

SOCIAL MEDIA MONITORING DASHBOARD FOR UNIVERSITY

BY

DICKEN TAN

A REPORT

SUBMITTED TO

Universiti Tunku Abdul Rahman

in partial fulfillment of the requirements

for the degree of

BACHELOR OF COMPUTER SCIENCE (HONS)

Faculty of Information and Communication Technology

(Kampar Campus)

JAN 2020

UNIVERSITI TUNKU ABDUL RAHMAN

REPORT STATUS DECLARATION FORM

Title: SOCIAL MEDIA MONITORING DASHBOARD FOR UNIVERSITY

Academic Session: JAN 2020

I
DICKEN TAN
(CAPITAL LETTER)

declare that I allow this Final Year Project Report to be kept in
Universiti Tunku Abdul Rahman Library subject to the regulations as follows:

1. The dissertation is a property of the Library.
2. The Library is allowed to make copies of this dissertation for academic purposes.



(Author's signature)

Verified by,



(Supervisor's signature)

Address:

146, JALAN PASIR PUTEH
31650 IPOH
PERAK

DR. PRADEEP ISAWASAN
(Supervisor's name)

Date: 24 APRIL 2020

Date: 24 APRIL 2020

SOCIAL MEDIA MONITORING DASHBOARD FOR UNIVERSITY

BY

DICKEN TAN

A REPORT

SUBMITTED TO

Universiti Tunku Abdul Rahman

in partial fulfillment of the requirements

for the degree of

BACHELOR OF COMPUTER SCIENCE (HONS)


Faculty of Information and Communication Technology

(Kampar Campus)

JAN 2020

DECLARATION OF ORIGINALITY

I declare that this report entitled “**SOCIAL MEDIA MONITORING DASHBOARD FOR UNIVERSITY**” is my own work except as cited in the references. The report has not been accepted for any degree and is not being submitted concurrently in candidature for any degree or other award.

Signature :  _____

Name : DICKEN TAN

Date : 24 APRIL 2020

ACKNOWLEDGEMENTS

First of all, I am grateful and thankful to have a final year project titled Social Media Monitoring Dashboard for University under the supervision of my supervisor, Dr. Pradeep Isawasan who has given me an opportunity to engage in this project.

During these 2 semesters, I have gained a lot of project experience and life experience. I have learnt to solve issues myself and with the guidance of my supervisor and friends. Hence, I wish the spirit and experience gained could be applied in future work or even daily life problem.

ABSTRACT

As social media has been penetrated our lives, analysis of data becomes essential in order to allow people to make well informed decisions. In this project, data related to education is important for the people mainly management of universities or higher education institutions. Hence, the collection of data to perform analysis for better management decision thinking is displayed on a dashboard by implementing a business intelligence. The project objectives of this project are to visualise all social media data at a glance in a dashboard and then monitor the data that is being extracted and analysed as these data is targeted to meet the requirements of users. The implementation of the dashboard is performed from scratch which starts with extraction of social media data from various platforms or sources and stored the data in database management system and finally the data is retrieved from the management system with SQL queries to display data at the dashboard. This university dashboard could show various types of tables and visualisation although there are some challenges and restrictions on overall design due to inadequate of time and data complexities. In short, the dashboard consists of few tabs which basically could be filtered based on name of university, type of university and date of the data being published. Some analysis will be displayed to allow users to view the sentiments and classification of content. Future work could mainly focus on collecting data from other social media platforms, performing trend analysis and more visualisation that is interesting to be shown.

TABLE OF CONTENTS

TITLE PAGE	i
DECLARATION OF ORIGINALITY	ii
ACKNOWLEDGEMENTS	iii
ABSTRACT	iv
TABLE OF CONTENTS	v
LIST OF FIGURES	viii
LIST OF TABLES	ix
LIST OF ABBREVIATIONS	x
CHAPTER 1 INTRODUCTION	1
1.1 Problem Statement	1
1.2 Motivation	2
1.3 Project Scope	3
1.4 Project Objectives	4
1.5 Impact, Significance and Contribution	5
1.6 Background Information	6
CHAPTER 2 LITERATURE REVIEW	8
2.1 Social Media	8
2.2 Social Media Data Providers	9
2.2.1 Data Access Via Tools	10
2.2.2 Data Access Via API	13
2.3 Social Media Analytics Platform	16
2.4 Sentiment Analysis	18
CHAPTER 3 SYSTEM DESIGN	19
3.1 System Overview	19

3.2 Methodology	21
3.2.1 Business Understanding	22
3.2.2 Data Understanding	23
3.2.3 Data Preparation	26
3.2.4 Modelling	29
3.2.5 Evaluation	33
3.2.6 Deployment	34
CHAPTER 4 DETAILED WORKFLOW	36
4.1 Data Collection	36
4.1.1 Twitter Data Collection.	37
4.1.2 Confessions Data Collection.	41
4.2 Challenges and Concerns	44
4.3 Tools to Use	45
4.4 Timeline	47
CHAPTER 5 SYSTEM IMPLEMENTATION	48
5.1 Dashboard Overview	48
5.2 Dashboard Features	49
5.2.1 Features in “Home” Tab	49
5.2.2 Features in “Twitter/ Facebook Data” Tab	50
CHAPTER 6 CONCLUSION	51
6.1 Project Review and Contributions	51
6.2 Novelties	51
6.3 Future Work	52
BIBLIOGRAPHY	53
APPENDIX A FINAL YEAR PROJECT BIWEEKLY REPORT	A-1

APPENDIX B POSTER	B-1
APPENDIX C PLAGIARISM CHECK RESULT	C-1
APPENDIX D FYP2 CHECKLIST	D-1

LIST OF FIGURES

Figure 2.1: Social Media Use.	8
Figure 2.2: Smart Mode and FlowChart Mode of ScrapeStorm.	11
Figure 2.3: Useful Utility Features of Mozenda.	12
Figure 2.4: Task Templates of Octoparse.	13
Figure 2.5: Active Users of Key Global Social Platforms.	14
Figure 3.1: Three Stage Processes.	19
Figure 3.2: CRISP-DM.	21
Figure 3.3: Business Understanding.	22
Figure 3.4: Data Understanding.	23
Figure 3.5: "See More" Link.	24
Figure 3.6: Published How Long Ago.	25
Figure 3.7: Data Preparation.	26
Figure 3.8: Data Cleaning Using Jupyter Notebook.	27
Figure 3.9: Part of Time Zone List.	28
Figure 3.10: Modelling.	29
Figure 3.11: Evaluation.	33
Figure 3.12: Deployment.	34
Figure 4.1: Steps to Extract Twitter Data.	37
Figure 4.2: URL Is Identified Based on Query.	38
Figure 4.3: Steps to Extract Confessions Data.	41
Figure 4.4: 'Post' Tab.	41
Figure 4.5: Date Format of Data.	42
Figure 4.6: Workflow Overview of Confessions Data Extraction.	43
Figure 4.7: Gantt Chart.	47
Figure 5.1: Tabs of The Dashboard.	48
Figure 5.2: Information and Visualisation in "Home" Tab.	49
Figure 5.3: "Facebook Data" Tab.	50

LIST OF TABLES

Table 3.1: Comparison of Data Preprocessing Methods.	30
Table 3.2: Accuracy of Content Category Based on Machine Learning Models.	31
Table 3.3: Accuracy of Content Category Based on Deep Learning Techniques.	31
Table 3.4: Accuracy of Content Type Based on Machine Learning Models.	32
Table 3.5: Accuracy of Content Type Based on Deep Learning Techniques.	32
Table 4.1: Public University with Their Respective Twitter Account.	38
Table 4.2: Private University with Their Respective Twitter Account.	40
Table 4.3: University with Their Respective Confessions Page.	44

LIST OF ABBREVIATIONS

AI	Artificial Intelligence
API	Application Programming Interface
CRISP-DM	Cross-industry Standard Process for Data Mining
CSV	Comma-separated Values
DT	Decision Tree
FYP	Final Year Project
GUI	Graphical User Interface
HTML	Hypertext Markup Language
HTTP	HyperText Transfer Protocol
ID	Identification
IDE	Integrated Development Environment
IP	Internet Protocol
JSON	JavaScript Object Notation
KNN	K-nearest Neighbours
LDJSON	Line- delimited JSON
LR	Logistic Regression
LSTM	Long Short-term Memory
NB	Naïve Bayes
POS	Part-of-speech
RF	Random Forest
SGD	Stochastic Gradient Descent
SQL	Structured Query Language
SVM	Support Vector Machine
TSV	Tab-separated Values
URL	Uniform Resource Locator
UTAR	Universiti Tunku Abdul Rahman
UTC	Universal Time Coordinated
XML	Extensible Markup Language
XLSX	Excel Microsoft Office Open XML Format Spreadsheet File

CHAPTER 1 INTRODUCTION

1.1 Problem Statement

Online social platforms have penetrated human's daily life within the last decade and those platforms had changed the lifestyle despite the fact that social media platforms are becoming one of the most essential things in the means of entertainment and communication. Hence, the days where companies need to spend months or years of time to collect and trawl over data consumer reports were gone.

Based on (The Economist, 2006) stated that as control of the unfiltered, direct and also brutally honest nature of much online discussion is gold dust to huge companies that want to find out what customers really think of them as well as identify the trends behind it. As a result, people nowadays always want to ensure that their view or opinion is heard by the companies they are targeted on social media. It is also relevant for other people, so that people could actively contribute to change or to improve for the services and products that they use or buy.

In addition, as the amount of data increases, it causes humans having difficulty to monitor and also draw conclusions based on all these metrics which are necessary and important in order to obtain useful views or insights so that to bring a brand's advertising tactic further to the next step. Hence, rather than making friendly and insightful user interfaces which data could be easily by users, it is also important to allow to draw conclusions and insights to make sense out of it. For instance, they do not have time to interpret those huge amounts of data such as those posts and comments are being said something positive or something negative about a company.

1.2 Motivation

When it comes to social media monitoring, the motivation comes in when European and American universities are not just schooling the masses by reserving for those with a university fund and the time for a degree. However, those higher education institutions also pay attention to branding. For instance, they implement tactics by monitoring at their social networks on the social media platforms and seeing how people market about themselves. It might seem weird for universities to be commodifying something as fundamental as education. Nevertheless, social media has now become a must by replacing the printed universities prospectuses. In fact, one of the studies by Drake University in 2013 shown that one out of five of their students enroll to their university because of the influence of the university's social media existence. As this university uses social media to improve their overall brand image so it is also possible for other universities to do so through social media monitoring. In order to further elaborate, Harvard University has accumulated more than 5.4 million of fans on their Facebook page and around 900 thousands of followers on Twitter and they are proved that good brand visualisation and good reputation on social media is rewarded. Then, Harvard uses social media monitoring to monitor everything being said about them online whether it is just news, research or footage that encouraging school spirit and brand message or positive and negative comments that they could not control so that they could response or react instantly to keep their brand image intact.

As a result, social media analytics and data are essential when it comes to the most important areas of a university's social media strategy due to the widespread popularity of social media. Lack of businesses opinions and decisions on where to invest their marketing resources and efforts will prevents them for having a greater competitive benefit.

1.3 Project Scope

A social media monitoring dashboard for university is to be delivered which could displays information or data through a single interface in the dashboard for any user who is interested in university data or particularly management of university.

The data will be collected from multiple sources from different platforms which are Twitter and Facebook. Multiple sources mean that both Tweets and Facebook data displayed will be scraped from different accounts and pages that are related to Malaysian University. As Twitter and Facebook are popular social media platforms that are very suitable in the manner of interacting with each other as well as expressing opinions within the online world. Those Tweets and Facebook data could give real data or information which is in the form of short texts such as clear-cut ideas, incidents as well as opinions captured within the moment. Eventually Tweets and Facebook data are appropriate and well-suited sources of streaming data which allow people to do sentiment analysis and text classification. Hence, the dashboard could also classify the collected data into three categories which are positive, negative or neutral sentiment. As a result, users do not need to trawl over every account to get and identify the content as well as sentiment of every Tweet and Facebook data.

Specifically, the data obtained will be Tweets tweeted by Malaysian or Malaysia-related University Accounts and Posts posted by Malaysian University Confessions Pages. For instance, accounts involved are 20 public universities and 75 private universities which inclusive of all public universities and one of the private universities included is UTAR which will definitely benefits the management of UTAR as well. Confessions Pages involved will be 12 universities which are active on their pages. Active in this context means that the particular page will post at least once a day.

1.4 Project Objectives

In order to develop a dashboard, the aims are as follow:

- To extract relevant data from multiple social media platforms automatically

As people all over the world will communicate and data can come from just about anywhere. Hence, it is important to obtain data from various type of social media platforms daily. Facebook and Twitter have been chosen as the platforms where the data should be obtained automatically because it is troublesome and impractical to manually extract data from those platforms. Hence, after deployment of dashboard, the manpower to maintain the dashboard could be reduced thanks to automation.

- To analyse social media data

This is exceptionally useful when comes in social media monitoring because it allows users to get an overview of the broader public opinion or idea behind certain topics. For instance, the dashboard will be able to identify and measure the level of attitudes, emotions, feelings based on the data collected which this information is subjective impressions rather than facts. It is because humans sometimes could not identify and find out clear conditions for evaluating the sentiment of a piece of text. It is mainly due to the influence by personal experiences, beliefs and thoughts which is a subjective task to measure the sentiment correctly. Hence, by performing those analysis on the correct data, more informed business decisions could be made easily, especially when planning social media marketing campaigns.

- To visualize and monitor all data at a glance in a dashboard

It should provide an interactive, centralised means of extracting, measuring monitoring, and also analysing an array of business insights from posts published by all university's accounts of Malaysia in Facebook and Twitter. The dashboard should be able to show that aggregated information to be displayed in a way that is both intuitive and visual in several key areas. By syncing or incorporating data from those social media platforms in a single dashboard could allow individuals to navigate and view the data easier. Eventually, the users who use the dashboard know what people are saying about the them or their products so that they can respond in an appropriate and quick manner in order to avoid the need of businesses to sift through colossal stacks of unstructured data, which is both inefficient and time-consuming.

1.5 Impact, Significance and Contribution

As social media monitoring is one of the ‘big’ things nowadays because when thinking about how large the social internet space has become, it is unimaginable. Every nanosecond, people are constantly communicating with each other online and expressing or broadcasting their opinions about pretty much everything, including businesses. Though monitoring the brand’s social media mentions, it allows people or companies to know their weaknesses and strengths so that people can improve their personalities, services and products based on the views or opinions of other people and other companies.

This dashboard allows any individual to develop a simple and cost friendly dashboard to keep track and monitor information on social media platform which is Twitter by querying over different keywords and settings. In addition, So, this dashboard allows users to stay on top of everything relevant to themselves and stay ahead of trends. This could save companies’ or users’ precious time to pay high amount of money to subscribe on those social media analytics software or dashboard to get notified about the news or information on each site. As a result, universities could see a detailed overview of their businesses in one quick glance and further reduce their amount of time too.

1.6 Background Information

Social media has been growing quickly and presenting a wide range of new opportunities and challenges for businesses to capture the attention of customers. As the number of users or consumers using social media rise exponentially over the years, the power of social media could not be ignored. A series of developments throughout the years ranging from new privacy regulations, major acquisitions and closures, increasing implementation of Artificial Intelligence (AI) as well as a rise of messaging apps have been disrupting and influencing the social media industry. Hence, new approaches and continuous learning are required to master new developments, technologies and analyse the increasing volume of various types of data efficiently such as online activity and customer interactions.

With new social practices and principles infusing business activities and processes, a lot of pressure has been placed on all parts of an organisation to capably manage, listen, monitor, engage and analyse within social media. Organisations have started to invest into social media programs recently since social media has become a typical business tool. One of the programs is social media monitoring.

