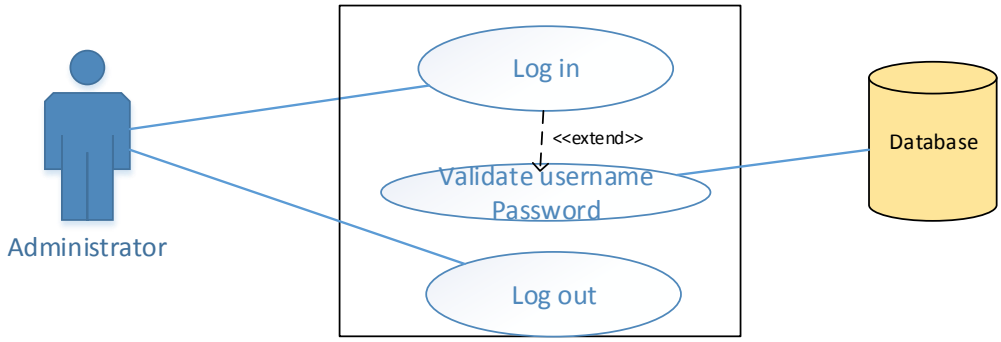


Figure 4-6 Administrator Login

4.1.7 Extract Tweets

4.1.7.1 Description and Priority:

This use case describes the process through which the Administrator extracts tweets from twitter website.

Priority = High

4.1.7.2 Stimulus and Response Sequence:

Stimulus: The admin requests to extract tweets.
Response: The system queries admin for the time period for which the tweets have to be extracted.

Stimulus: The admin enters the time period.
Response: The system extract all tweets in the given time period and stores them in database.

4.1.7.3 Functional Requirements (FR):

FR #1: The admin shall be able to specify the time period for the tweets.
FR #2: System shall extract all tweets within the given time period.

4.1.7.4. Use Case Description

USE CASE NAME	Extract Tweets
ACTOR	Administrator, Database(Secondary Actor)
NORMAL COURSE	1) The process is initiated when the admin specifies the time period for tweet extraction. 2) System shall establish a connection to Twitter. 3) System shall extract the geo-located tweets as per the time period specified. 4) Tweets shall be stored in the database.
ALTERNATE COURSE	3) (a) If no tweets exist in the given time period system shall give a warning: “No tweet found”
PRE CONDITION	✓ Administrator account had already been created. ✓ Admin has logged in. ✓ Geo located Tweets exist on twitter.
POST CONDITION	✓ The admin shall be notified that the tweets have been extracted.
ASSUMPTIONS	Geo located tweets exists on twitter.

Table 8 Extract Tweets

4.1.7.5. Use Case Diagram

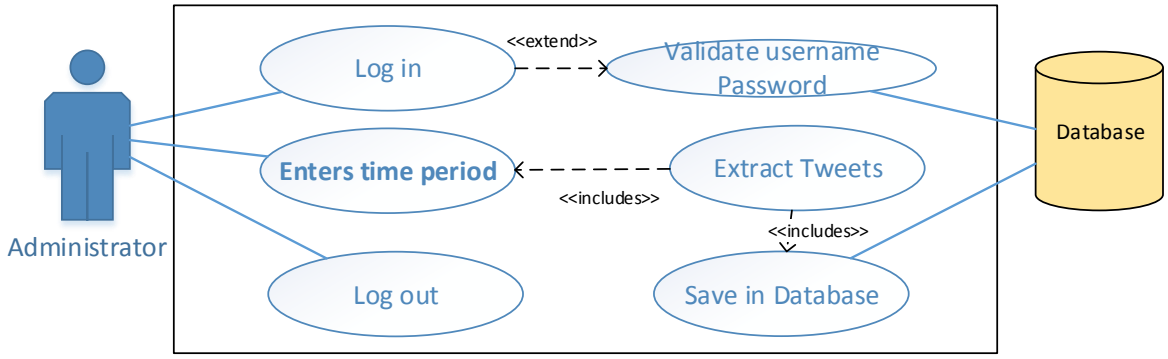


Figure 4-7 Extract Tweets

4.1.8 Extract News

4.1.8.1 Description and Priority:

This use case describes the process through which the Administrator extracts news from The Nation and Dawn News.

Priority = High

4.1.8.2 Stimulus and Response Sequence:

Stimulus: The admin requests to extract news.
Response: The system queries admin for the time period for which the news have to be extracted.

Stimulus: The admin enters the time period.
Response: The system extract all news (from the Dawn and the Nation website) for the given time period and stores them in database.

4.1.8.3 Functional Requirements (FR):

FR #1: The admin shall be able to specify the time period for the news.
FR #2: System shall extract all news within the given time period.

4.1.8.4. Use Case Description

USE CASE NAME	Extract News
ACTOR	Administrator, Database(Secondary Actor)
NORMAL COURSE	1) The process is initiated when the admin specifies the time period for news extraction. 2) System shall establish a connection to Dawn and The Nation website. 3) System shall extract the news as per the time period specified. 4) News shall be stored in the database.
ALTERNATE COURSE	3) (a) If no News exist in the given time period system shall give a warning: “No news found”

PRE CONDITION	✓ Administrator account had already been created. ✓ Admin has logged in. ✓ News exist on The Nation and Dawn news website.
POST CONDITION	✓ The admin shall be notified that the news have been extracted.
ASSUMPTIONS	Events are taking place in the country and are being reported on The Nation and Dawn news website.

Table 9 Extract News

4.1.8.5. Use Case Diagram

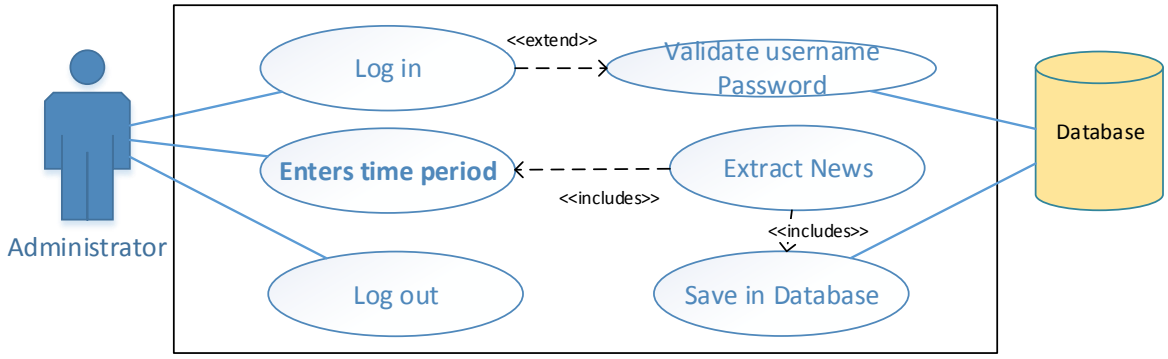


Figure 4-8 Extract News

4.1.9 Delete past Data

4.1.9.1 Description and Priority:

This use case describes the process with which the administrator deletes the past data (news and tweets) stored in the Database.

Priority = High

4.1.9.2 Stimulus and Response Sequence:

Stimulus: The admin requests to delete data.
Response: The system queries admin to select the data to be deleted.

Stimulus: The admin selects the past data to be deleted.
Response: The system deletes the selected data and notifies the admin.

4.1.9.3 Functional Requirements (FR):

FR #1: The admin shall be able to select data to be deleted from the database.
FR #2: System shall delete the past data(selected by admin).

4.1.9.4. Use Case Description

USE CASE NAME	Delete Past data
ACTOR	Administrator, Database(Secondary Actor)
NORMAL COURSE	1) The process is initiated when the admin specifies the data that he wants to delete. 2) System shall delete the data. 3) System shall notify the user that the data was deleted.
ALTERNATE COURSE	3) (a) If data is not deleted, system shall give a warning: "Error in deleting data"
PRE CONDITION	✓ Administrator account had already been created. ✓ Admin has logged in. ✓ Data (News and tweets) exists in the database.
POST CONDITION	✓ The admin shall be notified that the data had been deleted successfully.
ASSUMPTIONS	Extracted data from twitter and Nation and Dawn news exists in database.

Table 10 Delete Past data

4.1.9.5. Use Case Diagram

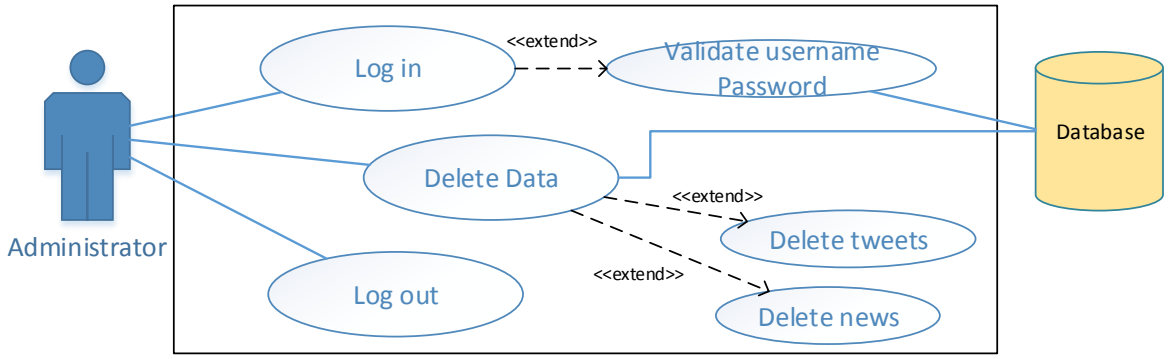


Figure 4-9 Delete Past data

4.2 Non Functional Requirements

4.2.1. Performance Requirements

4.2.1.1. Response Time

Since the system is web based plus mobile application, there shouldn't be much delay in processing and responding to the user's requests. The requests should be responded at most 2 seconds.

4.2.1.2. Request Handling

The system is designed to integrate information from multiple resources and lot of people is using news and social media now therefore there many users who will use the system. Therefore the web based system should be hosted on a server that is capable of handling a great amount of requests.

4.2.2 Safety Requirements

N/A

4.2.3. Security Requirements

- The system should not be accessed by any unauthorized user.
- The database should also be only be allowed to be manipulated by valid users.
- The application on the mobile device should also be secured.

4.2.4 Software Quality Attributes

4.2.4.1. Reliability

This system should be reliable and provide accurate data without errors. Since it is a web based system, the system is highly portable and can be accessed from anywhere with an internet connection. The system is easily extendible since we can add functionalities through web services. The benefit of web services is to make the functionalities reusable by other systems.

4.2.4.2. Usage Easiness

- 90% of a test panel of non-experienced users should be able to use the system within 5 minutes.
- After having used the application once, 95% of users are able to locate the experienced functionality within 1 minute.

4.2.4.3. Compatibility

Android client applications must be able to communicate with the rest of the system, and to handle all of the functionalities.

4.2.4.4. Overall Satisfaction

After conducting a survey, 80% of the users should keep using the application after a two week exploitation period.

4.2.4.5. Trust

After having used all the features three times, 90% of the users should feel confident about the reliability, robustness and be convinced that the product does what it is expected to do.

4.2.4.6. Understandability and Politeness

Our system uses only those symbols and words that are immediately understandable by users. All the technical details should be hidden from the user.

4.2.4.6. Learning

Any user without computer skills should be able to view list of desired events, plot them on map to see location and time of events within the first 5 minutes of usage without referring to the user manual.

4.2.4.7. Maintenance

The first version of the system is open source and is its maintenance is not provided. Future commercial versions of the software may offer maintenance.

4.2.4.8. Supportability

The required support level needed by the system should be low, even up to the point it could be handled by the service provider’s helpdesk.

4.2.4.9. Adaptability Requirements

The client application should be portable on Android version higher than 2.1 and all commonly used web browser.

4.2.5. Design Quality

4.2.5.1 Usability

4.2.5.1.1. Learnability

For users using system for the first time shall take no more than 40 minutes to get used to the interface. For users like Admin system shall be learnable after 20 minutes training.

4.2.5.1.2. Efficiency

The webpage shall be displayed in no more than 10 seconds using a 500-800 Kbps Bandwidth connection per Machine/Computer.

4.2.5.1.3. User satisfaction

At least 70% of candidates shall rate their satisfaction with the system After using it at 7 or more on a scale of 1 to 10.

4.2.6. Other Requirements

4.2.6.1. Database Requirements

A database is required to store all the information of the system. The database contains many tables and the information can be added/modified/deleted through the system as required.

4.2.6.2. Legal

This system is expected to be developed in accordance with the laws regulating communication, privacy and confidentiality.

4.2.6.3. Integration Requirements

In order to integrate news data and tweets from twitter, we have to offer the functionalities of as web services. So therefore it is not enough to only create functionalities in the programming language we choose, we also have to develop web services according to each functionality. We will use these web services later on android phone.

To create web services we can use automated tools that create services from the code.

The web services created can be used by systems that require their functionality without having to change their own systems architecture.

The only thing they need to create is a client that can access the web services and provide the appropriate output.

CHAPTER 5:

SYSTEM DESIGN &SPECIFICATIONS

5. System Design & Specifications

5.1. System Architecture

This section provides a detailed and comprehensive architectural overview of the system.

