

A MEDIA MONITORING DASHBOARD FOR UNIVERSITY

By

Su Jia Sen

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

REPORT STATUS DECLARATION FORM

Title: A MEDIA MONITORING DASHBOARD FOR UNIVERSITY

Academic Session: JAN 2020

I

SU JIA SEN

(CAPITAL LETTER)

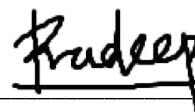
declare that I allow this Final Year Project Report to be kept in
Universiti Tunku Abdul Rahman Library subject to the regulations as follows:

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

Verified by,



(Author's signature)



(Supervisor's signature)

Address:

15, Lorong Dahlia 4,
Taman Bistari,
12300 Butterworth, Penang.

Dr. Pradeep Isawasan
Supervisor's name

Date: 24 April 2020

Date: 24 April 2020

A MEDIA MONITORING DASHBOARD FOR UNIVERSITY

By

Su Jia Sen

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 “**A 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 : SU JIA SEN

Date : 24 April 2020

ACKNOWLEDGEMENTS

I would like to express my sincere thanks and appreciation to my supervisor, Dr. Pradeep Isawasan, who has given me this bright opportunity to engage in a media monitoring dashboard design project. It is my first step to establish a career in data analyst field. A million thanks to you.

Finally, I must say thanks to my parents, family and friends for their love, support and continuous encouragement throughout the course.

ABSTRACT

A news media monitoring dashboard benefits university on handling brand reputation by listening to public's opinion. The automated process of extracting and displaying news article on data dashboard eliminates manually searching for news article to do self update daily. A simple and holistic media monitoring dashboard should become a trend for local universities on taking advantage to compete in the industry. CRISP-DM is the main guide to implement the system. Keywords have to be defined and fine-tuned to extract the most accurate news articles related to local higher education that is useful for target users. News API request for news media sources and returns JSON metadata with details of the updated news. All information collected will be analysed and presented with various visualisation that is useful and simple to understand. Short and precised information of news articles, and the analysis of the data extracted will be display to users as the final product developed by one of the R packages, ShinyDashboard. The main idea is to include artificial intelligence model to accurately categorise data collected, to provide better visualisation in the form of data table and charts.

TABLE OF CONTENTS

TITLE PAGE	i
DECLARATION OF ORIGINALITY	ii
ACKNOWLEDGEMENTS	iii
ABSTRACT	iv
TABLE OF CONTENTS	v
LIST OF FIGURES	viii
LIST OF ABBREVIATIONS	x
Chapter 1 Introduction	1
1.1 Problem Statement and Motivation	1
1.2 Background and Motivation	1
1.3 Objectives	2
1.4 Proposed Approach/Study	3
1.5 Highlight of What Have Been Achieved	4
1.6 Report organization	5
Chapter 2 Literature Review	6
2.1 News Media Monitoring	6
2.2 Data Dashboard	7
2.3 Existing Dashboard	8
2.3.1 Meltwater	8
2.3.2 LexisNexis Newsdesk	13
2.3.3 Mention for Education (used by University of Chester)	16
Chapter 3 System Design	18
3.1 Business Understanding	18
3.2 Data Understanding	19
3.3 Data Preparation	20
3.4 Modelling	23

3.4.1 Filter Model	24
3.4.2 Categorise Model	25
3.5 Evaluation	26
3.6 Deployment	27
3.6.1 Automation	27
3.6.2 Dashboard	28
Chapter 4 Methodology and Tools	29
4.1 Methodology	29
4.1.1 Web Scraping	30
4.2 Tools to Use	31
Chapter 5 System Implementation and Testing	33
5.1 Modelling	33
5.1.1 Filtration	33
5.1.1.1 Machine Learning Model	33
5.1.1.2 Deep Learning Model	33
5.1.2 Categorisation	35
5.1.2.1 Machine Learning Model	35
5.1.2.2 Deep Learning Model	35
5.2 Dashboard	36
5.3 Implementation Issues and Challenges	39
Chapter 6 Conclusion	40
6.1 Project Review	40
6.2 Novelties	41
6.3 Future Work	41
BIBLIOGRAPHY	42
APPENDIX A: Poster	A-1
APPENDIX B: Final Year Project Biweekly Report	B-1

LIST OF FIGURES

Figure Number	Title	Page
Figure 1.1	Overview Diagram of Propose System	3
Figure 2.1	Source of News Consumption by Malaysian from 2017 to 2019.	6
Figure 2.2	Domain name and Trademark of Meltwater.	9
Figure 2.3	Dashboard Features and Widgets.	10
Figure 2.4	Dashboard.	10
Figure 2.5	Search Engine.	11
Figure 2.6	List of Influencers.	12
Figure 2.7	Domain Name and Trademark of LexisNexis.	13
Figure 2.8	Analysis Tab.	14
Figure 2.9	Dashboard Tab.	14
Figure 2.10	New Newspaper / Alert to be Shared.	15
Figure 2.11	Domain Name and Trademark of Mention.	16
Figure 2.12	Extracted News Feeds.	17
Figure 3.1	Overview Flow.	18
Figure 3.2	News Scraping Flow.	19
Figure 3.3	Dataset Preparation.	20
Figure 3.4	Stored Data Attributes.	21
Figure 3.5	Filter Dataset Sample.	21
Figure 3.6	Categorise Dataset Sample.	22
Figure 3.7	Flow of Modelling.	23
Figure 3.8	Network Architecture of Filter Model.	24
Figure 3.9	Network Architecture for Categorise Model.	25
Figure 3.10	Automated Process.	27
Figure 3.11	Dashboard Formation Flow.	28
Figure 4.1	Phases CRISP-DM Process Model for Data Mining.	29
Figure 5.1	Accuracy Score of Filter Machine Learning Model.	33
Figure 5.2	Training Verbose and Accuracy of Filter Deep Learning Model.	33
Figure 5.3	Accuracy Score of Categorise Machine Learning Model.	35

Figure 5.4	Training Verbose and Accuracy of Categorise Deep Learning Model.	35
Figure 5.5	Home Page of Dashboard.	36
Figure 5.6	Side Bar Filter.	36
Figure 5.7	Sorting and Search.	37
Figure 5.8	News Information.	37
Figure 5.9	Reputation Analysis.	38

LIST OF ABBREVIATIONS

API	Application Program Interface
CRISP-DM	Cross-Industry Standard Process for Data Mining
CSS	Cascading Style Sheets
CSV	Comma Seperated Values
E.g.	for example
Etc.	Et cetera
FYP	Final Year Project
HTTP	Hypertext Transfer Protocol
IDE	Integrated Development Environment
IT	Information Technology
JSON	JavaScript Object Notation
KPI	Key Performance Indicator
LR	Logistic Regression
NB	Naïve Bayes Classifier
PDF	Portable Document Format
PR	Pubic Relation
PY	Python
RF	Random Forest Classifier
ROI	Return on Investment
SOV	Share of Voice
SVM	Support Vector Machine
TOS	Terms of Service
URL	Uniform Resource Locator
XML	Extensible Markup Language

Chapter 1 Introduction

1.1 Problem Statement and Motivation

The problem domain for this project is about Online Media Monitoring for local higher education which includes universities and colleges. In fact, marketing department of local higher institution practice on collecting articles from media manually to update themselves daily on trending issues about themselves and related to education industry. Most of them tend to spend excessive amount of time searching and collecting articles from different media. This would be due to the costly factor of existing monitoring software, or couldn't discover any convenient media monitoring tools that will benefits their monitoring practice.

1.2 Background and Motivation

Media monitoring is a practice involving reading and observing on a particular interested editorial content of media sources constantly. Media monitoring is traditionally introduced to capture editorial content, while nowadays is very useful on tracking the publication of commercials and social media. As the mediums in media is growing to more digitised in this developed society, media monitoring has adapted to the changes, which initially familiarized on media such as online newspaper, to the monitoring on different digital and social media such as Facebook and Twitter.

Media Monitoring captures varieties of information, including the location where the adverts posted, the timeline it was posted, the hard data and the number count of mention. By assembling all the information that is possible to get through online media, analysis can be proceed to grub new ideas for improvement of a particular brand and industry. Media monitoring is a mainly useful for industries and companies that have its target media audience. This can be a key tool for marketing and public relation industries, and also for a particular company that wish to monitor and observe the industry on the whole. The few examples of the purpose on implement media monitoring are to treasure the information of competitors and some hot issues that are relevant to the industry, track the performance of its competitors, discover new industry and business opportunities and also to manage a company reputation among media and societies.

Local higher education is a kind of service giving industry that continuously servicing different batches of adult people that wish to seek for knowledge. Opinion from the publics and the ways various media spreading information related to owns brand is important to be competitive in this industry. Without media monitoring dashboard and system, one might miss out certain news articles accidentally. Especially after a conducted campaign, marketing department will have a hard time on learning and collecting the responses from the public. Without the engagement, an institution can barely improve themselves on planning the next event to publicize their brand.

Media monitoring is helpful on lightening the workload of marketing department. It is important for a higher education institution to retrieve the opinions from the masses. Without feedback from public, it is hard for an industry and a brand to improve towards the expectation of their audiences. In contrast, higher education institutions can plan some strategies in advance on attracting more customers and grab more opportunities that can enhance its reputation and university ranking. In additional, media monitoring is a simple way to observe the latest developments in an industry by collecting the topics from competitor, follow their steps or analyse their action, in the same time avoiding the mistakes that is done. Moreover, media monitoring collects and analyses data. It saves the precious time for marketing department to focus only on the important part, and to have more time on planning for solutions or future actions to improve themselves rather than using their time to seek and filter the data all the time.

1.3 Objectives

The system is proposed to solve the mentioned problem statement.

1. To automate monitoring process for users convenient.

Targeted users can simplify the process of tracking updated news articles on various online media.

- To filter and extract updated news headline and link automatically.

This system automates the process of gathering and filtering news articles related to local higher education from news media, and will update itself daily whenever internet connection are available.

- To display important news information on data dashboard.

The article headlines will be arranged and displayed on a data dashboard, with the sentiment of each articles. Users can be guided to the related article with the address link attached to the article headline.

2. To understand news publishers' and public's opinions on extracted headlines.

Sentiment analysis will be done on all related articles to obtain the sentiment of each news. The sentiment result will be displayed on the data dashboard to let users easily determine positivity and negativity of the issue. An institution can get more insights and immediately take action to resolve negative issues before they get worse, while advertise the positive articles that are mentioned on media for self-reputation increment.

1.4 Proposed Approach/Study

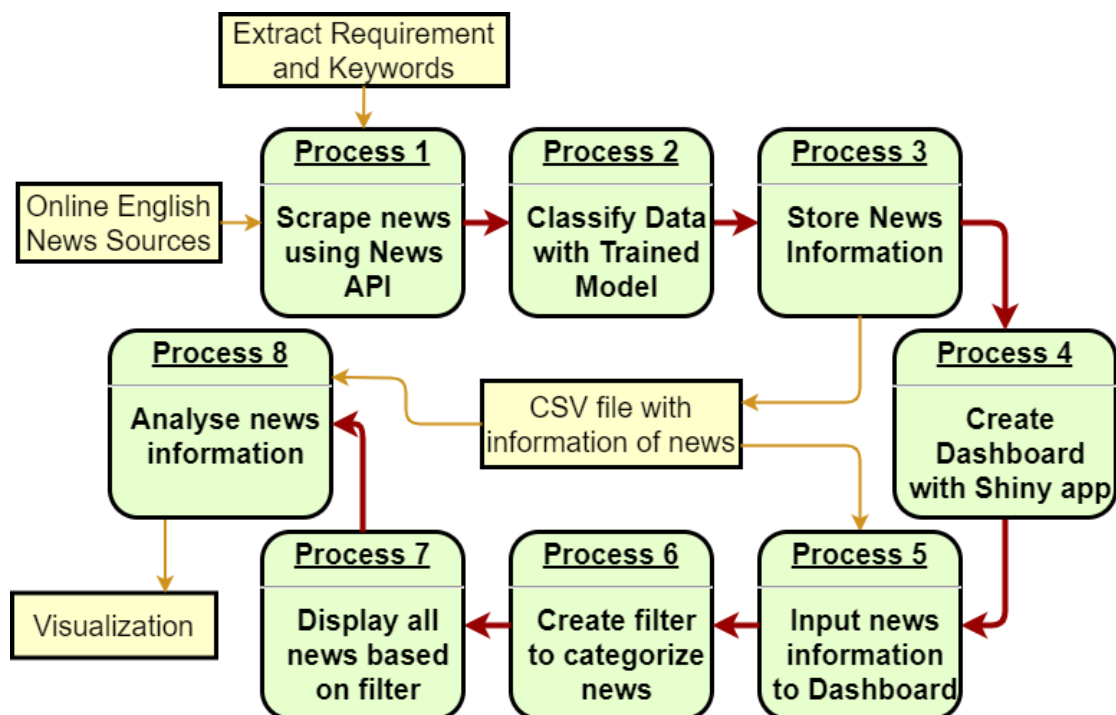


Figure 1.1: Overview Diagram of Propose System.

Targeted resources for the proposed system are English language news from The Star Online. News headlines and contents that are aimed to be extracted will be in the scope of higher education institutions issues. Keywords that is related to local higher education will be defined to filter and extract only relevant articles from the news media to collect information that is accurate to users concerned area. In this project, the keywords are stay in the range of all possible local higher education institution, and is

used to search in the article content. The collected information especially article content will go through sentiment analysis to obtain the positivity of every story collected. After keyword extraction, this projects implements artificial intelligence to filter out not relevant news scraped as well as categorize the news into 8 defined labels. Every news is labelled by trained model before storing for visualization.

List of article headlines extracted will be arranged and presented on a data dashboard with the sentiment of each article, to be easily recognize in a glance. Each headlines present in dashboard will attach with address link that can direct users to the original webpage of the article. Most importantly, visualization of the news information will also be displayed in the form of graph and charts, that can be easily glanced through by dashboard users. Hence, data dashboard will be generated as the end of the product as a report and a summary of daily news.

The targeted users for this proposed system is local higher education institution, especially the marketing department, where they are the group of people which can utilise the monitoring system the most to gain the knowledge of public opinions on own event and brand, as well as big events launched by competitors, in the same time able to advertise news articles related to own brand for reputation improvement.