In short, social media monitoring is a systematic and regular observation and analysis of social communities and social media data in the networks of consisting millions or billions of users. It is performed by tracking and gathering data or online information such as consumer sentiments, brand reputation, industry trends as well as competitors' actions of individuals, groups, companies or even organisations through social media that is relevant to your business. Basically, social media monitoring is the use of tools to listen what is being said across the web. Most of the social media monitoring tools work by continuously indexing and crawling sites, sometimes in real time. Once all those sites are indexed, they could then be searched to find sentiment, mentions as well as opinions on specific brands or products.

So, in order to incorporate or in line with the implementation of social media monitoring, dashboard is being used to perform monitoring. So, a dashboard basically is a user interface that manages and displays information in a way that is relatively easy to use and read. Most of the tools' graphical user interface (GUI) look a lot like a dashboard to some extent. In addition, dashboards are referred as some developers' tools' graphical user

CHAPTER 1 INTRODUCTION

interface as the dashboard is targeted to be able to incorporate information from various sources into one integrated display so that information could be viewed in one glance. But based on some other developers, they consciously make use of this comparison to allow user to recognise the similarity between an automobile's dashboard and the tool's user interface and instantly. Conversations should be able to be monitored and viewed in the dashboard by users through listening to the information which is followed by the users just in a easy and quick approach. At the same time, some graphical representation of data could be provided in dashboard in terms of listings, charts as well as past graphing of phrases and queries. Besides, some sophisticated settings or preferences for filtering are available in most of the tools such as region, language, type of media, or organise the findings uncovered. Hence, a dashboard that could be customised that suits and meets the requirements of the users by including wide-ranging of visualisation tools should be provided in those social media monitoring tools.

CHAPTER 2 LITERATURE REVIEW

2.1 Social Media

Given more number of consumers have immersed in the use of social media such as Facebook and Twitter, companies could not pay no heed to the influence that is weaved within its networks. (Kemp, 2018) presented a statistic which the usage of social media's persists to increase quickly as well as the number of active users who use the most active social media platform in every country has raised by nearly 1000 thousand new users every day during last year. In addition, the use of social media in each and every month is around 3 billion of people around the globe, with 9 in 10 of those people gain access to their preferred platforms through mobile devices. For instance, one of the crucial headlines which presented by digital in 2018 shown that the amount of active social media users in 2018 is 3.196 billion with 42% of active social media users being as a proportion of the total population in Figure 2.1.

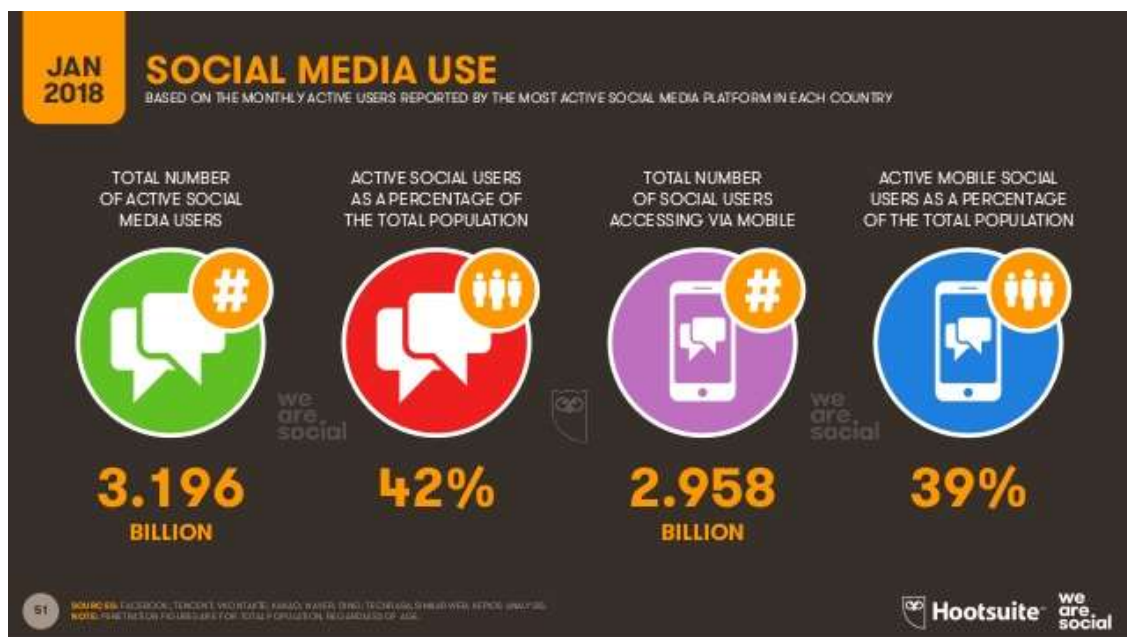


Figure 2.1: Social Media Use.

People nowadays are utilising the social media platforms to report everything that happens around that such as needs, complains as well as opinions regarding services and products that they used, and eventually to compare with other services and products which

are provided by other retailers. (Perdue, 2010) said that retailers tend to have a capability to communicate with consumers instantly and exponentially greater volume of audience because of the upsurge of social media platforms. In regard to that, (VanBoskirk, 2009) forecasted that social media marketing might achieve a growth rate of 34% annually in terms of spending which will outmatch all other forms of online marketing. Hence, social media is an important area for companies to take care of as shown as the statistics above.

2.2 Social Media Data Providers

Companies nowadays employ a wide range of traditional as well as modern approaches to listen to customers. However, (Brogan, 2007) stated that those traditional methods of collecting data cause a lot problems for survey researchers recently due to the decline in commitment of respondents to participate and also coverage of landline telephone. In addition, many industries and businesses are still vulnerable in the data realm. For instance, a survey had been conducted in 2017 reveals that 37.1% of the organisations do not have a Big Data strategy. While just a small percentage have achieved some success among the rest with data-driven businesses. One of the root reasons is due to companies are lack of or just have the minimal understanding of data technology. Furthermore, the nature of traditional survey researches which is not cost-friendly and also time consuming sustain the desirability of using free and unrestricted online sources of information further. This has led the companies to take advantage by started to use numerous data services, tools and analytics platforms which are available widely around the world. As a result, social media monitoring providers such as scraping tools and monitoring platforms have become known in recent years to deliver the need for customer listening or monitoring methods and also to exploit the abundance of information available online. For instance, social media data resources could broadly subdivide into 2 categories which are those providing data access via tools as well as data access via Application Programming Interfaces (API)s. Both methods are the same in terms of allowing users to crawl information online to be viewed or analysed later. However, they are different in terms of extracting the information as both have pros and cons.

2.2.1 Data Access Via Tools

ScrapeStorm

ScrapeStorm which is powered by AI is an online data scraping tool which comes in free and paid plans. It could identify and extract every data and information automatically in the form of list obtained from any website. It has an intuitive, straightforward and user friendly interface which just requires users to key in the URL of the first web page and it will start scouring and collecting data immediately from all pages without users dealing with programming or configuring the rules for scraping. This data scraping tool is designed and developed by individuals who worked on Google's crawlers. It is to ensure users that ScrapeStorm is capable of crawling and extracting tons of useful data from web pages as advertised. It is also capable running on all types of operating systems such as Windows, Linux, and Mac and could be utilised by individual users or by teams. One of the advantages of it is once data is extracted, users could save the data to any format of their choice such as CSV, Excel, and HTML or even move the data directly into their database. They could also opt to save all the data they extracted into a cloud-based hub for easy access, management and storage. Besides, ScrapeStorm allows users to crawl any web page using multiple crawling methods such as the Smart Mode and the Flowchart Mode as shown in Figure 2.2. However, there are still many problems with this software. For instance, some of the details are needed to be optimised on the operation page and some button settings are not very reasonable. The features of free version are also limited such as it just allows exportation of 100 data per day which is quite limited.

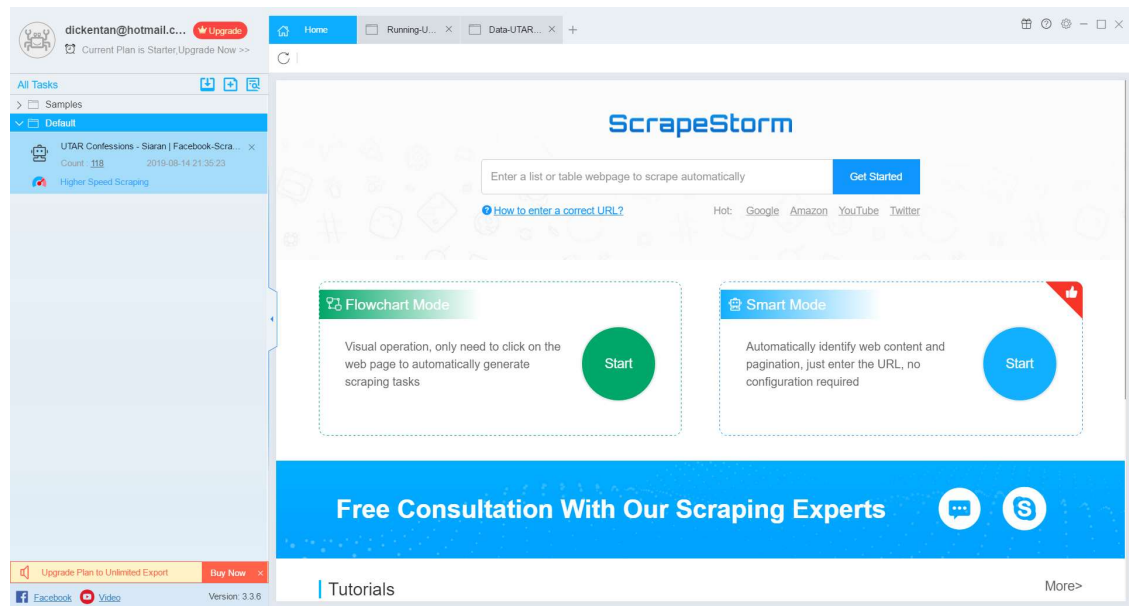


Figure 2.2: Smart Mode and FlowChart Mode of ScrapeStorm.

Mozenda

Mozenda is a cloud-based web scraping service that comes with paid and 30 days free trial versions with useful utility features for data extraction. The user-friendly interface of it is point-and-click user interface for scraping particular data from any website and enables automation to be performed as well as data export any file format such as XML, TSV and CSV. The Mozenda's scraper software consists of 2 parts which are Mozenda Agent Builder and Web Console. The Agents are to extract projects, view as well as arrange their results, and eventually export or publish the scraped data to cloud storage such as Amazon, Dropbox and Microsoft Azure whereas the web console is a web-based application that allows the users to the agents. In addition, a Windows application which is Agent Builder is used to build any data project. The optimised harvesting servers in Mozenda's Data Centers process the extracted data, therefore reducing the threats of IP-address banning of the client due to the loading of web resources if detected. One of the advantages of Mozenda is it supports image extraction and documentation extraction. For instance, the technology allows it to scrape all the significant information from the table structures rather easily shown in Figure 2.3 where other data extraction tools impossible to do it. Furthermore, it has also offered some more useful utility functions to ease the process for

data extraction. Besides smart data aggregation and multi-threaded extraction, Mozenda also provides Geolocation such as fake location to decrease the chance of IP banning, Test Mode and Error-handling to fix errors. However, this software is quite expensive as it costs from \$250 to more. Another disadvantage of it is logical functionality is inadequate and it requires Windows platform in order to run. Those extra-large websites will also cause stability issues to Mozenda.

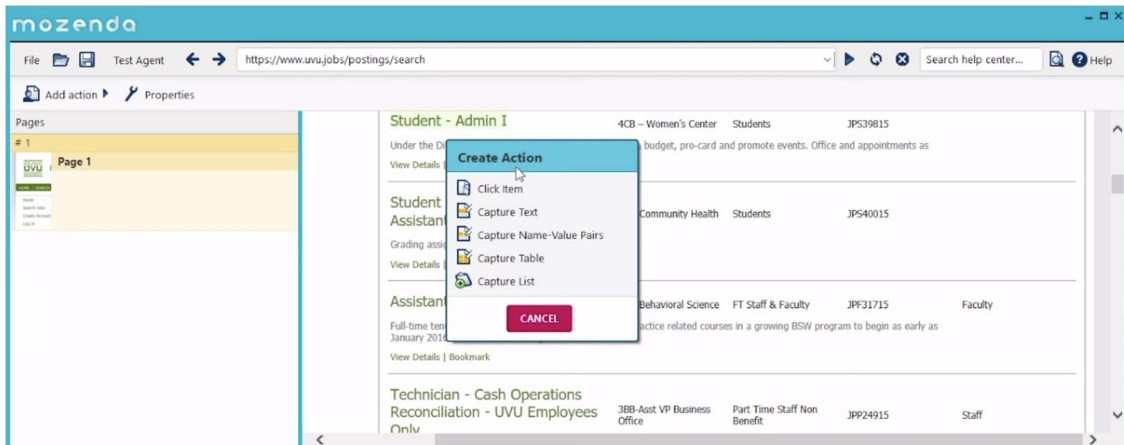


Figure 2.3: Useful Utility Features of Mozenda.

Octoparse

Octoparse is a strong web extractor with comprehensive and premium features used for scraping nearly all sorts of data that the user needs from the website. It comes in free version and paid for subscription version. Users could use Octoparse to grab information from any website with its extensive capabilities and functionalities. It is done by allowing users to extract all sorts of text from the website with its point-and-click interface, and therefore users could then obtain nearly any content from any website without any coding and save it by exporting it as a structured format such as CSV, Excel and HTML or even databases such as MySQL . One of the advantages of Octoparse is IP blocking by those websites is not a problem anymore as this software provides IP Proxy Servers which could automate IP's leaving without being spotted by uncompromising websites which comes with bot checker. In addition, this software offers task templates in its new version and some templates in paid version as shown in Figure 2.4, which contains ready-to-use tasks for scraping from different types of websites, such as Alibaba, Amazon, Tmall, eBay,

Facebook and many more. The free version provides a functionality which users could scrape unlimited pages as compared to other data scraping tools which have limited pages to be scraped with their free version. When using the task templates, the users just have to key in their intended parameters such as keywords or their intended page URL for searching. Then all the data will be extracted automatically by waiting a couple of minutes. Although the pricing of its premium version is cheaper and looks competitive as compared to others, some of the functions are physically broken. For instance, some of the templates could not extract any data at all and more. In addition, Octoparse could not extract data from PDFs unfortunately and also just allows users to extract the URs of the images but not download the images directly.

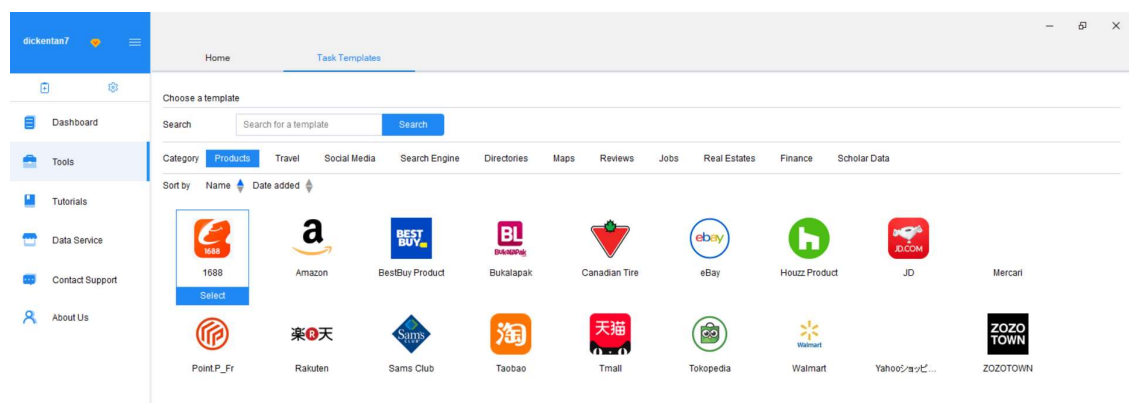


Figure 2.4: Task Templates of Octoparse.