5.1.1 High level design

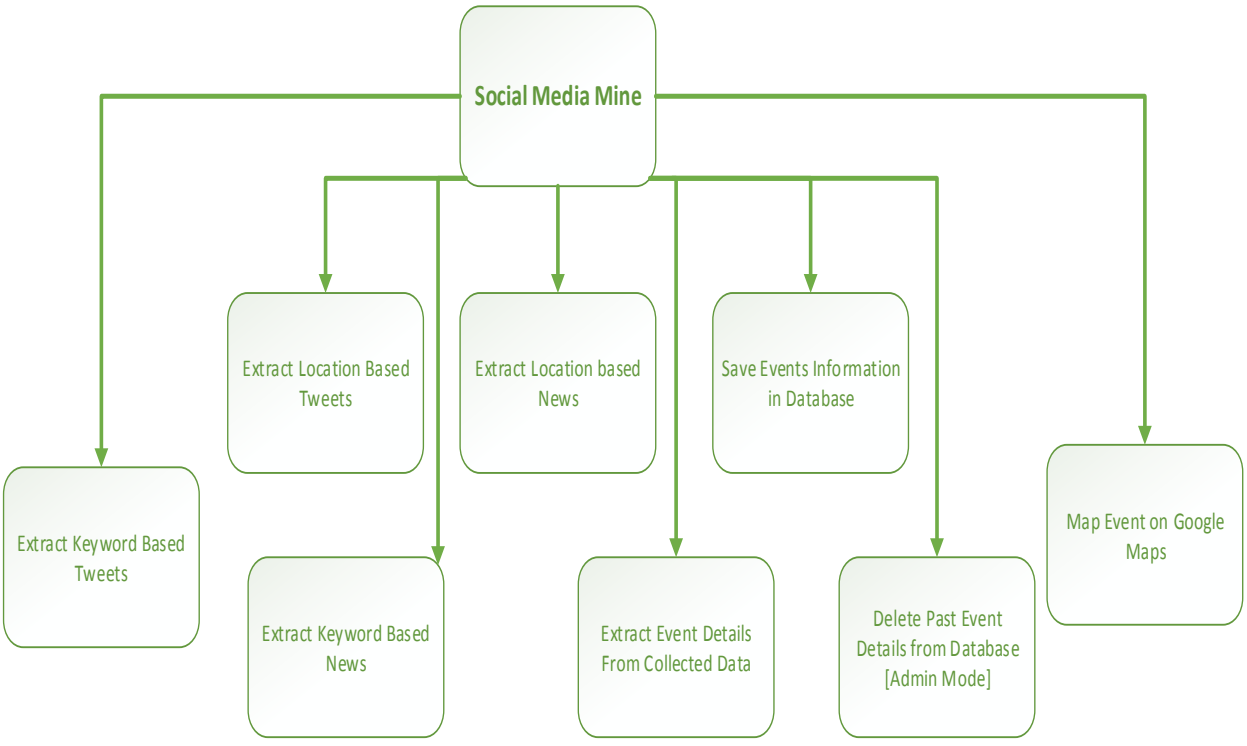


Figure 5-1 High level Design of SMM

5.1.2. Architecture of Web based Application

MVC has been used because of the project's nature which demands that view (i.e. the web app interface) needs to be separated from the back end app logic so that the back end complex logic is transparent from the user and he/she finds it easy to use the system. The user communicates with the **controller** when he makes a request through the web browser. The controller is responsible to instantiate the model. The controller selects the appropriate view for the user. The view then interacts with the model if it needs some data that is located in the external database (MySQL). When the **view** is ready, it is presented to the user.

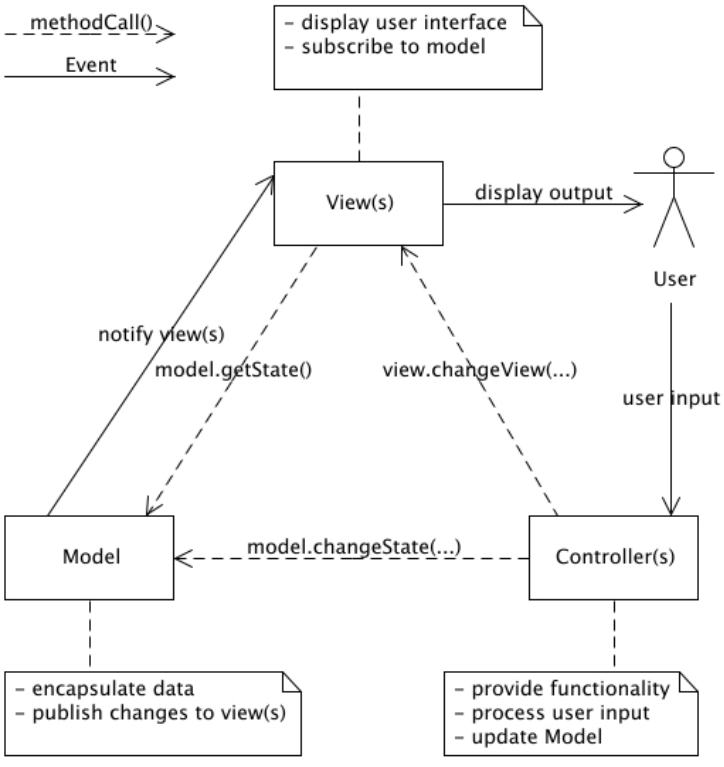


Figure 5-2 General MVC

5.1.2.1. Reasons for choosing MVC

5.1.2.1.1. Independence

We have divided our system into 3 parts. Now when we want to make changes to one part, the others remain unaffected and the change is isolated only in one part. This also reduces the complexity as compared to if there were no division. Now the complexity only lies within the 3 individual parts and the interdependence within the 3 divisions is kept to a minimum.

5.1.2.1.2. Manageability

This architectural style also makes our system more manageable. Because complexity is reduced, now we can identify any potential errors easily and can also fix and enhance any functionality without worrying about the effects it will have on the other parts.

5.1.2.1.3. Scalability

This also improves scalability. If we want to increase the database to cater for more users we can do that without having to rework the whole system. We can also add more views and interfaces according to our requirement without changing the model.

5.1.2.1.4. Reusability

Reusability is also a factor that comes along with this architecture. Since each of our three parts can exist independently, each one of them can be used later on in other projects without a lot of change. The internal logic can remain the same and the external connections may only need to be changed.

5.2.1.5. Testing

Testing can be much easier and efficient. Now that we can test each part separately from the others, we can reduce testing time, since testing one part won't be dependent on the other parts and if any fault is found, it will be easier to locate and fix it.

5.1.3. Architecture of Android Application

The android system also follows an **MVC** style. Whenever the user makes a request, the controller recognizes it and updates the model if required. The model then triggers the view that it needs to be changed and that the old view has expired. The view is then redrawn and displayed to the user. The advantages are similar to the ones mentioned with the architecture for the web system.

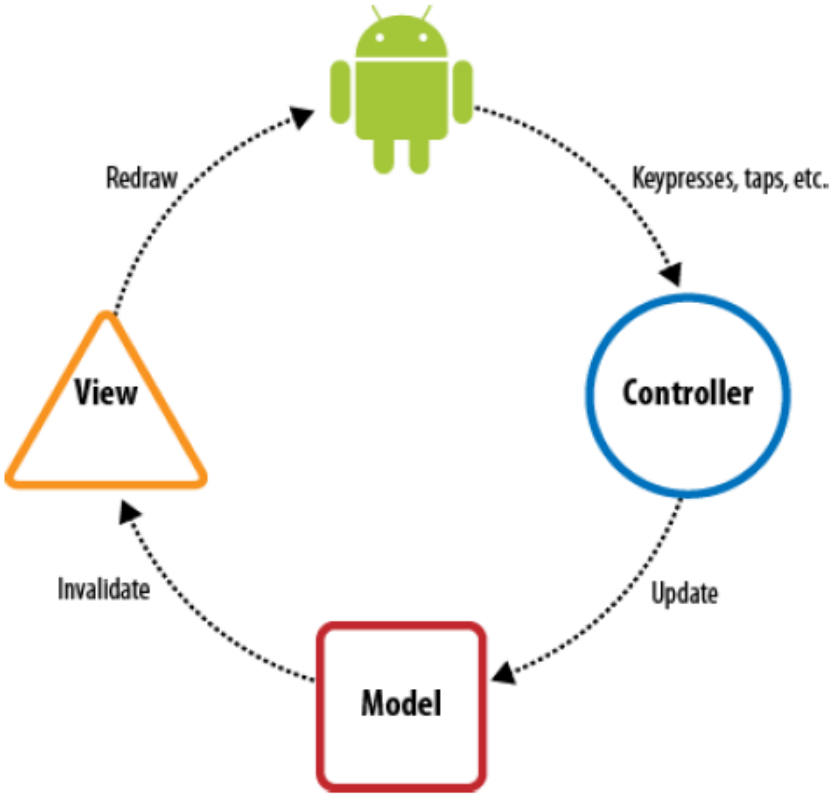


Figure 5-3 General MVC for Android

5.2. Design Details

5.2.1. Class Diagram

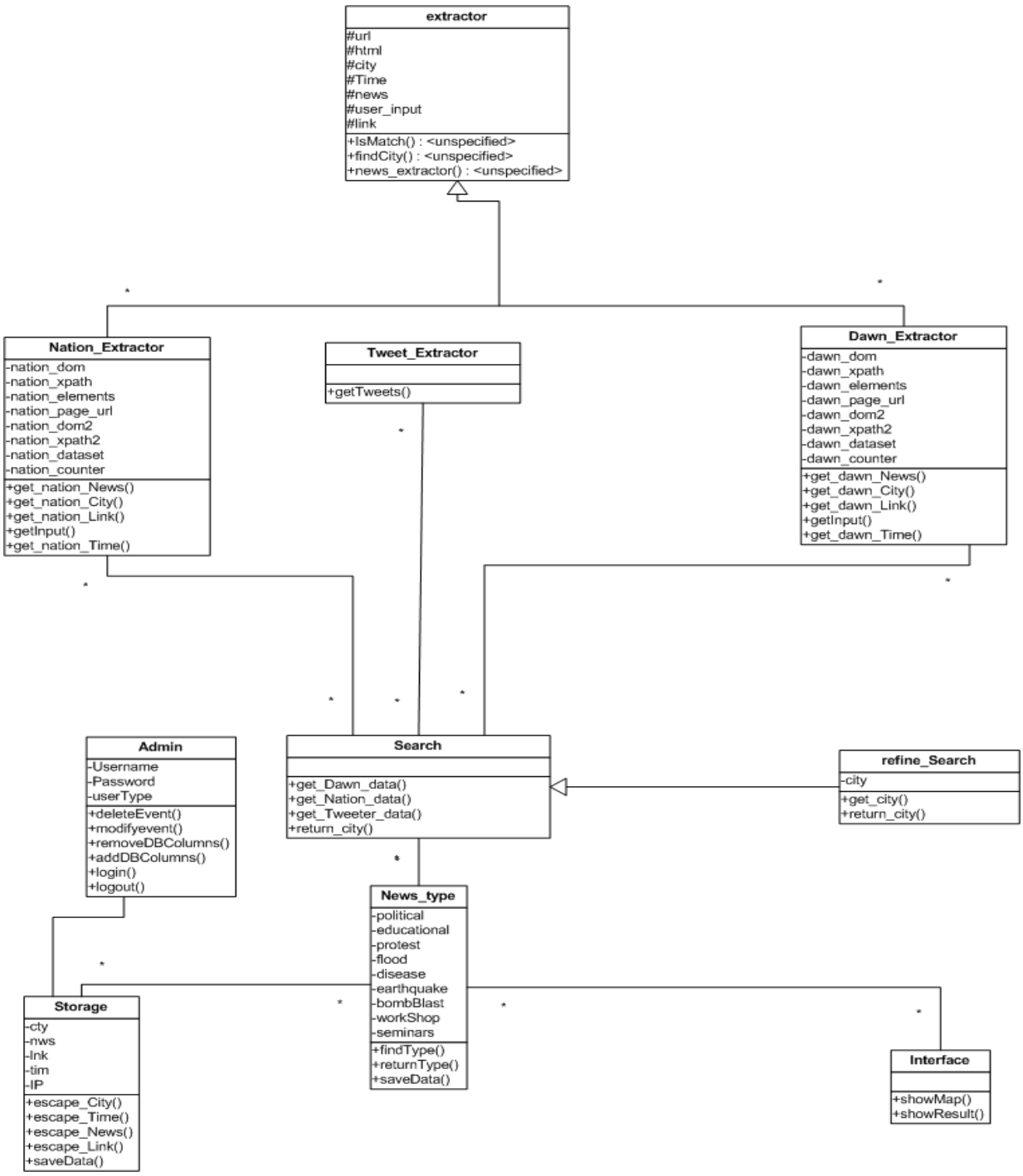


Figure 5-4 Class Diagram

The class diagram consists of classes such as:

- **extractor** which is abstract class.
- Two classes **Nation_Extractor** and **Dawn_extractor** are derived from extractor class. These two classes extract news, their time of occurrence of an event and location of a news/event based on keyword from *dawn.com.pk* and *nation.com.pk*.
- **Tweet_Extractor** extracts tweets of users of a specific location/city from their twitter profile based on keyword.
- Nation_Extractor, Dawn_extractor and Tweet_Extractor provide results/data to **Search** class.
- If a user wants to refine his search of events he/she can do this by providing the name of a specific city. So that system only searches news/events from that specific location. The class **refine_Search** will enable user to do this. This class takes a city name from a user and passes it to **Search** class. Then search class only gets data from that specific location/city.
- After getting data **Search** class passes it to **News_type** class. And this class will then perform analysis on this data for finding the type of news. Whether it is a constructive, destructive or political event.
- After performing analysis (finding the type of news) on this data, this data will send to database and interface from **Storage** and **Interface** class respectively. On interface the system will show the complete result of user search and mark the location of an event on Google map.
- **Admin** class enable an administrator to login and logout using **login()** and **logout()** method. An administrator can delete and modify an event and its details. Also he can add, modify and delete database columns.

5.2.2. Database Model

The database model describes the different tables that are included in the database. These tables are according to the classes mentioned in the class diagram. The database also identifies the primary keys of each table that uniquely identifies each record in that table.

- **Search** table shares primary key *ID* with each table in database as foreign key. It stores user keywords to be searched and user preference for search .i.e. whether a user want to search all resources or some of them.
- **News, Time and City** tables stores news, time and city from extracted data. Each of these tables has a foreign key *ID* and primary key *IDN, IDT* and *IDC* respectively.
- **Users** table stores username and passwords of administrators to the system.

