****

**CDS506 Research, Consultancy and Professional Skills**

**Semester 2, 2019/2020**

**Consultancy Project Proposal**

Mining social media data to understand pattern of social media postings regarding blood donation in Malaysia

|  |  |  |
| --- | --- | --- |
| **Author** | **Matric No.** | **Email Address** |
| **Chin Yi Xiang** | **P-COM0109/19** | **chinyixiang@student.usm.my** |

Table of Contents

[I) Abstract 1](#_Toc44072664)

[II) Background of Selected Company 1](#_Toc44072665)

[III) Introduction 1](#_Toc44072666)

[Problem Statement 1](#_Toc44072667)

[Research Question 2](#_Toc44072668)

[Objectives 2](#_Toc44072669)

[Benefit of the Project 2](#_Toc44072670)

[IV) Related Works 2](#_Toc44072671)

[Background 2](#_Toc44072672)

[Related Work of Data Science and Analytics Techniques 3](#_Toc44072673)

[Comparison between Data Science Techniques 4](#_Toc44072674)

[Related work and Comparison on Analytical Tools 4](#_Toc44072675)

[V) Proposed Methodology 7](#_Toc44072676)

[Proposed Solution 7](#_Toc44072677)

[Choice of Data Science and Analytic Techniques and Justification 8](#_Toc44072678)

[Choice of Analytical Tools and Justification 9](#_Toc44072679)

[Data Source 9](#_Toc44072680)

[Risks and Issues of the Project 9](#_Toc44072681)

[VI) Expected Outcome 10](#_Toc44072682)

[VII) Gantt Chart 10](#_Toc44072683)

[VIII) Conclusion 10](#_Toc44072684)

[IX) References 11](#_Toc44072685)

# I) Abstract

Blood Reservoir is one of the crucial resources in medical industry. Demand for blood supply is always high for various reasons including emergencies, surgeries, and treatment for blood-related diseases. The main source of blood reservoir comes from volunteer blood donors throughout the country. There are several ways to increase awareness of blood donation among Malaysia citizens. Social media postings play important role to reach most of the citizens. With close study of social media postings regarding blood donation in Malaysia, strategies can be designed to improve the efficiency of blood donation campaigns. In this research, two types of social media postings regarding blood donation will be examined: (a) postings by blood donation campaign organizers, (b) postings by general population. Webs Scraping techniques will be applied to obtain data on time and content of postings related to blood donation campaign in social media. The outcome of this study will provide a ground for analysts to perform analysis and design strategies based on current pattern of social media postings regarding blood donation in Malaysia.

Keywords: Blood donation campaign, Web Scraping, Social Media, Malaysia

# II) Background of Selected Company

The Advanced Medical and Dental Institute (AMDI) of USM is formed in 2002 with approved memorandum from Ministry of Education. AMDI focus on research and academics for novel and unconventional approaches and breakthroughs in medicine, dentistry, health sciences and tertiary healthcare services. AMDI consists of 6 clusters, namely Craniofacial & Biomaterial Science, Infectomics, Integrative Medicine, Lifestyle Sciences, Oncological & Radiological Science and Regenerative Medicine. Each cluster consists of specialists from respective fields for advanced research.

The Regenerative Medicine Cluster focus on development in regenerative medicine, a multidisciplinary field to establish applications for repairing, replacing and re-growing damaged tissues. The Regenerative Medicine Cluster excels in Stem Cell Therapy, Gene Therapy, Bioengineering and Immunology.

# III) Introduction

## Problem Statement

Blood donation is important for life-saving procedure at the hospital. Patients requiring blood transfusion ranges from those with hematological disease such as thalassemia, and those undergoing major operation. With an increasing demand for blood transfusion, hospital blood banks often face shortage of blood supplies. Apart from this, blood banks also need a continuous supply of blood donations because of the short lifespan of some blood components.