1.5 Highlight of What Have Been Achieved

Automation is the core of this project. From News scraping, filtering, labelling and storing is an automated process that will be ran daily. In this project, all information is retained just by running a Python file which is scheduled to be ran daily. In addition, filtering and labelling is done by artificial intelligence, which in this project is a deep learning model for each task. These models are trained with dataset that have been manually labelled with news category, and are saved to be used for daily scraping automated process.

The final product of this project is a data dashboard display with updated, filtered and labelled news information, as well as some visualisation according to information collected. Users can specify the information to be shown on the dashboard by controlling the filter created. And most importantly, all news information displayed on the final product refresh daily to keep users updated.

1.6 Report organization

The report is organized in 6 chapters, while each chapter consists of its own contents while interrelating with each other. The first chapter is concentrated on basic introduction of the project. Basically, this chapter purposes to deliver a brief explanation of the project to enables readers to preview the important details that will be discussing about in the report later.

After the introductory chapter, Chapter 2 reviews the previous related works that had been implemented. It consists of some literature reviews about the system developed or used by developers and users on how to implement more effective media monitoring. Each system is reviewed, compared and contrasted to analysis and interpret the pros and cons of the existing program, includes referring to reviews from the system or software users.

Moreover, Chapter 3 is mainly focused on the system design or overview which describes the flow of this project. Each process included in the project such as the model training and network architecture is clearly stated and explained to enable readers understand this project. Readers can obtain knowledge from this chapter and implement the project themselves.

Chapter 4 introduce about the methodology, at the same time presents the tools used in the project development, clearly stated where they are used in implementation. Furthermore, Chapter 5 talk about the system implementation and experiment especially for modelling and deployment phase. Generated result during modelling phase is displayed for comparison purpose and presents how well the model is trained. The final product deployed is shown in screenshot form so that reader will get the idea on how the data dashboard looks like. Issues and challenges met during system implementation will also mention in this chapter.

Lastly, a conclusion about the overall project review is in Chapter 6. The conclusion contains the review of the system, achievements relate to project objectives, challenges encountered in this project and the novelty of implementing this system. Future work is also discussed in this chapter to indicate improvements or further developments that can be made for this system in the future or by the reader.

Chapter 2 Literature Review

2.1 News Media Monitoring

In this era of global media and global culture, news consumption in Malaysia has been focused on online rather than printed news. Many current studies agreed that online newspaper is more widespread compared to printed newspaper, and this trend also followed by most industries in Malaysia that are interested in news media monitoring.

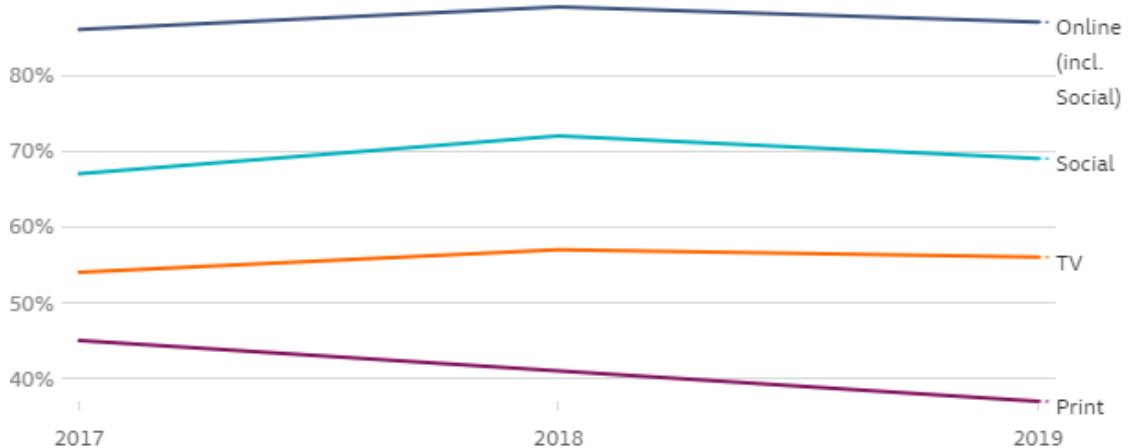


Figure 2.1: Source of News Consumption by Malaysian from 2017 to 2019.

From Figure 2.1, online media is the source that is used by Malaysian the most to get news article daily (Nain, 2019). Many individuals and groups have started to monitor their image from public's point of view. Various big companies started to monitor public's opinion of their brand and industry on media by paying for media monitoring job. These includes employs staff that expert in this field or subscribe for media monitoring software.

Most PR agencies or consultants offer media monitoring as part of their service packages. Glance over online news resources that is related to a brand news releases is vital to measure a company's influence and inspiration. Tracking opinion of mass about an organization, brands, influencers and executives absolutely brings a competitive advantage. Media monitoring process usually begins with establishing a search profile developed with terms or keywords phrases bring up within a news article's content including brand name, product of the brand or keyword related to any online or offline events and campaigns.

Local higher education institutions are having a hard time to understand its reputation and attention in public, in the same time curious about how the industry,

competitor and its brand is mentioned on online media. They hardly get up-to-date with the latest topics that is relevant to its industries, and even don't really understand customers' reaction on a certain action and event happened. Hence, there are actually plenty of solutions that is considerable and on-going to solve these problem. One of them is marketing department of a university has to keep on track of every online information that is posted on different online media. The staffs have to familiar with the sites to retrieve those important information and topics. Filtering must also be done manually such as the degree of importance of an information, category of the article, synopsis on the main point of the information etc. These steps are repeating daily and it would become a burden if there are still bunch of works to be settled. In addition, some institutions pay for media monitoring software that reduce the burden of monitoring news articles. Extraction of articles and analysation can be easily done by the modelling implemented in the software. The software mostly contains many features and users have to be trained before using the platform.

2.2 Data Dashboard

Dashboard is a data management tools that visually presents and analyses key performance indicator (KPI) on monitoring performance of a specific process, business or a department. Similar to the dashboard of a car, it stores, manages, and presents important data from various sources into one. The way of presenting on a dashboard is customisable that can meet the requirement of a business or a company. A dashboard connects to multiple files, attachments or API that contains bunch of information that is stored at the backstage, and displays visualised data on a screen including several types of graphs, gauges, table etc. A data dashboard provides a central location for business to track on large amount of data in order to monitor their performance comparing the trend and key topics in a particular industry. With the help of modern gadgets such as smartphones, tablets and projectors, data dashboard can easily be access any time to retrieve the key data and performance of certain project.

There are multiple advantages that effects in utilization of this tools. Data dashboards could easily be modified according to users' requirement and expectations. Each decision level dashboard can be personalized to display the most valuable set of information clearly. This allows different division of people having different level of

detail view that are needed in order to simplify their job and meet the goals sooner. Before dashboard appears, individual would spend excessive amount of time on analysing and reviewing different reports before getting a final conclusion. In this case, dashboard enable users to glance through an overall situation report to get the conclusion of the desired data. The time saved can be used on preparing plans to improve current situation.

Having all information in one screen doesn't mean that detail information will be loss in the presentation. Dashboards are technically advanced with the capability to show deeper information that is required by simply selecting the data point, variable or object. In addition, no exhaustive training should be done to reader of the dashboard before the information is presented. Dashboards are designed to be easy understanding and can be smoothly navigated by any individual. Last but not least, most dashboards platform are developed in the way that enable users to operate on mobile device. The idea is to outreach anywhere, to anyone, real-timed, with the most accurate information.

However, using data dashboard is still challenging. Users can only enjoy the positive impact of data dashboards when the right tools are operating in place. Since dashboards are developed to measure something meaningful, selecting the right metrics, specific metrics that suits the broader category, is the key issue of all. Dashboards are complicated to be set up, by requiring member of IT team and assistance from a developer, or a reporting tool that contains pre-loaded dashboards that don't consider suitable metrics relevant for every possible user.

2.3 Existing Dashboard

2.3.1 Meltwater

Meltwater is one of the most trending software for media monitoring in this decade. It tracks all relevant information, analyses them and enables their clients to get the important information from billions of online conversations. It extracts essential insights, and supports clients strategically on managing their brand and stay ahead compared to their competitors. It provides monitoring service by obtaining data from 4 main categories, which is online media, printed media, social media and mobile application. It inspects millions of information every day from blogs, social media platforms and online media such as news sites, filtering out the "not so important"

information or non-applicable ones and assigning sentiments to the ones its clients are interested in. It also serves an individual brand in finding key influencers, setting up searches and tracking online performance.

Meltwater allows clients to create dashboards for specific purposes and customize it with features and widgets that is preferred to insight for improving their brand. The default dashboard includes themes that monitor, benchmark or analyse activity.



Figure 2.2: Domain name and Trademark of Meltwater.

Features Rundown

Thorough social profiles on every community author that contain history of the brand and analysis of its influence and brand affinity. Profiles include complete public social memoirs, communication history and notes, brand engagement and sentiment analysis, a social graph with alternate social identities, activity analysis, Jitterater (a system engine that estimates the influence and assigns rating for each client based on the different elements of their social profile) influencer score, the individual's top 5 personal contacts, customizable tagging and contact segmentation (Glassman, 2011).

Social inbox provided can store inbound social communications for prompting response and action on online media. Social action can be done for all conversation found by Buzz's monitoring system. Buss monitoring system tracks the mentions of individual brand name throughout the Web including social media platforms. Clients can choose to share conversation, which is posting comments on social media platforms. Communication calendar is provided to review outbound of the message sent and scheduled (Glassman, 2011). A communication calendar is a tool that clearly display arranged new tasks that come along with running a business in order to let members of a company complete tasks on time. Social analysis and brand tagging that is provided enable a brand to segment individuals into groups of individuals that share common

characteristics. Social activities are measured and analysed including communication frequency analysis, brand impression analysis and most active or engaged commenter.

The dashboard is the very first thing that will be shown once log in is done into the Meltwater News platform. A simple scan overview of news and social media searches will be provided, including graphs and charts that shows information for the previous day, month or year. The metrics on the dashboard is different for social media and news searches, in order to coordinate with the difference of scope for the results.

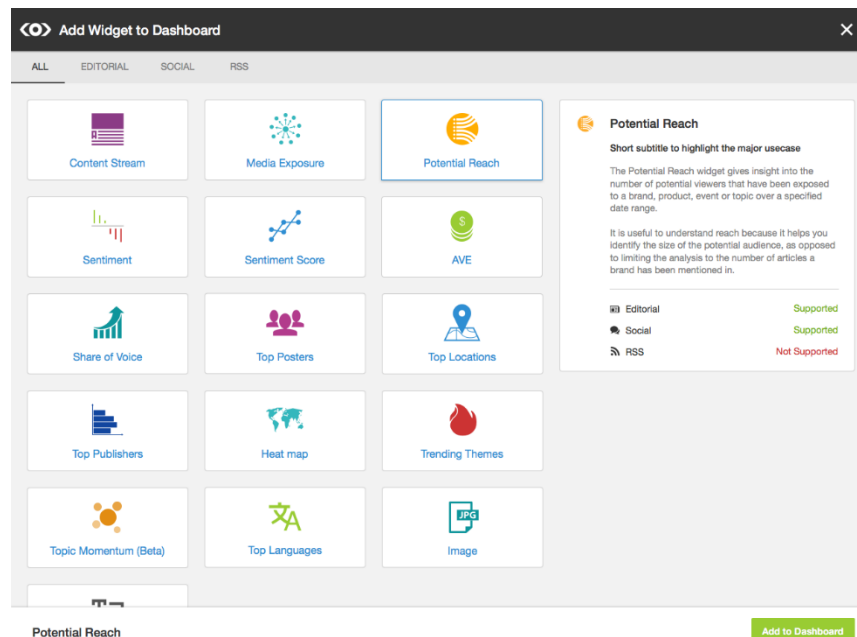


Figure 2.3: Dashboard Features and Widgets.

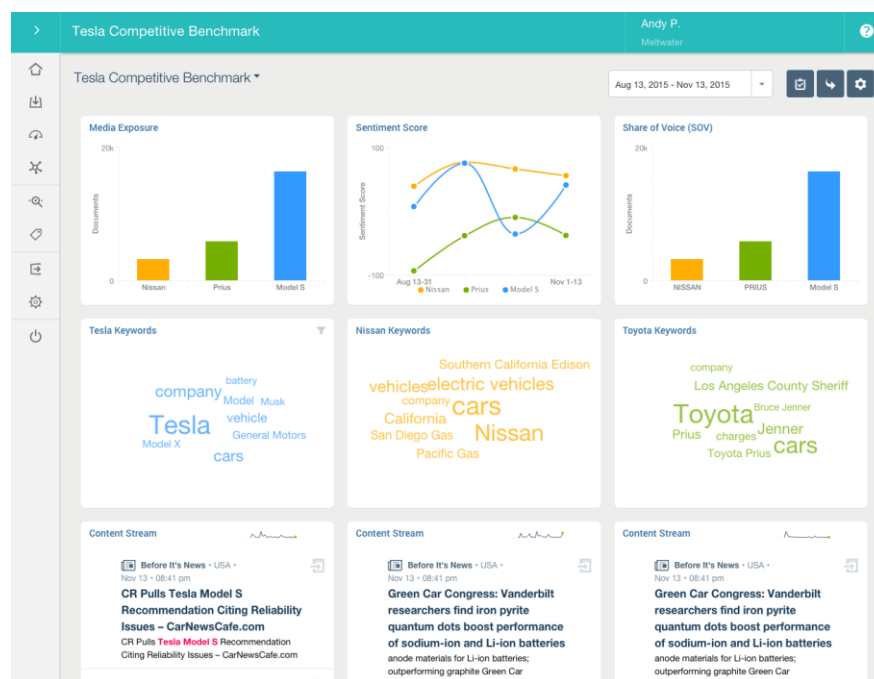


Figure 2.4: Dashboard.

Strength and Limitations

According to Business.com Editorial Staff, the search function is the key for Meltwater. It recommends phrases associated with the keywords that is chosen. Once keyword is selected, filters such as language, locations and source type also can be set. In this case, clients can have chance to control on filtering out noise, which can ignore the post or article that is not interested (Editor, 2019).