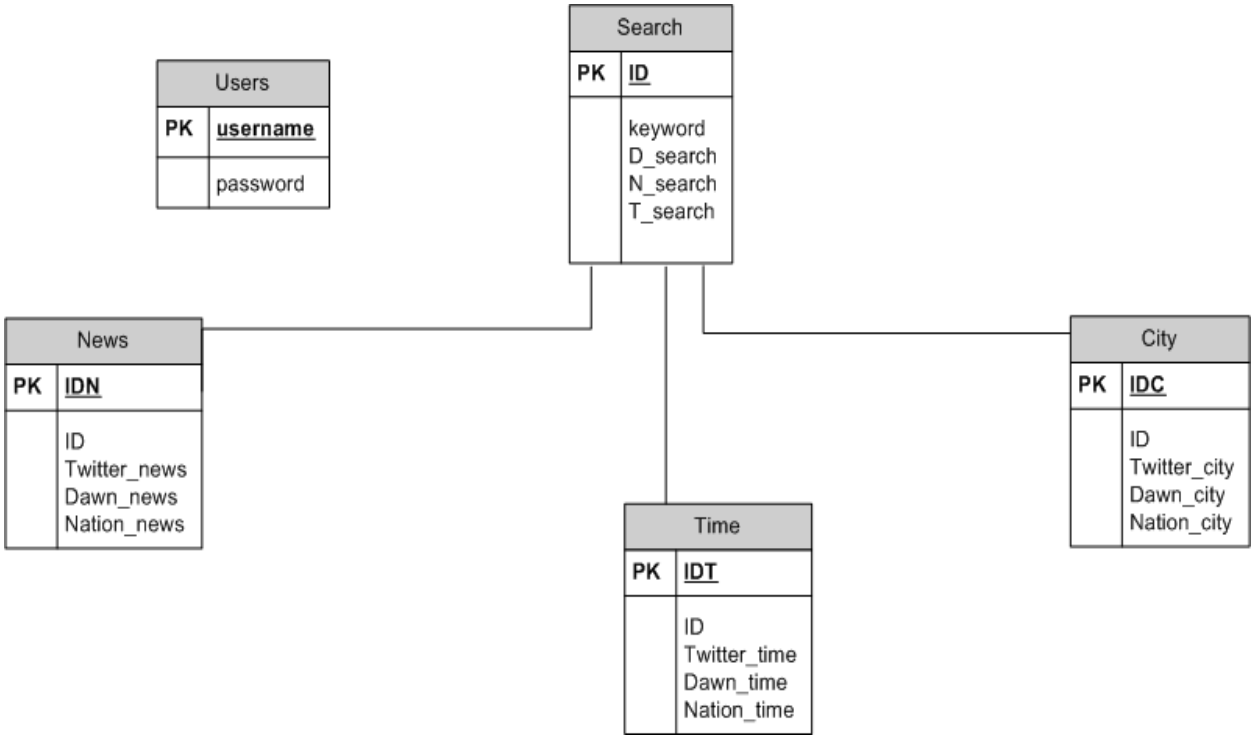


Figure 5-5 Database Diagram

5.2.3. Navigational Model

5.2.3.1. Access Model for user

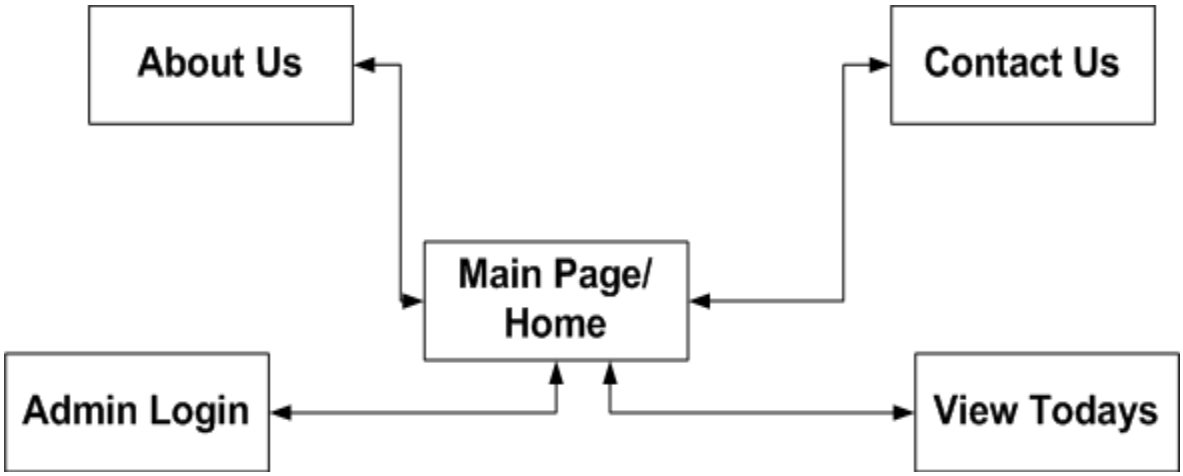


Figure 5-6 Access Model for user

The above hyperlink model shows the pages that can be visited from **the main page** or home that include the **Admin login** page, the **view today** page, **About Us** page and **Contact Us** page. All the pages also lead back to the main page.

5.2.3.1. Access Model for Admin

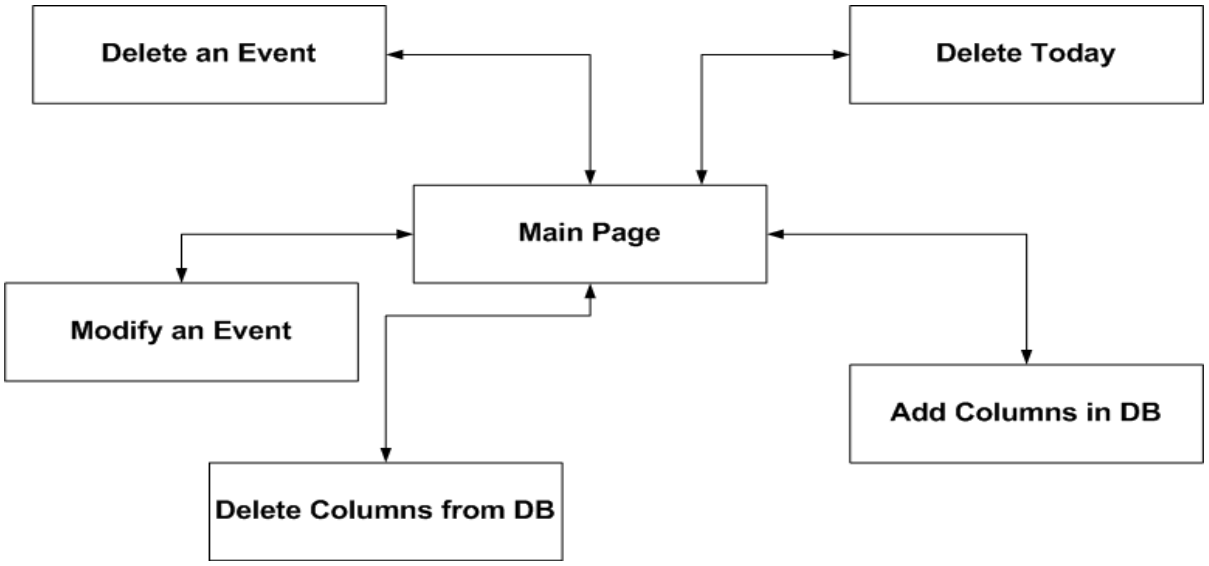


Figure 5-7 Access Model for Admin

The above access model is also for the admin and shows the number of other pages that the admin can visit from the main page of for admin and it also shows which navigational aid he can use.

5.3. UML Diagrams

5.3.1. Use Case Diagram

The complete use case diagram that covers all the use cases is given below.