The world blood donor day is celebrated on 14th June every year. During this time, blood donation campaigns are heavily promoted in social media by campaign organizers. However, the level of engagements in social media among the general population in Malaysia has never been quantified. Understanding the pattern of social media posts and level of engagements by the general population will aid in formulating strategies for future blood donation campaigns. By analyzing the pattern of social media posting regarding blood donation by campaign organizers and general population, this will help in improving online promotion strategies to encourage public to donate blood.

## Research Question

1. How frequent is social media posting regarding blood donation by campaign organizers?
2. How frequent is social media posting regarding blood donation by general population?
3. Is there any pattern in number of social media posting regarding blood donation by both campaign organizers and general population?

## Objectives

1. To identify the pattern (frequency, platform used, and content) of social media postings regarding blood donation throughout the year by blood donation campaign organizers
2. To identify the pattern (frequency, platform used, and content) of social media postings regarding blood donation throughout the year by general population
3. To determine the peak time (which month of the year) of blood donation promotion in social media by blood donation campaign organizers and assess the level of engagements among the public

## Benefit of the Project

With this project, the frequency and pattern of social media postings regarding blood donation campaign can be quantified and recorded. The text mining model built from this project can be used as a data source for future analysis on similar topics. Besides, the prediction model built can help to predict the peak season of blood donation campaigns and allocate resource to maximize efficiency of social media advertisements.

# IV) Related Works

## Background

Blood donation is an important source of replenishing blood bank as well as handling emergency blood demand. Blood bank usually requires continuous amount of blood supply as blood components have limited lifespan. Among the important components, red blood cells (RBC) can live up to 42 days and platelets can only live up to 5 days after blood donation. ([*Blood Components*](#_ENREF_4)).

In the recent Covid19 cases, around 2000 bags of blood helped to save around 1000 patients daily. ([SULAIMAN, 2020](#_ENREF_15)). In the interview, Dr. Noor Hisham Abdullah, the Director General of Health Malaysia stated the great significance of blood donation campaigns in achieving this result.

Among the various ways of promoting blood donation campaigns, social media stands to be one of the most efficient and cost saving channels. In recent researches, it is found that social media platforms like Facebook, Twitter and YouTube has evolved from platforms of information sharing into platforms for influence, bringing revolution to the marketing, advertising, and promotion industries ([Hanna et al., 2011](#_ENREF_9)).

The Malaysia Government especially Ministry of Health Malaysia has been taking various initiatives to promote awareness of blood donation campaign among public, including intensive advertising (via mainstream television and radio channels), mobile blood transfusion service center, giving incentives to blood donors, establishment of donation suites and collaboration with other government institution ([Nur Hairani et al., 2018](#_ENREF_11)). However, study on effect and patterns of social media postings are still lacking.

## Related Work of Data Science and Analytics Techniques

One of the biggest problems faced by blood supply chain is the fluctuation in blood demand. Due to the short lifespan of blood components, sudden spike in demand may cause blood bank shortage which might result in death of patients.

To overcome this, researches have been carried out to predict blood demand. In year 2004, three time series analysis methods are used to forecast the red blood cell transfusion demand ([Pereira, 2004](#_ENREF_13)). The three methods used are seasonal ARIMA, the Holt-Winters family of exponential smoothing methods and Neural Network. The performance is measured by the coverage rate and the outdate rate. The best-fit model is identified to be seasonal ARIMA (0,1,1)(0,1,1)12 model.

Table : Forecasting Performance of time series methods on RBC transfusion demand over 1-year horizon

|  |  |  |  |
| --- | --- | --- | --- |
| Forecasting Method | ARIMA (0,1,1)(0,1,1)12 | Exponential Smoothing | Neural Network |
| Coverage Rate (%) | 89 | 91 | 86 |
| Outdate Rate (%) | 8 | 11 | 13 |

In a more recent study, it is found that the Box-Jenkins methodology performs well in demand forecasts for total blood demand (TBD) as well as singled out blood types with the exception of type A- with Mean Percentage Error (MPE) as low as 0.0002 ([Fortsch & Khapalova, 2016](#_ENREF_6)).

