**Assignment (10%) – Data Science Solution**

**Due date: Week 4, April 4th 2022 (Group 1)**

**April 6th 2022 (Group 2)**

The job of a data scientist is to solve problems. As a problem-seeker and a problem-solver, data scientist must be aware that not every problem can be solved using data science.

For this assignment:

1. You need to think and look at several types of today’s world problems that can be solved using data science approach.

**State the characteristics of problems which data science can solved**.

For the current situation, covid 19 is considered the world’s significant problem. The covid tracking

Problem: Covid 19

Characteristics of Covid 19 (Problem)which can be solved by data science

* Evolution of virus through different variant

Problem =Rapid evolution

* Data science can identify patterns of spreading covid cases
* Identify hospitalised cases
* Fatality ratio
* Precise location of a regional outbreak
* Forecast the cases number, next variant nature
* Vaccine nature, behaviour,charactertics

1. Think and reflect the kind of dataset(s) that you deal at your work place, in life or in your environment, which are in abundance. Try to make sense of this dataset and find ways how it can be useful to generate insights. **Provide detail descriptions the dataset. Discuss the problems that can benefit from the dataset and the questions.**

**Provide detail descriptions the dataset (when ,duration,size)**

Title, Text, Subject, Category, Platform, date,author, issue, originality ( Creator, Forwarder),state

The dataset contains two types of articles fake and real News . This dataset was collected from real-world sources; the truthful articles were obtained by crawling articles from Astro Awani, Bernama, The Star,etc

As for the fake news articles, they were collected from unreliable sources which published on sebenarnya.my portal. the Sebenarnya.my portal is for the public to check the authenticity of news spread through social websites, curb the spread of false news and to ensure that they received genuine news.

The dataset contains different types of articles on different topics, however, the majority of articles focus on covid 19 topics.

**Discuss the problems that can benefit from the dataset and the questions**

* **Better way curb fake news, such as targeted platform, population state,**

**Differentiate which dataset dominant ( - targeted category)**

Prediction on the issues for future reference. **Awareness**

1. Propose a specific **data product** that you would like to create based on the problem discussed above. Name the data product, provide a short description of it, benefit, usefulness, and identify the beneficiaries.

**Data Product:** Fake News Detection.

**Description:**

The sharing of fake news is an offence under section 233 of the Communication and Multimedia Act 1998, which provides for a maximum fine of RM 50,000 or imprisonment not exceeding one year or both and a further fine of RM 1,000 for each day the offence is continued after conviction.

Countering fake news has been a big challenge for the Malaysian Communication and Multimedia Commission (MCMC). The MCMC complaint bureau team received all sorts of false news, especially those related to covid-19 , that were spread on social media and whatsapp application.

Thus, increasing the need for tools able to provide insights into the authenticity of online content. This data product aims to present an insight into the characterization of the news story detection approach based on the text-based analysis. The method is to generate result in form of percentage accuracy of the fake news circulated over the internet.

**Benefit -**

**Usefulness**

The portal sebenarnya.my developed by the MCMC, it relies solely on official verification and refutation made by the respective authoritative bodies. The data product will make this process more efficient and fast by identifying the news without liaise with the respective agencies.

**Beneficiary**

MCMC, PDRM, Public

**Soft Skill evaluation for this assignment:** Critical Thinking and Problem Solving (CT&PS). See CT&PS rubric below.

Length: 3 (minimum) – 6 (maximum) pages.

On the top right corner of your assignment, write your name (matric number) and a passport size photograph of yourself.

Name your file using your own name, e.g., “Salimah” and save as pdf file.

Never submit your work in compressed file, like zip file, rar file, etc.

Any **plagiarized** work will be given a zero.

Make sure you **cite every single reference** (following APA format) used in the assignment.

Submit your assignment in Spectrum for Assignment latest by the due date.

Submission after the due date is considered as late submission, and mark will be reduced by 5%.

--------\*\*\*\*\*--------

A person in a suit and tie

Description automatically generated with low confidence

SARAVANAN A/L SUKUMARAN

S2155503

1. You need to think and look at several types of today’s world problems that can be solved using data science approach.

**State the characteristics of problems which data science can solved**.

For the current situation, covid 19 virus evolution is considered the world’s significant problem. This virus continuously replicates or generate copies of itself. These changes are called “mutations.” A virus with one or several new mutations is referred to as a “variant” of the original virus.[1] Due to the existence of multiple variants, there is a requirement for clinical study about effective vaccine development for all types of variants. Data science helps the researchers to develop a drug or vaccine by using datasets like antiviral data, anti-host protein data, toxicity data and vaccine data. Combining all these data into a target analytical biomolecule model will help the researchers drive insights to identify potential vaccines based on patterns within the data.[2]

Unemployment is seen as the biggest problem faced by graduates today. The root cause of this problem is the imbalance of the supply of graduates with the demand in the labour market. The job market is fast-changing, with new industries and skillsets generating demand for labour. With the data science approach, we can identify structural unemployment, like unemployment caused by a mismatch of skills of graduates out work and the skills required for the existing job opportunities. The data collection process consists of the latest employment datasets and graduates unemployment datasets. With the data, authorities can generate insight to create effective worker-management strategies by focusing to produce more graduates in demanding job markets.

1. Think and reflect the kind of dataset(s) that you deal at your work place, in life or in your environment, which are in abundance. Try to make sense of this dataset and find ways how it can be useful to generate insights. **Provide detail descriptions the dataset. Discuss the problems that can benefit from the dataset and the questions.**

The dataset I deal with at my workplace is MCMC Enforcement Case Management Dataset. This dataset consists of detailed information about social media crimes investigated by MCMC under section 233 Communication and Multimedia act 1998. The dataset is in XLS format, and size of 880Kb, duration 2012-2022.