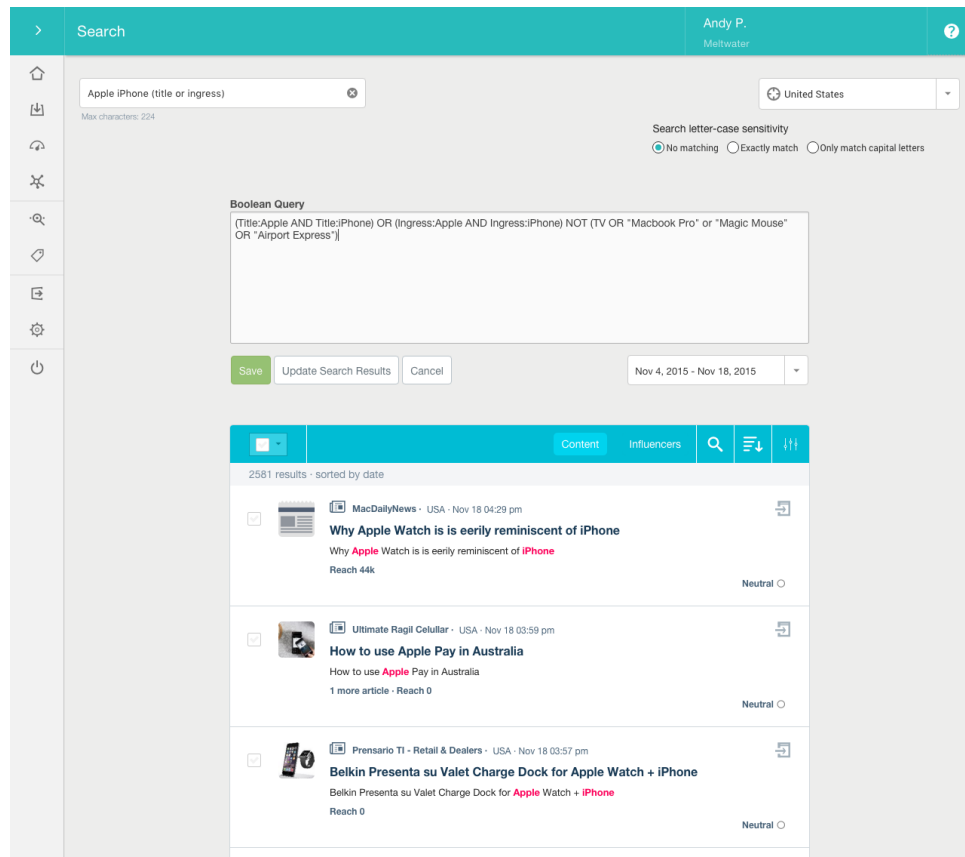


Figure 2.5: Search Engine.

Meltwater also assigns sentiments to the articles and posts that match a search. Sentiment analysis and the reports concludes whether the information found is positive, negative or neutral.

Engagement feature from Meltwater is also a strength. Other than posting at the dashboard, it enables client to set up key influencers as contacts. This contact list enables a company to easily monitor posts and comments from its influencer on a company page (Editor, 2019).

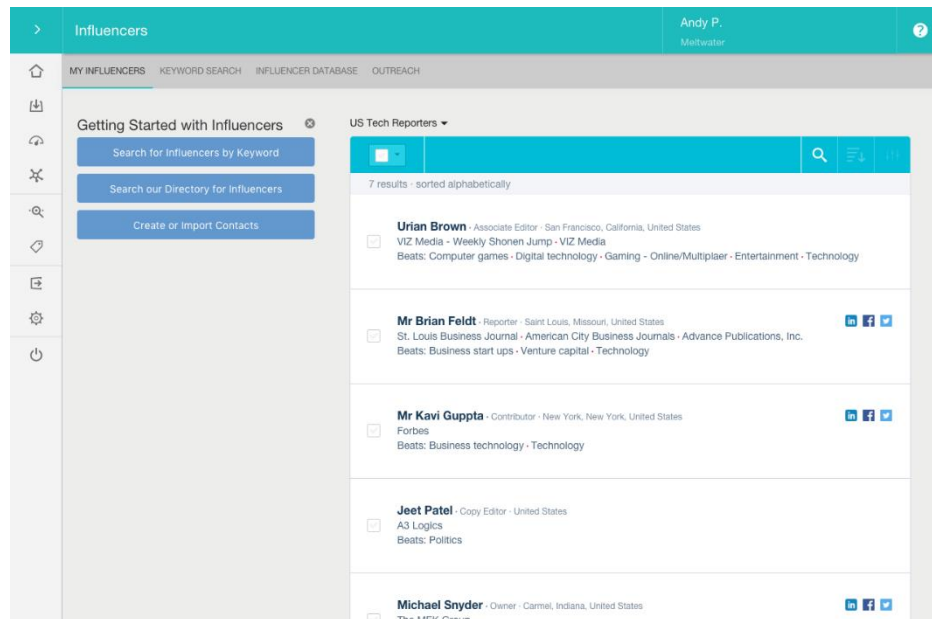


Figure 2.6: List of Influencers.

Meltwater works on desktop interface as well as mobile. With the mobile feature, clients are able to immediately respond to alerts, such as specific discussions of his company, especially when he is away from office.

The main limitation of Meltwater that is most mentioned from the reviews given from public user is Meltwater is not really user-friendly. The features of this tool are complicated and hard to apply even after training is done. More time has to be spent by an individual to get used to using the features it in the best way in order to have the best monitoring experience.

Clients

There are some famous companies that are using Meltwater to monitor their brand. One of them is Watsons, a health and beauty care chain store. According to Watsons, Meltwater helps them track their campaign ROI, using PR value to precisely measure their success. As their PR agencies use various metrics, having Meltwater as their main PR tool helps them to access the output of their agencies fairly. The second company is Nando's, a South African restaurant chain. According to Nando's, Meltwater is user-friendly and detailed. The layout of the tools is simple for clients to browse and obtain insights. Meltwater team is responsive and helpful for them to get important information.

2.3.2 LexisNexis Newsdesk

From setting up alerts to effortlessly generating charts for visualisation, as well as uncover media insights and sharing data through custom newsletter with copyrights, Lexis Newsdesk is a media monitoring and analytics solution platform that provide services including searching, analysing and sharing new coverage. It monitors over 2.4 million new articles a day. Nearly real-time media monitoring analysis is done by obtaining resources across 85 thousand online news, broadcast and print sources, including resources that provided by a company. It then filters the unimportant information get and analyse the sources for its client.



Figure 2.7: Domain Name and Trademark of LexisNexis.

Features Rundown

LexisNexis Newsdesk is simple that makes users easy to move from monitoring media and discussions to understanding the resources obtained. Advanced analysis and filtering capabilities are enhanced by built-in data visualisation tools. A simple control which drag-and-drop one or multiple media feeds to the analysis tab can easily turn raw data into near real-time awareness. In addition, there are various templates in Newsdesk, that combine multiple charts and graphs to provide a quick and clear industry, competitive, market or social media analysis of the information extracted. In this case, LexisNexis Newsdesk empowers various analysis including media coverage by journalist or source, brand mentions by location, SOV and engagement of the brand. Users can click on a particular data point of these interactive information visualisations to read the articles behind the data and lead to superior insight.

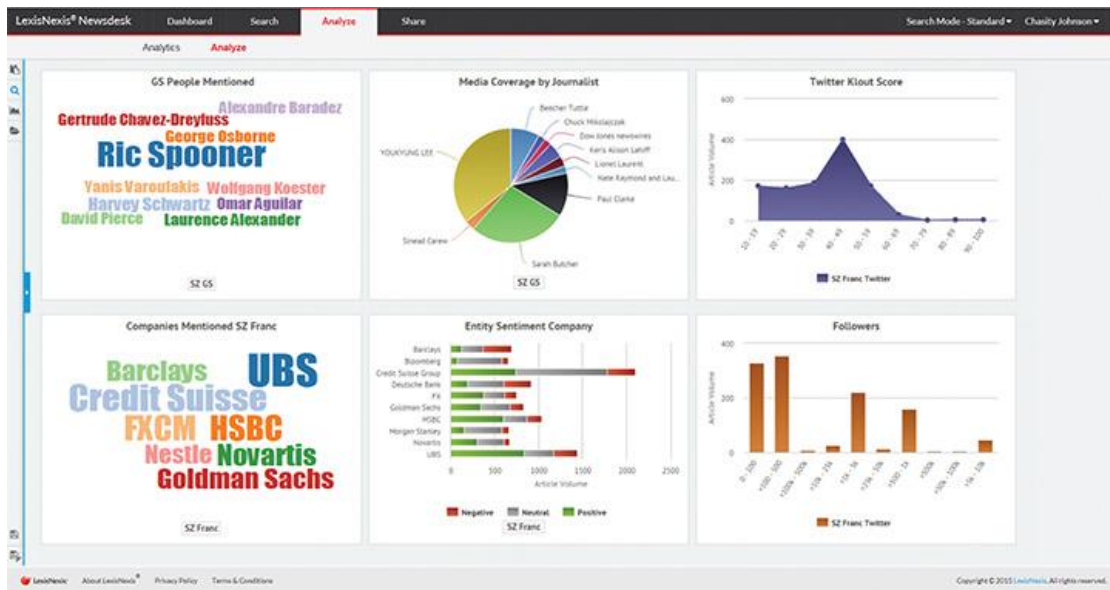


Figure 2.8: Analysis Tab.

Dashboard is a personalized homepage for every client. It provides the views of latest headlines from favourite news, social feeds from the searches, analysis charts that can be modified and videos from various websites. The searches and analysis that is saved at analysis and search tab will automatically turn into widgets that can be choose to display on the dashboard. Similar to the analysis tab, the data point in the charts and graphs can be clicked to find out more information or articles related to the story behind it.

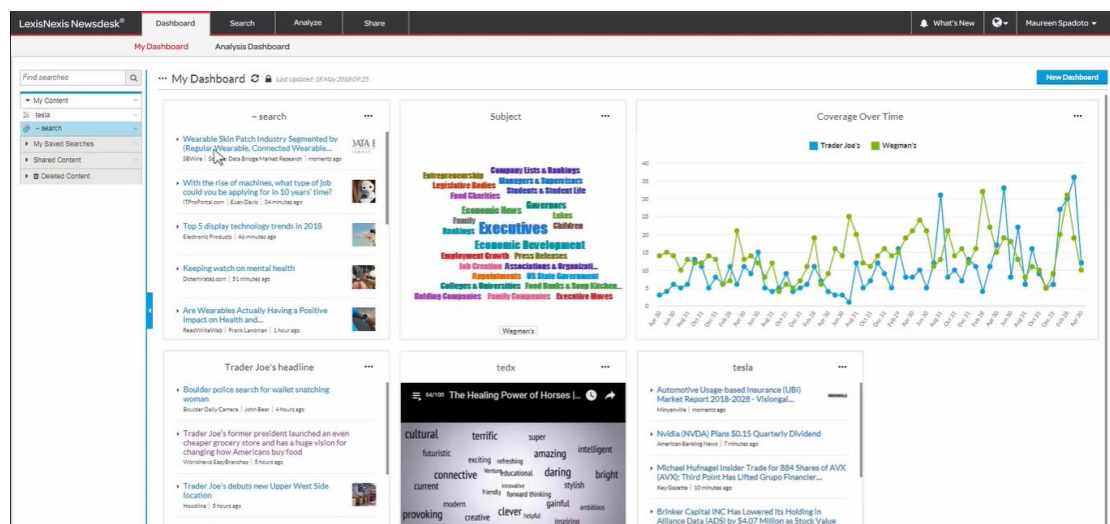


Figure 2.9: Dashboard Tab.

With analysis and monitoring that is done, a company still need to push out relevant media intelligence to decision makers, which is the team across the company, clients of the company and the stakeholder. LexisNexis has a share function that would simplify this task and accomplish it in a professional way. With the built customized dashboard with multiple featuring charts, data visualisation that can update automatically, and the created embeddable interactive charts that can drive awareness on a website, share tab can set up branded alerts and create newsletters that is copyrighted to share relevant insights of company with multiple audiences. Administrator can also share group dashboard. Group dashboard can be shared to the team of the whole organization.

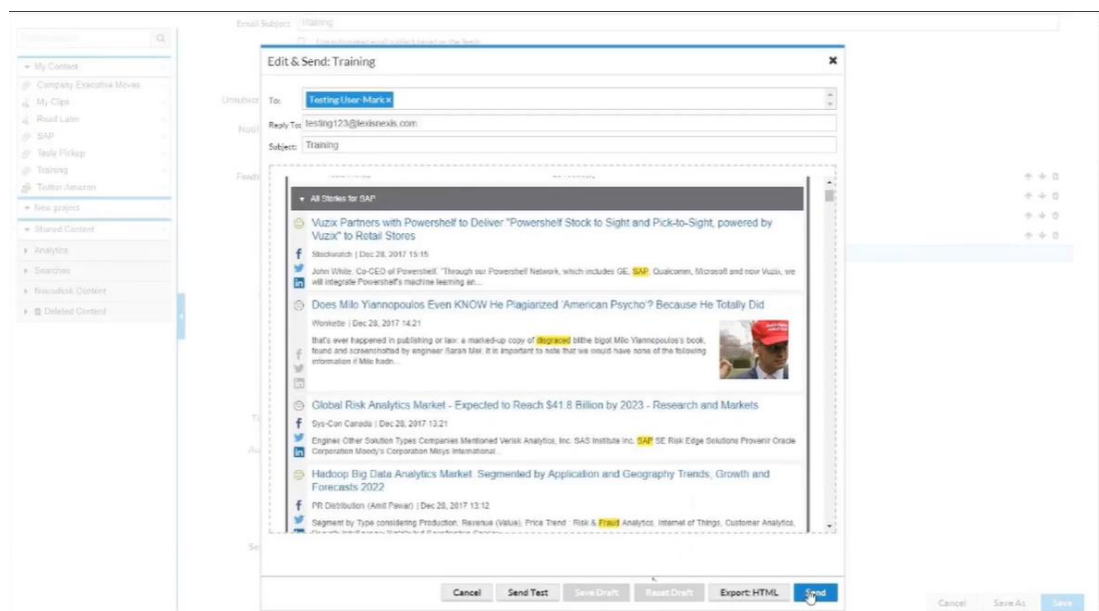


Figure 2.10: New Newspaper / Alert to be Shared.