In another study, supply and demand data of blood banks in Ontario is collected for analysis and long term forecasting ([Drackley et al., 2012](#_ENREF_5)). This analysis gives a big picture of blood supply and demand relationship but not helpful in regulating blood supply in short term fluctuations.

In Malaysia, the Blood Action Team, formed in 2011 under directive of Director of National Blood Centre have been implementing some proactive approaches to overcome seasonal nature of blood demand ([Wooi Seong et al., 2014](#_ENREF_16)). The measures taken includes using new measures to recruit and retain blood donors, building a blood forecast system and collaborating with blood collection centers.

## Comparison between Data Science Techniques

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Research topic | Published Year | Best performing model | Accuracy Measure | Limitation |
| Three time series analysis methods on RBC transfusion demand | 2004 | ARIMA (0,1,1)(0,1,1)12 | Coverage 89%, Outdate Rate 8% | Accuracy of model drops drastically when applying to two-year horizon |
| Demand forecast of total blood demand and each blood type demand | 2016 | Box-Jerkins | MPE 0.0002 | Exception occurs for A- blood type. No supply analysis |
| Long term forecasting of blood supply and demand in Ontario | 2012 | Not mentioned | Not mentioned | Not much detail on models used. Unable to account for high frequency fluctuation |
| Malaysia Blood Action Team | 2014 | Not mentioned | Not mentioned | No supply analysis |

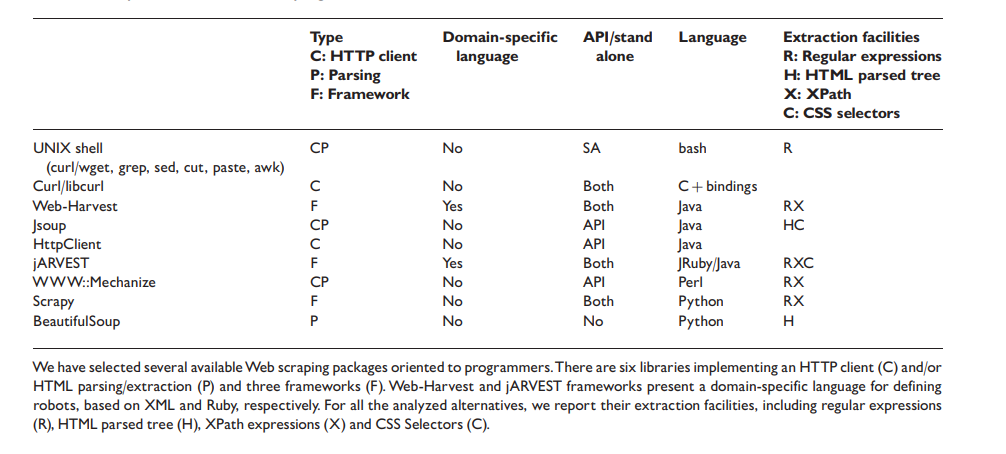
## Related work and Comparison on Analytical Tools

As a text mining project on social media platforms, web scraping plays a significant role in this project.

This book section summarizes some common issues faced in social media text mining as well as some examples of social media text mining ([Gundecha & Liu](#_ENREF_8)). Some of the issues highlighted in this section are (i) community analysis, (ii) sentiment analysis and opinion mining, (iii) influence modelling, (iv) information diffusion and provenance, and (v) privacy, security and trust.

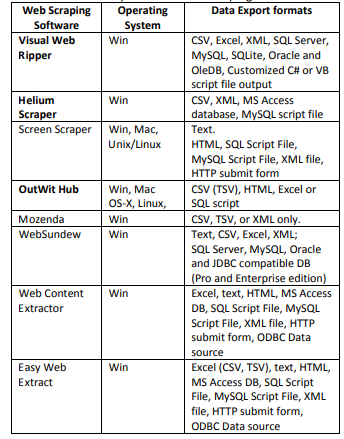
A review paper summarized the open-source web scraping libraries and frameworks in terms of type, domain-specific language, API compatibility, Programming Language used, and Extraction facilities ([Glez-Pena et al., 2014](#_ENREF_7)).