2.2.2 Data Access Via API

Questionably the most helpful sources of social media data for researchers and developers are applications or webpages that offer programmable access through their APIs, especially using HTTP-based protocols. For those popular social networks, such as Twitter and Facebook allow their data to be accessible via APIs given their importance to academics. Although there are plenty of social media sites provide APIs, however, there is still a minority of sites that does not provide APIs access for scraping data such as Bing, LinkedIn and Skype. In addition, many top and prominent networks are limiting their free access features, even to researches since an increasingly number of social networks are moving to widely available content. More social networks mean that more information will be available widely and due to privacy and security concerns, restrictions and security

proactive measures are made. For instance, Foursquare made an announcement in December 2013 that it will not provide private check-ins on iOS 7 anymore. Hence, in order to offer an endless stream of anonymised check-in data, it has now partnered with Gnip instead. The data is then available and provided in two packages which are a filtered version through Gnip’s PowerTrack service and a full Firehose access level.

The APIs offered by Facebook and Twitter were chosen to be discussed below because Facebook and Twitter are two of the widely used social platforms are shown as the statistics below. According to (Boor & Grunwald, 2011), in a sample consisting of 3 thousand students in United States of America (USA), 9 out of 10 use Facebook, whereas 37% of them use Twitter as a communication platform. At the same time, about 71% of universities’ or colleges’ students are Facebook users based on another study in this country as stated by (Duggan & Brenner, 2013). Therefore, among the social networks, Facebook is possibly the most popular for education and personal purposes. Besides, social networks are being used by academic institutions for internal management to solve educational issues as stated by (Forkosh-Baruch & Hershkovitz, 2012). In addition, based on (Kemp, 2018), the global social landscape is dominated by Facebook’s core platform, with the total users up 15% year-on-year to achieve roughly 2.17 billion whereas Twitter just has around 300 million of users at the start of 2018 as shown at Figure 2.5.

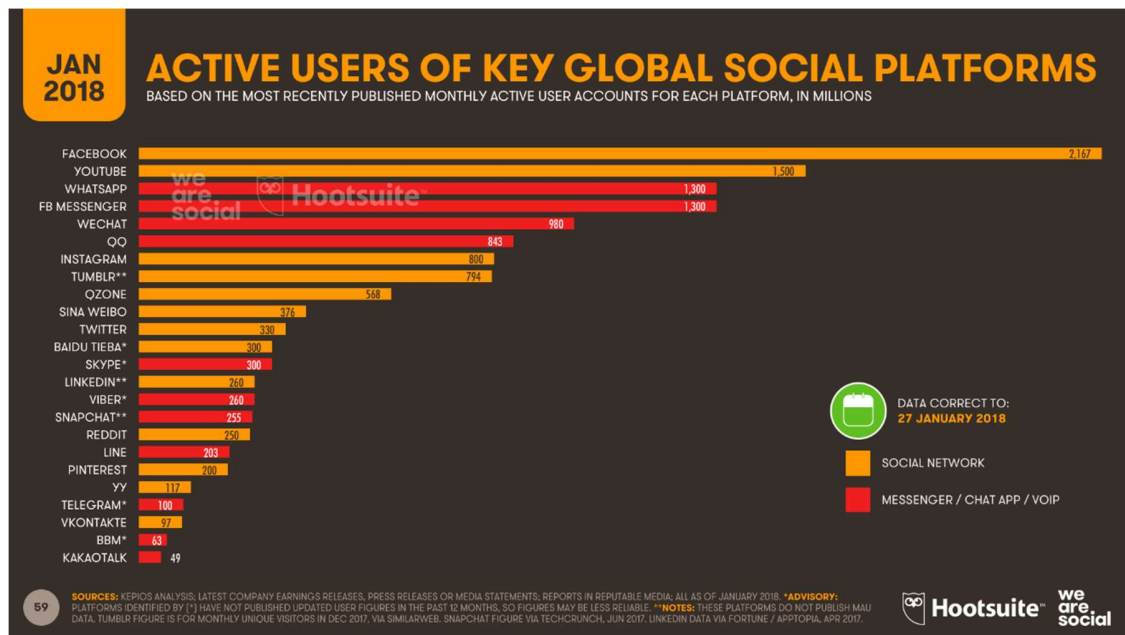


Figure 2.5: Active Users of Key Global Social Platforms.

For instance, Twitter allows data to be collected with those widely accessible and available APIs. The search and filter APIs are the most common APIs which grant users to place certain search instructions to just filter information that users need, especially keywords. The access to the data is available for everyone for free, however, it is not able to access to all potential data and just a minor proportion of Tweets are able to access. In the effort to prevent data leakage and also to protect the privacy of users, the developers had also modified their API access model in 2018 which allow users to choose from those options which comprise of different subscription levels. In order to get more data, paid access has always been available, with some research groups and universities have to expand their accesses through subscriptions so that researches could be done stated by (Kate & Megan, 2015). Although the access model to those data is probably will not change and the developers always just tend to continue to adjust the rates and offer different degrees of access to data. Besides, the underlying access model of Twitter in searching for location, user or a particular keyword and getting a rate-restricted response is expected to continue. Hence, this situation causes other researchers eventually work with a various dataset which is depending solely on the keywords, conditions or terms with respect to their search queries as well as depending on which subscription levels that the researchers could pay for.

The user timeline API of Twitter could extract a user's most recent Tweets up to 3200 Tweets as of June 2018. Each only request that calls to the API could just extract 200 Tweets. Hence, in order to acquire all 3200 Tweets would likely to take 16 requests. However, the request of this public API endpoint is rate-restricted up to 1500 requests for every 15 minutes. This also imposes a highest number of extraction of 93 users for every 15 minutes or 6 users per minute. However, this number is not relevant in reality as it might be a little higher due to not every user has tweeted for 3200 times, therefore their entire timelines could be retrieved in less than 16 requests. As a result, this information will be kept at the cloud storage bucket in a line- delimited JSON (LDJSON) file which is one tweet per line after a user's full available timeline is downloaded.

As compared to Tweets, the privacy issues of Facebook are more complicated and complex than Twitter's which mean that most of the status messages are tougher to be extracted than Tweets as Facebook requires 'open authorisation' status from users.

Currently, Facebook places all data as objects. It also has a series of APIs which ranges from the Keyword Insight APIs to Public Feed API. Hence, the API call can be made with its unique ID to be known in order to gain access to the properties of an object. But the key problem is by what means to scrape the information that is obtainable on Facebook and by what means it could be used in order to make useful conclusion or insight. The way to achieve this is firstly, a data analyser for text data is a must to be built. So, through public Facebook APIs, practitioners and researchers could get the text data which is available on Facebook. Then developers are allowed by Facebook to scrape data via Facebook REST API and the Streaming API. In simple words, user is needed to create an account using Facebook developer first if the user is interested in doing analysis on Facebook data and then just click on a button which is 'new application' which is provided by Facebook to create an application easily. Next, the access tokens are needed to be created after creating the new application so that users are required to provide their authentication details information. Also, the user is able to get one consumer keys to get access to that application for acquiring Facebook data right after creating the application. However, Facebook, it has rolled out additional API changes which place more restrictions on developer from accessing data for applications by deprecating and restricting some API products since last year such as Graph API, Media Solutions, Profile Expression Kit, Marketing API, Pages API and Live Video APIs as well as Lead Ads Retrieval. These were due to the Cambridge Analytica data misuse scandal as well as the consequence of an application that had been leaking data on millions of users. These restrictions caused Facebook required advanced developer permission on these APIs. These changes are planned to continue helping developers create applications while protecting the privacy and data of users using Facebook. Hence, it is very difficult to obtain Facebook data since 2018.

2.3 Social Media Analytics Platform

The growing number of commercial services in recent years for commercial sources which could extract social media data as well as then offer paid-for access through simple monitoring tools are important for companies to do analysis easily. For instance, two examples of social media analytics platforms are Brandwatch and Meltwater. They not just could measure influential topics, demographics and sentiments but those platforms also

comprise of text analytics and sentiment analysis on users' conversations which are widely available online. Most of the social media analytic platforms also offer interfaces which are user-friendly by customising query like search query, reports, dashboards as well as file export options such as to CSV or Excel format. By using a distributed crawler which is usually widely available and used by most of the platforms to scrape an array of social media data that aims at micro-blogging like blogs such as Blogger, Twitter through full Twitter Firehose and WordPress as well as other social networks such as MySpace and Facebook, for those images sites, forums and news sites like Flickr and corporate sites. In addition, multi-language support for widely used languages are even supported and provided by some of the platform. Although those professional social monitoring platforms are quite impressive and great to be used. However, those tools usually cost up to several thousands of dollars a month.

Meltwater

For instance, Meltwater offers a social dashboard, a personal social advisor and multi-user workflow to assist the user study about the advantages of the tool and also to offer support and guidance for the platform's features. Firstly, the dashboard provides a brandometer. This brandometer provides a visualisation of sentiment for users, a graphical view of the overall sentiment, themes cloud and also search results. In addition, the user's campaign search outcomes are also provided in graphical view of which is categorised by social channel. Next, a glimpse of those common search result themes is offered by the theme cloud. The theme could help in identifying possibly harmful conversations. Moreover, the "campaign" searches which could be enabled by the tool are committed to a certain brand, topic or market and recognises relevant key influencers and conversations. Last but not least, Meltwater Buzz helps consumers to find out the overall sentiment of community conversations about the consumer' brand and identify social trends such as sentiment, the volume and media spread of conversations.

Brandwatch

Brandwatch is another famous platform available that able to operate almost in real time by collecting data. The search query is very useful which could search information

ranging from blogs, social networks and micro-blogging sites such as Twitter. It is also undeniable that news information which is in different types of category such as in regional, national and international could be searched in the query. Others include those image sites, video sites, corporate sites and finally discussion forums. The dashboard simply just needed a browser and nothing else, so no software is required to be installed. In addition, API is not available for consumers as this function is not included by developers of Brandwatch. Hence, consumers who hope to integrate their data in Brandwatch with their own system could not be done. In addition, both API and dashboard work by sending queries to the Brandwatch's data-centre and by using a distributed, huge and redundant collection of servers, the data is being stored and being guaranteed in terms of availability and performance. In contrast, the restrictions of this application are the accuracy spam filtering and also sentiment classification. However, results could be enhanced and improved with the involvement of human to correct the filtering spam and sentiment as stated by Brandwatch.

2.4 Sentiment Analysis

Sentiment analysis over social media platforms such as Twitter provides organisations to perform a quick and efficient way in monitoring feelings of publics towards their brand, directors as well as business. As compared to the traditional text such as review documents, sentiment analysis of Tweets is judged as troublesome or problematic due to Tweets are in the form of short text. In addition, the utilisation of irregular and informal words as well as the swift development of language in Twitter also contributed in the problem of sentiment analysis. Following the approaches which are feature-base, there has been a lot of works have been performed in Twitter sentiment analysis during the past decade. As a result, three types of features which are part-of-speech (POS), lexicon and microblogging features are most of the existing works focus on. However, multiple discoveries have been discovered with some said that the significance of POS tags with or without word prior polarity included, while others emphasised the use of microblogging features.

CHAPTER 3 SYSTEM DESIGN

3.1 System Overview

A development of social media monitoring dashboard for university basically involves a three stage processes which are capture, understand, and present as shown in Figure 3.1.

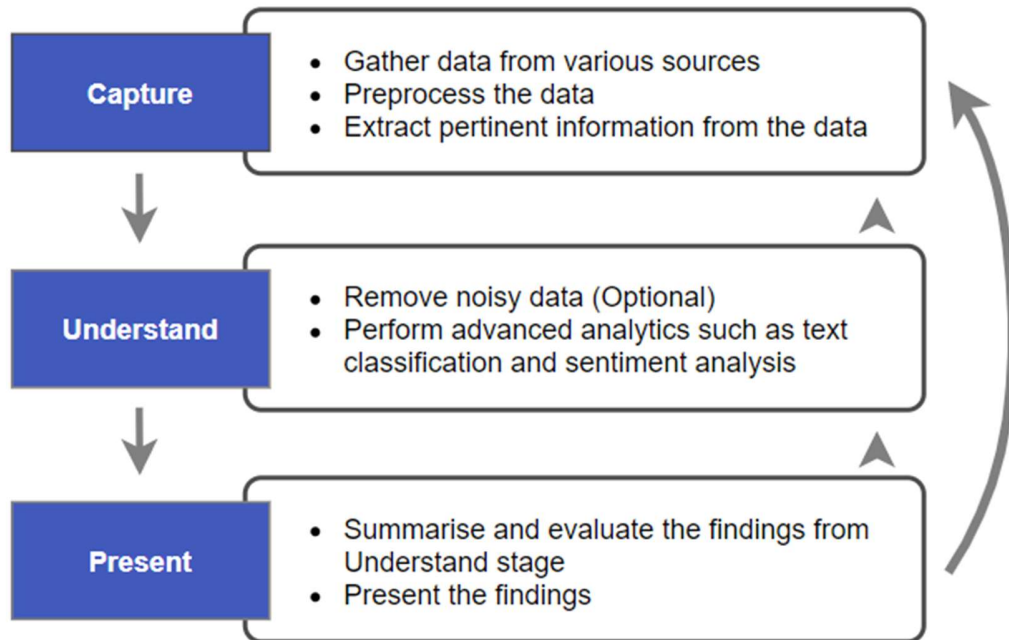


Figure 3.1: Three Stage Processes.

Firstly, the capture stage includes acquiring related social media data. It is done by monitoring or listening to different social media sources, which then gathering those related data and scraping appropriate data. This method usually could either be accomplished by a business itself or via a third-party vendor. As not all the social network data that are scraped will be helpful. Hence, the next stage which is the understanding stage is performed to select related social media data through filtering or sorting for modelling. Different advanced data analytic ways are employed to gain insights from the data that is being collected and analyse earlier. However, before analysing to gain insights, data should be cleaned by removing noisy and low-quality data. Lastly, the findings from the understand stage will be used by displaying them in the present stage in a way that is meaningful to

users. In addition, this final stage also enables users to listen, monitor and analyse the activities that is in the understand stage. It also allows users to summarise all the information and visualise activities which all fall under the present stage.

However, some stages could be overlap among themselves. For example, the understand stage builds models that could assist the capture stage so that less work is required to be done in the capture stage. Similarly, visual analytics which assist in judgement of human beings will add more workload during the understand stage but then reduce the workload in the present stage. So, the capture, understand and present stages are not performed strict and linear manner, but they are ongoing and iterative manner. Hence, extra data is needed again if the models in the understand stage have failed to discover useful trends or insights and improvement is needed in order to strengthen their analytical power. Likewise, if the obtainable results that are shown on the dashboard are not interesting which they could not capture the attention of users or the results are having low analytical power, it might be needed to revert back to the capture stage or even the understand stage so that data could be adjusted or parameters could be further finetuned so that to produce powerful insights that could be used in analytics. In a nutshell, dashboard development is at least inclusive of these three simple stages in order to produce a system that is able to support social media analytics which may run through several repetitions or enhancement before it suits the users. Hence, data analysts and statisticians play an important role in helping to develop and perform testing on systems before they are released and deployed to the users.

3.2 Methodology

The methodology or model proposed is Cross-industry standard process for data mining which also known as CRISP-DM. It provides a well-defined approach to plan a data mining project. This project will basically produce a social media monitoring dashboard for university by collecting data from popular social media platforms. The life cycle of a model as shown in Figure 3.2 splits CRISP-DM project into 6 phases, and some stages are allowed to go back and forth between them.