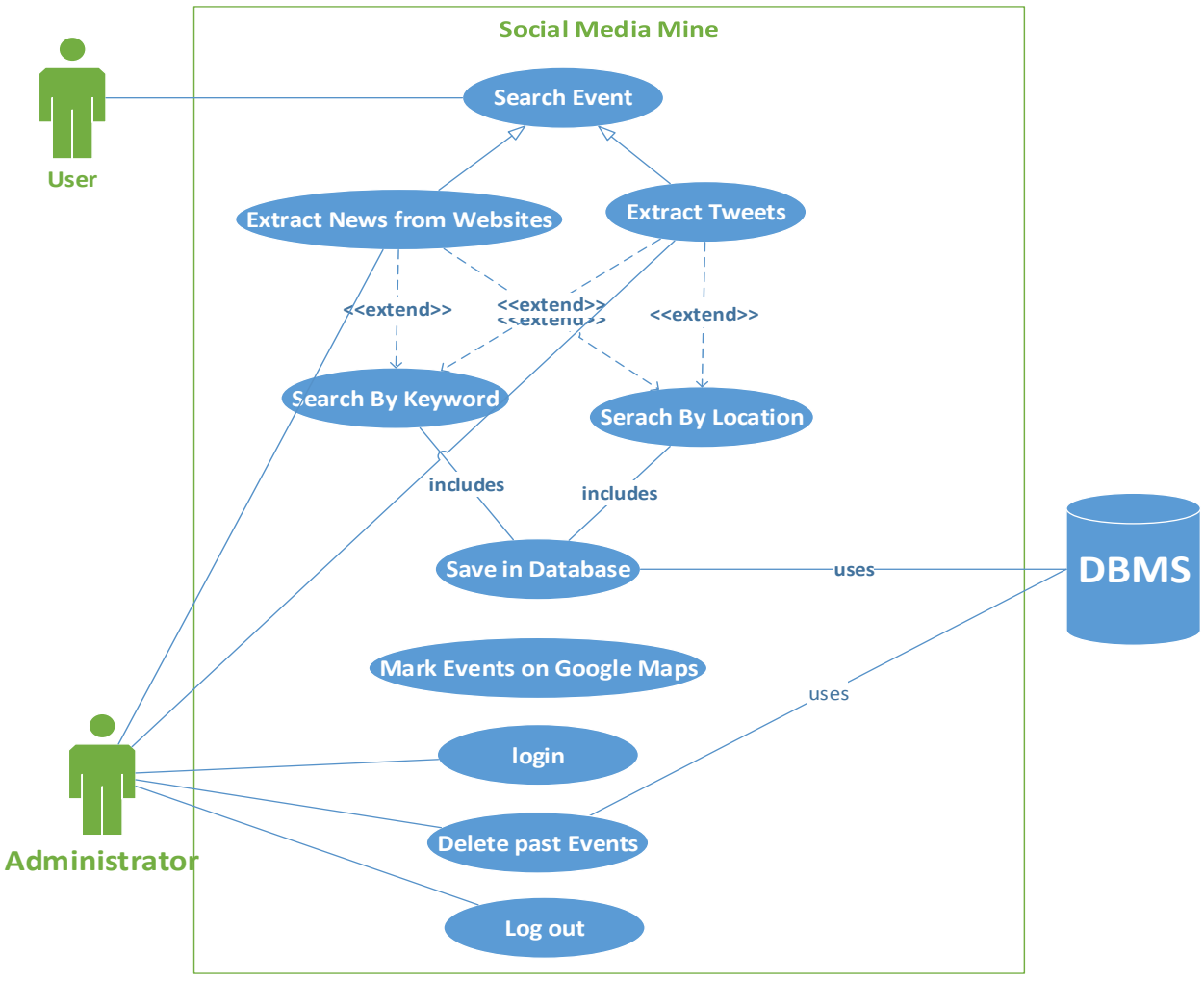


Figure 5-8 Use case Diagram

5.3.2. Sequence Diagrams

5.3.2.1. Refine News Search

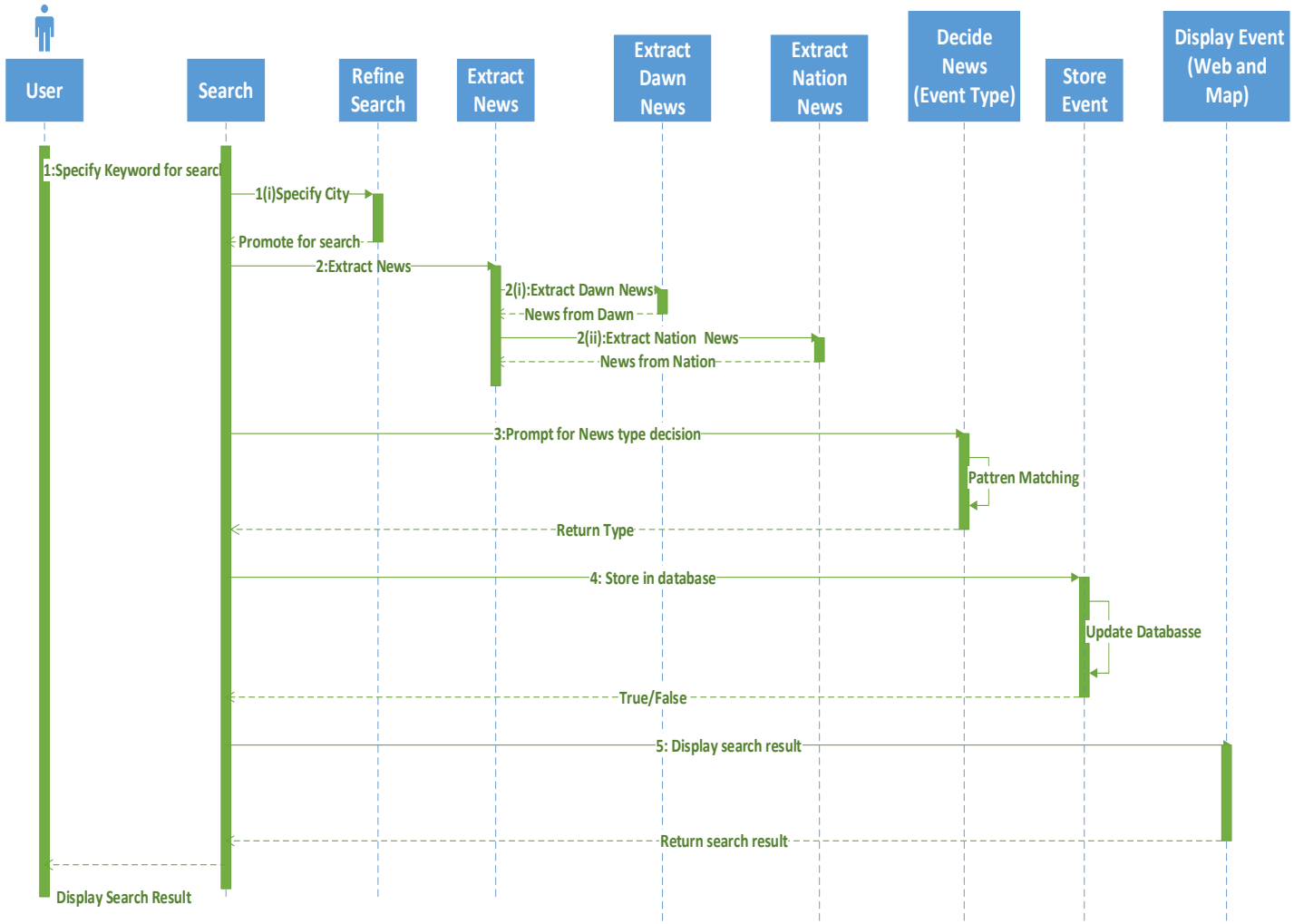


Figure 5-9 Sequence Diagram-Refine News Search

5.3.2.2. Refine Tweet Search

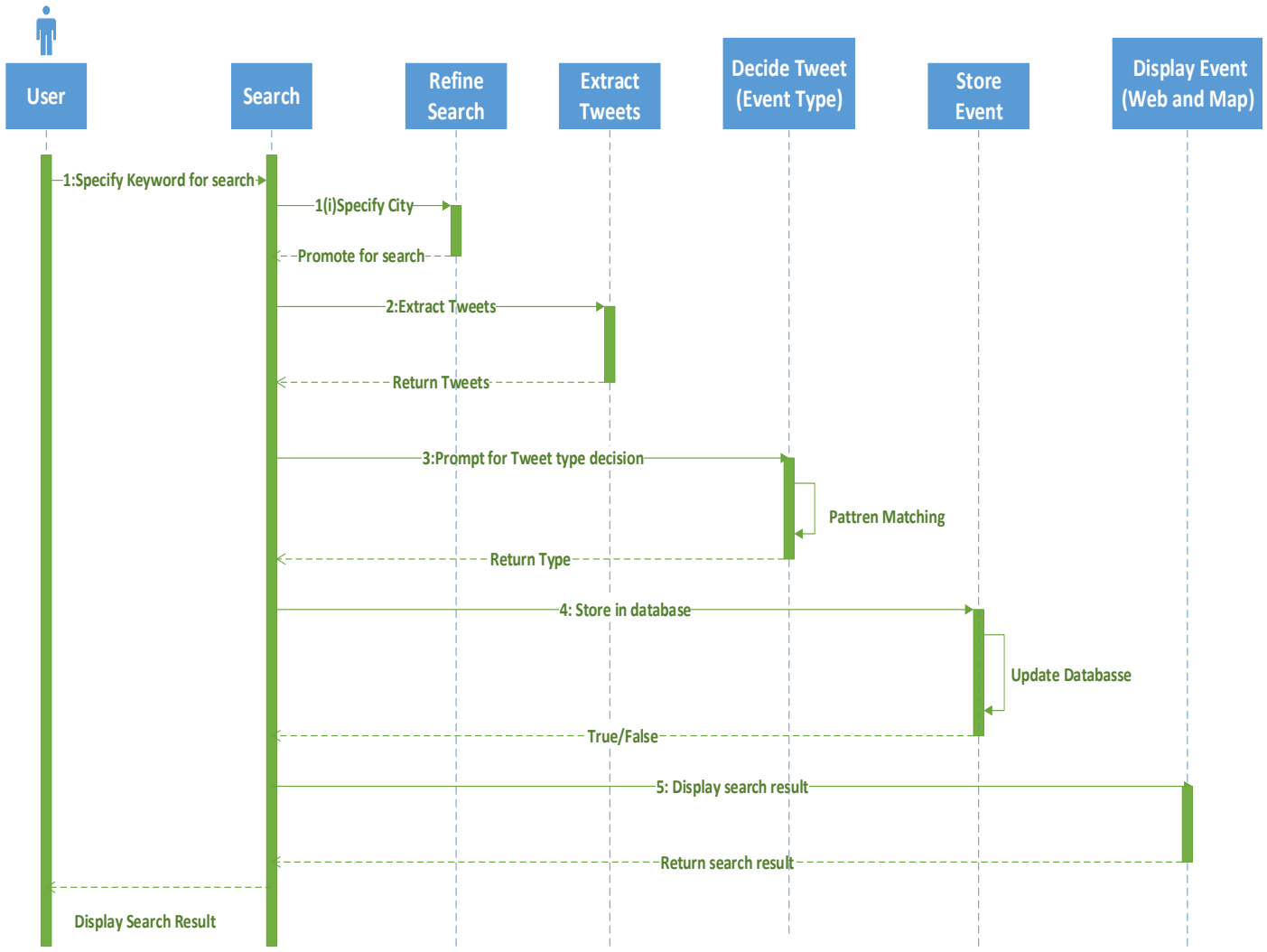


Figure 5-10 Sequence Diagram - Refine Tweet Search

5.3.2.3. Search News:

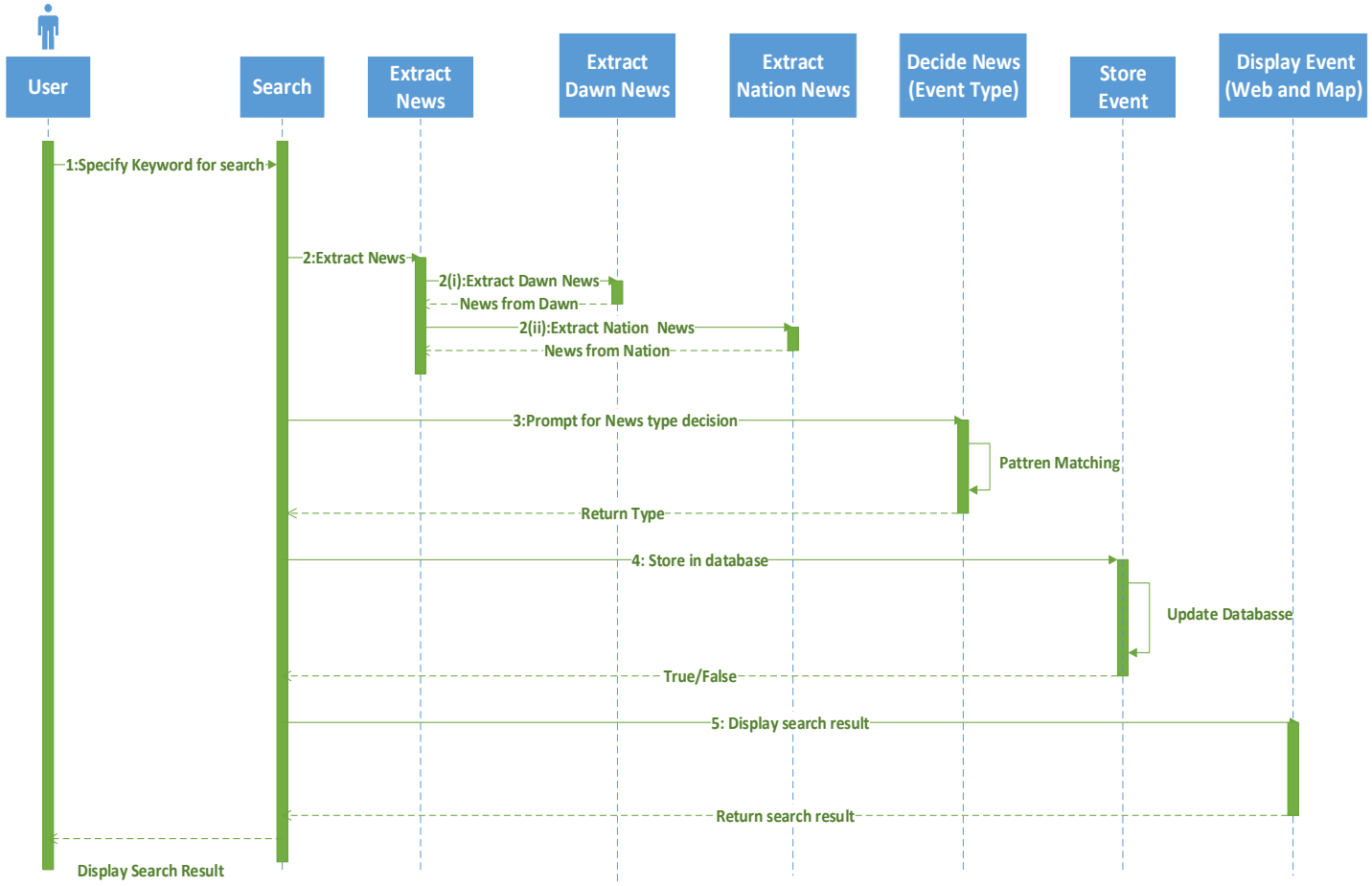


Figure 5-11 Sequence Diagram - Search News

5.3.2.4. Search Tweet

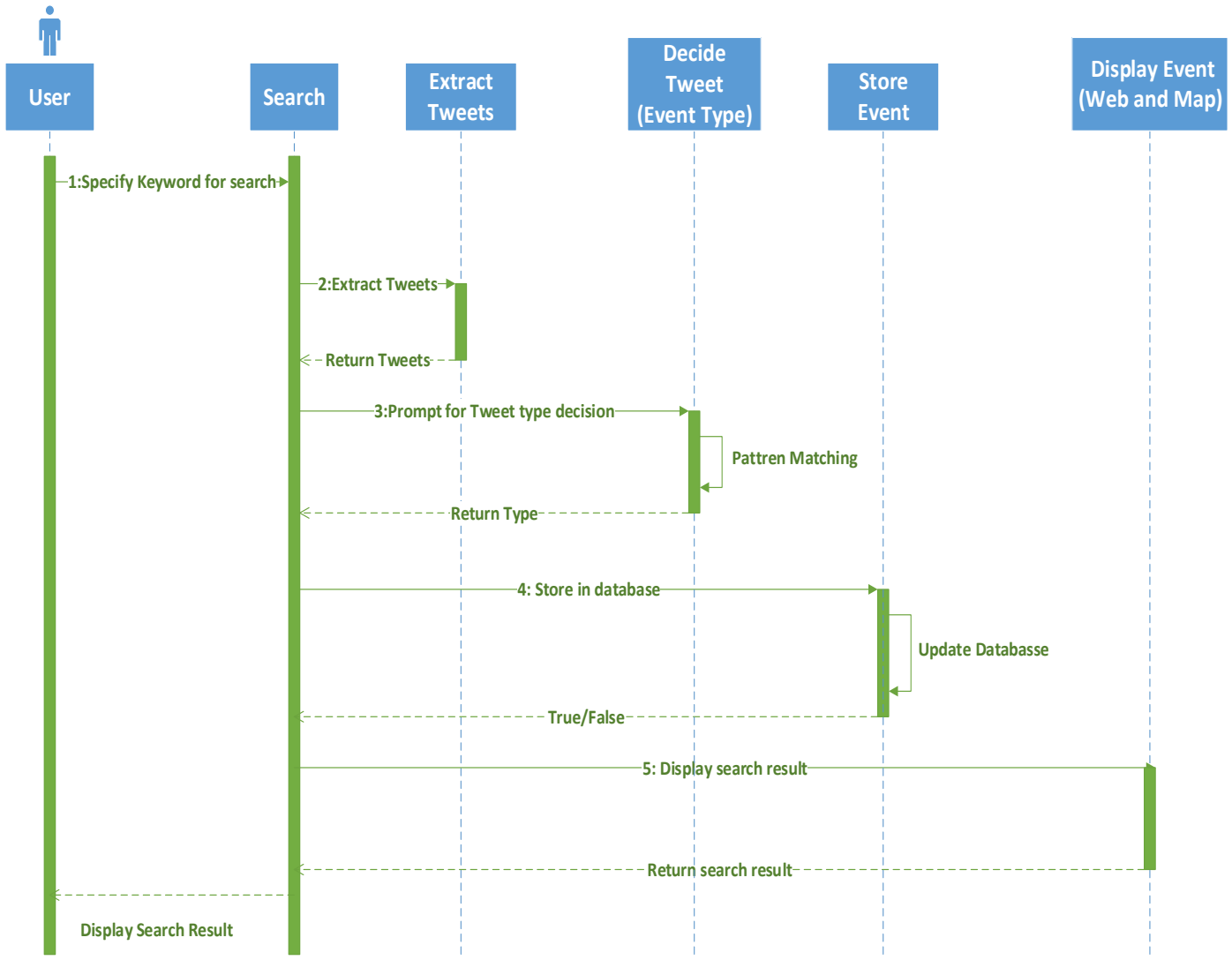


Figure 5-12 Sequence Diagram - Search Tweet

5.3.3. Activity Diagram

This section demonstrates the different activities that can be performed. Only the representative diagrams are shown to allow space for other diagrams.

5.3.3.1. Searching news by refining search query

This diagram describes how a user can search news by refine the search query (by providing the name of a city). By refining search query the system will find news of that particular location and mark them on Google map.

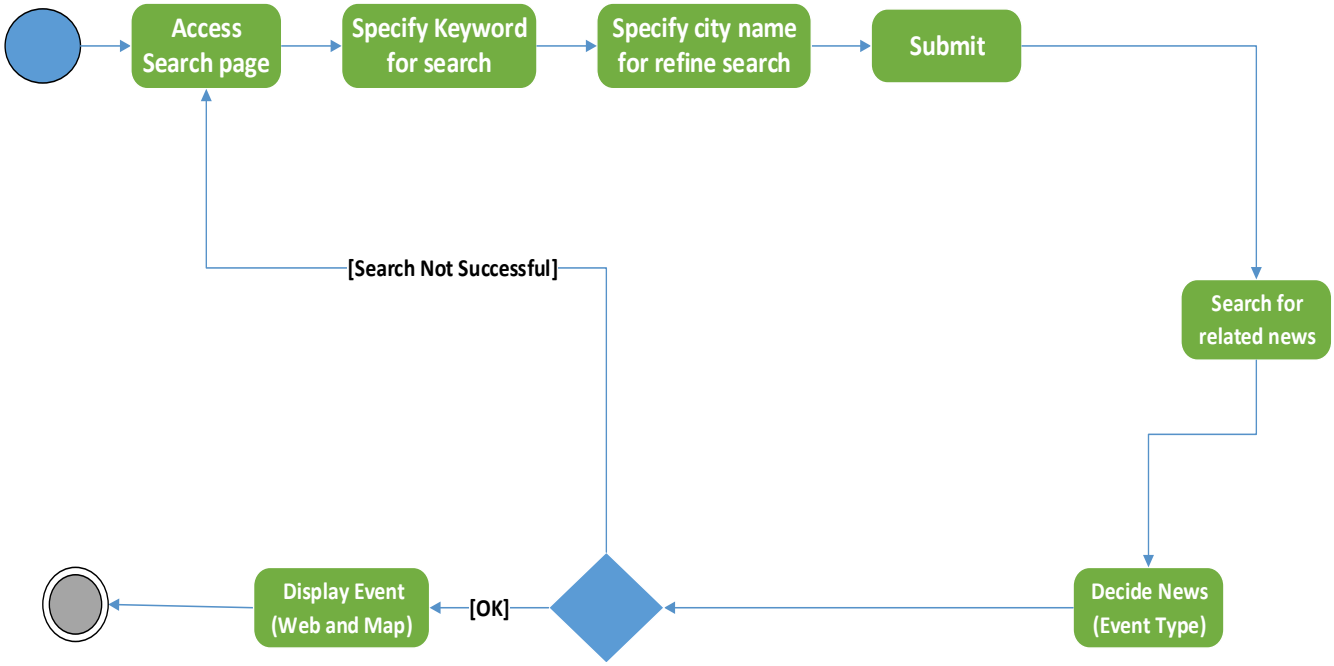


Figure 5-13 Activity Diagram - Searching news by refining search query

5.3.3.2Searching tweets by refining search query:

This diagram describes how a user can search tweets by refine the search query (by providing the name of a city). By refining search query the system will find tweets of that particular location and mark them on Google map.