Table : Comparison of web-scraping libraries and platforms. Extracted from ([Glez-Pena et al., 2014](#_ENREF_7)) at 2020-06-26



In another review paper, the author compared different web-scraping software in terms of operating system and data export formats ([Sirisuriya, 2015](#_ENREF_14)).

Table : Comparison of Web Scraping Software. Extracted from ([Sirisuriya, 2015](#_ENREF_14)) at 2020-06-26

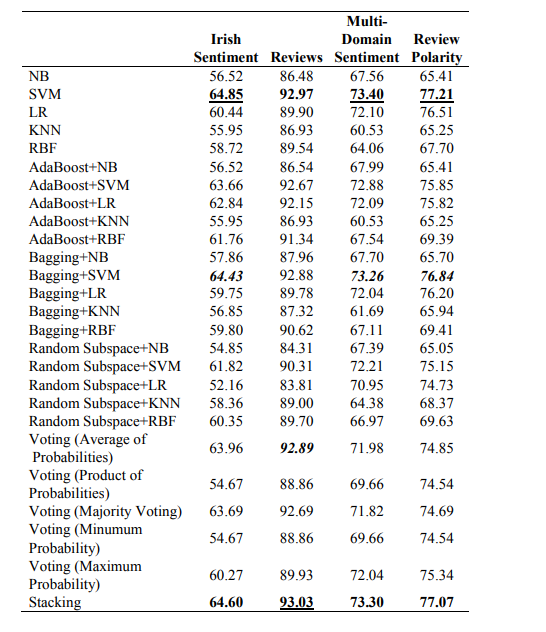


The other part that is crucial in this project is text mining, or more specifically, topic modelling. Topic modelling is a machine learning technique in natural language processing (NLP) which automatically assign labels to text documents.

One of the well-known algorithms of topic modelling is Latent Dirichlet Allocation (LDA). This algorithm is used to assign sets of latent topics to a document ([Blei et al., 2003](#_ENREF_3)).

In this paper, an empirical analysis is carried out on different combinations of LDA-based topic modelling with ensemble learning to improve accuracies ([Onan et al., 2016](#_ENREF_12)).

Table : Classification Accuracies of Sediment Analysis Dataset Using Different Combination of LDA-based Topic Modelling. Extracted from (Onan et al., 2016)at 2020-06-26.



However, one of the problems with LDA topic modelling is it suffering from “order effects”, where outcome of training might differ due to shuffling of training data. To fix this, LDADE, a search-based tool using Differential Evolution is introduced ([Agrawal et al., 2018](#_ENREF_1)). According to the author, the additional tuning dramatically increases the model stability.

There are also other algorithms proposed for topic modelling. In this paper, the author proposed a technique called Clustering-based Topic Modelling (ClusTop) which forms word network and determines the topics using community detection approach ([Lim et al., 2017](#_ENREF_10)). Using this algorithm, the number of parameters to be tuned is less as well as appropriate number of topics can be automatically determined.

# V) Proposed Methodology

## Proposed Solution

A close up of text on a black background

Description automatically generated

Figure : Methodology Flow Chart

The overall process is broken down into 4 stages. First stage is data collection. In this stage, Facebook pages/groups that are managed by blood donation campaign organizers or actively promoting blood donations such as: Kementerian Kesihatan Malaysia, Pusat Darah Negara Kementerian Kesihatan Malaysia, Persatuan derma darah Malaysia (Penderma), Pertubuhan Komuniti Penderma Darah, Persatuan Penderma Darah Kuantan, Persatuan Penderma Darah (S.P.U), and Blood Bank Hospital Umum Sarawak will be accessed and extracted postings in the time interval of 2015-01-01 till 2019-12-31. The posts extracted will be transformed and stored in JSON format where each “document” represents one post.