Strengths

The main strength of LexisNexis is it has a share function which enable the analysis of information gathered related to a company to be shared by email, alert or newspaper. These alerts and newspapers are fully-branded that the company owns the copyright of the documents. Automatic sharing can also be set on the sharing function to schedule a fix daily, weekly or monthly to create email and send to internal and external audience. Users can customize their alert and searches specifically with specific need. This makes the platform a worthy choice for corporations where customisation of searches are different in various departments, to enables each department quickly edit the searches,

alerts and feeds in order to react to real-time changing of trend in the industry or company (Knight, 2016).

In addition, cost and time savings also able to achieve by using Newsdesk because users can do media monitoring by monitoring bunch of licensed content, open web sources and social media, rather than need to purchase multiple separate services for resources (Knight, 2016). This is not only valuable to administrators but same to users while they don't have to learn multiple platform and have more time to skill on this platform.

The graph and charts that is analysed and displayed on dashboard is the visualisation of information extracted from media. The information resources are linked to the data point on the displayed graphs and charts. This features enable users to read the articles or media to effectively find out the origin factor of the showed results.

2.3.3 Mention for Education (used by University of Chester)

Mention is good media monitoring platform for its users on managing crisis. The software can be connected to mobile to let users react immediately before the problem spreads. Mention software can connect to different media platform, in this case, users can immediately engage and respond on the community discussions within the application. If a company is mentioned, sentiment analysis can quickly be measured by the public opinion so this would be easy for users to identify the outcomes and responses of an event.



Figure 2.11: Domain Name and Trademark of Mention.

The University of Chester is one of the eldest higher education providers in the United Kingdom. They encourage various personalities, and treasure an extensive range of approaches to learning. The University of Chester needed a monitoring tool that monitored beyond media platform. They used Mention to improve their reputation and crisis management. They were finding a tool that monitors websites forum, not just social media. Going through testing various media monitoring platform, Mention is the

best choice of all (Shai Vure, 2018). Mention platform helps them find and sort out the updated and trend news that were happening in the local area with their students. For managers in the university who aren't really apprehend with computers, the information can be quickly exported to PDF and sent to them, so that the receiver can easily access the information topic. This is extremely useful especially when there are issues that need to be handle before the crisis is spread.

According to Shai (2018) from Digital Marketing Manager of the university, Mention is more than a good choice for them. With this software, they can effortlessly get more trended information compared to before.

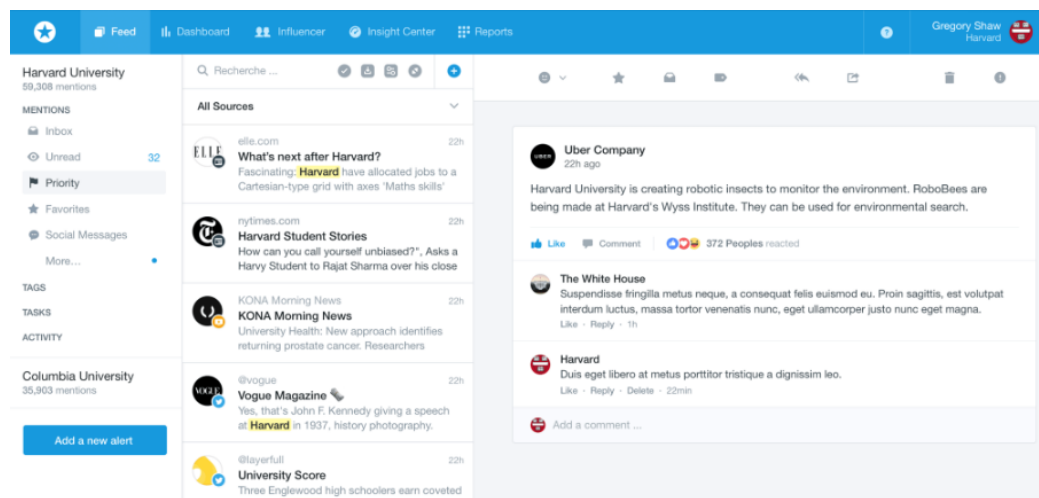


Figure 2.12: Extracted News Feeds.

Strengths and Limitations

Users can get notify by Mention about any discussion that is related to the campus. The discussion can come from many platform including social media and all sources through the web. Mention ensures that users always get up-to-date on the discussion that is related to its clients to respond on it. According to the users of Mention, most of them reviewed that Mention is an easy software to be used. In addition, Mention shows the article that is alerted in the application itself. Users have time to decide that whether the content is worth for attention and to be responded.

However, the limitation that have been mentioned the most is about the pricing part. According to users' review, it is a little expensive to get the upgrade version of Mention to have better service. In fact, the free version still makes users satisfied with the service and simple control, even users that paid for more features feels value for it.

Chapter 3 System Design

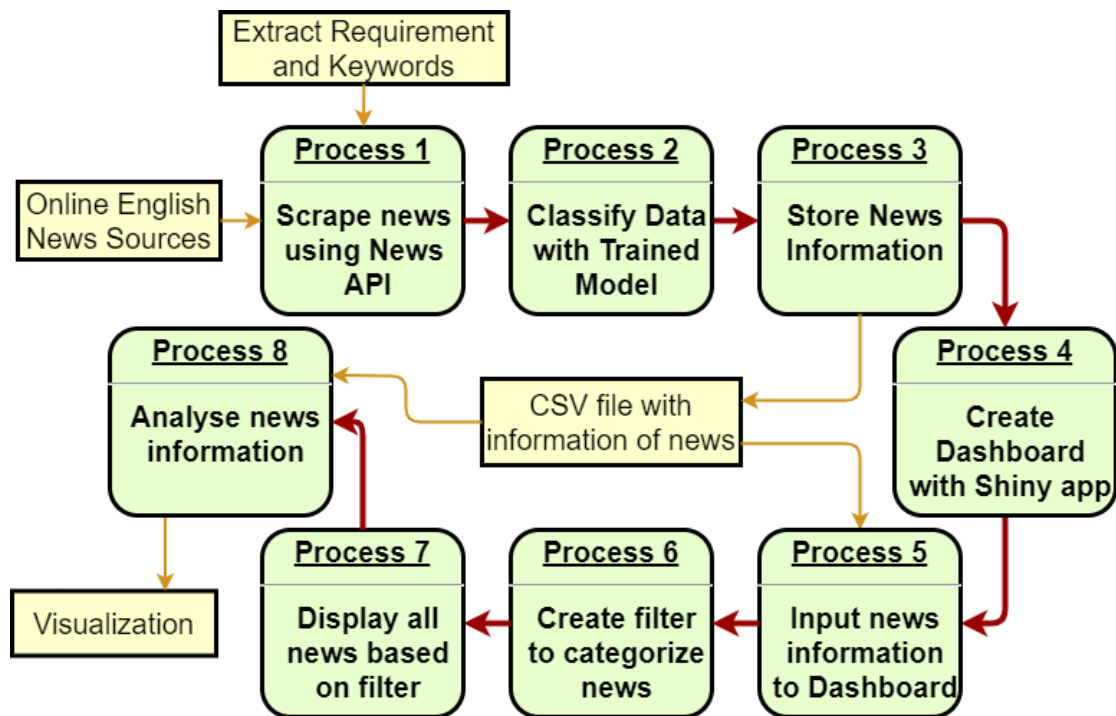


Figure 3.1: Overview Flow.

3.1 Business Understanding

In the initial phase, project objectives and data mining goals from a business perspective should first be determined. The system is proposed to automate the process of media monitoring, be convenient and simple for its users. The business objectives are as mentioned in Section 1.3 Objectives, which more focus to ease the collection and presentation of news articles headlines, no more manually extracting news to be up-to-date, as well as understanding the opinion from the mass and publishers of the articles. These objectives will definitely attract the targeted users, marketing department of institutions, to have a try on such simple and convenient system.

The motivation for this project is to spread the importance of media monitoring among local higher education. News monitoring should become the trend of taking advantages to be competitive in the industry. In comparison, overseas higher education institutions have already considered media monitoring as an intense weapon to benefit them on their marketing planning. In addition, manually extracting news articles should be fall into disuse as it is a burden for an individual to keep in touch with updated news

from various news media every single day. Automated extraction and filtration of relevant data with related keyword would be simple and ease the monitoring process.

3.2 Data Understanding

There is a close relationship between Data Understanding and Business Understanding. The design of the proposed problem and project plan require more or less understanding of the data that should be used.

In this project, the data that have to be focused on is the information from news articles on Malaysia online news media. The data to be extracted from news media is the article headline, address link, published date, author, article description and the article content. These data are sufficient to build the system dashboard, where only headlines have to be displayed attached with each address link, while the use of article content is for sentiment analysis to show the positivity of the news, data labelling such as campus name and their respective category that will be implemented in model part. Other than news articles, keyword that should be define for extracting news articles also the data that need exploration, where should be completely in the scope higher education. This attribute is purposely defined to search for similar keywords in article content from news media, to extract only relevant information to develop final product dashboard.

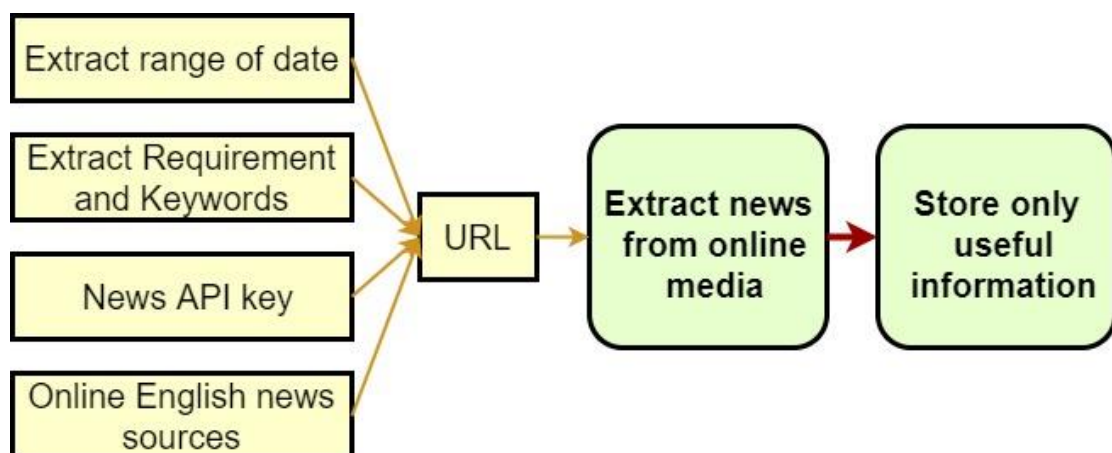


Figure 3.2: News Scraping Flow.

The very first step to implement the system is to web scrape news articles from online media. A URL is created with the range of the date of news, extracting keyword for

article title, online English news media and most importantly the News API key got by registering for the tool. Request is made by the URL to get the related article information as respond. Among all the information collected from the news sources, only useful attribute will be remained and stored. The news collected is of purpose to be displayed on final product, as well as to be a dataset for training model to perform categorization on daily updated and extracted news.

During data understanding phase, the extracted data is mostly high quality, most of the instance does not contain unknown value or complex symbol. News information from the news source is almost complete and data format is exactly consistence. Every attribute of the data extracted is relevant to its attribute description, which will be very helpful during data preparation as only simple data cleaning has to be perform.

3.3 Data Preparation

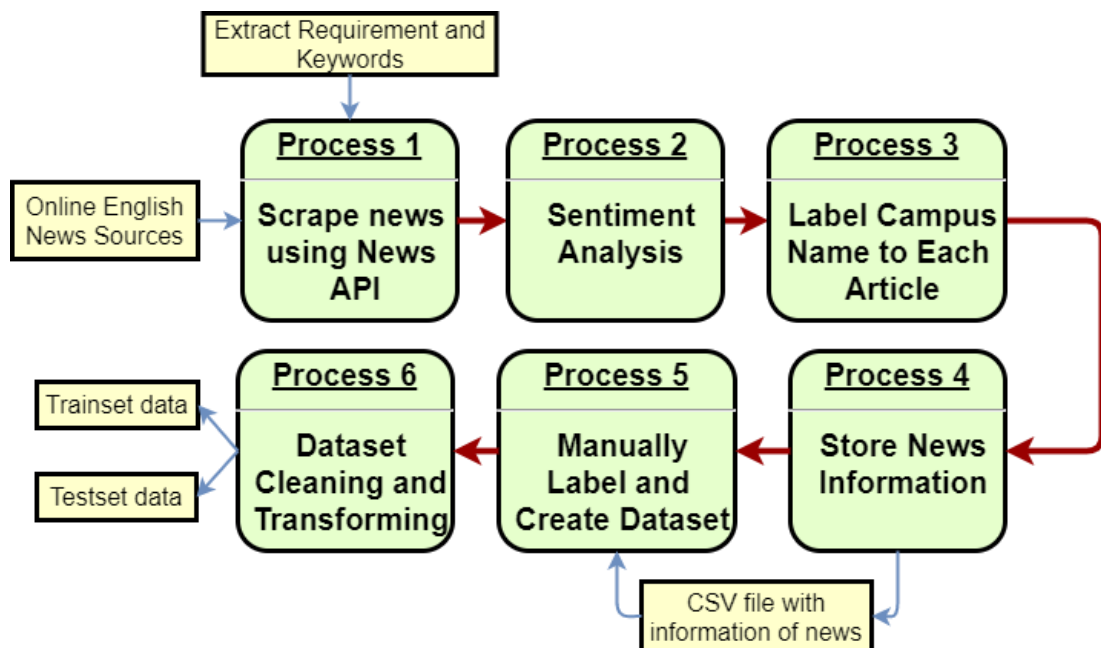


Figure 3.3: Dataset Preparation.