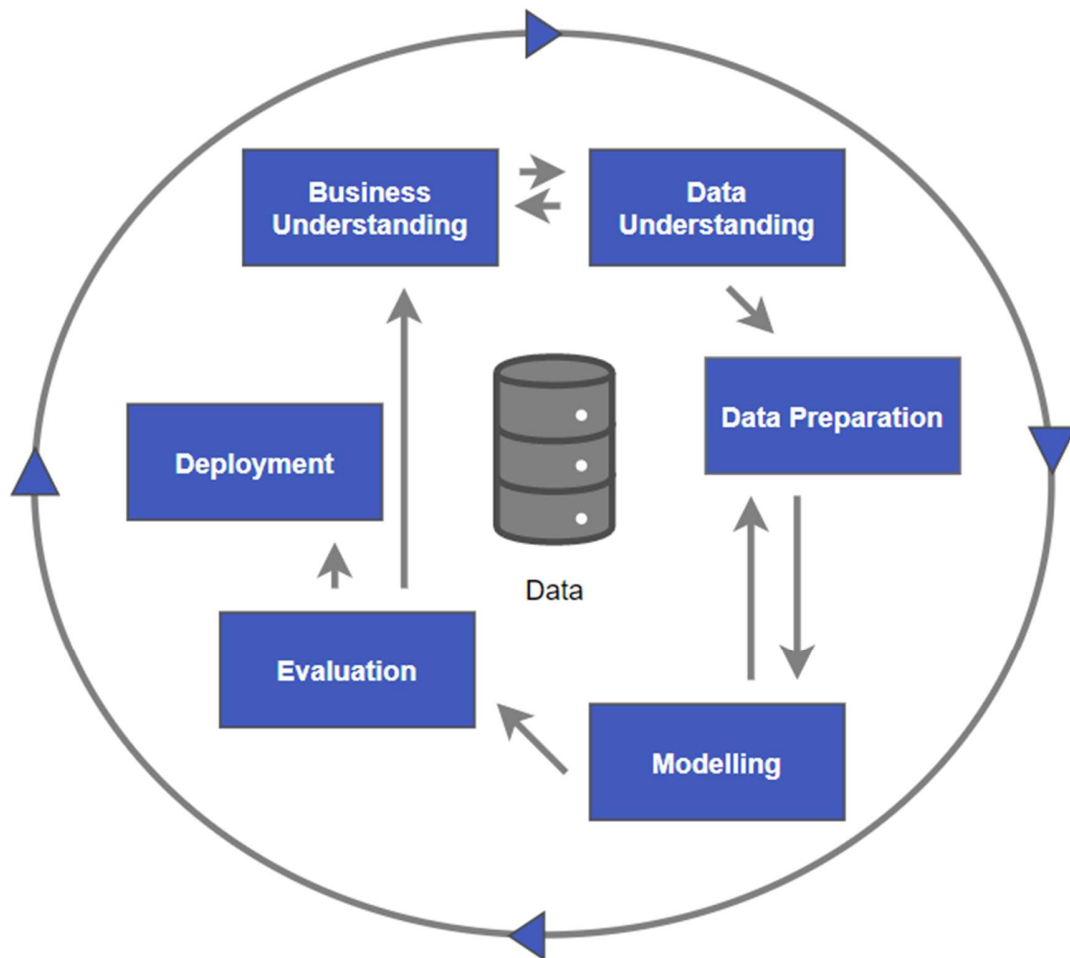


Figure 3.2: CRISP-DM.

3.2.1 Business Understanding

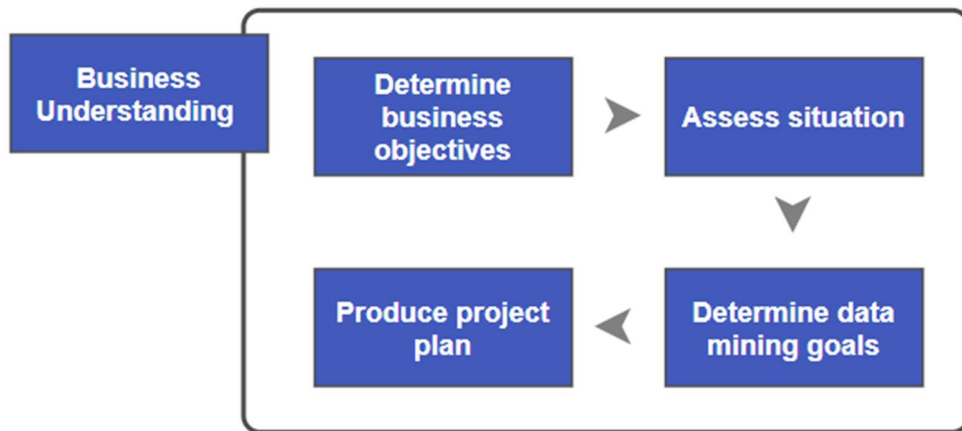


Figure 3.3: Business Understanding.

Determine business objectives

The business objectives are as stated earlier in Chapter 1 which are to extract data from different accounts across multiple social media platforms which are Twitter and Facebook. It is then to be visualised at a glance in a dashboard but before displaying all data in the dashboard, analysis and measurements are performed on data by analysing the social media data which are sentiment analysis and text classification. In addition, those data extracted are from different university's accounts in Twitter and Facebook. For the business success criteria is users able to find out useful insights according to the information provided in the dashboard.

Assess situation

This project is proposed because most of the universities' management currently using social media analytics platform which is quite costly. Hence, a simple, cost friendly as well as user friendly dashboard is needed to overcome this problem. However, due to the limitation of time and computer system, not all university's accounts in Malaysia will be extracted and displayed in the dashboard. Hence, just some data of the university's accounts are chosen to be extracted just to show that this project could be done if there are more computer systems available.

Determine data mining goals

Goals of data mining in this project are to extract relevant data and perform accurate analysis that are extracted from Posts or Tweets posted or tweeted by Malaysian universities' account. For instance, the analysis is text classification and sentiment analysis, so the goals are not just able to extract data but to perform sentiment analysis and text classification accurately.

Produce project plan

For the project plan, the detailed timeline will be shown at the last section of this chapter.

3.2.2 Data Understanding

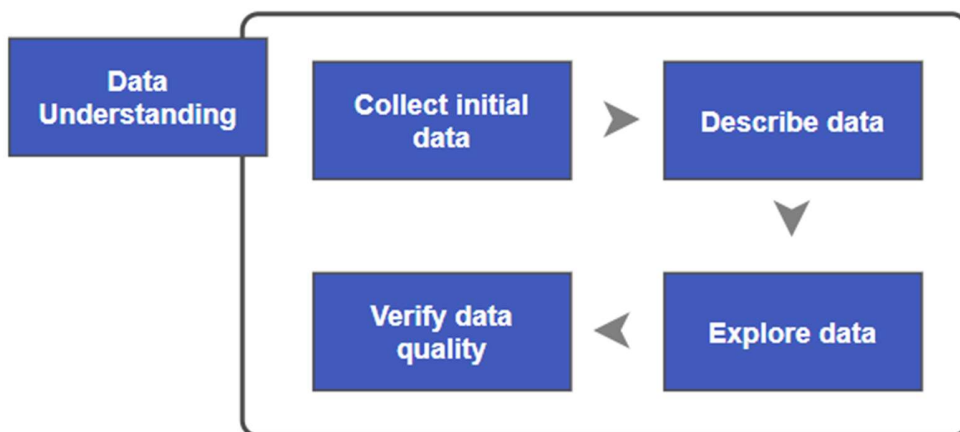


Figure 3.4: Data Understanding.

Collect initial data

For the collect initial data phase, the data acquired are Twitter's Tweets tweeted by university's accounts (Twitter data) and Facebook's Posts posted by Confessions' pages (Confessions data). Those data are collected by using third-party tool which are Octoparse and UiPath respectively. By using Octoparse, the extraction workflow is built upon task template. Hence, the workflow to build it is not difficult. For UiPath, the workflow is built step by step so that data could be extracted according to the flowchart built. The detailed

steps of building the workflows and problems encountered in Octoparse and UiPath will be further discussed in Chapter 4.

Describe data

For the describe data phase, the data acquired are in the format of XLSX. The attributes obtained from Twitter data Octoparse are Category, Keyword, Web_Page_URL, Tweet_Website, Author_Name, Author_Web_Page_URL, Tweet_Timestamp, Tweet_Content, Tweet_Image_URL, Tweet_Video_URL, Tweet_Number_of_Likes, Tweet_Number_of_Retweets, Tweet_Number_of_Reviews and Tweet_Number_of_share. The attributes extracted from UiPath are Name, Date and Content which represent the name of confession page posted, time of the post posted and the content of the post.

Each time the extraction starts, the task template in Octoparse scrapes according to the total number of page scrolls set and obtains all Twitter data within the page range. In contrast, the flowchart in UiPath does not scrape all Confessions data but just limited to the rows of data set to be scraped. It is due to some of the posts are too lengthy in Confessions page. Hence, the posts are hidden and in order to view the full content of those posts, “See More” link has to be clicked to view the post in full as shown in Figure 3.5. In order to avoid posts with “See More” link being scraped. The automated scraping process is set to click the “See More” link whenever the link is found and hence, this “See More” link has limited the total number of posts to be scraped.



Figure 3.5: "See More" Link.

As a result, the data obtained satisfies the relevant requirements as the main components of those data attribute is obtained which are the name, content and time of published.

Explore data

From Twitter data, all the values from Tweets_Timestamp attribute is in Unix timestamp. The Unix timestamp is total number of seconds that have been spent since the first day of 1970, without the inclusion of leap seconds. So, it is essential to convert the timestamps to read datetime which could be using the second or the milliseconds or microseconds methods to allow users could understand easily. On the other hand, from Confessions data, the values in Date attribute are in the format of “published how long ago” as shown in Figure 3.6. In addition, the Content attribute of Confessions data is having data in Chinese, so it is better to be transformed into English to allow better analysis. Besides, the Name attribute from Twitter data and Confessions data could be categorized into public or private university to allow more analysis. Both data could also be further categorise into its category by implementing text classification.



Figure 3.6: Published How Long Ago.

Verify data quality

The Name, Date and Content attributes are important attributes to show that the existence of data with publisher, published date and published content. Hence, in this case, there is some null values from Tweet_Content and Author_Name of Twitter data. So if the Tweet_Content is null, the whole record should be removed while if the Author_Name is null with Tweet_Content is not null, the Author_Name could be identified through Author_Web_Page_URL to fill the Author_Name. For Confessions Data, since it is built handcrafted, it might have some flaws. For instance, the Content attribute is having data that is link to other pages which is “Continue Reading” which is similar to “See More” link as mentioned above. However, this problem could not be resolved as “See More” link could be solved by clicking and expanding it to read the whole information but for “Continue Reading”, it could not be expanded but redirect the user to other page. However, this does not affect the analysis of the data.

3.2.3 Data Preparation

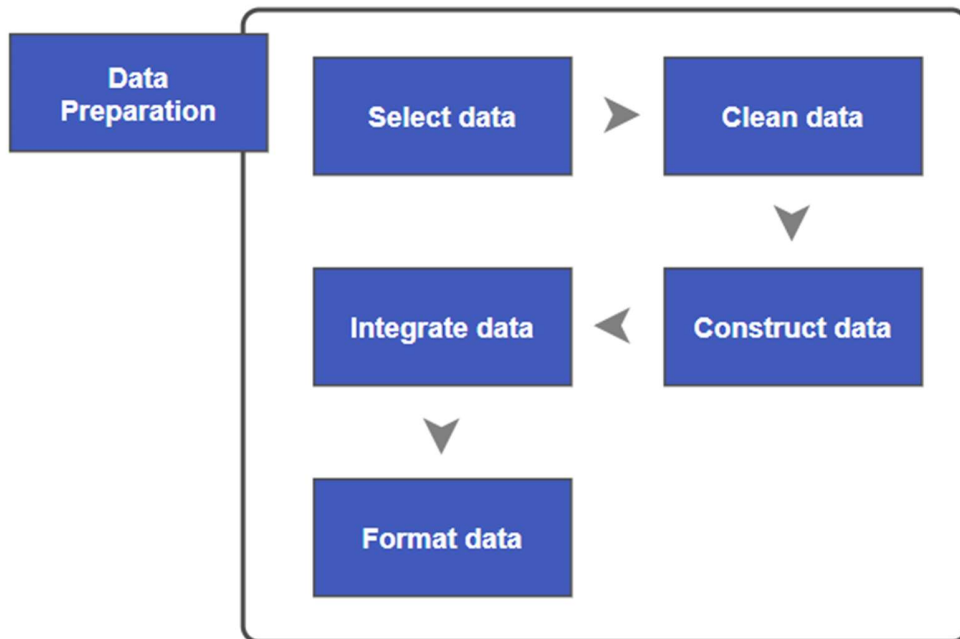


Figure 3.7: Data Preparation.

Select data

The attribute selected in Twitter data to be used for analysis are just 3 attributes out of 14 attributes which are Author_Name, Tweet_Timestamp and Tweet_Content whereas in Confessions data, all attributes are selected which are Name, Date and Content. The reasons these 3 attributes are included because at the later stage, Tweet_Content and Content attributes are needed to perform sentiment analysis and text classification whereas the other two attributes are helped in generating visualization to be displayed on the dashboard. The other attributes from Twitter data are not selected because those attributes are not used in both sentiment analysis and text classification. Hence, those attributes could be used when this dashboard is being further developed in future work. However, for the next phase, which is data cleaning phase, some attributes are needed to be used to assist in data cleaning. So, attributes will be removed once data cleaning is done.

Clean data

In this project, Python in Jupyter Notebook are used to perform data cleaning. The verified data quality problems from Twitter data mentioned in Section 3.2.2 which the Tweet_Content is null, the whole record should be removed while if the Author_Name is null with Tweet_Content is not null, the Author_Name could be identified through Author_Web_Page_URL to fill the Author_Name as shown in Figure 3.8. To be specific, the Author_Name of @msiauni_kd and @INCEIF are referring to the university or institution named Malaysia University of Science and Technology and The Global University of Islamic Finance. In addition, Tweet_Timestamp is added with the value of 28800000 because if Tweet_Timestamp is converted into datetime directly without adding the value, it is converted into Universal Time Coordinated (UTC)'s time zone. Hence, the value added to change it from UTC's time zone to Malaysia's time zone. Those unselected attributes from Twitter data will also be removed before performing analysis.

```
# Remove empty rows
twitter = twitter.dropna(subset=['Tweet_Content'])
twitter = twitter.reset_index(drop=True)

# Name empty university name
for i in range(len(twitter)):
    if ("msiauni_kd" in twitter['Author_Web_Page_URL'].loc[i]):
        twitter['Author_Name'].loc[i] = "@msiauni_kd"
    if ("INCEIF" in twitter['Author_Web_Page_URL'].loc[i]):
        twitter['Author_Name'].loc[i] = "@INCEIF"

# Switch time from unix to normal datetime
twitter['Tweet_Timestamp'] = twitter['Tweet_Timestamp'] + 28800000
twitter['Tweet_Timestamp'] = pd.to_datetime(twitter['Tweet_Timestamp'],
                                             unit='ms')

# Drop irrelevant columns
twitter = twitter.drop(columns = ['Category', 'Keyword', 'Web_Page_URL',
                                  'Tweet_Website', 'Author_Web_Page_URL',
                                  'Tweet_Image_URL', 'Tweet_Video_URL',
                                  'Tweet_Number_of_Likes',
                                  'Tweet_Number_of_Retweets',
                                  'Tweet_Number_of_Reviews',
                                  'Tweet_Number_of_share'])
```

Figure 3.8: Data Cleaning Using Jupyter Notebook.