The second stage is data preprocessing. This stage is crucial for NLP to work. This stage includes the following processes:

1. Tokenization   
   Tokenization is the process where long streams of text are converted into tokens which are words stored separately. In this process, special characters and punctuations are also removed.
2. Removing words less than 3 characters  
   This step is basically to remove words that are not useful in the analysis. Short terms like “a”, ”an”, ”by”, ”on”, ”at” are removed from the tokens.
3. Removing Stopwords  
   Stopwords usually consists of determiners (the, a, an, another), coordinating conjunctions (for, an, yet, but, so) and prepositions (in, under, towards, after, before).
4. Lemmatizing words  
   Third person word changed to first person, verbs in past and future tenses are changed into present tense
5. Stemming words  
   Words are reduced into their root form. Eg. (management -> manage, donation -> donate)

The third stage is the topic modelling stage. In this stage, LDA-based algorithms including modified LDA will be used. The outcome of this stage will be a list of “documents” or posts with their corresponding topics. By doing this, we have successfully made the unstructured data from social media into structured data.

In the final stage, time series analysis will be performed on each topics and keywords as well as the total number of posts related to our interest. It can then be paired with blood transfusion data from National Blood Centre to see if the campaign serves the purpose of blood demand well.

## Choice of Data Science and Analytic Techniques and Justification

Time series analysis and forecasting will be done using a few different algorithms.

1. ARIMA or seasonal ARIMA.   
   ARIMA have been proven to be consistent and highly accurate model in time series analysis. Using ARIMA gives at least a reliable model to start with.
2. Box-Jerkin model  
   From the literature review, one of the paper found that Box-Jerkin model performs best ([Pereira, 2004](#_ENREF_13)). In this project, the model will be tested again to see how it performs in this situation.
3. Windowing and Machine Learning method  
   By windowing the time series, regression or ML models can be built to do forecasting of a time-series value.

The model built will be evaluated by their Mean Average Error (MAE) as well as Root Mean Squared Error (RMSE).

## Choice of Analytical Tools and Justification

1. Programming Language: Python  
   The whole project will be carried out in Python as it is the most common programming language used in data science projects. There are a lot of libraries available for web scraping, NLP, and time series analysis.
2. Web Scraping: Beautiful Soup and Scrapy  
   From the review paper ([Glez-Pena et al., 2014](#_ENREF_7)), both Beautiful Soup and Scrapy are python based and not domain-specific. This is good because it is easier to lean and use. Beautiful Soup covers HTML DOM extraction while Scrapy covers Regex and XPath extraction.

## Data Source

The data source used in this project are social media postings coming from mainly facebook pages such as Kementerian Kesihatan Malaysia, Pusat Darah Negara Kementerian Kesihatan Malaysia, Persatuan derma darah Malaysia (Penderma), Pertubuhan Komuniti Penderma Darah, Persatuan Penderma Darah Kuantan, Persatuan Penderma Darah (S.P.U), and Blood Bank Hospital Umum Sarawak.

Additional data source might be added depending on needs and availability. Some of the possible additional data sources are:

1. Blood Transfusion demand statistics
2. Actual Donor statistics – to see the conversion rate of campaigns
3. Blood donor infographics – for target segmentation of marketing campaign

## Risks and Issues of the Project

As the project involves web scraping data from social media pages, some of the risk are likely to happen:

1. Platform policy and privacy  
   The platform used in this project is predominantly Facebook. Although there is no legal consequences of web scrapping from Facebook, the latest terms and conditions on automated data collection requires written permission in order to do web scrapping ([*Automated Data Collection Terms*, 2010](#_ENREF_2)).
2. Platform dependency  
   Web scraping from social media platforms using XPath, HTML DOM and regex also suggest future problems as the web scraping algorithm will be designed to fit current design of the webpages. In any cases of change in User Interface of platforms used, the web scrapping model might not be functioning.