In this phase, news data are scraped using keyword which has more relationship with higher education, including the name of local higher institutions, scanned through news content to ensure that no important news are left behind. The collected data are labelled to their respective category, which in this project consists of 3 types of labelling that will be displayed to user in final product, including sentiment, campus name and news category according to article content. Another important process is to filter the news,

whether are they related to the target information for this project, which means whether the news higher education related.

During data cleaning process, the instance with unknown value in any of the attribute will be removed. This is to make sure the news display in final product is completely useful to target user. Sentiment is labelled right after data cleaning, by using a build-in package in the notebook, run with the news content extracted, which consists of 3 labels include 'Positive', 'Negative' and 'Neutral'. While for campus name labelling, every single word in the content will be scan through to find out if there is any word similar to listed 108 local institution name. These 2 labelling is done before news filtration, which is to decide whether all the news are related to target information and should it be shown on final product. The filtration and categorisation of news data will be implemented during modelling phase.

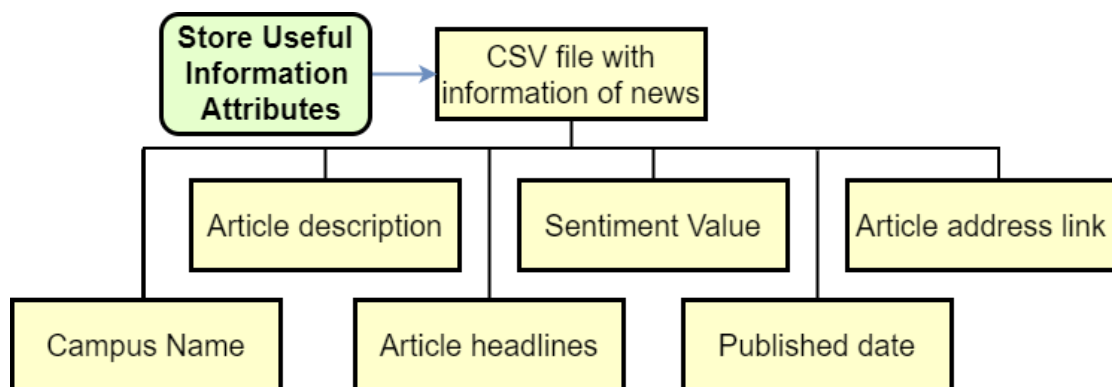


Figure 3.4: Stored Data Attributes.

All the information stated in Figure 3.4 is then stored in CSV file as a database.

Another process in data preparation is for modelling phase, which is to prepare training and testing dataset for model training. 2 model should be trained for news data filtration and categorisation problem, and thus, 2 types of dataset have to be prepared.

CONTENT	FILTER
IPOH: Police will not investigate the report lodged by	HighEdu
China Daily Editor's note: Tsai Ing-wen has been re-elected	Other
KUALA LUMPUR: Universiti Pendidikan Sultan Idris (UPSI)	HighEdu
COMPILED BY NG ZHE QUN , C.ARUNO and R.ARAVIN	HighEdu

Figure 3.5: Filter Dataset Sample.

CONTENT	CATEGORY
IPOH: Police will not investigate the report	Incident
KUALA LUMPUR: Universiti Pendidikan Sultan	Politic
COMPILED BY NG ZHE QUN , C.ARUNO and	Credit
KUALA TERENGGANU: The body of the second	Incident

Figure 3.6: Categorise Dataset Sample.

Dataset is prepared by using the information stored. Every news articles are labelled with respective category. The dataset to train the filter model is labelled into 2 categories as shown in Figure 3.5, which ‘HighEdu’ indicates related articles, while ‘Other’ represents the article content is not related to target information for this project. Meanwhile, the dataset to train the categorise model is labelled into 8 categories as shown in Figure 3.6, which includes ‘Branded’, ‘Career’, ‘Credit’, ‘Education’, ‘Event’, ‘Incident’, ‘International’ and ‘Politic’.

During modelling phase, only two attributes of the data is used, category and article content. News content is cleaned by removing punctuations, symbols and stop words that is not really helpful on model training, as well as vocabulary lemmatisation in the news content, which indicates the process of transforming a word to its basic form. Every sentence and paragraph in news content is tokenised, so that model can identify the most used keyword or vocabulary for each label. All labels are encoded as models train well with and numbers, which they just have to learn on grouping data according to existing number of categories.

3.4 Modelling

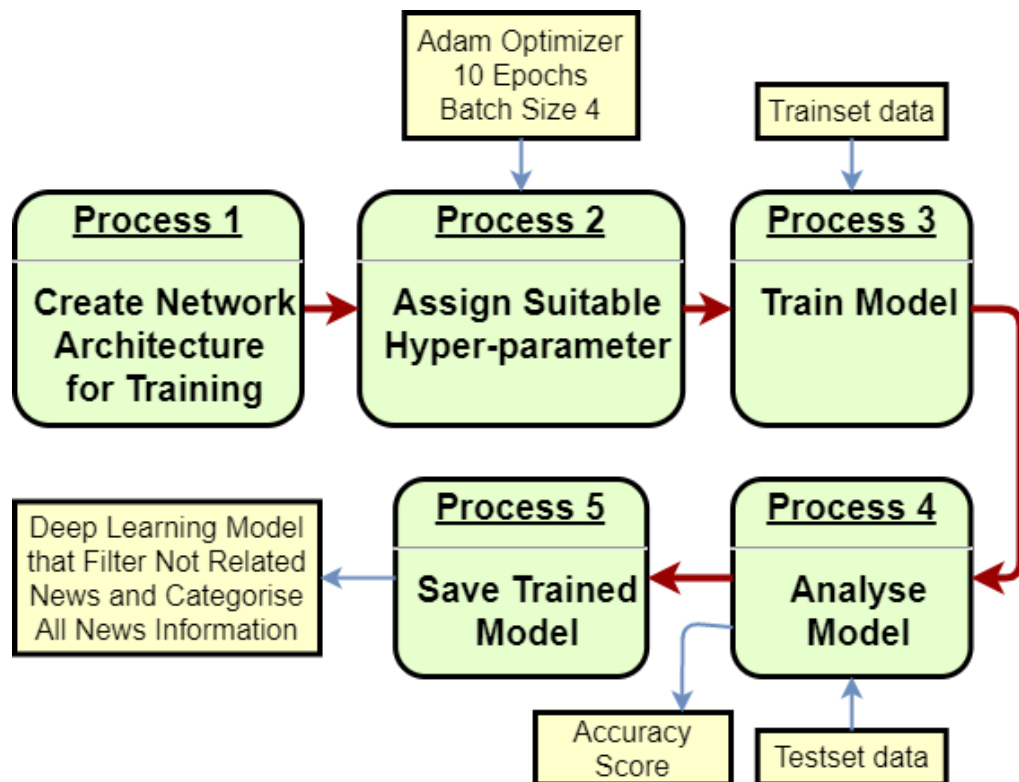


Figure 3.7: Flow of Modelling.

Figure above shows the flow of implementing model for this project. In this phase, two models are implemented, a model that can filter the extracted news as if they are related to the targeted information, and a model that can predict and categorize all news into 8 self-defined labels. Both model is implemented through the same flow as figure above. All dataset used for training in this project is manually labelled with their category referring to each news content, cleaned and prepared as in data preparation phase.

In this project, 2 models are implemented with 3 layers deep learning network architecture, and each model is stacked with different types of layer. The training and testing dataset are split from the dataset prepared, with the ratio of 9:1.

3.4.1 Filter Model

Layer (type)	Output Shape	Param #
embedding_3 (Embedding)	(None, 350, 100)	1397000
global_max_pooling1d_6 (Glob	(None, 100)	0
dense_38 (Dense)	(None, 10)	1010
dense_39 (Dense)	(None, 1)	11
Total params: 1,398,021		
Trainable params: 1,398,021		
Non-trainable params: 0		

Figure 3.8: Network Architecture of Filter Model.

Figure above shows network architecture created for the filter model. The Sequential model is a linear stack of layers, where can be applied with varieties of available layers. The most common layer, Dense layer, a regular densely connected neural network layer is the best choice to start off the training. The Embedding layer is added to compress the input feature space into smaller dimension, which will be computationally effective. Lastly a max pooling layer, to down-sample the input from Embedding layer to Dense layer. The layer is added in the same time to help over-fitting case. The network is implemented in Keras, trained 10 epochs with a batch size of 4, Adam optimizer and Global Vector word embedding method.

3.4.2 Categorise Model

Layer (type)	Output Shape	Param #
=====	=====	=====
dense_87 (Dense)	(None, 256)	3840256
activation_75 (Activation)	(None, 256)	0
dropout_53 (Dropout)	(None, 256)	0
dense_88 (Dense)	(None, 128)	32896
activation_76 (Activation)	(None, 128)	0
dropout_54 (Dropout)	(None, 128)	0
dense_89 (Dense)	(None, 8)	1032
activation_77 (Activation)	(None, 8)	0
=====	=====	=====
Total params: 3,874,184		
Trainable params: 3,874,184		
Non-trainable params: 0		

Figure 3.9: Network Architecture for Categorise Model.

Similar to the Filter model, Dense layer is the regular layer to connect whole neural network for this model. This model is rather simpler as the layer stacked are 3 Dense layer without any embedding layer, which the total parameter is obviously more than Filter model. This network is built as the dataset prepared to train this model is not too big which won't result to computational heavy case, as well as going through experiment, it results with better accuracy compared to start off with embedding layer. Dropout layer is added in the network to help over-fit case on the limited dataset. The network is also implemented in Keras, trained 10 epochs with a batch size of 4 and Adam optimizer.

3.5 Evaluation

Before deployment of the final product, the steps executed to develop the model will be evaluated and reviewed, to be definite it appropriately achieves the business objectives. Depending on the result model, decision to any phase should be made for improvement of the proposed system. It is decided whether all objectives are achieved and any enhancement can be done to develop better system.

For this project, evaluation is mostly focus on the model built to solve the filter and classify problem, where to evaluate whether they are suitable to the problem assigned to each model. Initially, the system is implemented with experimenting few machine learning model, while with the change of time and experiments, deep learning model results on doing better compared to machine learning model. Multiple trial and error in this phase to decide how to improve the implemented model, which includes deciding layer of neural network, number of neuron in each layer or neural network and the number of epoch. And lastly, both model is trained with manually labelled dataset and saved for later use. The result accuracy of training both models is mentioned in Chapter 5 System Implementation, to compare their performance with machine learning models.

During deployment phase, the objective mentioned will be achieve by automating the media monitoring system, as well as to display arranged and updated news information, and provide reputation analysis on news article according to interest range of target users.

3.6 Deployment

3.6.1 Automation

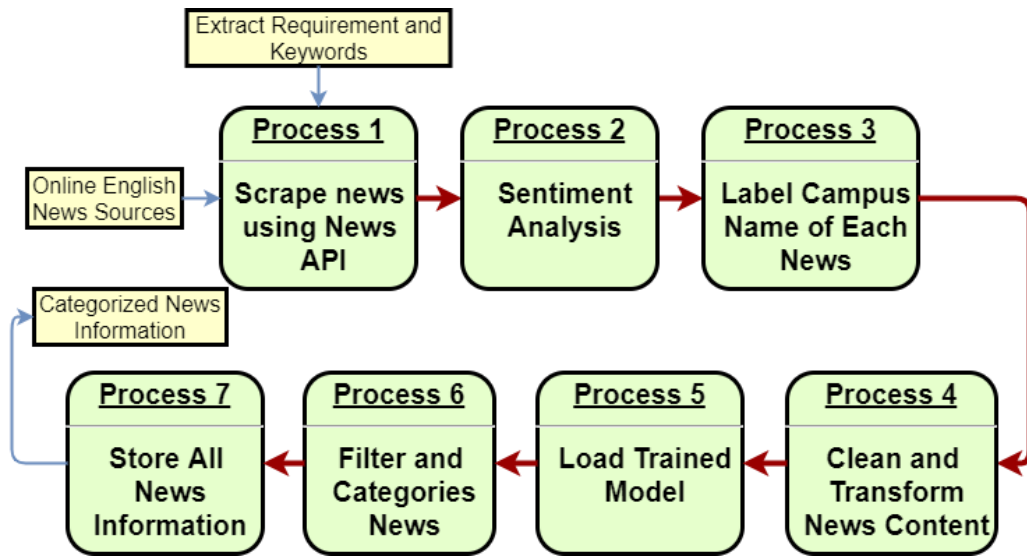


Figure 3.10: Automated Process.

One of the objectives deploying this system is to create automated media monitoring system, where can scrape and do categorisation daily to make the final product always updated. Process 1, 2 and 3 in the figures above is similar to data preparation phase but conversely, they are all deployed automated. Before prediction process happens, the news content extracted will be clean and transformed as how the data transform to train the models. For filtering and categorising news data, saved model is load to do prediction on the new news content extracted. The source code of these process is written and scheduled to be run daily whenever internet connection exists.

3.6.2 Dashboard

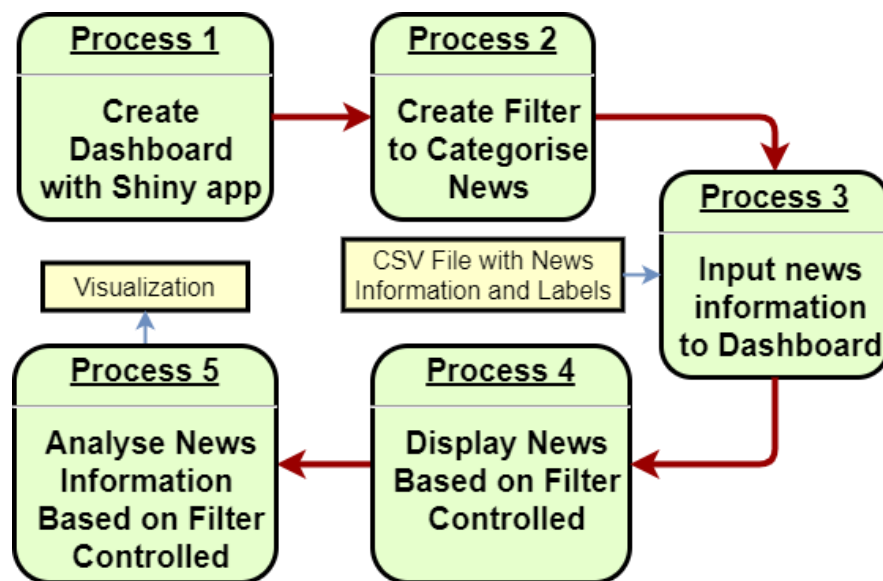


Figure 3.11: Dashboard Formation Flow.

