

**Faculty of Information and Communication Technology**

BITI 2513

Introduction to Data Science

TASK 2

Fake News Detection Using Python

|  |  |
| --- | --- |
| Name | Matric No |
| Thitthima A/P Wat | B031810361 |
| Elle Aliz Binti Aminuddin | B031810295 |
| Muhammad Harraz Bin Harun | B031810426 |

Tutorial Group: 2BITI S1G2

Lecturer’s Name: Prof. Dr. Sharifah Sakinah Binti Ahmad

**Data Management Plan**

This is a project to determine either the news that we found is real or fake by using data science. The data that we will collect is fake news and real news. For this project we will use two type of data which are the data that we found it in the internet as a training data and the data we scrap it from many source website, Facebook, twitter and other social media. The first data is around 6000 data and the data for testing is around 300 data. For the first data, we took it on 23rd May 2020. For the second data, we used more time to get it since we need to scrap it. We find the second data and scrap it from 28th May 2020 until 8th May 2020.

In first data that we will use as training data set, there are only 3 variables which are title, text, and label. In second data, there are 5 variables which are title, text, subject, date and label. The reason why we take 5 variables for the second data is for us to do data analysis. If we us first data, we can only analysis the about text like the frequency of words in fake and real news and the pattern of sentences. So, to make the analysis more interesting, we use more variables.

For the first step, we will do the web scraping using Anaconda, Spyder and Jupyter Notebook. We use python as the language. We only scrap the title, it explanation date and category of the news. After that we will do data wrangling using python language. The software that we will use are Anaconda, Spyder, and Jupyter Notebook.

When we do the web scraping, we will save the data into txt format. After that we will import the data into excel (csv format) in data wrangling process. We choose to use the format cvs because it is easy for us to manage it. When other people want to use the data, they also easily to understand about the data. CSV format can be open in many kind of software like MS excel and WPS. So with this format, the data still can be use in future and will not out-of-date. For our backup of data, we plan to upload it at Google Doc. By using this we can retrieve the data anywhere and any times that we want. Other than that we can easily share the data with other people if we save in Google Doc.

To make our data more organize, we will undergo data wrangling process. We will handle the missing value and reduce the bias as much as possible to have a good data for our project. After that we will documented it in excel, so that we can easily use the data.

For storing, we will store the data in our machine(laptop). This is for short term and for the long term, we will store it in Google Doc. The reason why we choose to do like this because sometimes our machine lost data without any notification. So if we store it in Google Doc , so we can always retrieve the data in it and we can share the data with other people. In the data, we will have 3 variables which are title, text and label.

There are no data sensitive in our data. Since some of our data we took it from news website and and other source that all people can read it. For the data storage, it does not need any security since there are no sensitive data in it. Anyone can use our data. Since we will make the data by our self, so we will be the owner of the data. We will be the audience of the data since we need to use it for our project. All people can see and use the data apart from the creators. There will no an embargo period for our data since we are not selling the data. We find the data from open source so there will no embargo period.

We will ensure that the plan is followed. If there are any problems or any data updated, we need to review the data and update our plan. For example, if there are data that contain sensitivity, we need to get the policy and the permission to the owner to use their data.