Region (time zone)	Converted timestamp 1587108388	Relative to UTC/GMT	Date in DST	Offset In seconds
Asia/Krasnoyarsk (+07)	Apr 17 2020 14:26:28	GMT +07:00		+25200
Asia/Kuala Lumpur (+08)	Apr 17 2020 15:26:28	GMT +08:00		+28800
Asia/Kuching (+08)	Apr 17 2020 15:26:28	GMT +08:00		+28800
Asia/Kuwait (+03)	Apr 17 2020 10:26:28	GMT +03:00		+10800

Figure 3.9: Part of Time Zone List.

Construct data

For Confessions data, since it is scraped from UiPath, new data will be constructed directly from UiPath. For instance, Date attribute which has values in the format of “published how long ago” as mentioned in Explore data from Section 3.2.2. Hence, all values have been modified through nested if-else flowchart using the current time deduct with the values in Date attribute in order to identify the content is published on which date instead of how long ago. For example, the regular expression of “Convert.ToDouble(System.DateTime.Now.ToString(“HH”))” is applied to obtain current time and deducted by “Convert.ToInt32(Regex.Replace(item.ToString.Trim, “\D”, “”))” which could obtain the digit from the how long ago so that could categorise the time into today’s date or yesterday’s date. Besides, Content attribute is further derived into Translated attribute by automatically translating the content from Chinese to English due to Chinese characters are harder to be used to perform modelling.

On the other hand, both Confessions data and Twitter data are then categorise into university type which is public university or private university as well as categorise according to their social media platforms to support visualisation later. In addition, both data are being used to apply sentiment analysis with the library provided by Python which is TextBlob where when the polarity value is bigger than zero, the specific row of data is positive sentiment, zero is neutral and finally smaller than zero is negative sentiment.

Integrate data

After data are being categorise with its respective university type and labelled with its respective sentiment label. The university category and sentiment label will be integrated by appending into the original dataframe.

Format data

Both Confessions data and Twitter data are reformatted to perform data cleaning again, but these are for modelling. Hence, the data cleaning here is different than the previous data cleaning phase. These data are cleaned to remove all irrelevant symbols, punctuations, HTML coding, stopwords, digits and finally standardise the wordings. These data are saved in a new attribute in order to perform modelling but will be abandoned once modelling is done.

3.2.4 Modelling

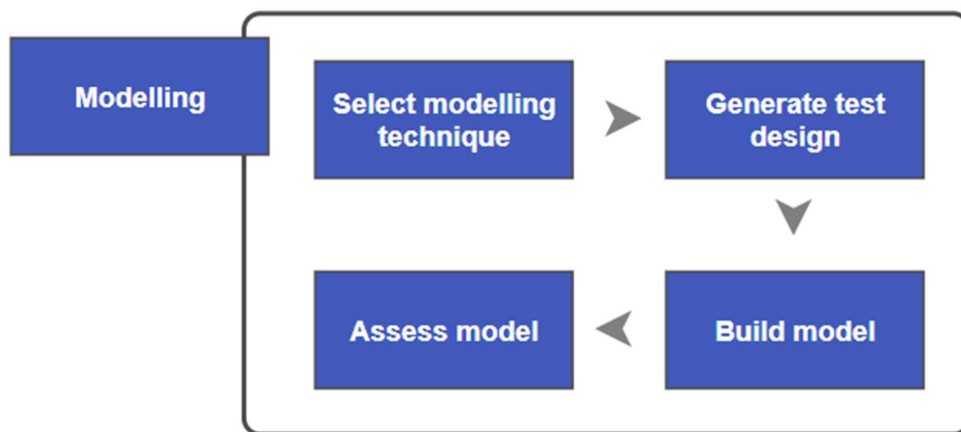


Figure 3.10: Modelling.

Select modelling technique

Multiple modeling techniques and models have been approached to perform text classification on Confessions data. There are two ways in performing text classification which are by supervised machine learning and deep learning. For machine learning, models selected are Naïve Bayes (NB), Support Vector Machine (SVM), Logistic Regression (LR), K-nearest Neighbours (KNN), Decision Tree (DT) and Random Forest (RF), SVM with SGD as well as LR with SGD. Whereas for deep learning, techniques chosen are Keras and Long Short-term Memory (LSTM). In this project, the models and techniques are then being evaluated and chosen to categorise content according to its category whether the content is related to relationship, connection, education, or story. On the other hand,

they are also evaluated and chosen to categorise content according to its statement type whether the content is a normal statement, a complaint, or a question.

Generate test design

As mentioned, there are multiple modelling techniques and models that have been approached. Hence, when using machine learning, the testing will be involved by tuning the test size and using 2 different data preprocessing methods on each model. Whereas, when using deep learning, the testing will be involved by tuning the batch size and test size using Keras and Long Short-term Memory (LSTM). Both different data preprocessing methods involve changing texts to lower case, removing stop words. However, both still have areas that are different from each other.

	Method 1	Method 2
Change Text to Lower Case	Yes	Yes
HTML Decoding	No	Yes
Remove Bad Characters/ Symbols	No	Yes
Remove Non Alpha Text	Yes	No
Remove Stop Words	Yes	Yes
Word Lemmatisation	Yes	No
Word Tokenisation	Yes	No

Table 3.1: Comparison of Data Preprocessing Methods.

Build model

All the models and techniques are then being built for both content category and content type. For both machine learning models and deep learning techniques, they are built with different test size of data which ranges from 0.05 to 0.3. However, for deep learning techniques, batch size is also being tuned. As a result, all the models are being built based on different hyperparameter to find out which model performs the best.

The difficulties encountered when building the models and techniques are the guidelines of the models and techniques proposed sometimes are being meant to perform in binary text classification and not on multiclass text classification. Hence, the performance might be affected due to those guidelines are not optimized for multiclass text classification. In addition, the in adequate of data is the main issue that resulted in poor

accuracies produced by both machine learning models and deep learning techniques. For instance, the data is being trained on 1000 labelled data which is hard for models and techniques to learn.

Assess model

The accuracies resulted from data trained on content to produce content category and content type are being shown in Table 3.2, Table 3.3, Table 3.4 as well as Table 3.5 which Table 3.2 and Table 3.3 are accuracy results trained from content category whereas Table 3.4 and Table 3.5 are accuracy results trained from content type.

Content Category		Test Size (%)			
Models		0.05	0.1	0.2	0.3
Method 1	NB	62.00	62.00	59.50	60.67
	SVM	60.00	63.00	64.00	64.67
	LR	58.00	62.00	61.50	60.33
	KNN	52.00	52.00	49.00	49.00
	DT	56.00	50.00	40.50	47.33
	RF	56.00	66.00	60.50	60.33
Method 2	NB	62.00	59.00	64.00	59.33
	SVM	72.00	67.00	66.00	66.67
	LR	70.00	69.00	66.50	65.33
	KNN	58.00	62.00	59.00	55.33
	DT	58.00	54.00	56.00	53.33
	RF	54.00	64.00	58.50	51.00
	SVM with SGD	76.00	67.00	67.00	67.33
	LR with SGD	66.00	64.00	67.00	64.67

Table 3.2: Accuracy of Content Category Based on Machine Learning Models.

Content Category	Batch Size 32				Batch Size 64			
	Test Size (%)				Test Size (%)			
Techniques	0.05	0.1	0.2	0.3	0.05	0.1	0.2	0.3
Keras	60.00	60.00	59.50	63.33	62.00	58.00	59.50	61.33
LSTM	60.00	61.00	55.50	55.67	52.00	56.00	55.50	44.67

Table 3.3: Accuracy of Content Category Based on Deep Learning Techniques.

Content Type		Test Size (%)			
Models		0.05	0.1	0.2	0.3
Method 1	NB	64.00	56.00	57.50	60.67
	SVM	60.00	53.00	56.00	58.67
	LR	64.00	56.00	58.50	59.67
	KNN	62.00	59.00	56.50	54.67
	DT	50.00	47.00	51.00	51.67
	RF	64.00	57.00	60.50	60.67
Method 2	NB	68.00	66.00	56.50	57.33
	SVM	66.00	64.00	56.00	57.00
	LR	68.00	66.00	58.50	57.67
	KNN	62.00	56.00	55.50	52.67
	DT	56.00	60.00	52.50	56.00
	RF	56.00	53.00	49.50	53.00
	SVM with SGD	68.00	64.00	54.50	57.67
	LR with SGD	68.00	64.00	54.50	57.67

Table 3.4: Accuracy of Content Type Based on Machine Learning Models.

Content Type	Batch Size 32				Batch Size 64			
	Test Size (%)				Test Size (%)			
Techniques	0.05	0.1	0.2	0.3	0.05	0.1	0.2	0.3
Keras	62.00	55.00	52.50	56.67	60.00	55.00	53.00	58.67
LSTM	66.00	63.00	57.50	52.33	50.00	57.00	50.50	52.67

Table 3.5: Accuracy of Content Type Based on Deep Learning Techniques.

As a result, SVM with SGD with the test size hyperparameter of 0.5 is chosen to predict the content category whereas LR with SGD with the test size hyperparameter of 0.5 is chosen to predict the content type as both of the models have the highest accuracies among other models and techniques which are 76% and 68% respectively.

3.2.5 Evaluation

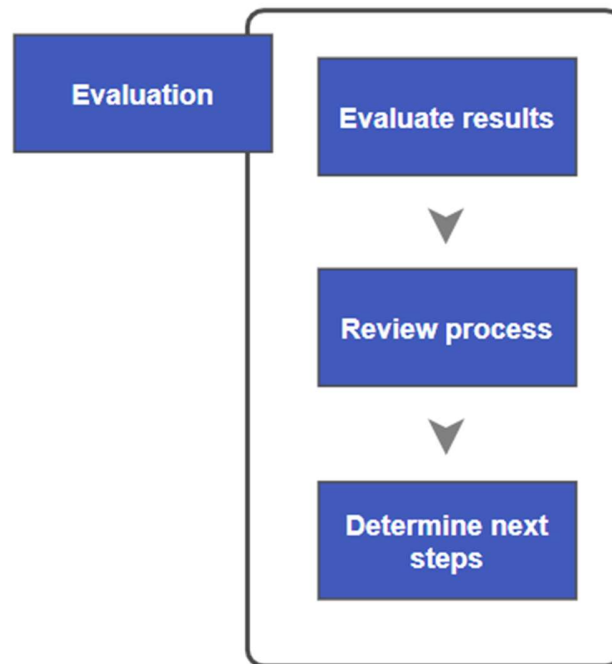


Figure 3.11: Evaluation.

Evaluate results

Based on the results from modelling stage, the correctness of predicting content category and content type of the machine learning tasks are yet to be improved but due to time limit, the results are acceptable. Two of the project objectives have met which are extracting relevant data from multiple social media platforms automatically as well as analysing social media data which is performed from the modelling stage earlier.

Review process

The process from business understanding to modelling are well performed and there is no essential factor that has been left or overlooked.

Determine next steps

All the findings should now be deployed into a dashboard. However, before deployment, all data is better to being stored in a database system as it could manages data efficiently and handles large amount of data within this software application.

3.2.6 Deployment

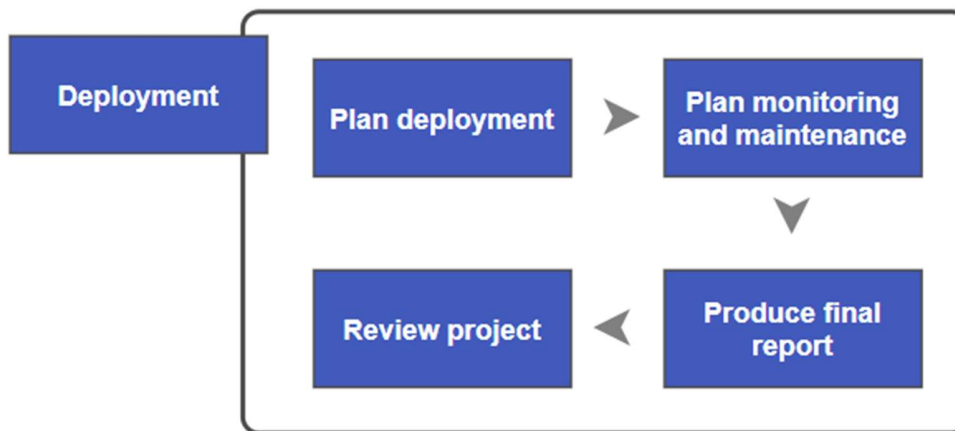


Figure 3.12: Deployment.

Plan deployment

As mentioned from previous phase, all the findings are now being imported into a centralised database management system with a scheduler using UiPath. This software will perform all steps from capturing data, performing analysis to exporting those data into the dashboard automatically. Hence, all data could now be shown in a dashboard, specifically shiny dashboard which is developed using RStudio. In the dashboard, all data are being visualised and viewed in different tabs with different settings for users to view them easily and the user interface will be further discussed in Chapter 5. However, before all data could be viewed and visualized, linkage between database management system and RStudio should be built. Lastly, the dashboard is published to the web after all visualisation and data are set up appropriately.

Plan monitoring and maintenance

The data extraction is the most critical area to be aware of as it is performed with a computer system which is recommended not to be interrupted by other tasks. For instance, a computer system is to be switched on 24/7 so that data extraction would run automatically everyday during a specific time. The computer system must be switched on because data extraction is done through screen scraping which means if any user is using the computer system to perform other tasks, the process will be interrupted and failed. In addition, if the

name of the university's account is being changed, the information from that university account will not be extracted anymore which will results in null values since then.

Product final report and Review project

As a result, a social media monitoring dashboard for university is developed and the flows from top to down is described and stated from previous stages.

For the pitfalls and difficulties encountered during the whole flow, the data extraction is the most tedious stage as the interface of Facebook might change over time. Hence, the whole flowchart in UiPath might need to be reconfigured again. However, it should be accepted as this extraction is free to use and does not require API that is hard to get developer approval to extract data legally.

CHAPTER 4 DETAILED WORKFLOW

4.1 Data Collection

Data collection of Twitter data could be done through 3 different methods. Two of the methods are by using Octoparse. It started with creating a workflow to build the data extraction process step by step from navigating to the page to extracting the data through screen scraping in order to get tailored data. On the other hand, another method is started with using a task template to obtain the data which is an easier approach because the data extraction will be quicker, and it is not scraped through screen scraping but cloud scraping. The last method is by using UiPath which also started with creating a workflow similar to Octoparse which it requires people to build the data extraction process step by step too.

In this project, the method used is using a task template through Octoparse. It is because the first method which is by creating a workflow is heavily dependent on the user interface of the social media platform, Twitter. Hence, if Twitter updated its user interface, the workflow is needed to be reconfigured again due to the data could not be detected and extracted anymore through screen scraping. Although customer support is available, they could not solve the problem of zero data being extracted, the customer support recommended task templates instead which are prepersonalised templates which allow extraction of data to be done easily without any configuration of algorithm. In addition, the user interface of Octoparse has flaw which extraction of data is limited to a very small amount of data, so the scraper needs to be run more frequently so that no data will be missed in data collection. On the other hand, the last method which is creating a workflow through UiPath might has better result but due to time constraint, collection of data through UiPath will be done with Confessions data only.

4.1.1 Twitter Data Collection.

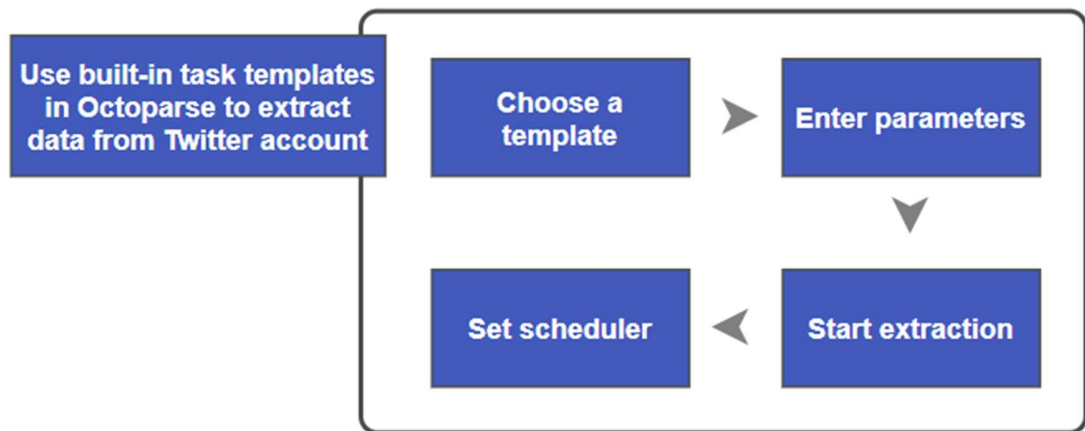


Figure 4.1: Steps to Extract Twitter Data.