The final product of this system is a media monitoring dashboard, which target users directly interact to. The dashboard is built with interactive filters, which can categorise and only display information that interested by target users.

The news information will be display in two form, which are list headlines and graph visualisation. For listing the headlines and description of news articles, only news in range of 30 days will be displayed. Meanwhile, analysation is done on the news information, showing in the form of simple and easy understanding visualisations such as charts and graphs. These visualisation is also filter controllable, where user can directly get the information needed.

The final product dashboard is published online by registering an account through the development environment which is further explained in Chapter 5 System Implementation.

Chapter 4 Methodology and Tools

4.1 Methodology

The most suitable methodology to develop this project system is Cross-Industry Process for Data Mining (CRISP-DM).

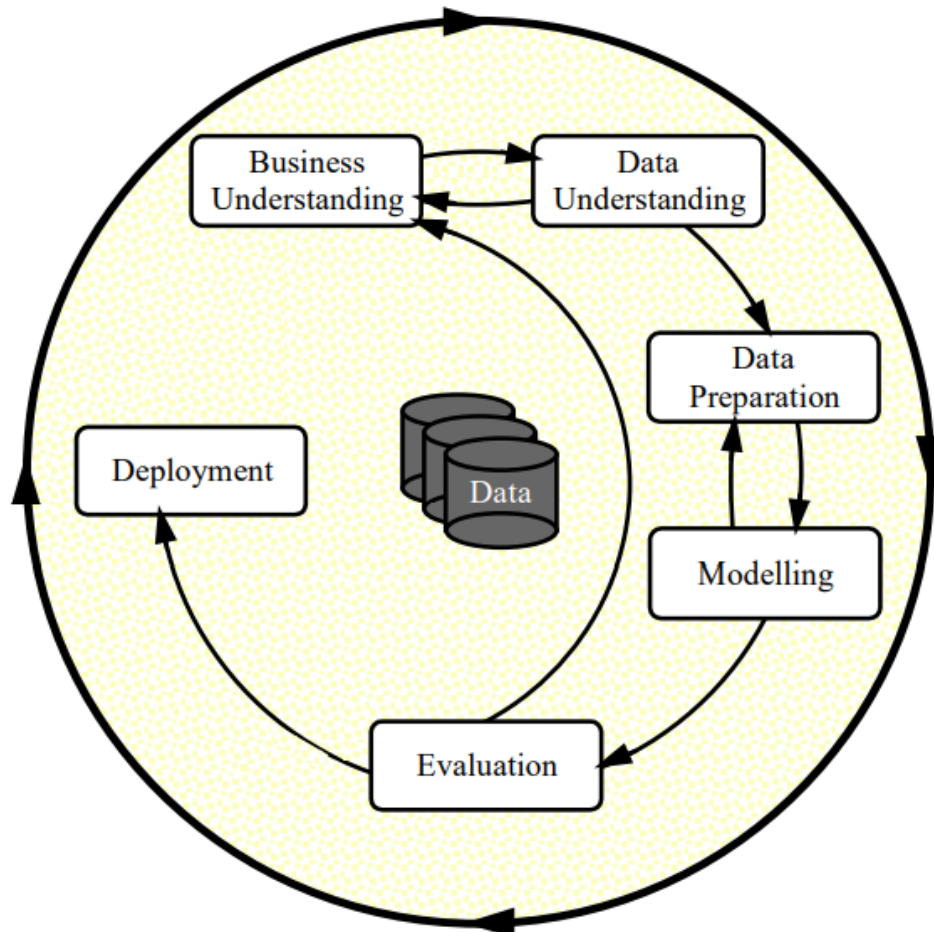


Figure 4.1: Phases CRISP-DM Process Model for Data Mining.

Figure 3.1 shows the life cycle of a data mining which is distributed into 6 phases (Wirth, 2000). The order of the phases is not restricted. The arrows show the most frequent reliance between each phases, backtrack to previous tasks will regularly be necessary and repeated for certain actions. Data mining process is not ended as soon as a solution is developed. The lessons digested during the practice and from the developed solution can generate new and more focused business questions.

4.1.1 Web Scraping

In order to practice media monitoring, online media and websites is one of the platform for data extraction. Copy and paste the data manually is feasible especially on the website that contain tons of information. Hence, web scraping comes into play. Web scraping is a practice of computerizing the extraction of important information efficiently. Web scraping is also named as Web harvesting, where web data extraction will be done. There are multiple methods can be used to collect information that is posted across the internet. Generally, all the information extracting is done with program or software that simulate human action on surfing internet to retrieve particular information from different online media websites. The main purpose of using web scraping software may be looking for certain information to be traded to other user, or use it for commercial purpose online. The extracted web data can be saved in multiple format such as CSV for further use and analysis.

In this era of technologies and modern gadgets, human started to practice real-time analytics with the data available in the internet. Real-time simply means that analysis could be done once the data is able to be extracted online. This is different from batch-style analytics that will delay the analysis insights.

Since web scraping mean extracting information that is posted by other, is it legal or illegal to do web scraping? Bernard (2017) stated that web scraping itself is legal, anyone can scrape or crawl information through the web to educate and learn. It's a cheap and great method to gather tons of data without the need of cooperate with more partnership. Many companies use web scrapers to obtain resources for their own achievement but in the same time refuse on letting others use bots against them (Essaid, 2018). The trouble is the terrible widespread unawareness on the lawful aspect of web scraping. Legitimacy is totally relying on the legal jurisdiction. Publicly accessible resources crawling or scraping is not illegal. This doesn't include gathering personal information without the knowledge of the owner, that could infringe data protection law.

According to Bernard (2017), the problem arises when scraping and crawling is done without gaining the resources owner prior written permission, or even ignore their TOS. If the data downloaded is for personal use or analysis, it is always ethical to obtain the data to learn new things. In comparison, if the information get is planned to be

published on a new website and own it as self, in the way where no permission is get from the resource owner, it is obviously unethical and illegal (Bernard, 2017).

In addition, there are cases where web scraper tools try to excessively access the server of the resources that is aim to be obtained. This will absolutely effect on the performance and bandwidth of the server itself to work normally, where this would be a burden for the platform owner to service their customers. In this case, the platform will action on denied any access from particular origin.

Web scraping is a method that is greatly used in this project since data understanding phase until the end of the system usable lifetime. Every phase in the methodology, even after deployment and the data collection to update users in the media monitoring dashboard involve web scraping. The target news media in this project is The Star Online, and hence the scraping method is configured in the way to scrape news from the site.

4.2 Tools to Use

1. News API

News API is an easy-to-use and simple API that make request on multiple online sources and returns JSON metadata for the news information recently published on a range of news sources and blogs. In this system, News API is the tool for scraping news data from online news source. It extracts news data according to specific date-time, scraping source and keyword in the content.

2. Python 3

Python is a general-purpose programming language. It is an object-oriented, high-level programming language and incorporated dynamic semantics mainly used for website and application development. Both the standard library and the interpreter are accessible without extra charges, in source and binary form. This language is mainly used for data preparation and modelling in this project, where from news scraping to model training, even for the automated news extracting process.

3. Python Script (PY File)

A PY file is a script or program file written in Python language, an interpreted object-oriented programming language. It can be generated and edited with a text editor such as notepad application, but requires a Python interpreter to run. PY files are often used to program web servers and other administrative computer systems. PY notebook is used for automation source code in this system. The automation is coded in this format and scheduled to run by python.exe, in order to scrape updated news articles daily.

4. RStudio

RStudio is an IDE for coding R language. It supports direct code execution by including syntax-highlighting editor, a console, as well as tools for history, plotting visualisation and graphs, workspace management and debugging. This development environment is mainly used during development phase, developing the data dashboard and plotting visualization that can directly interact by users. It allows developer to publish their project or system online and accessed the program by URL.

5. ShinyDashboard

Shiny is an R package that is convenient to develop interactive web apps directly from R. A dashboard or standalone application on a webpage can be developed by using Shiny, with the support of CSS themes. In this project, all data dashboard development and design is done through this application, using the build-in library to help displaying data table and plotting visualisation graphs and charts.

6. R Language

R is a language and environment for graphics and statistical computing. R offers a wide variety of graphical techniques and statistical for example linear and nonlinear modelling, classical statistical tests, etc., at the same time is highly extensible. This language is mainly used in RStudio IDE, to develop and design user interactive data dashboard.

Chapter 5 System Implementation and Testing

5.1 Modelling

5.1.1 Filtration

5.1.1.1 Machine Learning Model

The system is initially implemented with machine learning model. Hence, few machine learning model is trained and compared with their result accuracy on predicting the testing dataset. The model chosen includes Naïve Bayes Classifier, Logistic Regression Classifier, Support Vector Machine and Random Forest Classifier. Their accuracy of training with prepared dataset is as figure below.

```
NB: 0.9069767441860465
LR: 0.9069767441860465
SVM: 0.5116279069767442
RF: 0.8372093023255814
```

Figure 5.1: Accuracy Score of Filter Machine Learning Model.

Naïve Bayes Classifier and Logistic Regression Classifier performs well in binary classification of filtering news by content, with accuracy score of 90.67%.

5.1.1.2 Deep Learning Model

```
Train on 227 samples, validate on 57 samples
Epoch 1/10
- 3s - loss: 0.6891 - acc: 0.5330 - val_loss: 0.6574 - val_acc: 0.7544
Epoch 2/10
- 2s - loss: 0.6680 - acc: 0.5330 - val_loss: 0.6315 - val_acc: 0.7544
Epoch 3/10
- 2s - loss: 0.6281 - acc: 0.5727 - val_loss: 0.5875 - val_acc: 0.7895
Epoch 4/10
- 2s - loss: 0.5538 - acc: 0.7269 - val_loss: 0.5184 - val_acc: 0.8070
Epoch 5/10
- 2s - loss: 0.4381 - acc: 0.9471 - val_loss: 0.4488 - val_acc: 0.8246
Epoch 6/10
- 2s - loss: 0.3003 - acc: 0.9736 - val_loss: 0.3958 - val_acc: 0.8596
Epoch 7/10
- 2s - loss: 0.1822 - acc: 0.9912 - val_loss: 0.3635 - val_acc: 0.8596
Epoch 8/10
- 2s - loss: 0.1061 - acc: 1.0000 - val_loss: 0.3439 - val_acc: 0.8772
Epoch 9/10
- 2s - loss: 0.0635 - acc: 1.0000 - val_loss: 0.3386 - val_acc: 0.8772
Epoch 10/10
- 2s - loss: 0.0405 - acc: 1.0000 - val_loss: 0.3290 - val_acc: 0.8772
Accuracy: 0.9754
```

Figure 5.2: Training Verbose and Accuracy of Filter Deep Learning Model.

Figure above shows the training verbose and accuracy of deep learning model trained with prepared dataset. The model is trained with dataset batch size 4, 10 epochs and

Adam optimizer. Both of the training loss and validation loss keep decrease throughout the training process as shown, indicates that the model is actually learning the training dataset fed to it. The test accuracy shown at the bottom part of the figure, 97.54%, is higher than the accuracy score of any machine learning model, shows that deep learning model can learn train dataset better. Hence, deep learning model is implemented to solve the binary classification problem.

5.1.2 Categorisation

5.1.2.1 Machine Learning Model

Similar to Filter model, 4 machine learning models are initially chosen to train and compare the accuracy score perform on prepared dataset with 8 defined labels. Figure below shows the accuracy scores of those models.

```
NB:  0.4444444444444444
LR:  0.5
SVM: 0.3888888888888889
RF:  0.2777777777777778
```

Figure 5.3: Accuracy Score of Categorise Machine Learning Model.

Machine learning don't really perform well as shown above as the highest accuracy score among 4 models is just 50%. This might due to limited dataset prepared for multi-classification problem lead to the fact that the model doesn't really learn well.

5.1.2.2 Deep Learning Model

```
Train on 95 samples, validate on 11 samples
Epoch 1/10
- 4s - loss: 2.2675 - acc: 0.2632 - val_loss: 2.0121 - val_acc: 0.1818
Epoch 2/10
- 2s - loss: 1.5844 - acc: 0.6316 - val_loss: 1.7148 - val_acc: 0.5455
Epoch 3/10
- 2s - loss: 0.9196 - acc: 0.7895 - val_loss: 1.7985 - val_acc: 0.5455
Epoch 4/10
- 2s - loss: 0.5915 - acc: 0.8000 - val_loss: 1.6320 - val_acc: 0.5455
Epoch 5/10
- 2s - loss: 0.6827 - acc: 0.7789 - val_loss: 1.5937 - val_acc: 0.4545
Epoch 6/10
- 2s - loss: 0.3176 - acc: 0.8632 - val_loss: 1.6660 - val_acc: 0.4545
Epoch 7/10
- 2s - loss: 0.3324 - acc: 0.8632 - val_loss: 1.6666 - val_acc: 0.4545
Epoch 8/10
- 2s - loss: 0.4671 - acc: 0.9158 - val_loss: 1.6428 - val_acc: 0.4545
Epoch 9/10
- 2s - loss: 0.5515 - acc: 0.8737 - val_loss: 1.4128 - val_acc: 0.4545
Epoch 10/10
- 2s - loss: 0.4824 - acc: 0.8947 - val_loss: 1.4925 - val_acc: 0.4545
Accuracy: 0.5000
```

Figure 5.4: Training Verbose and Accuracy of Categorise Deep Learning Model.