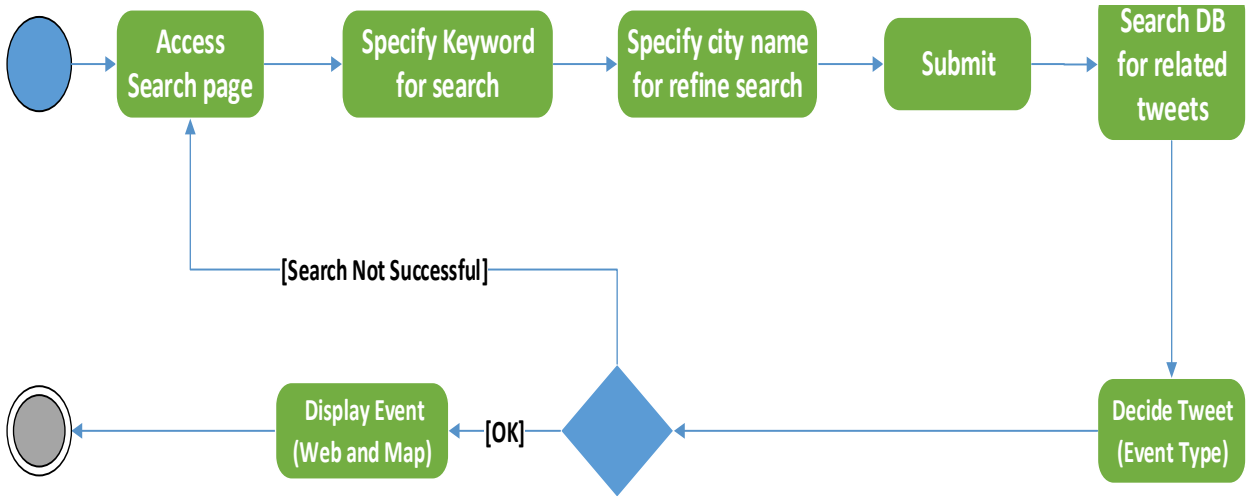


Figure 5-14 Activity Diagram - Searching tweets by refining search query

5.3.3.3. Searching news without refining search query:

This diagram describes how a user can search (news websites) for an event from all over Pakistan. System will search for provided event from all cities. After finding news related to provided search query system will mark it on Google map.

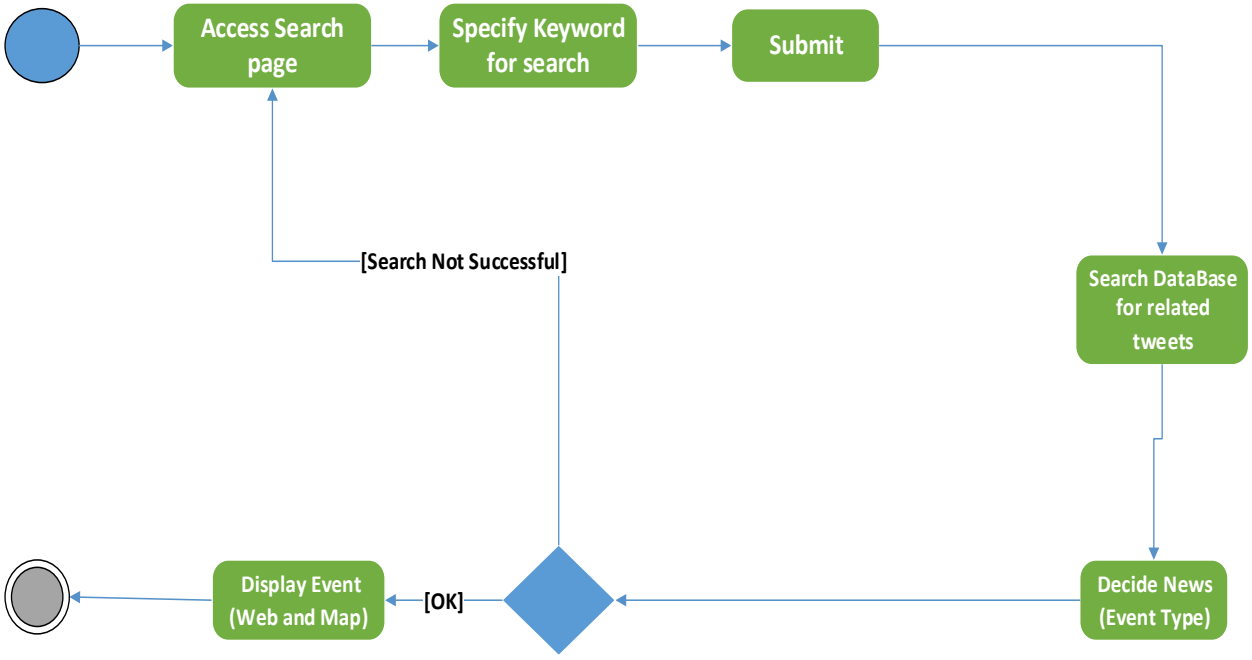


Figure 5-15 Activity Diagram - Searching news without refining search query

5.3.3.4. Searching tweets without refining search query:

This diagram describes how a user can search (twitter) for an event from all over Pakistan. System will search for provided event from all cities. After finding news related to provided search query system will mark it on Google map.

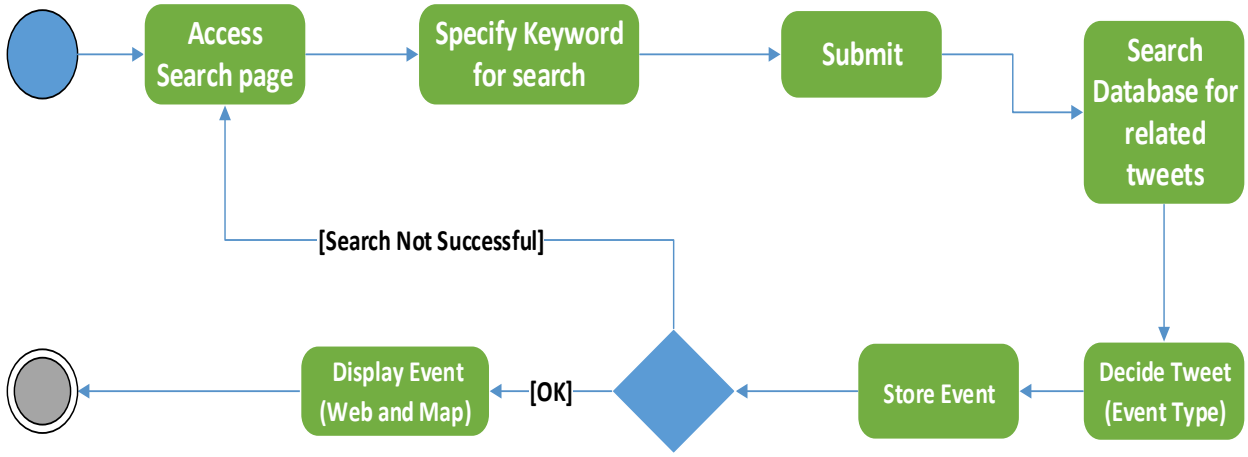


Figure 5-16 : Activity Diagram - Searching tweets without refining search query

5.3.3.5. Extract and Delete Data:

This diagram describes how admin will extract news from the Nation news and Dawn News website and tweets from twitter for events from all over Pakistan. It also describes process of deleting the past data. System will search these sources and extract all news, tweets related.

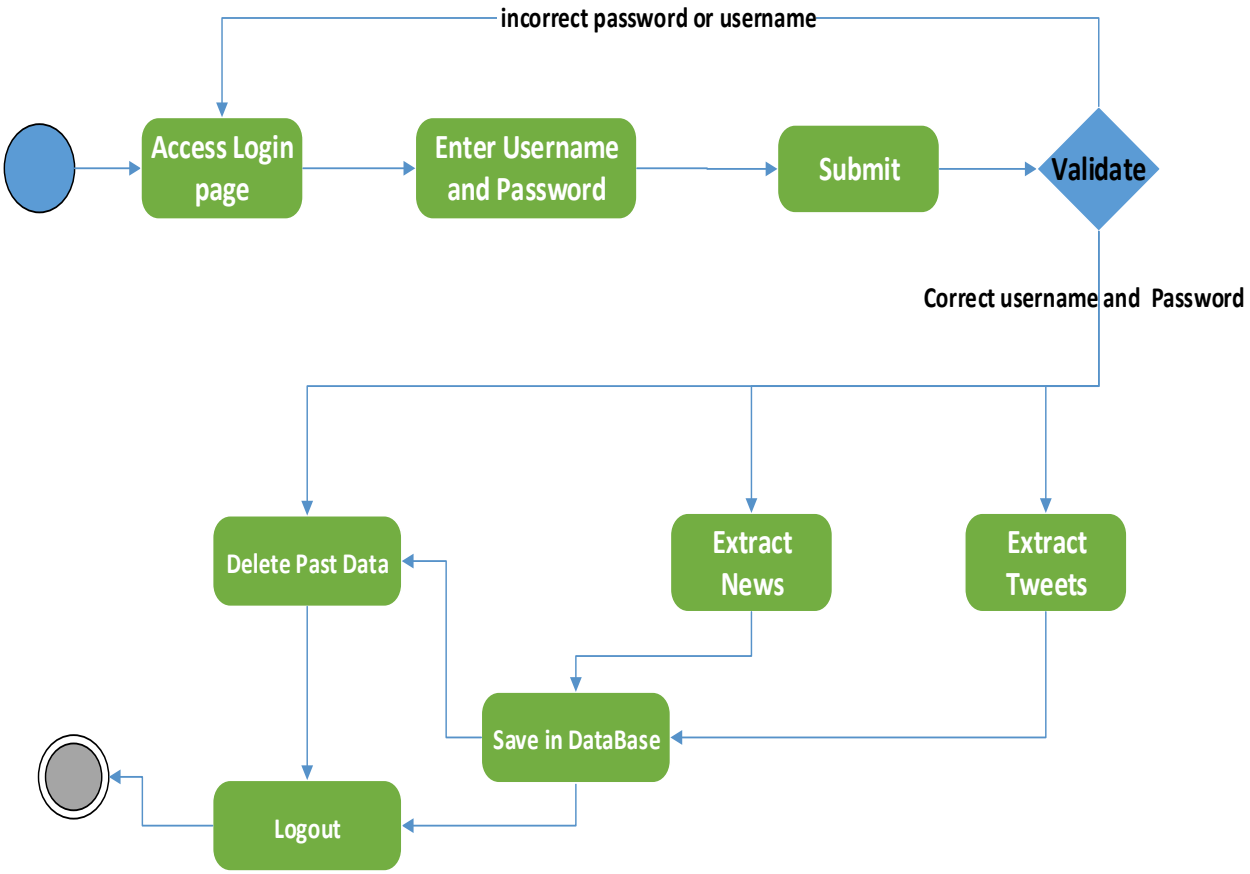


Figure 5-17 Activity Diagram – Extract and Delete data

5.3.4. Collaboration Diagram

This section demonstrates the collaboration diagrams. Only the representative diagrams are shown to allow space for other diagrams.