# VI) Expected Outcome

The expected outcomes of the project include:

1. A web scraping model to extract data from Blood Donation related social media postings.
2. A topic modelling model to identify topics for each post collected.
3. A time series model to analyze and forecast future trend of blood donation campaign postings on social media.
4. A detailed report of steps taken in the project to achieve the objectives as well as discussion on any new findings or potential development of the project.

# VII) Gantt Chart

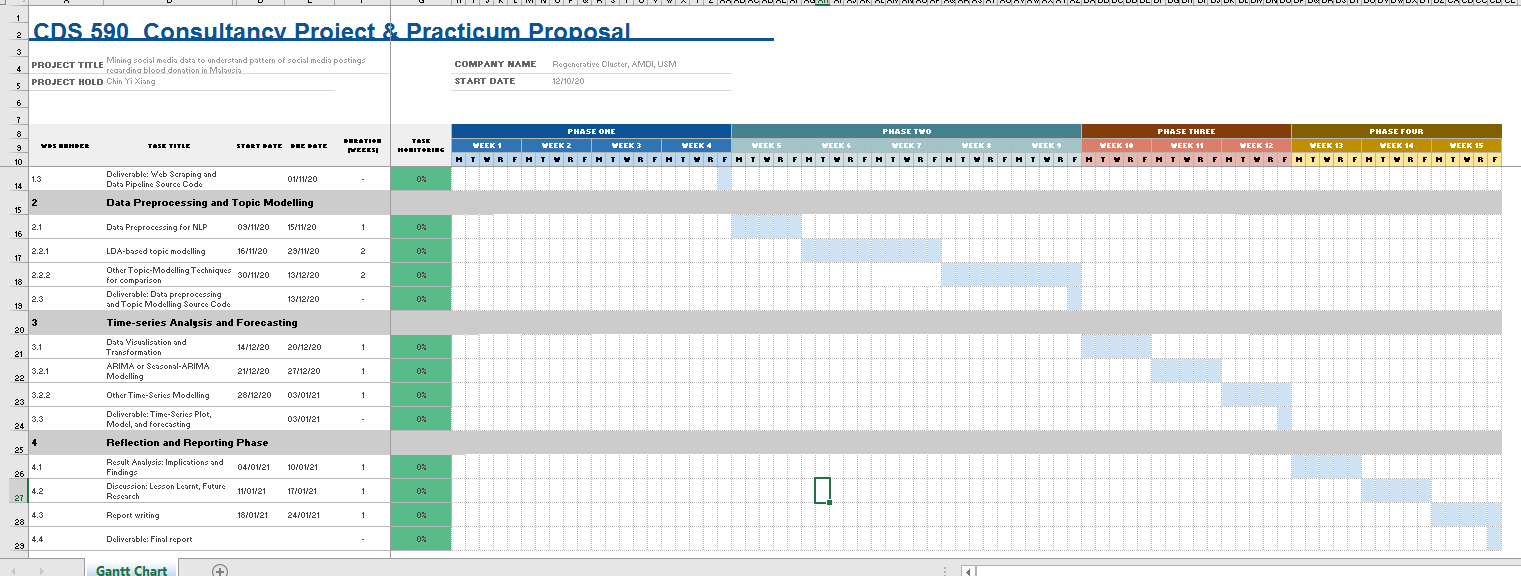
[](CDS590%20Documents/Gantt%20chart%20CDS590.xlsx)

Figure : Gantt Chart

Please refer to attached file “Gantt Chart CDS590.xlsx” for detailed description

# VIII) Conclusion

In order to understand the questions:

1. Frequency of social media postings regarding blood donation by campaign organizers,
2. Frequency of social media postings regarding blood donation by general population and
3. Peak time and any existing patterns in social media postings regarding blood donation

in Malaysia, a solution combining

1. web scrapping using Beautiful Soup and Scrapy,
2. topic modelling using LDA-based models, and
3. time-series analysis using ARIMA, Machine Learning, and Box-Jerkin model

is proposed. The project aims to provide an overview on usage of social media in promoting blood donation campaign in Malaysia so that further research can be done on optimizing resources and efficiency of the campaigns.