Firstly, a template named “Tweets (URLs)” is chosen by navigating into the Social Media category. Then, the criteria of extraction is allowed to be set by filling in the URLs of each Tweeter as well as the total number of scroll-down times. For instance, an URL will be looked like:

“[https://twitter.com/search?q=\(from%3Astudyunimalaya%20OR%20from%3AUPSI_Malaysia\)&src=typed_query&f=live](https://twitter.com/search?q=(from%3Astudyunimalaya%20OR%20from%3AUPSI_Malaysia)&src=typed_query&f=live)”

where “%3” indicates a colon, “%20” indicates a blank space and “&f=live” indicates information to be searched by latest instead of top Tweets. The URL could be identified easily by typing the query into the search bar and the URL will be generated as shown in Figure 4.2. Hence, plenty of URLs are used as the maximum accounts to be searched in a single query is 25 accounts. Hence, at least a total number of 4 URLs are needed to extract Tweets tweeted from 20 public and 75 private university’s accounts. For instance, the accounts that are selected to obtain information from are listed in Table 4.1 and Table 4.2. After that, data extraction could be performed through cloud extraction and a XLSX file will be generated. For instance, there is two data extraction templates, so two XLSX file will be generated. Lastly, a scheduler is set so that data could be obtained automatically daily during a specific time. When data is obtained automatically, it will undergo data preparation, modelling and finally being visualised at the dashboard.

CHAPTER 4 DETAILED WORKFLOW

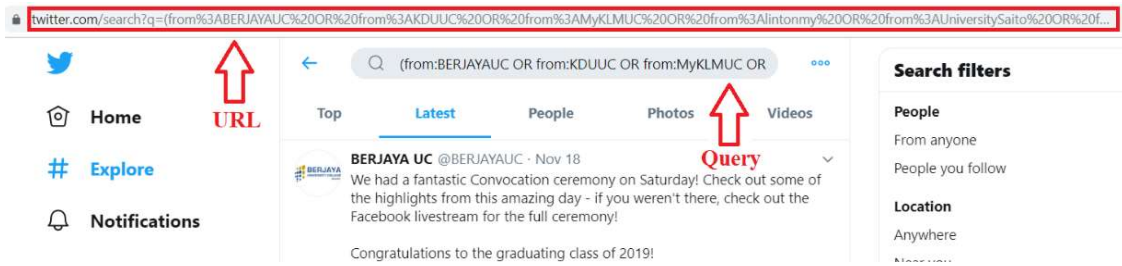


Figure 4.2: URL Is Identified Based on Query.

No.	Public University Name	Twitter Account
1.	University of Malaya	studyunimalaya
2.	Sultan Idris Education University	UPSI_Malaysia
3.	Putra University Malaysia	uputramalaysia
4.	MARA University of Technology	uitmofficial
5.	University of Science Malaysia	USMOfficial1969
6.	National University of Malaysia	ukm_my
7.	University of Technology Malaysia	utm_my
8.	Sultan Zainal Abidin University	uniswaedumy
9.	International Islamic University Malaysia	OfficialIIUM
10.	Northern University of Malaysia	uumnews
11.	University of Malaysia Sarawak	UNIMASofficial
12.	Tun Hussein Onn University of Malaysia	uthmjohor
13.	University of Malaysia Sabah	UMS_EcoCampus
14.	National Defence University of Malaysia	upnm
15.	University of Malaysia Terengganu	UMT_Official
16.	Islamic Science University of Malaysia	usim
17.	Technical University of Malaysia Malacca	MyUTeM
18.	University of Malaysia Perlis	PROUniMAP
19.	University of Malaysia Pahang	umpmalaysia
20.	University of Malaysia Kelantan	OfficialUMK

Table 4.1: Public University with Their Respective Twitter Account.

No.	Private University Name	Twitter Account
1.	AIMST University	Aimst2U
2.	Albukhary International University	OfficialAIU
3.	Al-Madinah International University	mediumalaysia
4.	Asia e University	asiaeuniversity

CHAPTER 4 DETAILED WORKFLOW

5.	Asia Metropolitan University	asiametropolit
6.	Asia Pacific University of Technology & Innovation	AsiaPacificU
7.	Berjaya University College	BERJAYAUC
8.	Binary University	Binary_U
9.	City University Malaysia	CityUniPj
10.	Curtin University	CurtinUni
11.	Curtin University Malaysia	CurtinMalaysia
12.	DRB-HICOM University of Automotive Malaysia	drbhicom_u
13.	GlobalNxt University	GlobalNxt
14.	Han Chiang University College of Communication	HanChiangColl
15.	HELP University	HELPUni
16.	Heriot-Watt University	HeriotWattUni
17.	Heriot-Watt University Malaysia	HWUMalaysia
18.	Infrastructure University Kuala Lumpur	myIUKL
19.	International Centre for Education in Islamic Finance	INCEIF
20.	International Medical University	IMUMalaysia
21.	International University of Malaya-Wales	IUMWKL
22.	INTI International University	INTI_edu
23.	Islamic University of Malaysia	UIOfficialMY
24.	Kuala Lumpur Metropolitan University College	MyKLMUC
25.	Limkokwing University of Creative Technology	Limkokwing_MY
26.	Linton University College	lintonmy
27.	MAHSA University	MAHSAedu
28.	Malaysia Institute of Supply Chain Innovation	misiedu
29.	Malaysia University of Science and Technology	msiauni_kd
30.	Management & Science University	MSUmalaysia
31.	Manipal International University	manipalmy
32.	Meritus University	mymeritus
33.	Monash University	MonashUni
34.	Monash University Malaysia Campus	MonashMalaysia
35.	Multimedia University	mmumalaysia
36.	National Energy University	uniten
37.	Newcastle University	StudentsNCL
38.	Newcastle University Medicine Malaysia	NUMedMalaysia
39.	Nilai University	NilaiUniversity
40.	Open University Malaysia	OpenUniMalaysia

CHAPTER 4 DETAILED WORKFLOW

41.	Perdana University	perdana_univ
42.	Petronas University of Technology	UTPOfficial
43.	Quest International University Perak	qiup_edu
44.	University College Dublin	ucddublin
45.	Royal College of Surgeons in Ireland	RCSI_Irl
46.	Saito University College	UniversitySaito
47.	SEGi University	SEGi_tweets
48.	Sunway University	SunwayU
49.	Swinburne University of Technology	Swinburne
50.	Swinburne University of Technology Sarawak Campus	Swinburne_Swk
51.	Taylor's University	Taylors_Uni
52.	Tun Abdul Razak University	myunirazak
53.	Twintech International University College of Technology	twintechedumy
54.	UCSI University	ucsiuniversity
55.	UNITAR International University	UNITARofficial
56.	University College of Islam Melaka	officialkuim
57.	University College of Technology Sarawak	uctsofficial
58.	University College Sabah Foundation	ucsfabahtwt
59.	University College TATI	uctati
60.	University Malaysia of Computer Science & Engineering	UniMyOfficial
61.	University of Cyberjaya	UniCyberjaya
62.	University of Kuala Lumpur	UniKLOfficial
63.	University of Nottingham	UniofNottingham
64.	University of Nottingham Malaysia Campus	UoNMalaysia
65.	University of Reading	UniofReading
66.	University of Reading Malaysia Campus	UoRMalaysia
67.	University of Selangor	UNISELOFFICIAL
68.	University of Southampton	unisouthampton
69.	University of Southampton Malaysia Campus	Southampton_MY
70.	University of Wollongong	UOW
71.	UOW Malaysia KDU	KDUUC
72.	University Tunku Abdul Rahman	UTARnet
73.	Wawasan Open University	WawasanOU
74.	Widad University College	Twt_Widad
75.	Xiamen University Malaysia	xmumalaysia

Table 4.2: Private University with Their Respective Twitter Account.

4.1.2 Confessions Data Collection.

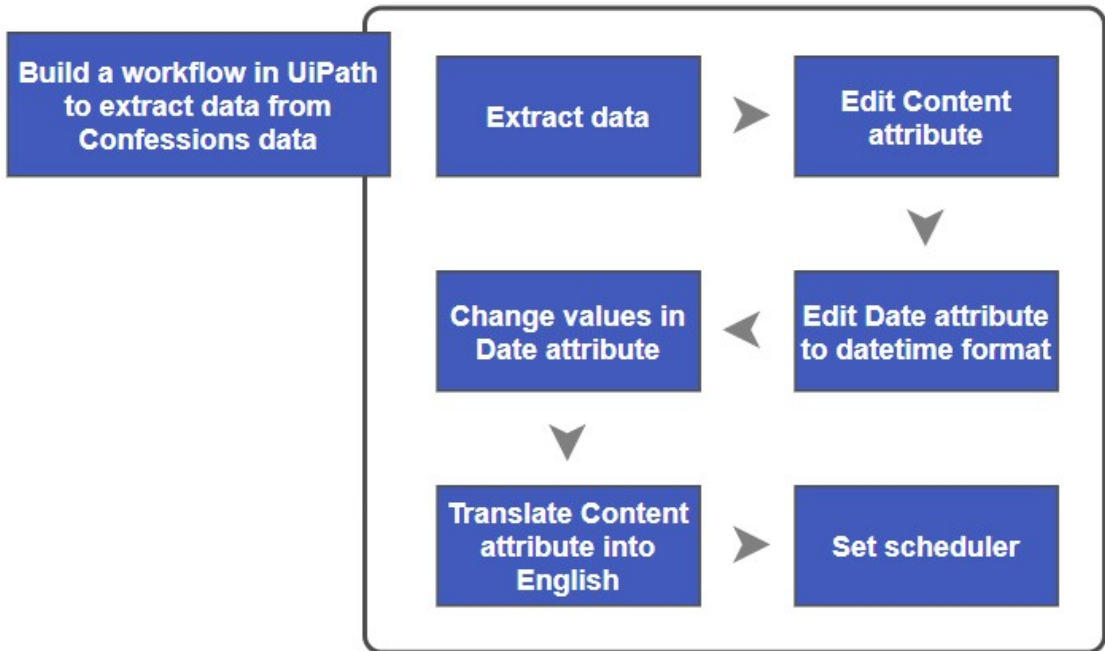


Figure 4.3: Steps to Extract Confessions Data.

Firstly, data extraction is started with opening a Facebook page that its data is wished to be extracted. It is then required to navigate to the ‘Posts’ tab as all the relevant contents will be shown there as shown in Figure 4.4. The page is then set to scroll down automatically for a specific number of times. As some content of posts are hidden in “See More” link as mentioned previously, it is set to click the “See More” link to expand the content so the content could be viewed. Then, all posts will be extracted in a XLSX file.

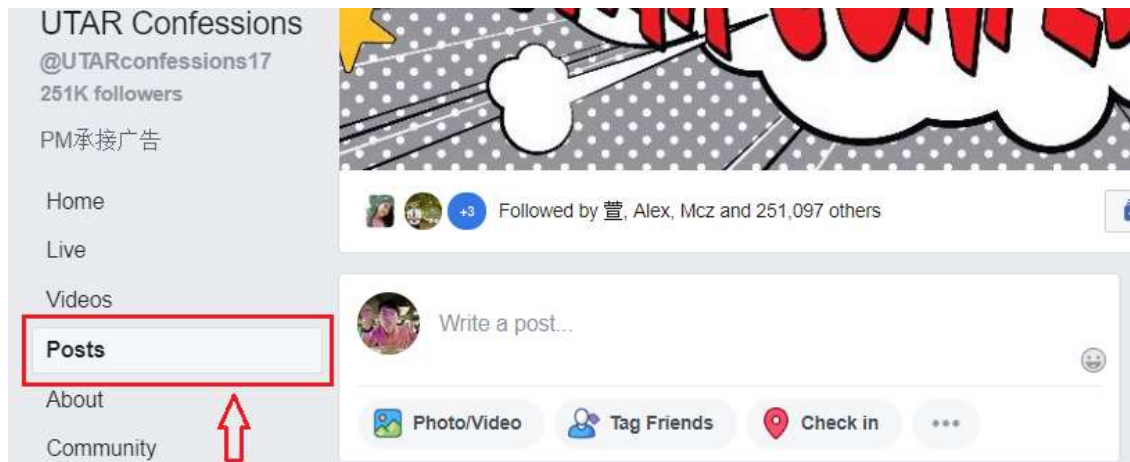


Figure 4.4: 'Post' Tab.

However, the Content attribute is having unnecessary blank spaces which will affect the viewing of users if these data are presented in the dashboard. Hence, these data is being reformatted with a regular expression of “Regex.Replace(item.ToString, “\s+|\n”, “ ”)” to replace unnecessary blank spaces to just one blank space.

Then, the category of Date attribute in XLSX file is changed from General to Short Date so the date will be projected correctly in datetime format.

Next, all values in Date attribute is reedited as they are in the format of “published how long ago” as shown in Figure 4.5. Hence, when the data contains “hr” or “min” it will be converted to today’s date or yesterday’s date by calculating it with the current time deduct with data’s value. When the data contains “Yesterday”, it will be converted to yesterday’s date and so forth.



Figure 4.5: Date Format of Data.

The Content attribute is having data in Chinese characters which is not suitable for some users to read as not all users understand Chinese. In addition, the data in Chinese characters is less suitable for modelling too. Hence, the data is being translated into English characters and saved in a new attribute which is Translated attribute.

Lastly, a scheduler is set so that data could be obtained automatically daily during a specific time. But before scheduler is set, the data is obtained automatically from each individual page rather than stacks of URL in the data collection of Twitter data. As shown in Figure 4.6, the extraction for each confessions page is done before proceeding to data preparation, modelling and finally being visualised at the dashboard with the help of scheduler. For instance, the list of confessions page that undergoes data extraction is shown in Table 4.3.

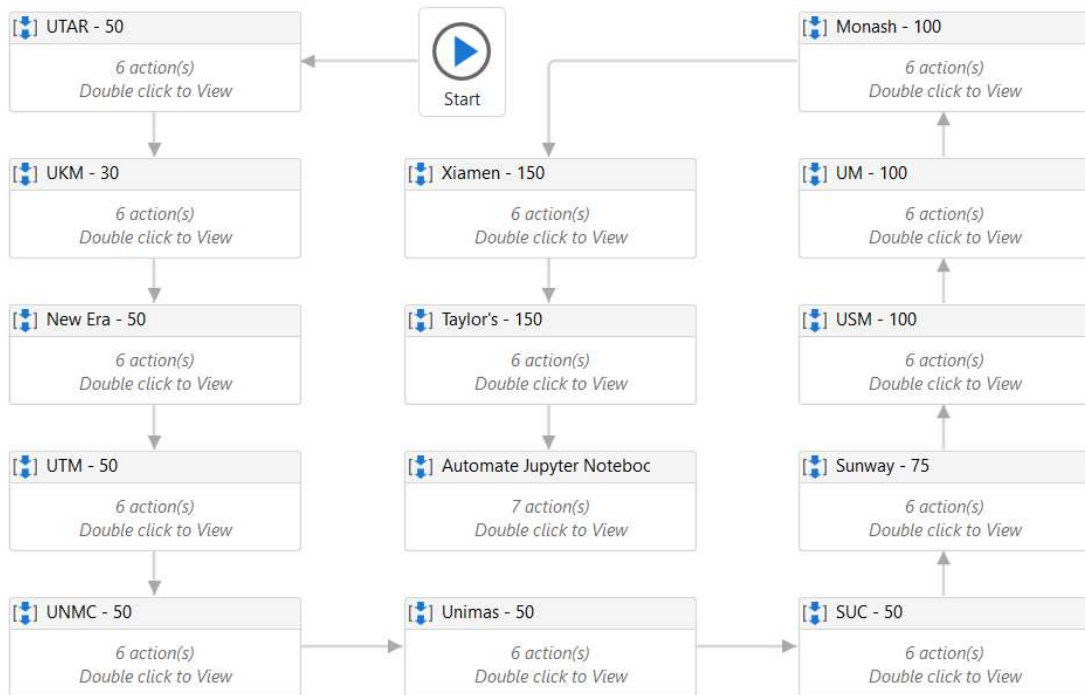


Figure 4.6: Workflow Overview of Confessions Data Extraction.