5.3.4.1. Refine Search for News

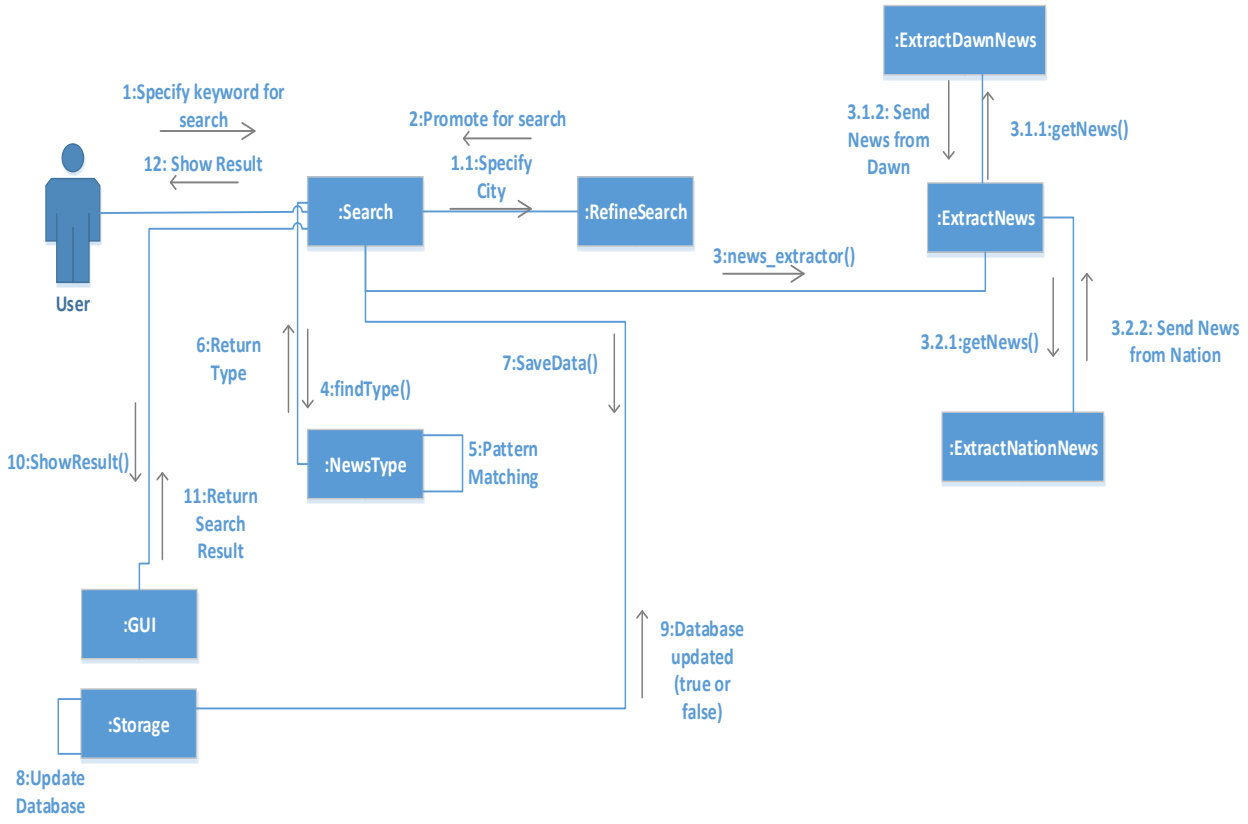


Figure 5-18 Collaboration Diagram - Refine Search for News

5.3.4.2. Refine Search for Twitter

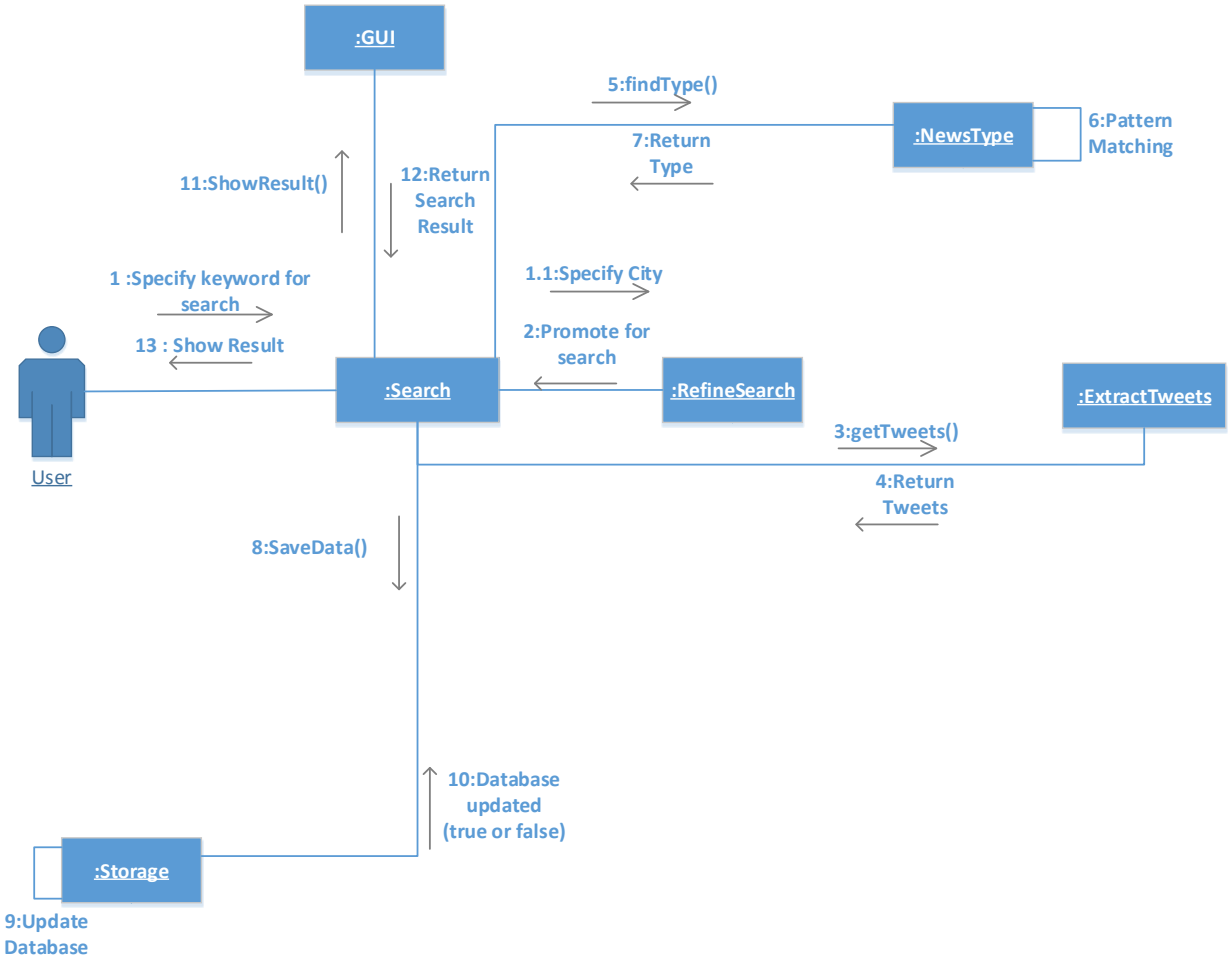


Figure 5-19 Collaboration Diagram - Refine Search for Twitter

5.3.4.3. Searching News without Refine Search

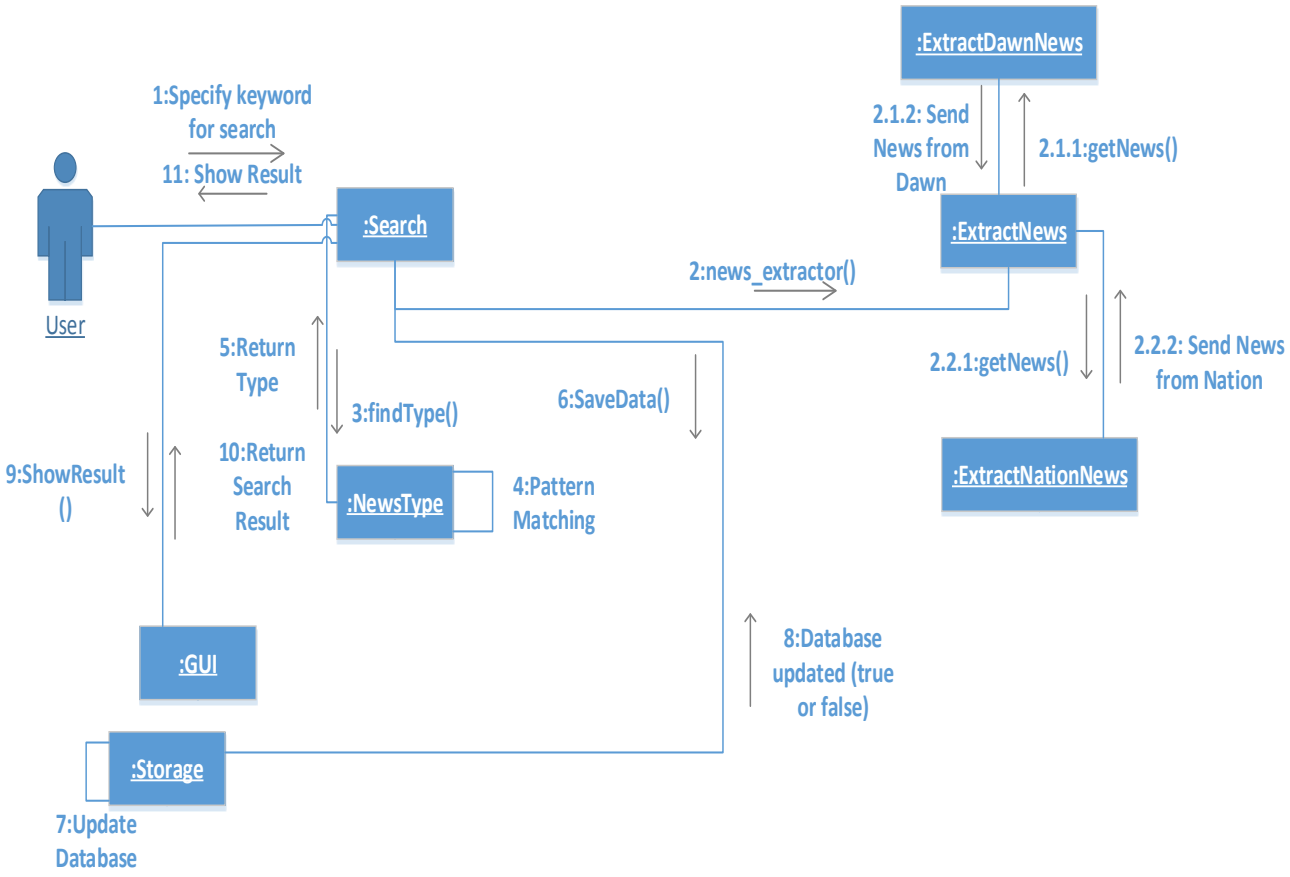


Figure 5-20 Collaboration Diagram - Searching News without Refine Search

5.3.4.4. Searching Tweets without Refine Search

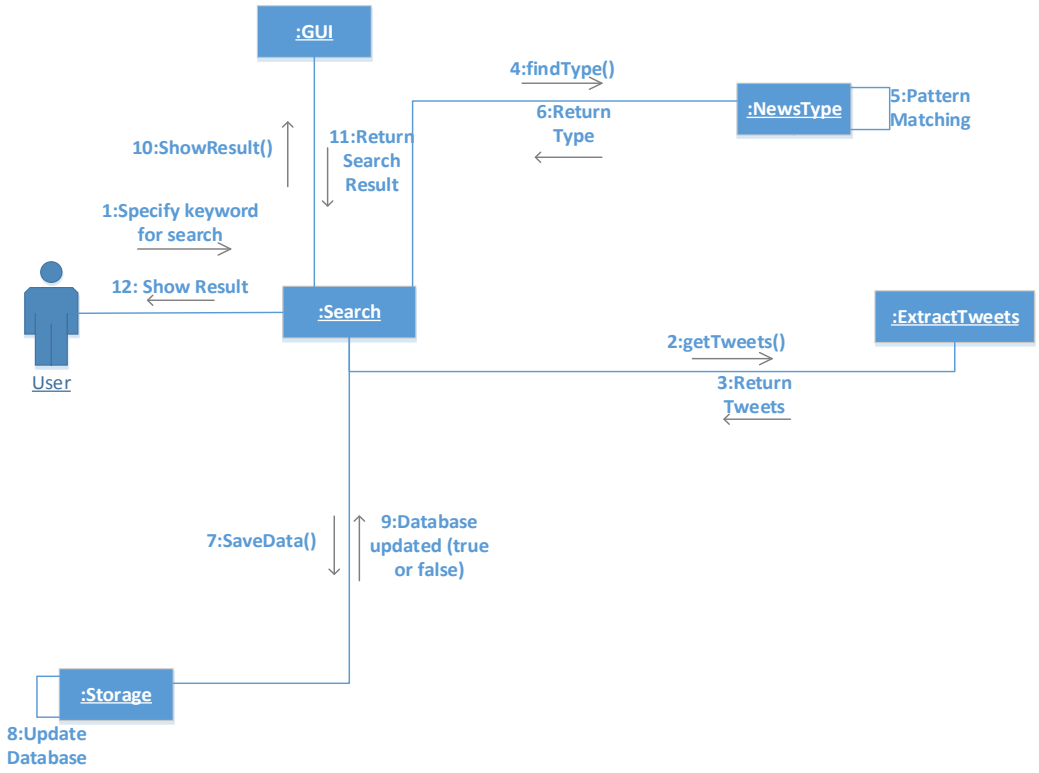


Figure 5-21 Collaboration Diagram - Searching Tweets without Refine Search

CHAPTER 6:

SYSTEM IMPLEMENTATION

6. System Implementation

6.1. Technologies Used

6.1.1. Programming Language

PHP is used as programming language to develop Web pages. Java Script is used as

Client side scripting language, mySQL language is used for managing data held in

Database. While the app has been developed in Android for phones.

6.1.2. Development Tools

Website has been designed and implemented in Zend Studio 10.10 and Notepad++.

Android app has been developed in the Eclipse studio.

6.1.3. Database

Database was developed and managed in mySql .

6.1.4. Operating System

Website application is tested for IE6 and above all browsers on Microsoft

Windows. Android app is tested for android 2.3.5 and above.

CHAPTER 7:

TESTING AND RESULTS ANALYSIS

7. Testing and Results Analysis

7.1 Introduction

Testing involves checking whether the system meets the requirements that were established in the beginning of this software's lifecycle.

In this section we will use functional testing[21] to check whether our system performs as expected.

7.2 Testing Technique

Functional testing typically involves five steps:

1. The identification of functions that the software is expected to perform.

2. The creation of input data based on the function's specifications.

3. The determination of output based on the function's specifications.

4. The execution of the test case.

5. The comparison of actual and expected outputs.

Functional Test Cases & Their Execution: The following test cases were created and executed against the application.