The performance of deep learning model is similar to machine learning model in multi-classification case. Observing the training sample, it is much more lesser compare to training Filter model. The training and validation loss traced rise and drop indicates that the model is over-fit by the dataset prepared. The model is finally implemented to the system with the accuracy of 50%

5.2 Dashboard

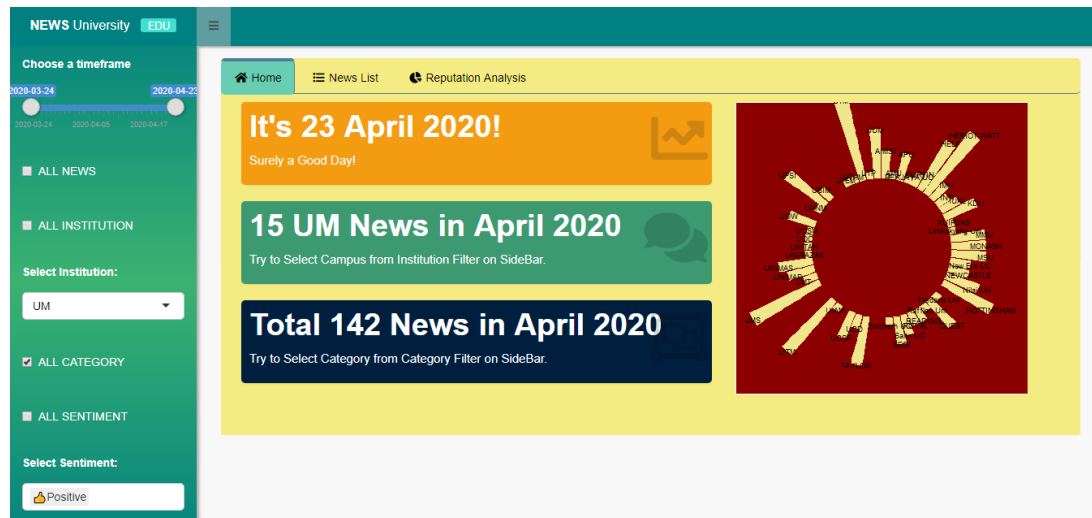


Figure 5.5: Home Page of Dashboard.

Dashboard is the final product of this project. The homepage dashboard is shown figure above and it is how user will see once they access to the dashboard. The value boxes display number of news categorised by each filter mention in the captions, by controlling the filter on the side bar. The visualization on the right hand side is for design purpose, which will vary and react to the total news stored according to each campus.

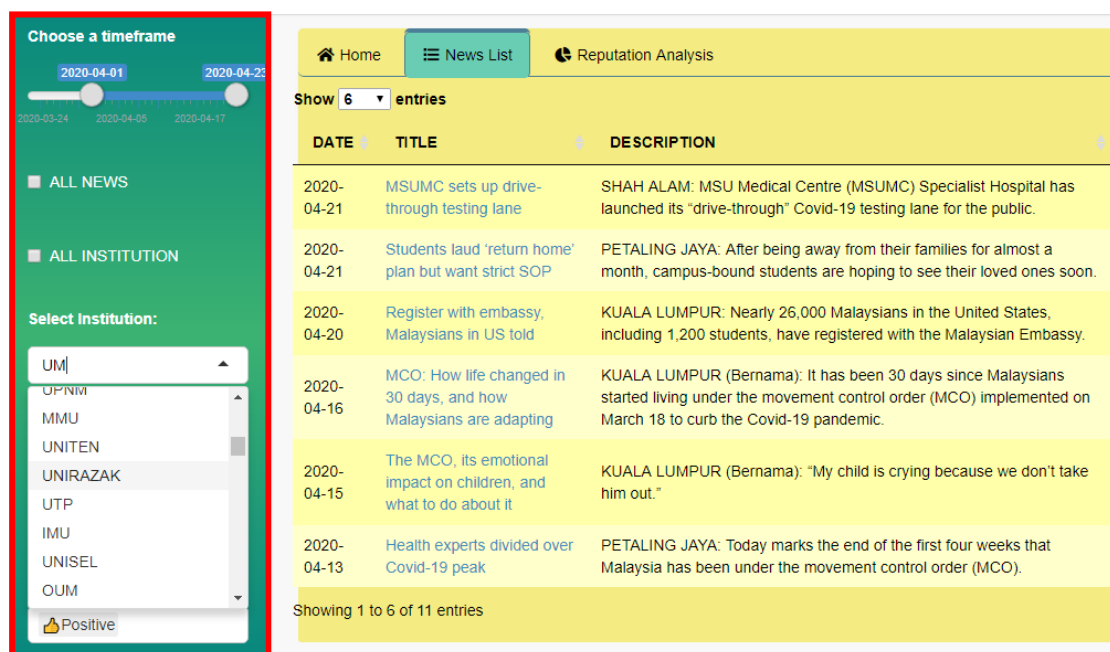
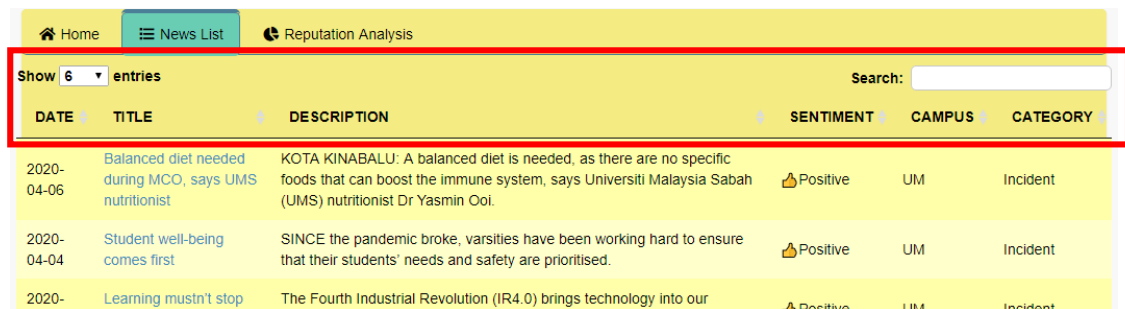


Figure 5.6: Side Bar Filter.

Red box in the figure above shows the sidebar menu of the dashboard which able to control by user. The item in the sidebar menu can determine how the dashboard body

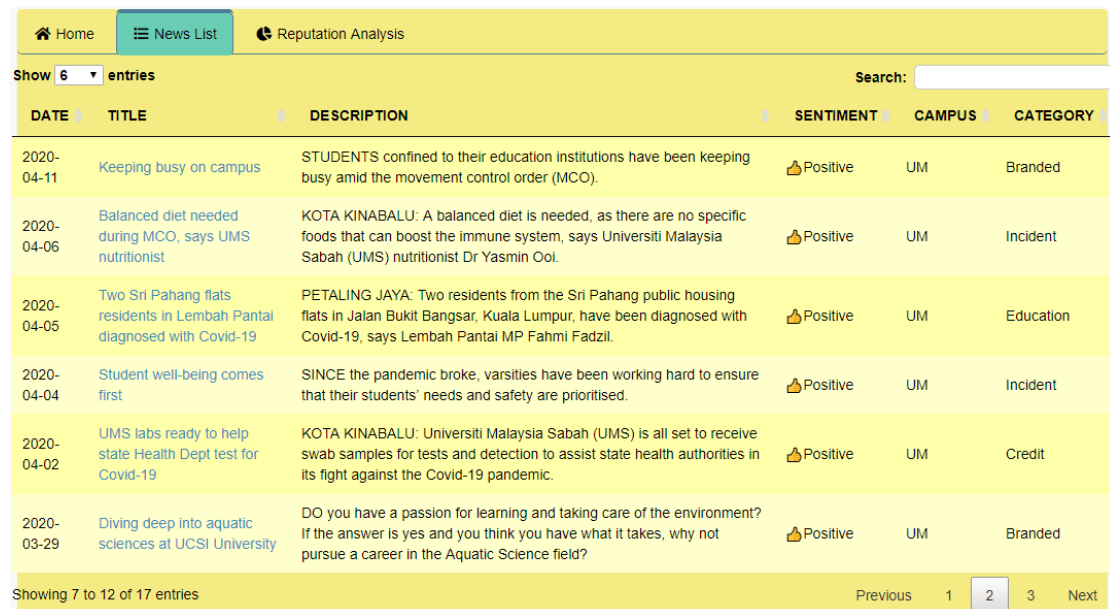
display the articles information. From figure 5.5, there are 4 main filter to be controlled, which is timeframe, institution name, category and sentiment. The select box for each category with 'All' if selected, will collapse the filter option so that users can ignore the filter that is not important to them, and hence the news information displayed in data table will not be filtered out by the particular filter option. User can manage any or all the inputs to search for the news they are most interested to at the ease.



DATE	TITLE	DESCRIPTION	SENTIMENT	CAMPUS	CATEGORY
2020-04-06	Balanced diet needed during MCO, says UMS nutritionist	KOTA KINABALU: A balanced diet is needed, as there are no specific foods that can boost the immune system, says Universiti Malaysia Sabah (UMS) nutritionist Dr Yasmin Ooi.	Positive	UM	Incident
2020-04-04	Student well-being comes first	SINCE the pandemic broke, varsities have been working hard to ensure that their students' needs and safety are prioritised.	Positive	UM	Incident
2020-	Learning mustn't stop	The Fourth Industrial Revolution (IR4.0) brings technology into our	Positive	UM	Incident

Figure 5.7: Sorting and Search.

Besides customizing information display by the input on the sidebar menu, users can choose to use the search function on the top right of dashboard body, to directly find for certain news. The search function is applicable on every column of the data table. The other feature of the dashboard is the column name above the data table. Every column can be sorted ascending or descending by clicking the particular column name.



DATE	TITLE	DESCRIPTION	SENTIMENT	CAMPUS	CATEGORY
2020-04-11	Keeping busy on campus	STUDENTS confined to their education institutions have been keeping busy amid the movement control order (MCO).	Positive	UM	Branded
2020-04-06	Balanced diet needed during MCO, says UMS nutritionist	KOTA KINABALU: A balanced diet is needed, as there are no specific foods that can boost the immune system, says Universiti Malaysia Sabah (UMS) nutritionist Dr Yasmin Ooi.	Positive	UM	Incident
2020-04-05	Two Sri Pahang flats residents in Lembah Pantai diagnosed with Covid-19	PETALING JAYA: Two residents from the Sri Pahang public housing flats in Jalan Bukit Bangsar, Kuala Lumpur, have been diagnosed with Covid-19, says Lembah Pantai MP Fahmi Fadzil.	Positive	UM	Education
2020-04-04	Student well-being comes first	SINCE the pandemic broke, varsities have been working hard to ensure that their students' needs and safety are prioritised.	Positive	UM	Incident
2020-04-02	UMS labs ready to help state Health Dept test for Covid-19	KOTA KINABALU: Universiti Malaysia Sabah (UMS) is all set to receive swab samples for tests and detection to assist state health authorities in its fight against the Covid-19 pandemic.	Positive	UM	Credit
2020-03-29	Diving deep into aquatic sciences at UCSI University	DO you have a passion for learning and taking care of the environment? If the answer is yes and you think you have what it takes, why not pursue a career in the Aquatic Science field?	Positive	UM	Branded

Figure 5.8: News Information.

Figure above shows the details of news articles displayed News List tab, which includes published date, news headlines, news description, sentiment of news content, institution

name and the category of news data. News headlines are attached with address link of the news, which can directly link user to the webpage of the news just by a click.

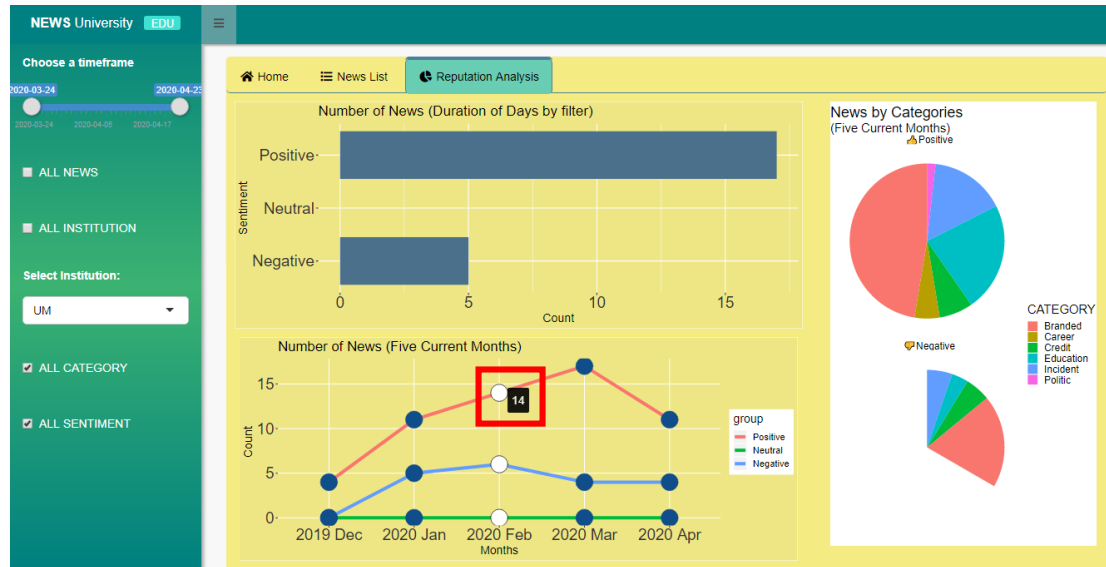


Figure 5.9: Reputation Analysis.

The above figure shows the visualisation in ‘Reputation Analysis’ tab. The top left bar chart shows the number of news according to each sentiment. This chart is reactive to all of the filter on the side bar except for sentiment filter as all number of news according to each sentiment is display to users. The bottom left line charts represents the number of news in the duration of 5 months. It is also displayed according to each sentiment label. This visualisation is reactive to institution and category filter. The last pie chart on the right shows the number of news according to each category, and is separated with sentiment label for each pie chart. This visualisation is only interactive to institution filter, as information of all other label is already plotted on the charts. The actual number or count of each plot can be shown as the mouse hover to the exact position of the point as shown in the red box.

The overall dashboard is built to gather news information for target user of this project. The deployment of this dashboard save so much time for them to quickly respond to the trend on media, in the same time keep themselves updated to new issues. The live dashboard for this project can be reached [here](#).

5.3 Implementation Issues and Challenges

Among the issues met during development, one of them is about the news sources. For current system implementation, by using News API to extract articles, only articles from The Star Online can be collected. The News API is actually mainly built for many other online news media from other countries but not really focus on online new source in Malaysia. Thus, Malaysia online news media is not in the scope of the sources provided by News API. However, The Star Online is successfully reached by using the feature domain, 'thestar.com.my', for URL to be requested and this method of requesting for news article is only success for The Star Online media. In this case, the development of this system will only focus on articles from The Star Online and this could effectively inhibit the practise of collecting redundant news articles with the same story from multiple news media.