No.	University Name	Confessions Page ID
1.	University of Malaya	umconfession.universityofmalaya
2.	University of Science Malaysia	usmconfessionmain
3.	National University of Malaysia	UKMCNConfession
4.	University of Technology Malaysia	Utmcv2point0
5.	University of Malaysia Sarawak	unimasconfessionchinese
6.	Monash University Malaysia Campus	monashconfessions1
7.	New Era University College	neucconfessions2020
8.	Southern University College	confessionsSUC

9.	Sunway University	Sunway-University-Confessions-4-1189114867892695
10.	Taylor's University	taylorconfessions2019
11.	University of Nottingham Malaysia Campus	UNMCCONFESSION
12.	University Tunku Abdul Rahman	UTARconfessions17
13.	Xiamen University Malaysia	XMUMConfession

Table 4.3: University with Their Respective Confessions Page.

4.2 Challenges and Concerns

One of the challenges is the approach to collect social media data is via APIs. As mentioned earlier, although social media data is able to be accessible through APIs, but due to the business value and security concerns of the data, most of the main sources like Facebook and Google implement algorithms to avoid researchers and academics to gain full access to their raw data. However, social data access is still available to academia and researchers as some sources do provide affordable data access plans. This causes data could not be easily extracted and alternative ways have to be considered in order to extract data. In addition, the workflow in UiPath is self-developed which data is extracted through screen scraping is time consuming. For instance, 1000 records of data require at least 3 hours for the process to be completed. Lastly, the challenge in data collection is Tweets extraction via Octoparse. It is because data is extracted using task templates which requires standard plan or professional plan in Octoparse. These plans cost money, hence free trial is applied in order to extract data for visualisation. However, this issue could be eliminated if extraction of Tweets is performed via UiPath which the workflow requires days to be developed.

Besides, there is challenge in the translation of content from Chinese characters to English characters which is derived from the Content Attribute. In fact, it is better to be performed through the help of API instead of UiPath. However, as mentioned, data is difficult to be processed using it due to the API to translate the content has daily restriction on translating. Although it is recommended and better to be used as it is relatively quick and easy to build it. For instance, Google Translate API has a default limit of 2 million characters to be translated per day. Hence, UiPath is used to translate the content attribute in this project as the alternative for it.

During FYP 1, Tweets are extracted are not solely on Tweets tweeted by university's accounts, but Tweets are also extracted from keywords. Hence, Tweets tweeted by normal user like us will be scraped. However, this type of extraction has been removed in this project due to the consideration of ethical issues. Web scraping is not an illegal activity, but there is rules in scraping data which publicly shared content is allowed to be scraped but not personal information such as Tweets and Facebook Posts from personal accounts. As a result, the scope of data just covers data of universities but not personal information.

4.3 Tools to Use

Octoparse

It is a data extraction third party software which allows users to extract information from websites easily with minimal or no coding at all. In this context, Octoparse is used to extract Tweets from Twitter using task template by providing the relevant URLs and these data could be set to be extracted daily at a particular time.

UiPath

It is a Robotic Process Automation (RPA) software which allows users to configure their computer software with regards to UiPath to simulate the actions of individual interacting automatically in digital systems to accomplish a business process. However, in this context, it is used to extract Posts from Facebook by building a workflow. Then, the process is automated by undergoing data preparation, modelling and finally parse the data into the database system which is Microsoft SQL Server Management Studio.

Jupyter Notebook

It is an open-source web application that could allow users to create and share documents. It is supported with multiple programming languages and in this context, Python language is used. Data preparation and modelling is done in Jupyter Notebook to reprocess all data by cleaning, constructing, integrating and formatting them and finally performing, sentiment analysis as well as text classification on them before it is being imported into a database system.

Microsoft SQL Server Management Studio

It is a database system to manage the data that is extracted via Octoparse and UiPath by keeping data of Tweets and Posts imported from Jupyter Notebook. Linkage is then developed between this system and RStudio to visualise data in Shiny Dashboard. Most people use spreadsheets as a kind of database, but mostly because they do not understand why they would use a database instead. One of the reasons is databases could collect huge amounts of data which could be stored with better capability for safety and security on a central repository.

RStudio

It is an integrated development environment also known as IDE for R language. Basically, it is a code editor as well as a development environment. In this context, Shiny which is an R package is also used to build interactive web apps. So, scripts are written to connect to the database and perform necessary works such as cleaning and sentiment analysis. Then, an interactive web apps is being built with Shiny as a dashboard for the user.

4.4 Timeline

During the last semester, just one project objectives is achieved which is to visualise the data on the dashboard. However, all project objectives are now achieved by completing the stages ranging from initial data collection to deployment. For instance, enhance data collection is a stage where besides from Twitter data, Facebook data will be tried to be collected although there are many restrictions. Facebook data is chosen to be scraped because most of the university students in UTAR rather to use UTAR Confessions Page on Facebook to communicate than using Twitter. Hence, it is a good approach to get involved in Facebook data to perform analysis. Then, machine learning is to be involved to train the datasets in order to perform text classification. The overall sentiment could also be performed to identify each content is pages that is having positive, negative, or neutral sentiments. Lastly, all findings are being deployed in the social media monitoring dashboard.

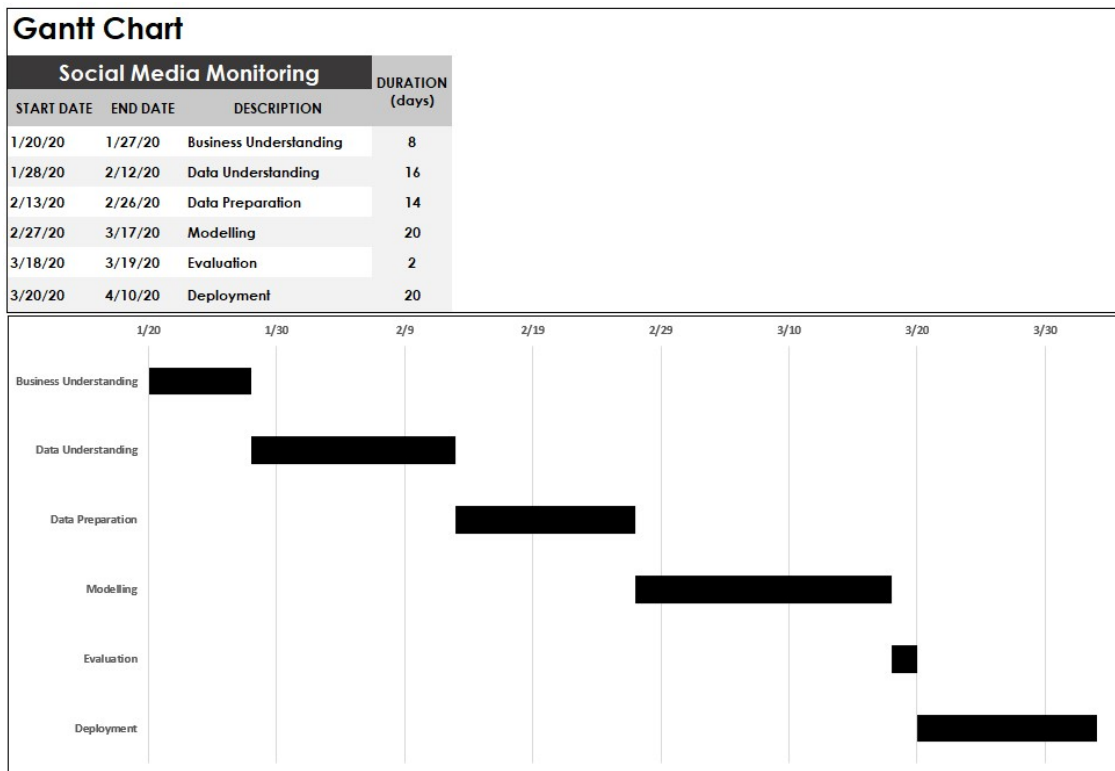


Figure 4.7: Gantt Chart.

CHAPTER 5 SYSTEM IMPLEMENTATION

5.1 Dashboard Overview

After analysis and modelling, all data is uploaded into a database management system which is then connected to a dashboard. The dashboard will retrieve the data by establishing a connection between them.

The user interface of this dashboard is divided into four tabs, “Home” tab, “Twitter Data” tab, “Facebook Data” tab and “Reputation Analysis” tab as shown in Figure 5.1. The “Home” tab is to display overall information that integrated all universities information across both Twitter and Facebook platforms. For “Twitter Data” and “Facebook Data” tab, both will show raw data which is processed after performing data preparation stage and modelling stage. Lastly, “Reputation Analysis” tab will show information about each individual university.

For the data distribution in each tab, “Home” tab and “Reputation Analysis” tab will include data from Twitter and Confessions. Whereas, “Twitter Data” tab will just contain data from Twitter while “Facebook Data” tab will be data from Confessions.

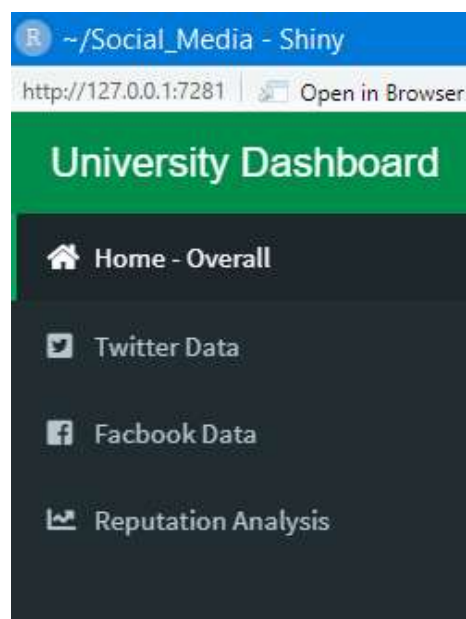


Figure 5.1: Tabs of The Dashboard.

5.2 Dashboard Features

5.2.1 Features in “Home” Tab

The data displayed in the dashboard as shown in Figure 5.2 is very flexible and customisable as all the data could be filtered according to the criteria which users like. The filter options are interactive as the changes in those options will affect the three visualisation which started with bottom left corner, sentiments based on all the Tweets and posts from Twitter and Confessions, respectively. The map in top right corner is locations of all universities in Malaysia and by zooming out, overseas universities that have branches or campuses in Malaysia will also be included in the map. Lastly, bottom right corner is total posts and Tweets which posted and tweeted within a date range which could shows that how active a university is.

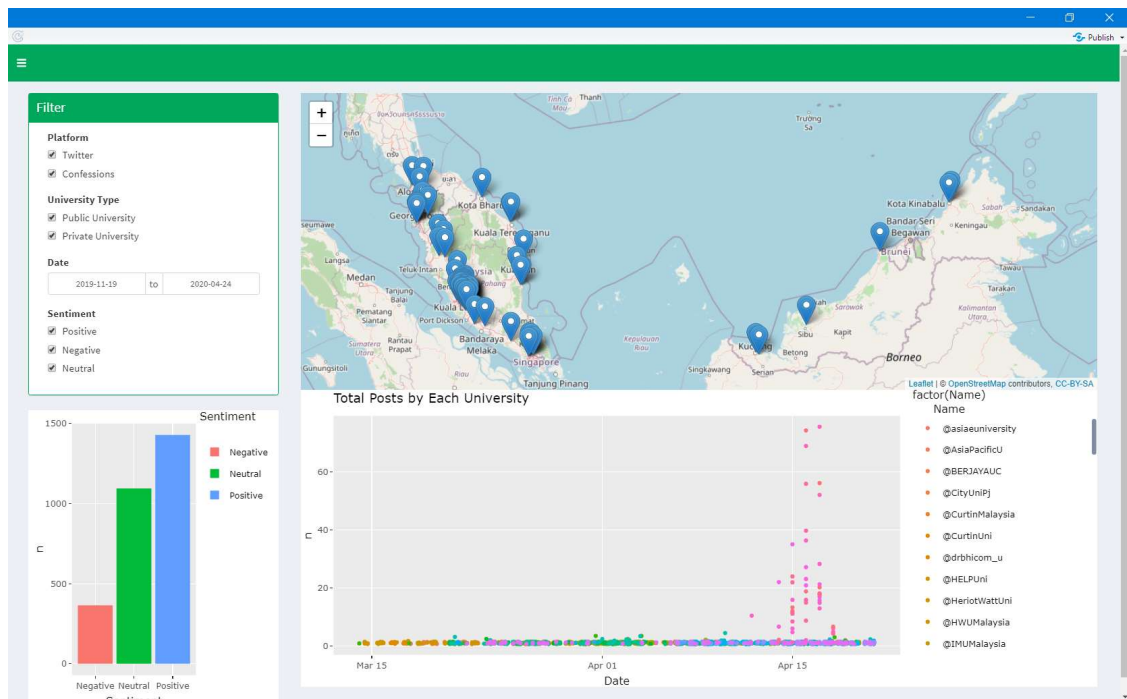


Figure 5.2: Information and Visualisation in "Home" Tab.

5.2.2 Features in “Twitter/ Facebook Data” Tab

For the filter options in “Twitter/ Facebook Data” tab are pretty similar. Both could filter the data by university type, name of the university, date, sentiment. However, “Facebook Data” tab could further filter content category and content type based on the content posted by anonymous person. The top part will be the filter options available where bottom part will be all the raw data presented.

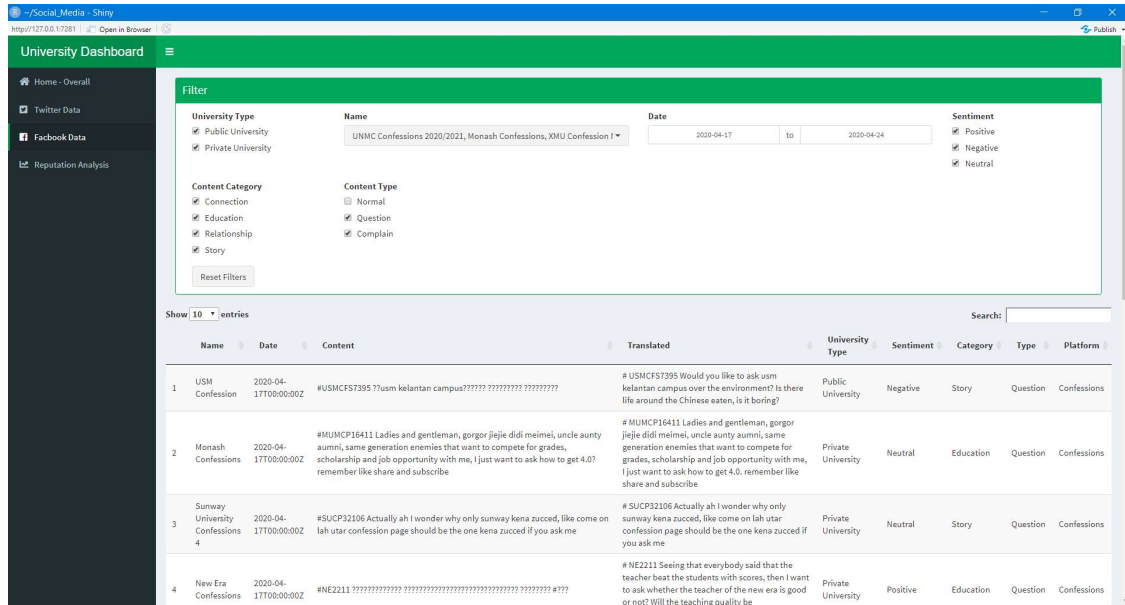


Figure 5.3: "Facebook Data" Tab.