7.3 Extract Tweets by Keyword

TESTCASE NAME	Extract Tweets By Keyword
TEST CASE ID	TC_01
DESCRIPTION	This feature allows the user to get event information from Twitter websites
TESTING TECHNIQUE USED	Black Box Testing
PRE CONDITION	System is running, has access to internet.
INPUT VALUES	Key word to search.
VALID INPUT	A keyword in string format
SEQUENCE	1. Go to Search Page 2. Enter keyword to search 3. Select Twitter as Search option 4. Press Search Button

EXPECTED OUTPUT	1. Tweets containing keyword provided by user or information message in case no tweet has been found.
ACTUAL OUTPUT	1. Tweets were displayed, that contained the keyword.
STATUS	**PASS**

Table 11 Extract Tweets by Keyword

7.4 Extract News by Keyword

TESTCASE NAME	Extract News By Keyword
TEST CASE ID	TC_02
DESCRIPTION	This feature allows the user to search Dawn & The Nation website for update about event taking place as per the provided keyword.
TESTING TECHNIQUE USED	Black Box Testing
PRE CONDITION	System is running, has access to internet and has updated

	news in database via continuous and automatic crawling.
INPUT VALUES	1. Key word to search.
VALID INPUT	1. A keyword in string format.
SEQUENCE	1. Go to Search Page 2. Enter keyword to search 3. Select Dawn or Nation as Search option 4. Press Search Button
EXPECTED OUTPUT	1. Links to news blog containing event information as per specified keyword or information message in case no news has been found. 2. Event marker set on Google map in a pop-up window
ACTUAL OUTPUT	1. Hyperlinks were displayed, that contained the keyword and event location. 2. Marker was set on Google maps on the location where event is taking place.
STATUS	**PASS**

Table 22 Extract News by Keyword

7.5 Refine News Search by Location

TEST CASE NAME	Refine News Search By Location
TEST CASE ID	TC_03

DESCRIPTION	This feature allows the user to refine Dawn or Nation news search by providing either city/province to check if there is any event taking place in them or not.
TESTING TECHNIQUE USED	Black Box Testing
PRE CONDITION	System is running, has access to internet and has updated news in database via continuous and automatic crawling.
INPUT VALUES	1. Key word to search. 2. City or Province to refine search.
VALID INPUT	1. A keyword in string format. 2. City or Province name from dropdown.
SEQUENCE	1. Go to Search Page 2. Enter keyword to search 3. Select Dawn or Nation as Search option 4. Select City or Province Name 5. Press Search Button
EXPECTED OUTPUT	1. Links to news blog containing event information as per specified keyword and location or information message in case no news has been found. 2. Event marker set on Google map in a pop-up window
ACTUAL OUTPUT	1. Hyperlinks were displayed, that contained the keyword and event location. 2. Marker was set on Google maps on the location where event is taking place.
STATUS	**PASS**

Table 34 Refine News Search By Location

7.6 Mark Location on Google Maps Test Case

TEST CASE NAME	Mark Location on Google Maps
TEST CASE ID	TC_04
DESCRIPTION	This feature allows the user to view the location of on-going event on Google Maps.
TESTING TECHNIQUE USED	Black Box Testing
PRE CONDITION	System is running, has access to internet and has updated event information in database via continuous and automatic crawling.
INPUT VALUES	1. Any Search method
VALID INPUT	1. A keyword in string format. 2. or City/Province name from dropdown.
SEQUENCE	1. Go to Search Page. 2. Perform any of the search options. 3. Press Search button.
EXPECTED	1. Event marker set on Google map in a pop-up window. 2. Links to news blog containing event information as

OUTPUT	per specified keyword and location or information message in case no news has been found.
ACTUAL OUTPUT	1. Marker was set on Google maps on the location where event is taking place.
STATUS	**PASS**

Table 45 Mark Location on Google Maps

7.7 Delete past Event Information Test Case

TESTCASE NAME	Delete Past Event Information
TEST CASE ID	TC_05
DESCRIPTION	This feature provides administrator the right to delete information of all those events that have occurred in past.
TESTING TECHNIQUE USED	Black Box Testing
PRE CONDITION	1. System is running and has access to internet. 2. Administrator is logged in.
INPUT VALUES	1. Database entries to be deleted

VALID INPUT	1. Select all or few entries of events database
SEQUENCE	1. Go to Admin login Page. 2. Provide credential for successfully login. 3. Select the entries to be deleted. 4. Click Delete button.
EXPECTED OUTPUT	Entries deleted and message of deletion is displayed.
ACTUAL OUTPUT	Entries deleted from database and message of deletion is displayed
STATUS	**PASS**

Table 16 Delete Past Event Information

7.8 Admin Login

TEST CASE NAME	Admin Login
TEST CASE ID	TC_06
DESCRIPTION	This feature provides administrator to login to the system in order to access administrative rights.
TESTINGTECHNIQUE	Black Box Testing
PRE CONDITION	System is running.

INPUT VALUES	1. Username 2. Password
VALID INPUT	1. Valid username. 2. Valid password.
SEQUENCE	5. Go to Admin login Page. 6. Enter username 7. Enter Password 8. Press Login button.
EXPECTED OUTPUT	Admin is successfully logged in
ACTUAL OUTPUT	Admin logged in.
STATUS	**PASS**

Table 17 Admin Login

7.9 Integration Testing

Social Media Mine's different modules which were developed and tested independently were also tested during integration to ensure system stability. Integration testing helped in ensuring that different modules when combined give complete functionality and nothing is missed or some functionality doesn't give error when integrated with other modules. Integration testing gave us more than 95% results ensuring that most modules were integrated with others as well as compatible. This shows that errors were minimized during integration testing.

7.10 System Testing

System testing was performed at the end of development and integration of Social Media Mine. Complete system was tested using sample data. News Extraction, Tweets Extraction, Refine Search, Administrator roles and thus all sub modules were tested as a whole using sample data. Almost 95% of test cases were successful ensuring that most of errors and bugs in the system were removed and system was stable enough to perform optimally.

CHAPTER 8:

CONCLUSION AND FUTURE WORK

8. Conclusion and Future Work

8.1 Introduction

So far, important system requirements and features and important design decision related to development of Social media mine have been discussed. Moreover, it has been shown how this system has been developed and its usage has been briefly explained. Testing techniques used and the result have also been shown in chapter 7. This chapter concludes this report by highlighting our reflections about this project and future work to improve this work.

8.2 Conclusion

There are a wide range of monitoring platforms available, but all of them mainly focus on a specific area, either news websites or social media. Our aim here was to develop a system that provide search and monitoring functionalities from both news websites and social media. We have used our own algorithms and our own techniques when developing the functions, at the same time not deviating from the standard protocols. This system has been implemented on multiple platforms and its scalability has been tested by developing mobile application on Android platform. This provided multiple accessibility modes to different types of users.

8.3 Future Work

Despite all these advantages, there are still a lot of opportunities to make this system better and expand it so it may be able to cater for other needs as well.

The system is divided into modules and therefore it is very easy to add a new module and integrate it with the existing features if needed. We can add further operations that can be called in between organizations via web services to automate more functions that are currently done manually.

Examples of further expansion include adding a new algorithm for getting data from facebook and other sources for getting more and accurate results and perhaps moves from now-casting to fore-casting, as in the case of predicting risk of flood or future evolution of an epidemic. These are all examples of further enhancements to this system.

9. REFERENCES:

[1] Mining Social Media: A Brief Introduction Pritam Gundecha, Huan Liu Arizona State University, Tempe, Arizona 85287 fpritam@asu.edu, huan.liu@asu.edug

Available: http://tweettracker.fulton.asu.edu/

[2] S. Kumar, R. Zafarani, and H. Liu. Understanding user migration patterns across social media. Twenty-Fifth International Conference on Artificial Intelligence. Association for the Advancement of Artificial Intelligence, Palo Alto, CA, 2011

Available:http://ibnlive.in.com/news/mumbai-blasts-twitter-joins-hands-to-help/167345-3.html

[3] Open Domain Event Extraction from Twitter Alan Ritter, Mausam, Oren Etzioni, Sam Clark University of Washington Computer Sci. & Eng Seattle, WA.

Available: http://statuscalendar.com

[4] ***Informing the Curious Negotiator: Automatic News Extraction from the Internet Debbie Zhang and Simeon J. Simoff***

[5] ECON: An Approach to Extract Content from Web News Page Yan Guo 1#, Huifeng Tang 3, Linhai Song 12, Yu Wang 12, Guodong Ding 11key laboratory of Network Science and Technology Institute of Computing Technology, Chinese Academy of Sciences Beijing, China

[6] An Effective and Efficient Web News Ex traction Technique for an Operational NewsIR System Javier Parapar and ´Alvaro Barreiro IRLab, Department of Computer Science , University of A Coru˜na, Campus de Elvi˜na s/n, 15071, A Coru˜na, Spain

[7] Nowcasting Events from the Social Web with Statistical Learning VASILEIOS LAMPOS and NELLO CRISTIANINI Intelligent Systems Laboratory University of Bristol, UK

[8] Real-time Spatio-temporal Analysis of West Nile Virus Using Twitter Data Shamanth Kumar, Fred Morstatter, and Huan Liu. "Twitter Data Analytics", Springer 2013 Ramanathan Sugumaran, Jonathan Voss Department of Geography University of Northern Iowa

[9] Fred Morstatter, Shamanth Kumar, Huan Liu, and Ross Maciejewski. "Understanding Twitter Data with TweetXplorer"(Demo), KDD 2013

[10] Shamanth Kumar, Geoffrey Barbier, Mohammad Ali Abbasi, and Huan Liu. "TweetTracker: An Analysis Tool for Humanitarian and Disaster Relief"(Demo), ICWSM 2011

Available: http://cordis.europa.eu/fp7/ict/netmedia/docs/publications/social-networks.pdf

APPENDIX A:

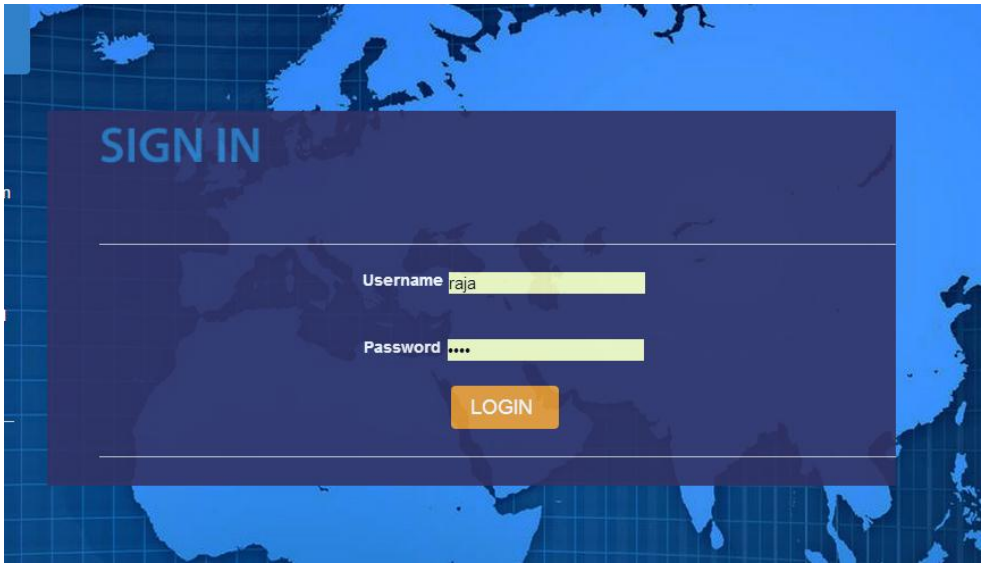
USER MANUAL

10. USER MANUAL

User Manual for web based system

Admin Login Screen:

1. Enter User Name
2. Enter Password
3. Click “Login” button



Search News Website:

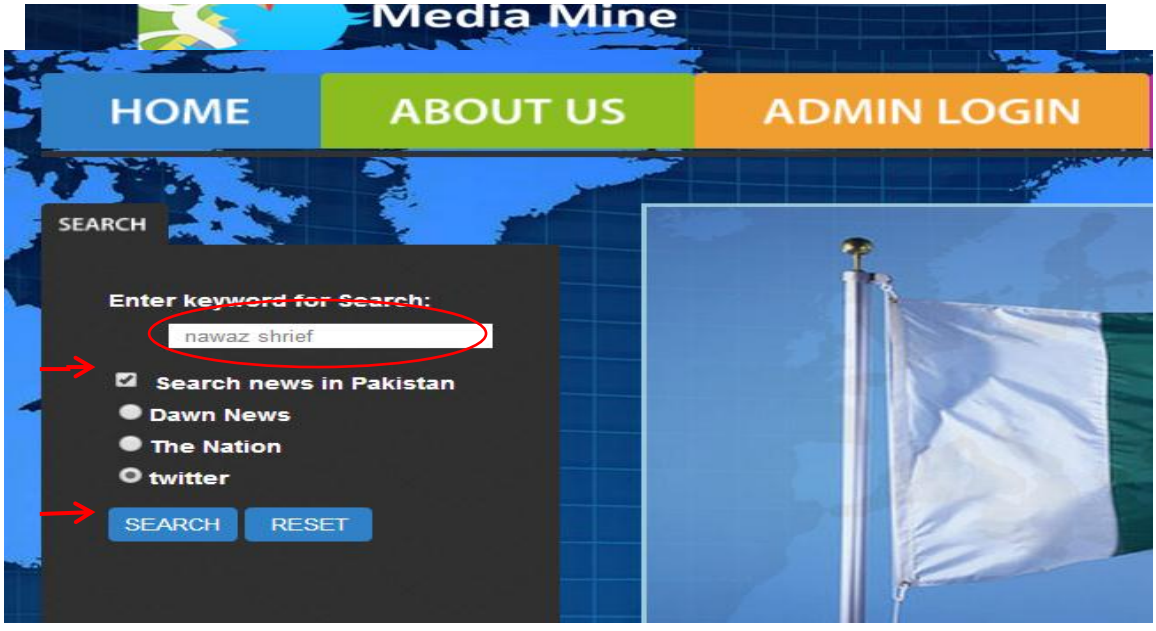
1. Input the keyword for which you want to search event from news website.
2. Select “Search News in Pakistan” option.
3. Then, select any one from Dawn News or The Nation
4. Click Search button.

5. A screen would appear displaying the reference link to the news of event.
6. Click on the “Click to see complete story” for details of event.
7. A pop-up screen would appear, displaying the details of event from the source.

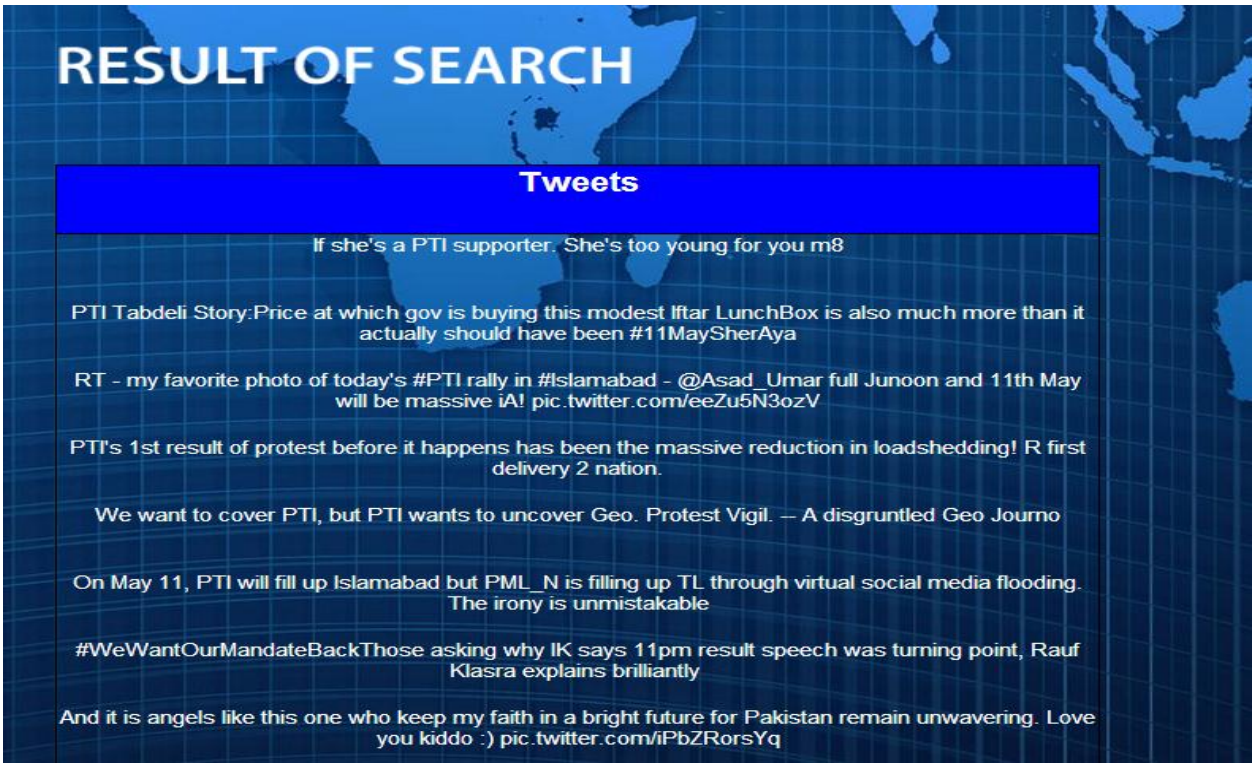
8. A marker will be set on the location where the event is happening in Google maps.
9. Zoom in the map to see the exact location of event.