Moreover, the deep learning model implemented in this project is not good enough by observing the accuracy scores and training process. The accuracy is high for Filter model binary classification but in real prediction in automated updating process, there is still many news that is not really related to or scoped in local higher education. Especially for categorisation model with multi-classification, the model accuracy is as low as 50%, with over-fitting case. This is due to limited dataset extracted during the implementation of this system, as well as the imbalance label on the dataset. For instance, Incident is one of the 8 labels defined to categorise news data, and it stands almost 50% after the dataset is manually labelled. Thus, model learn less to predict on other categories, at the same time over-fit on given dataset and will not predict accurately on new incoming data.

Chapter 6 Conclusion

6.1 Project Review

Implementation of the media monitoring dashboard is not really a new idea but should soon become a trend among higher education institution in Malaysia. Media monitoring benefit every business including higher education. There are many media monitoring software out there but most of them are too costly to subscribe, and difficult to control as even users have to be trained before using it. Therefore, local higher education institutions tend to employs staff to manually collect news article daily to update themselves with latest trends and events. In this case, the implementation of the automated media monitoring system and dashboard will effectively contribute on solving the problem.

The current progress on the project is already helpful for collecting news articles. The information will be updated daily in an automated process and user will be at the ease to only view all arranged data on the dashboard developed. The news data displayed on the dashboard can be categorised by filter, where users can interact with all designed filter so that only essence data will be display at a time.

For the analysis part, useful visualisation is designed and created so that users can glance through and get a holistic insight of the trends recently, without having to go through every single story uploaded by news media. Most importantly, users will not be left out from important insight when they have to respond or take action in urgent. The visualisation in final product dashboard is also reactive to the filter design, to let user have quick and multiple insight of the issues happened.

All objectives mention in Chapter 1 is achieved. This system is automated as the coded file is scheduled to run daily, extract, label and store the data itself to display all news data and visualisation on final product for users. Sentiment analysis also implemented right after the news is extracted, and hence visualization also includes graph and charts relate to sentiment and reputation of certain institution or in a whole.

6.2 Novelties

As observed, most of the local institution hired human to manually do media monitoring, or subscribe to a plan which hire online people update themselves with the news related to its industry. The final product they got might be much better than this implemented system in term of the accuracy of news given, but all of these practice takes times and is costly. Even some important news might be lost out if are extracted manually.

With this system, news articles is always updated whenever the coded script is executed, which means news will be updated all the time if the automated process runs frequently. In addition, interesting and useful visualisation created benefits target user to enlarge their insight of the trend, which this function is difficult and rarely found on human-hiring news updating system. But if only they did, this will be more work as they have to categorise and label every single news article all the time.

6.3 Future Work

Modelling part of this system has the most room of improvement in term of the accuracy works on their problem. The model should be fine-tuned with more data as possible to prevent overfitting and increase the accuracy. In the other perspective, network architecture of the model and hyper-parameters can be improved to suit the dataset and problem assigned to them. To conclude, more experiments can be made with the rise of amount of dataset to be trained.

BIBLIOGRAPHY

- Bernard, 2017. *Web Scraping and Crawling Are Perfectly Legal, Right?.* [Online]
Available at: <https://benbernardblog.com/web-scraping-and-crawling-are-perfectly-legal-right/>
[Accessed 4 August 2019].
- Editor, 2019. *Meltwater Review.* [Online]
Available at: <https://www.business.com/reviews/meltwater/>
[Accessed 2 August 2019].
- Essaid, R., 2018. *Is Web Scraping Illegal? Depends on What the Meaning of the Word Is*. [Online]
Available at: <https://resources.distilnetworks.com/all-blog-posts/is-web-scraping-illegal-depends-on-what-the-meaning-of-the-word-is-is>
[Accessed 15 August 2019].
- Glassman, N., 2011. *What Every Social Media Marketer Should Know About Meltwater Buzz Engage.* [Online]
Available at: <https://www.adweek.com/digital/socialmedia-apps-meltwater-buzz-engage/>
[Accessed 2 August 2019].
- Kaulback, M., 2016. *A brief history of media monitoring (and analysis).* [Online]
Available at: <https://www.agilitypr.com/pr-news/analysis/a-brief-history-of-media-monitoring-and-analysis/>
[Accessed 2 August 2019].
- Knight, J., 2016. *Product Review of LexisNexis Newsdesk.* [Online]
Available at: https://www.lexisnexis.com/documents/pdf/20160424041922_large.pdf
[Accessed 3 August 2019].
- Nain, Z., 2019. *Malaysia.* [Online]
Available at: <http://www.digitalnewsreport.org/survey/2019/malaysia-2019/>
[Accessed 13 August 2019].

BIBLIOGRAPHY

- Shai Vure, 2018. *How the University of Chester Boosts Exposure with Mention*. [Online]
Available at: <https://mention.com/en/customers/university-of-chester/>
[Accessed 4 August 2019].
- Wirth, R., 2000. *CRISP-DM: Towards a Standard Process Model for Data*. s.l.,
Semantic Scholar.

APPENDIX A: Poster



APPENDIX B: Final Year Project Biweekly Report
FINAL YEAR PROJECT BIWEEKLY REPORT
(Project I / Project II)

Trimester, Year: Y3S3	Study week no.: 2
Student Name & ID: Su Jia Sen 16ACB04566	
Supervisor: Dr Pradeep a/l Isawasan	
Project Title: A Media Monitoring Dashboard for University	

1. WORK DONE

-

2. WORK TO BE DONE

Define new keyword to extract news data, not restrict to institution name only so that won't miss out any important news.

3. PROBLEMS ENCOUNTERED

Many extracted news not related to local higher education.

4. SELF EVALUATION OF THE PROGRESS

Able to defined some new keyword that more relate to local higher education.



Supervisor's signature



Student's signature

FINAL YEAR PROJECT BIWEEKLY REPORT

(*Project I / Project II*)

Trimester, Year: Y3S3	Study week no.: 4
Student Name & ID: Su Jia Sen 16ACB04566	
Supervisor: Dr Pradeep a/l Isawasan	
Project Title: A Media Monitoring Dashboard for University	

1. WORK DONE

New keywords defined for news scraping

2. WORK TO BE DONE

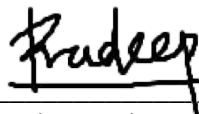
- Label all news article with their respective category including those collected during trimester break.
- Data preparation for model.
- Label all news with institution name if mentioned in article content.

3. PROBLEMS ENCOUNTERED

Huge amount of news content to be read, thus slow progress.

4. SELF EVALUATION OF THE PROGRESS

Have to spend more time on news labelling.



Supervisor's signature



Student's signature

FINAL YEAR PROJECT BIWEEKLY REPORT

(*Project I / Project II*)

Trimester, Year: Y3S3	Study week no.: 6
Student Name & ID: Su Jia Sen 16ACB04566	
Supervisor: Dr Pradeep a/l Isawasan	
Project Title: A Media Monitoring Dashboard for University	

1. WORK DONE

- All news able to labelled with campus name.
- Dataset preparation

2. WORK TO BE DONE

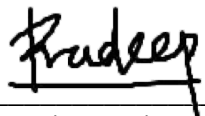
Build, train and evaluate machine learning model.

3. PROBLEMS ENCOUNTERED

Accuracy and performance of model is not ideal

4. SELF EVALUATION OF THE PROGRESS

Have to consider go into deep learning network architecture.



Supervisor's signature



Student's signature

FINAL YEAR PROJECT BIWEEKLY REPORT

(*Project I / Project II*)

Trimester, Year: Y3S3	Study week no.: 8
Student Name & ID: Su Jia Sen 16ACB04566	
Supervisor: Dr Pradeep a/l Isawasan	
Project Title: A Media Monitoring Dashboard for University	

1. WORK DONE

Few machine learning model trained.

2. WORK TO BE DONE


- Apply best machine learning model into system.
- Build, train and evaluate deep learning model architecture.

3. PROBLEMS ENCOUNTERED

- Deep learning seems to be not learning existing dataset at all.
- Insufficient dataset caused overfit.

4. SELF EVALUATION OF THE PROGRESS

Should put more time understand how deep learning works



Supervisor's signature



Student's signature

FINAL YEAR PROJECT BIWEEKLY REPORT

(*Project I / Project II*)

Trimester, Year: Y3S3	Study week no.: 10
Student Name & ID: Su Jia Sen 16ACB04566	
Supervisor: Dr Pradeep a/l Isawasan	
Project Title: A Media Monitoring Dashboard for University	

1. WORK DONE

- Machine learning model deployed in system.
- Deep learning model is trained.

2. WORK TO BE DONE

- Compare performance of deep learning model and machine learning model.
- Decide which model should be deployed in final system.
- Design final product dashboard with existing data.
- Prepare for reports

3. PROBLEMS ENCOUNTERED

Models doesn't seem performing well with limited dataset.

4. SELF EVALUATION OF THE PROGRESS

Should speed up progress.



Supervisor's signature



Student's signature

FINAL YEAR PROJECT BIWEEKLY REPORT

(*Project I / Project II*)

Trimester, Year: Y3S3	Study week no.: 12
Student Name & ID: Su Jia Sen 16ACB04566	
Supervisor: Dr Pradeep a/l Isawasan	
Project Title: A Media Monitoring Dashboard for University	

1. WORK DONE

- Dashboard is designed with interactive data and visualization.

2. WORK TO BE DONE

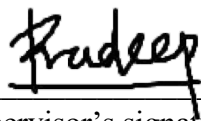
- Deploy deep learning model to system.
- Automate news scraping and labelling with use of trained model.
- Prepare for reports

3. PROBLEMS ENCOUNTERED

-

4. SELF EVALUATION OF THE PROGRESS

Should speed up in order to complete everything in time.




Supervisor's signature



Student's signature

APPENDIX C: Plagiarism Check Result



Originality Report

Processed on: 23-Apr-2020 16:50 +08
ID: 1305391315
Word Count: 9533
Submitted: 1

A Media Monitoring Dashboard for University

By Su Jia Sen

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

Document Viewer

exclude quoted exclude bibliography exclude small matches

mode: show highest matches together Change mode

ABSTRACT A news media monitoring dashboard benefits university on handling brand reputation by listening to public's opinion. The automated process of extracting and displaying news article on data dashboard eliminates manually searching for news article to do self update daily. A simple and holistic media monitoring dashboard should become a trend for local universities on taking advantage to compete in the industry. CRISP-DM is the main guide to implement the system. Keywords have to be defined and fine-tuned to extract the most accurate news articles related to local higher education that is useful for target users. News API request for news media sources and returns JSON metadata with details of the updated news. All information collected will be analysed and presented with various visualisation that is useful and simple to understand. Short and precise information of news articles, and the analysis of the data extracted will be display to users as the final product developed by one of the R packages, ShinyDashboard. The main idea is to include artificial intelligence model to accurately categorise data collected, to provide better visualisation in the form of data table and charts. Chapter 1 Introduction Problem Statement and Motivation The problem domain for this project is about Online Media Monitoring for local higher education which includes universities and colleges. In fact, marketing department of local higher institution practice on collecting articles from media manually to update themselves daily on trending issues about themselves and related to education industry. Most of them tend to spend excessive amount of time searching and collecting articles from different media. This would be due to the costly factor of existing monitoring software, or couldn't discover any convenient media monitoring tools that will benefits their monitoring practice. Background and Motivation Media monitoring is a practice involving reading and observing on a particular interested editorial content of media sources constantly. Media monitoring is traditionally introduced to capture editorial content, while nowadays is very useful on tracking

- < 1% match (student papers from 21-Apr-2009)
[Submitted to Republic Polytechnic](#)
- < 1% match (student papers from 19-Apr-2016)
[Submitted to University of Warwick](#)
- < 1% match (student papers from 18-Sep-2006)
[Submitted to University of Illinois, Chicago](#)
- < 1% match (student papers from 11-May-2015)
[Submitted to The University of Manchester](#)
- < 1% match (student papers from 26-Aug-2018)
[Submitted to University of Oxford](#)
- < 1% match (Internet from 16-Dec-2018)
<https://www.probytes.net/blog/top-100-python-interview-questions-and-answers/>
- < 1% match (publications)
[J. M. Mesa, C. Menendez, F. A. Ortega, P. J. Garcia, "A smart modelling for the casting temperature prediction in an electric arc furnace", International Journal of Computer](#)

A Media Monitoring Dashboard for University

ORIGINALITY REPORT

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

PRIMARY SOURCES

1	Submitted to Republic Polytechnic Student Paper	<1%
2	Submitted to University of Warwick Student Paper	<1%
3	Submitted to CSU, San Jose State University Student Paper	<1%
	Submitted to University of Illinois, Chicago	1

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 & COMMUNICATION TECHNOLOGY

Full Name(s) of Candidate(s)	Su Jia Sen
ID Number(s)	16ACB04566
Programme / Course	Bachelor in Computer Science (Hons)
Title of Final Year Project	A Media Monitoring Dashboard for University

Similarity	Supervisor's Comments (Compulsory if parameters of originality exceeds the limits approved by UTAR)
Overall similarity index: <u>2</u> % Similarity by source Internet Sources: <u>0</u> % Publications: <u>1</u> % Student Papers: <u>2</u> %	
Number of individual sources listed of more than 3% similarity: <u>0</u>	
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: _____



UNIVERSITI TUNKU ABDUL RAHMAN


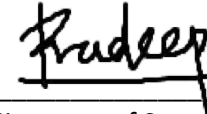
FACULTY OF INFORMATION & COMMUNICATION TECHNOLOGY (KAMPAR CAMPUS)

CHECKLIST FOR FYP2 THESIS SUBMISSION

Student Id	16ACB04566
Student Name	Su Jia Sen
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>(Signature of Student) Date: 24 April 2020</p>	<p>Supervisor verification. Report with incorrect format can get 5 mark (1 grade) reduction.</p>  <p>(Signature of Supervisor) Date: 24 April 2020</p>
---	---