CHAPTER 6 CONCLUSION

6.1 Project Review and Contributions

As social media is now a thing for businesses to draw useful insights out of it, in this context is the business refers to education institutions. It is because the motivation which is education sectors do pay a lot of attention in social media as it will affects the ranking and the enrolment rate for the universities. Hence it is essential for the management of universities in precise to know its branding and reputation. So, by having this project, all the information could now be controlled in the management's fingertips as all those problems which were stated earlier at problem statement are solved. The tedious way of collecting and viewing the data over consumer reports are now gone with the unified display dashboard. Hence, all of the view or opinion will be harder to miss as all of the data is not being stored in database and with the huge amount of data presents, conclusions and insights could be drawn easier based on sentiments of the data to perform business tactics in order to improve its universities as the competition between each university is very high.

In a nutshell, the process is done by scraping data through and accounts using third party software and the data is being imported into a database management system. Then, it is being displayed in the dashboard with additional information which is derived from the data which fully satisfies the general work procedures which are capture, understand, and present.

6.2 Novelties

The novelty in this project is mainly in the initial data collection phase from data understanding stage. The usual method will be scraping data through APIs or third party software, but in this project, the data is extracted through screen scraping using UiPath. Some of the third-party software like Octoparse does not allow extraction from Facebook unless users pay for subscription plan. Yet, the workflow could be built on Octoparse, but the software will crash due to much bugs in this software. On the other hand, Facebook and some other big companies have blocked users like us to scrape data through APIs due to the concern of privacy of personal information. UiPath in this case is a free software with stable version which does not has any restriction on scraping Facebook data. However, workflow must be built in order to scrape.

6.3 Future Work

There are plenty of improvements could be done which is in initial data collection from data understanding stage. The workflow could be further optimised so that it would take less than 3 hours to scrape 1000 records of data. In addition, by learning to use more complex library in RStudio, it could allow better visualisation and more interactive user interface to be displayed to users. Animations and better representations or interactive of diagrams would be best as they could capture the attention of users.

In the modelling stage, just 1000 records are labelled and trained to perform text classification which is totally inadequate to have high accuracy of labels to be predicted. Hence, more records should be labelled and trained to improve the accuracy of labels to be predicted. For instance, most of the guidelines available on web perform predictions on 20000 records of data.

BIBLIOGRAPHY

BIBLIOGRAPHY

- Boor, T. d. & Grunwald, P., 2011. *ECAR National Study of Undergraduate Students and Information Technology*, s.l.: ECAR.
- Brogan, C., 2007. *Marketing is NOT Social Media-Social Media is NOT Marketing*. [Online]
Available at: <http://chrisbrogan.com/marketing-is-not-social-media-social-media-is-not-marketing/>
[Accessed 1 August 2019].
- Duggan, M. & Brenner, J., 2013. *The Demographics of Social Media Users — 2012*, Washington, D.C.: Pew Research Center.
- Forkosh-Baruch, A. & Hershkovitz, A., 2012. A case study of Israeli higher-education institutes sharing scholarly information with the community via social networks. *The Internet and Higher Education*, 15(1), pp. 58-68.
- Kate, C. & Megan, F., 2015. The limits of crisis data: analytical and ethical challenges of using social and mobile data to understand disasters. *GeoJournal*, 80(4), pp. 491-502.
- Kemp, S., 2018. *Digital in 2018: World's internet users pass the 4 billion mark*, England and Wales: We Are Social.
- Octoparse, 2019. *Best Data Scraping Tools for 2019 (Top 10 Reviews)*. [Online]
Available at: <https://medium.com/@octoparsejerry/best-data-scraping-tools-for-2019-top-10-reviews-4a686061a184>
[Accessed 1 August 2019].
- Perdue, D. J., 2010. *Social Media Marketing: Gaining a Competitive Advantage by Reaching the Masses*, Lynchburg: Liberty University.
- The Economist, 2006. Listening to the internet. *Technology Quarterly*, 11 March.
- VanBoskirk, S., 2009. *Interactive Marketing Nears \$55 Billion; Advertising Overall Declines*, Cambridge: Forrester.

APPENDIX A FINAL YEAR PROJECT BIWEEKLY REPORT

FINAL YEAR PROJECT BIWEEKLY REPORT

(Project II)

Trimester, Year: Y3S3	Study week no.: 2
Student Name & ID: Dicken Tan 16ACB05731	
Supervisor: Dr. Pradeep Isawasan	
Project Title: Social Media Monitoring Dashboard for University	

1. WORK DONE

Initial data collection is done by scraping data from Facebook.

2. WORK TO BE DONE

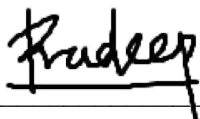
Enhancement on the workflow so that Facebook data could be collected smoothly.

3. PROBLEMS ENCOUNTERED

The workflow on scraping Facebook data is not well performed. Hence, enhancement on the workflow is needed.

4. SELF EVALUATION OF THE PROGRESS

Facebook data which is hard to be scraped is finally able to be collected.



Supervisor's signature



Student's signature

FINAL YEAR PROJECT BIWEEKLY REPORT

(Project II)

Trimester, Year: Y3S3	Study week no.: 4
Student Name & ID: Dicken Tan 16ACB05731	
Supervisor: Dr. Pradeep Isawasan	
Project Title: Social Media Monitoring Dashboard for University	

1. WORK DONE

The workflow is built in a better way.

2. WORK TO BE DONE

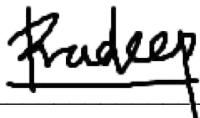
Data preparation is needed to perform data cleaning and data construction before modelling could take place.

3. PROBLEMS ENCOUNTERED

The data collected is not clean and organised to be used in modelling.

4. SELF EVALUATION OF THE PROGRESS

Although the workflow is built in a better way, improvement is still needed to reduce the extraction time.



Supervisor's signature



Student's signature

FINAL YEAR PROJECT BIWEEKLY REPORT*(Project II)*

Trimester, Year: Y3S3	Study week no.: 6
Student Name & ID: Dicken Tan 16ACB05731	
Supervisor: Dr. Pradeep Isawasan	
Project Title: Social Media Monitoring Dashboard for University	

1. WORK DONE

Data is being cleaned and constructed.

2. WORK TO BE DONE

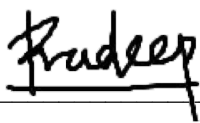
Text classification and modelling are to be done.

3. PROBLEMS ENCOUNTERED

Data is not labelled with required category and type. Hence, text classification could not be performed.

4. SELF EVALUATION OF THE PROGRESS

The data is cleaned easily with multiple data preprocessing methods guided online.



Supervisor's signature



Student's signature

FINAL YEAR PROJECT BIWEEKLY REPORT

(Project II)

Trimester, Year: Y3S3	Study week no.: 8
Student Name & ID: Dicken Tan 16ACB05731	
Supervisor: Dr. Pradeep Isawasan	
Project Title: Social Media Monitoring Dashboard for University	

1. WORK DONE

Sentiment analysis is being performed.

2. WORK TO BE DONE

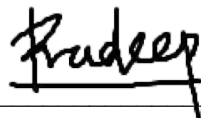
Text classification is yet to be done when adequate data is available for training.

3. PROBLEMS ENCOUNTERED

Data labelled is too less to perform text classification.

4. SELF EVALUATION OF THE PROGRESS

The correctness of data being labelled in the correct category is uncertain.



Supervisor's signature



Student's signature

FINAL YEAR PROJECT BIWEEKLY REPORT

(Project II)

Trimester, Year: Y3S3	Study week no.: 10
Student Name & ID: Dicken Tan 16ACB05731	
Supervisor: Dr. Pradeep Isawasan	
Project Title: Social Media Monitoring Dashboard for University	

1. WORK DONE

Text Classification is done by choosing the best performed model among the others.

2. WORK TO BE DONE

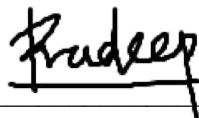
Connection of data to database is required to visualise data in the dashboard.

3. PROBLEMS ENCOUNTERED

The accuracy of text classification is a bit low due to inadequate amount of data being used to train.

4. SELF EVALUATION OF THE PROGRESS

The work of labelling is tedious as the content of the data is very long.



Supervisor's signature



Student's signature

FINAL YEAR PROJECT BIWEEKLY REPORT

(Project II)

Trimester, Year: Y3S3	Study week no.: 12
Student Name & ID: Dicken Tan 16ACB05731	
Supervisor: Dr. Pradeep Isawasan	
Project Title: Social Media Monitoring Dashboard for University	

1. WORK DONE

Connection between the database system and dashboard is established.

2. WORK TO BE DONE

All data is yet to be deployed and visualised in the dashboard.

3. PROBLEMS ENCOUNTERED

Tedious coding is required to display a well-organised and interactive dashboard to users.

4. SELF EVALUATION OF THE PROGRESS

The dashboard is able to display and visualise most of the data effectively.



Supervisor's signature



Student's signature

APPENDIX B POSTER

UTAR
UNIVERSITI TUNKU ABDUL RAHMAN

Universiti Tunku Abdul Rahman
Faculty of Information and Communication Technology

Manual Report

- Time consuming?
- Huge amount of data?
- Opinion or view is missed out?

Why not utilise?

Social Media Monitoring Dashboard for University

Just capture, understand visualise data at a glance on dashboard

APPENDIX C PLAGIARISM CHECK RESULT

Social Media Monitoring Dashboard for University

ORIGINALITY REPORT

2%	0%	1%	2%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Bogdan Batrinca, Philip C. Treleaven. "Social media analytics: a survey of techniques, tools and platforms", AI & SOCIETY, 2014 Publication	1%
2	Submitted to University of Ulster Student Paper	<1%
3	Submitted to Universiti Teknologi Malaysia Student Paper	<1%
4	Submitted to Modern College of Business and Science Student Paper	<1%
5	irep.ntu.ac.uk Internet Source	<1%
6	Submitted to Higher Education Commission Pakistan Student Paper	<1%
7	towardsdatascience.com Internet Source	<1%

APPENDIX C

Document Viewer

Turnitin Originality Report

Processed on: 24-Apr-2020 09:43 +08
ID: 1306062531
Word Count: 11683
Submitted: 2

Similarity Index	Similarity by Source
2%	Internet Sources: 0% Publications: 1% Student Papers: 2%

Social Media Monitoring Dashboard for Univers... By Dicken Tan

exclude quoted	exclude bibliography	exclude small matches	mode: <input type="text" value="quickview (classic) report"/>	Change mode	print	download
1% match (publications) Bogdan Batrinca, Phillip C. Treleaven. "Social media analytics: a survey of techniques, tools and platforms". AI & SOCIETY, 2014						
<1% match (student papers from 02-Dec-2019) Submitted to Universiti Teknologi Malaysia on 2019-12-02						
<1% match (student papers from 15-Apr-2020) Submitted to Modern College of Business and Science on 2020-04-15						
<1% match () http://irep.ntu.ac.uk						
<1% match (student papers from 02-May-2018) Submitted to University of Ulster on 2018-05-02						
<1% match (Internet from 21-Jan-2020) https://towardsdatascience.com/end-to-end-python-framework-for-predictive-modeling-b8052bb96a78?gclid=632b41016522						
<1% match (student papers from 27-Dec-2016) Submitted to International Hellenic University on 2016-12-27						
<1% match (student papers from 20-Jul-2018) Submitted to Higher Education Commission Pakistan on 2018-07-20						
<1% match (student papers from 20-Feb-2019) Submitted to Upper Iowa University on 2019-02-20						
<1% match (Internet from 01-Jul-2015) http://www.epa.wa.gov.au						

APPENDIX C

Universiti Tunku Abdul Rahman			
Form Title : Supervisor's Comments on Originality Report Generated by Turnitin for Submission of Final Year Project Report (for Undergraduate Programmes)			
Form Number: FM-IAD-005	Rev No.: 0	Effective Date: 01/10/2013	Page No.: 1 of 1



FACULTY OF INFORMATION AND COMMUNICATION TECHNOLOGY

Full Name(s) of Candidate(s)	DICKEN TAN
ID Number(s)	16ACB05731
Programme / Course	BCS (HONS) COMPUTER SCIENCE
Title of Final Year Project	SOCIAL MEDIA MONITORING DASHBOARD FOR UNIVERSITY

Similarity	Supervisor's Comments (Compulsory if parameters of originality exceeds the limits approved by UTAR)
Overall similarity index: 2 % Similarity by source Internet Sources: 0 % Publications: 1 % Student Papers: 2 %	
Number of individual sources listed of more than 3% similarity: 0	
Parameters of originality required and limits approved by UTAR are as Follows: (i) Overall similarity index is 20% and below, and (ii) Matching of individual sources listed must be less than 3% each, and (iii) Matching texts in continuous block must not exceed 8 words <i>Note: Parameters (i) – (ii) shall exclude quotes, bibliography and text matches which are less than 8 words.</i>	

Note: Supervisor/Candidate(s) is/are required to provide softcopy of full set of the originality report to Faculty/Institute

Based on the above results, I hereby declare that I am satisfied with the originality of the Final Year Project Report submitted by my student(s) as named above.

Signature of Supervisor

Name: DR. PRADEEP ISAWASAN

Date: 24 APRIL 2020

Signature of Co-Supervisor

Name: _____

Date: _____

APPENDIX D FYP2 CHECKLIST




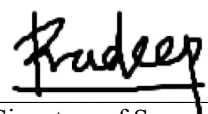
UNIVERSITI TUNKU ABDUL RAHMAN
FACULTY OF INFORMATION & COMMUNICATION
TECHNOLOGY (KAMPAR CAMPUS)

CHECKLIST FOR FYP2 THESIS SUBMISSION

Student Id	16ACB05731
Student Name	DICKEN TAN
Supervisor Name	DR. PRADEEP ISAWASAN

TICK (√)	DOCUMENT ITEMS
	Your report must include all the items below. Put a tick on the left column after you have checked your report with respect to the corresponding item.
√	Front Cover
√	Signed Report Status Declaration Form
√	Title Page
√	Signed form of the Declaration of Originality
√	Acknowledgement
√	Abstract
√	Table of Contents
√	List of Figures (if applicable)
√	List of Tables (if applicable)
√	List of Symbols (if applicable)
√	List of Abbreviations (if applicable)
√	Chapters / Content
√	Bibliography (or References)
√	All references in bibliography are cited in the thesis, especially in the chapter of literature review
√	Appendices (if applicable)
√	Poster
√	Signed Turnitin Report (Plagiarism Check Result - Form Number: FM-IAD-005)

*Include this form (checklist) in the thesis (Bind together as the last page)

<p>I, the author, have checked and confirmed all the items listed in the table are included in my report.</p> <p style="text-align: center;"></p> <p>(Signature of Student) Date: 24 APRIL 2020</p>	<p>Supervisor verification. Report with incorrect format can get 5 mark (1 grade) reduction.</p> <p style="text-align: center;"></p> <p>(Signature of Supervisor) Date: 24 APRIL 2020</p>
--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------