Search Tweets:

1. Input the keyword for which you want to search event from news website.
2. Select “Search News in Pakistan” option.
3. Then, select Twitter
4. Click Search button

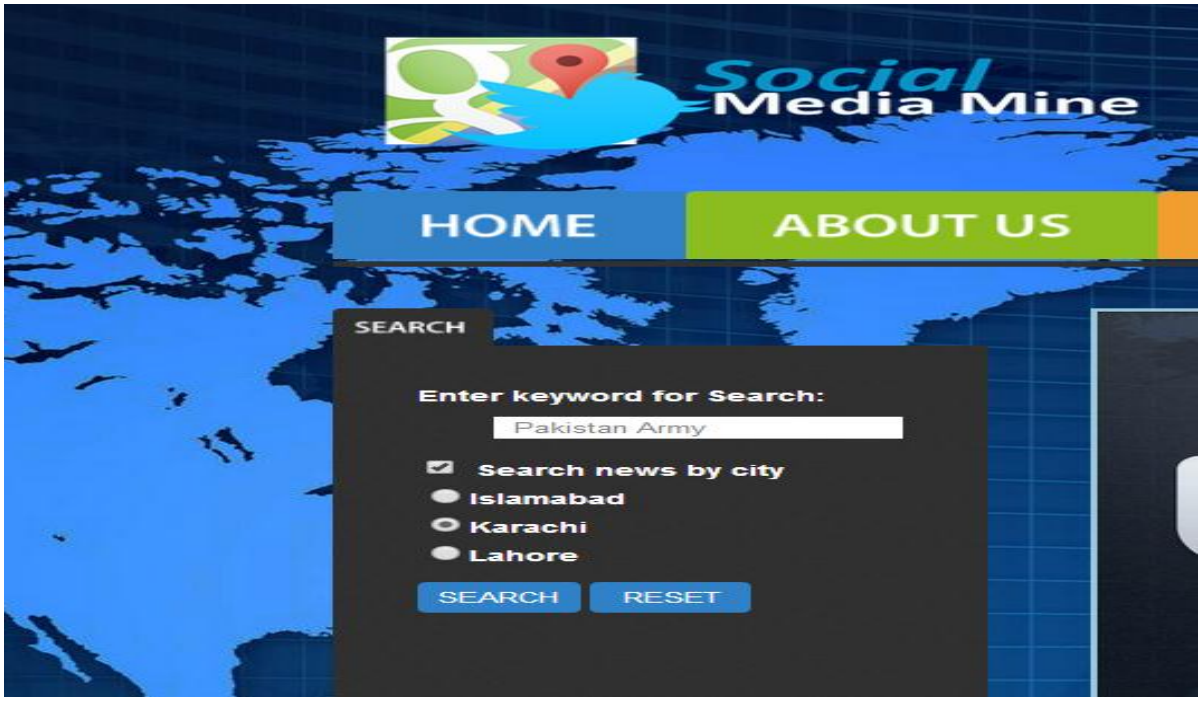


5. A screen would appear displaying the resulted tweets.



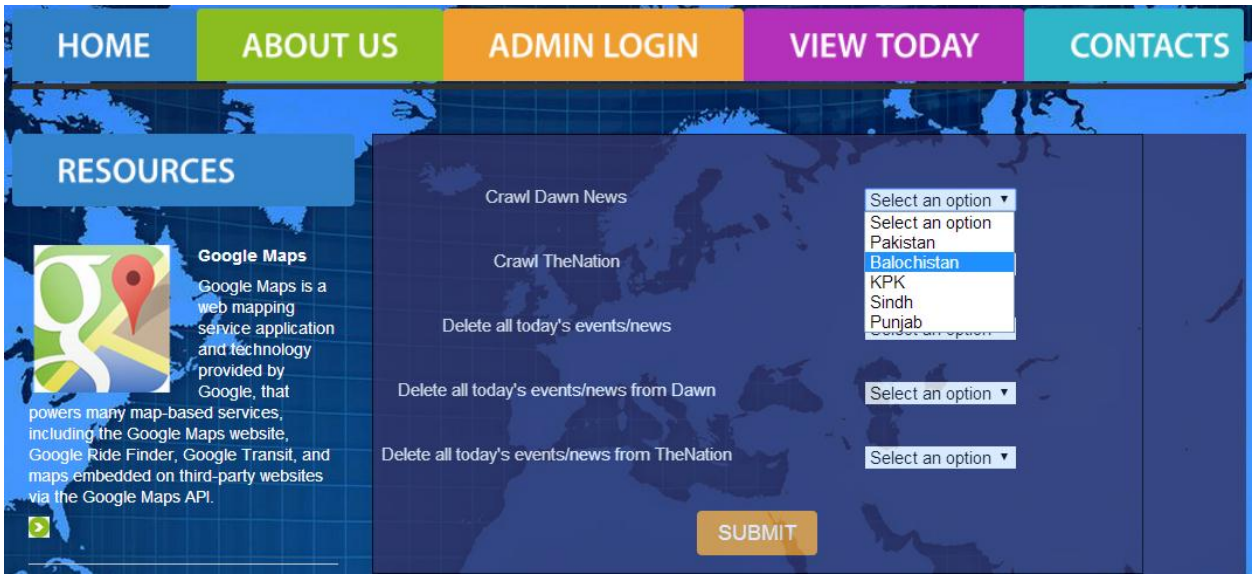
Refine Search:

1. Input the keyword for which you want to search event from news website.
2. Select “Search news by city” or “Search news by province” option.
3. Then, select City name
4. Click Search button



Crawl News Websites:

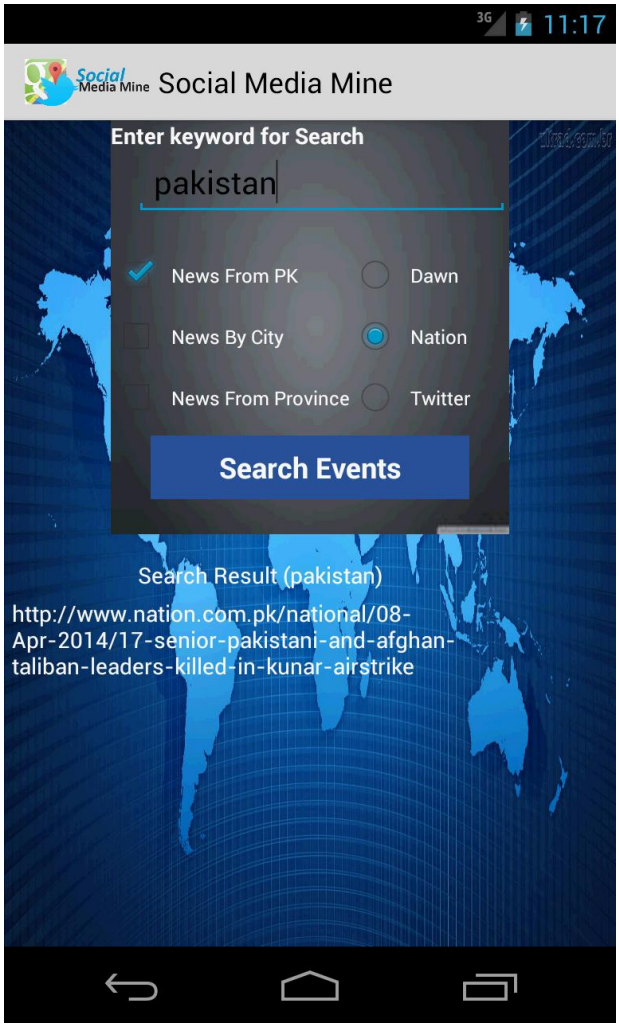
1. Select option to crawl from.
2. Click “Submit” button.



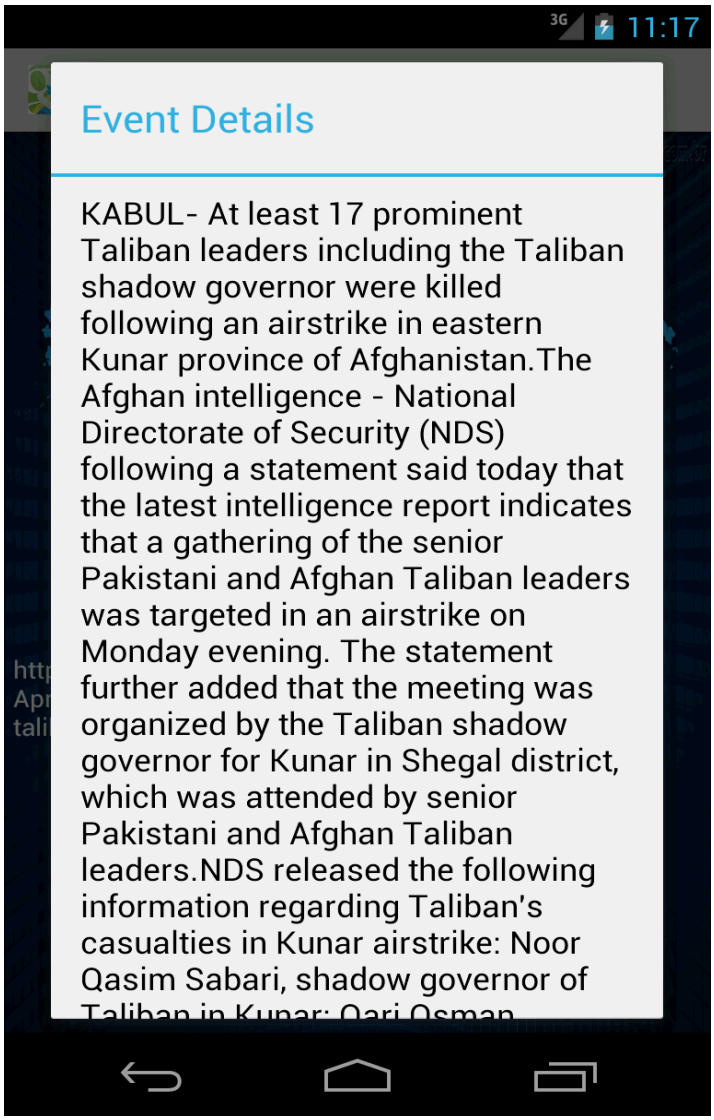
User Manual for Android Application

Search News Website on Android Application:

1. Input the keyword for which you want to search event from news website.
2. Select from different options (News from Pakistan, News by city and News by province).
3. Then, select any one from Dawn News or The Nation
4. Click Search button
5. Result will be displayed in the forms of links.

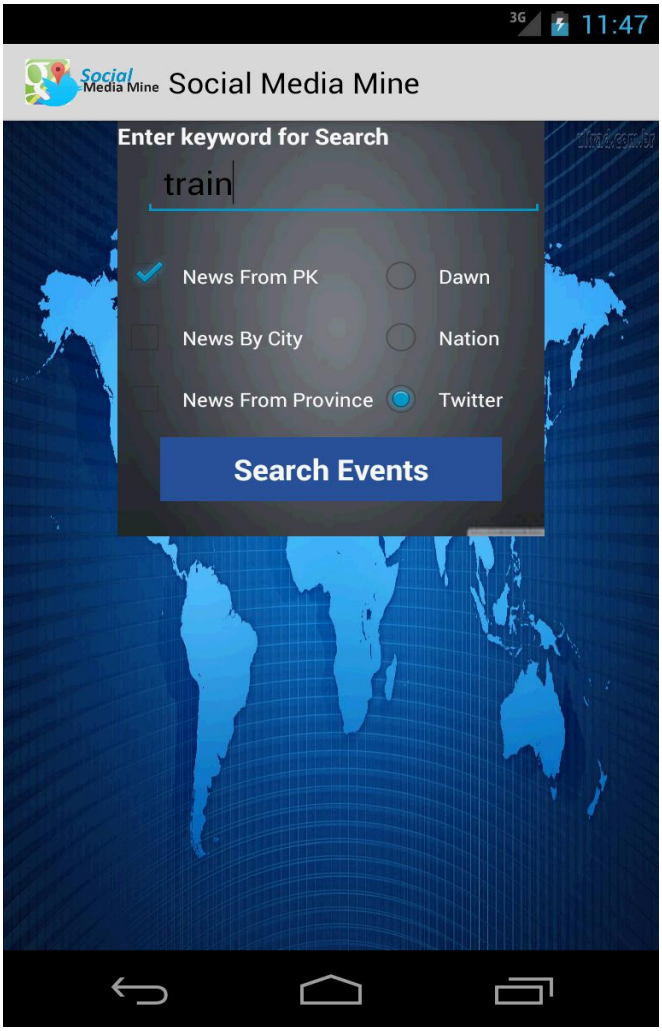


6. A screen would appear displaying the reference link to the news of event.
7. Click on the link for details of event.
8. A pop-up screen would appear, displaying the details of event from the source.



Search Tweets on Android Application:

1. Input the keyword for which you want to search event from twitter.
2. Select “Search News in Pakistan” option.
3. Then, select Twitter
4. Click Search button.



Refine Search on Android Application:

1. Input the keyword for which you want to search event from news website.
2. Select “Search news by city” or “Search news by province” option.
3. Then, select City name
4. Click Search button