This dataset consists of 18 columns of variable and 2014 rows of records. Every column describe a particular variable. And each row corresponds to a given member of the data set, as per the given question. The dataset describes variables such as

|  |  |  |
| --- | --- | --- |
| **Variables** | **Data Type** | Description |
| Year | Numeric | Crime committed year (2012-2022) |
| Case Number | Categorical | Enquiry or Investigation paper number |
| FIR Received Date | Numeric | Date started for investigation |
| Complainant Name | Categorical | Name of the person who raised complaint |
| Investigation Officer Name | Categorical | The person assigned for investigate the case |
| Suspect Name | Categorical | Suspect name based on profiling |
| Suspect Gender | Categorical | Suspect gender based on profiling |
| Suspect Race | Categorical | Suspect race based on profiling |
| Suspect IC Number | Categorical | Suspect ID based on profiling |
| Suspect Age | Numeric | Suspect age based on profiling |
| Suspect Occupation | Categorical | Suspect occupation based on profiling |
| Offence Location | Categorical | 14 states of Malaysia |
| Offence Element | Categorical | 5 elements (Indecent, Obscene,Menacing, False, Offensive) |
| Platform | Categorical | Social Medias like FB, Twitter,TikTok, IG,etc |
| Issue | Categorical | Royal, Religious, Racism, Covid 19, Politics |
| Offence Section | Categorical | Section 233 AKM 1998, Section 505(b) penal code |
| Timeline | Numeric | Timeline 3 months to complete a case |
| Case Status | Categorical | Final decision ( Charged, Compounded, Warning Letter, Ongoing, NFA) |

Each row of the dataset collects data point of an observation at a given time. Based on this

The insights that can draw from this dataset are

* Identify the number of cases by years
* Identify common platforms used for this social media crimes
* Performance Tracking Investigation Officer based on timeline
* Status of cases

Problems Benefit

* Using different types of data variables such as the suspect’s age, gender, race, occupation and location can help in predicting the target group. Based on the obtained data, the regulator can approach the target group of people and raise awareness of the misuse of social media.
* Adding variables like platforms in the dataset can help in identifying the dominant platforms linked with their activities. This would encourage the regulator to engage with the service provider for immediate actions like removing the post, user profiling, and terminating the user account.
* Providing with the data variable called the issue, would help to predict which one is low priority issues and high priority issues. For high-priority cases, the regulator should act fast to curb the spreading of the content before it gets viral. For example, if the content is fake news, the regulators should verify the news and announced it to the portal for public reference.

Prediction on the issues for future reference. **Awareness**

1. Propose a specific **data product** that you would like to create based on the problem discussed above. Name the data product, provide a short description of it, benefit, usefulness, and identify the beneficiaries.

**Data Product** Social Media Crime Prediction

**Description:**

The misuse of social media is an offence under section 233 of the Communication and Multimedia Act 1998, which provides for a maximum fine of RM 50,000 or imprisonment not exceeding one year or both and a further fine of RM 1,000 for each day the offence is continued after conviction.

Countering social media crimes has been a big challenge for the Malaysian Communication and Multimedia Commission (MCMC). The MCMC complaint bureau team received all sorts of complaints regarding social media misuse, which were categorized as indecent, obscene, false, menacing, and offensive.

Thus, the increasing number of social media crimes requires a special tool able to provide insights so that the regulators can prepare with strategies to overcome the problems. As a result, a data product is proposed to make better predictions and decisions based on the insights.

This data product aims to predict social media crime rates, target suspect groups, the platform used, and focus issues.

* Social media crime rates – The data product can forecast the numbers of cases in the upcoming so that the Human resource department can plan for recruitment strategies.
* Target suspect groups – The data product is able to classify the suspects by age, gender, race, occupation, and location. Based on this categorization, identify which group has a high chance to do misuse social media in the future, then educate them about social media ethics.
* The Platform used – The data product can predict the most popular platform used for social media crimes. Based on the predictions, the regulators should sign the agreement with platform service providers to remove any content against the provision under section 233 CMA 1998 in order to operate in our country.
* Focus issue- The data product can determine which issues will be viral. The viral issue can be classified as high priority or low priority. High priority means publishes or circulates any content with intent to cause fear or alarm to public. The regulator will set guidelines on how to tackle the spreading of viral issues.

When discussing the benefits of this data product, this product helps to identify the forecast of social media crime rate. So that, the regulator is well prepared with the strategic plan like organizing awareness campaigns.

**The data product usefulness is mainly for the regulators to understand the current and future crime situations involving social media. With the complete data product, the regulator can improve collaboration with Law Enforcement Agencies (LEAs) which will make the process more efficient and fast to identifying the suspects.**

**The beneficiaries of this data product are MCMC, and PDRM. Beside MCMC, RMP also investigesting social media crimes. This data product can help both agencies to come out with better decision making in order to reduce social media crime cases.**

.

**REFERENCES**

**[1]** World Health Organization. (n.d.). *SARS-COV-2 evolution*. World Health Organization. Retrieved April 1, 2022, from <https://www.who.int/news> room/questions-and-answers/item/sars-cov-2-evolution

[2] Keshavarzi Arshadi Arash, Webb Julia, Salem Milad, Cruz Emmanuel, Calad-Thomson Stacie, Ghadirian Niloofar, Collins Jennifer, Diez-Cecilia Elena, Kelly Brendan, Goodarzi Hani, Yuan Jiann Shiun (2020). Artificial Intelligence for COVID-19 Drug Discovery and Vaccine Development.2624-8212.doi: 10.3389/frai.2020.00065