# IX) References

1. Agrawal, A., Fu, W., & Menzies, T. (2018). What is Wrong with Topic Modeling? (and How to Fix it Using Search-based SE). *Information and Software Technology*. <https://doi.org/10.1016/j.infsof.2018.02.005>
2. *Automated Data Collection Terms*. (2010, 2010-04-15). Facebook. Retrieved 2020-06-26 from <https://www.facebook.com/apps/site_scraping_tos_terms.php>

1. Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research, 3*(Jan), 993-1022.
2. *Blood Components*. American Red Cross. Retrieved 2020-06-26 from <https://www.redcrossblood.org/donate-blood/how-to-donate/types-of-blood-donations/blood-components.html>

1. Drackley, A., Newbold, K. B., Paez, A., & Heddle, N. (2012). Forecasting Ontario's blood supply and demand. *Transfusion, 52*(2), 366-374. <https://doi.org/10.1111/j.1537-2995.2011.03280.x>
2. Fortsch, S. M., & Khapalova, E. A. (2016). Reducing uncertainty in demand for blood. *Operations Research for Health Care, 9*, 16-28. <https://doi.org/10.1016/j.orhc.2016.02.002>
3. Glez-Pena, D., Lourenco, A., Lopez-Fernandez, H., Reboiro-Jato, M., & Fdez-Riverola, F. (2014). Web scraping technologies in an API world. *Brief Bioinform, 15*(5), 788-797. <https://doi.org/10.1093/bib/bbt026>
4. Gundecha, P., & Liu, H. Mining Social Media: A Brief Introduction. In *New Directions in Informatics, Optimization, Logistics, and Production* (pp. 1-17). <https://doi.org/10.1287/educ.1120.0105>
5. Hanna, R., Rohm, A., & Crittenden, V. L. (2011). We’re all connected: The power of the social media ecosystem. *Business Horizons, 54*(3), 265-273. <https://doi.org/10.1016/j.bushor.2011.01.007>
6. Lim, K. H., Karunasekera, S., & Harwood, A. (2017). Clustop: A clustering-based topic modelling algorithm for twitter using word networks. 2017 IEEE International Conference on Big Data (Big Data),
7. Nur Hairani, A., Muhammad Asri, M., & Mohammed Farhan, M. (2018). Blood Donation Program in Malaysia: Government Initiatives towards Attracting Volunteer Blood Donors. *International Journal of Engineering & Technology, 7*(4.15), 240-243.
8. Onan, A., Korukoglu, S., & Bulut, H. (2016). LDA-based Topic Modelling in Text Sentiment Classification: An Empirical Analysis. *Int. J. Comput. Linguistics Appl., 7*(1), 101-119.
9. Pereira, A. (2004). Performance of time-series methods in forecasting the demand for red blood cell transfusion. *Transfusion, 44*(5), 739-746. <https://doi.org/10.1111/j.1537-2995.2004.03363.x>
10. Sirisuriya, D. S. (2015). A comparative study on web scraping.
11. SULAIMAN, R. M. S. Z. (2020). Derma darah bantu 1,000 pesakit setiap hari. *Sinar Harian*. <https://www.sinarharian.com.my/article/87778/BERITA/Nasional/Derma-darah-bantu-1000-pesakit-setiap-hari>

1. Wooi Seong, K., Raffeal, V., & Ayob, Y. (2014). Adopting a proactive approach to blood shortages: experience from the National Blood Centre, Malaysia. *ISBT Science Series, 9*(1), 189-192. <https://doi.org/10.1111/voxs.12104>