**Example of web Scraping code**

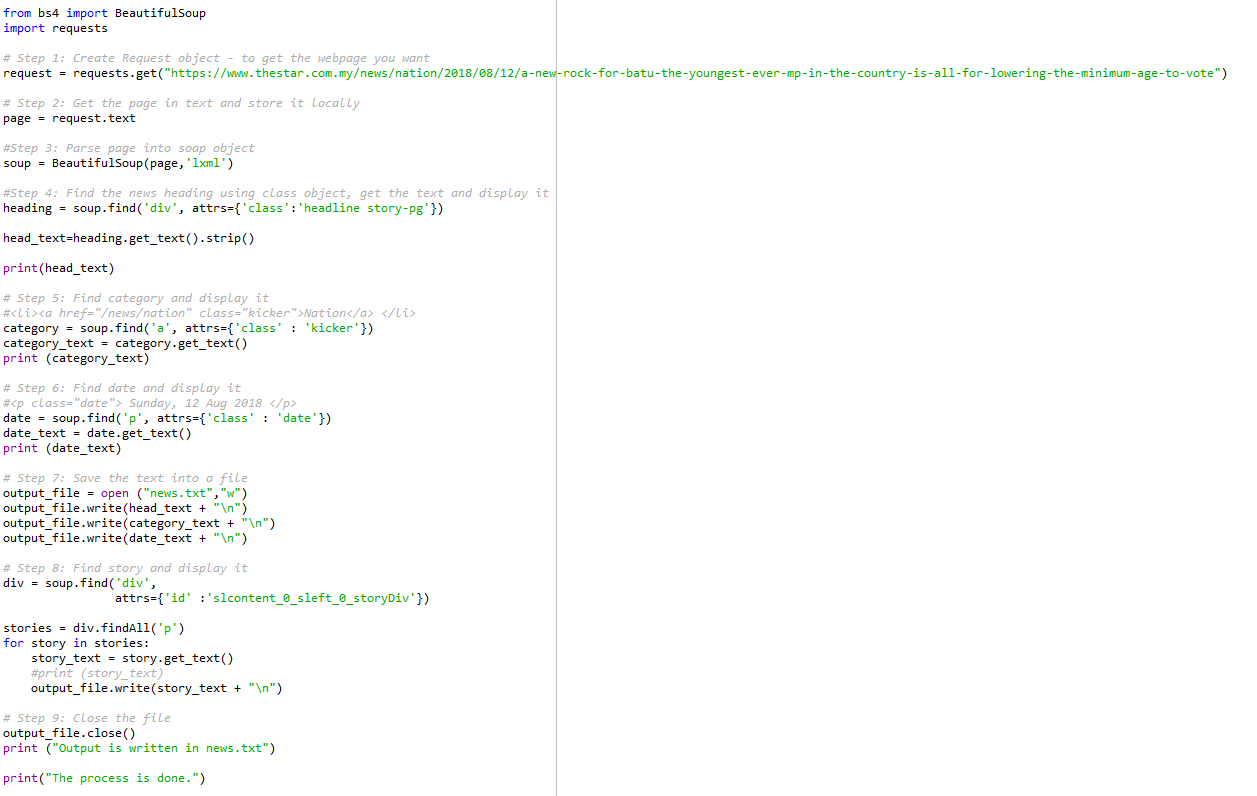


Figure 1: Example of Web Scraping Code

This is an example of the coding for web scraping. We use BeautifulSoup for the process. First, we need to import it. After that we need to request from the link. After that, we find the data that we want and scraping it. At the end we save it in txt form. We keep doing this with other data and some of them we are scraping it from a table.

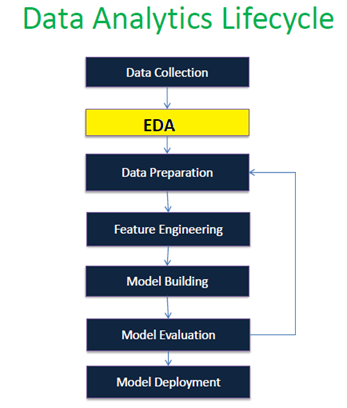


Figure 2: Flow of Project

This is the flow of this project. We start from collecting the data. After that we will undergo EDA process and data preparation process. After that we will build the model and evaluate it.

**How this plan will facilitate the needed analysis:**

Our project is about to predicting either the news is fake or real. We need a lot of data to help us to do the analysis and project. We need to plan how to handle with this data. We need to reduce the bias as much as possible and we need to do something with the data with the missing value. It is either remove it or just replace it with other value. Based on our data management plan, it will help us on which step should do first and what should we do.

When we follow the plan, we will get a quality data using a short period of times. We will waste more time during web scraping part but at least we know what next step we should do. So, for the analysis part, we just reduce the bias as much as possible and handle with the missing values in our data. We will undergo Exploratory Data Analysis for the data that we get. The process of exploratory data analysis will be more easily since the data that we have is already clean after following the plan.

For exploratory data process, we will do a basic statistical analysis for the numeric data like the date. For the text data, we will use the frequency of the word used to using word cloud to check it.

Below is the raw data statistical analysis:-

For first data:

Figure 3: Chart of data before cleaning process.

For second data:

Figure 4: Graph of real news.

Figure 5: Graph of fake news.

Above is the data before we do any process which looks quite messy. We can see that we have a bias data. Thus, we will handle it in the data wrangling part. After we clean the data, we will use it for the testing process.

**Data Wrangling**

Data wrangling refers to the process of cleaning, restructuring, and enriching the raw data available into a more usable format. This will help the scientist quicken the process of decision making, and thus get better insights in less time. This practice is being followed by many top firms in the field, partly owing to the benefits it has and partly because of large amounts of data which is supposed to be analysed. Organizing and cleaning data before analysis has been shown to be extremely useful and helps the firms quickly analyse larger amounts of data. Our data take mostly from website, twitter, Facebook and, etc.

The Steps of Data Wrangling:

1) Discovering

In this step, the data is to be understood more deeply. Before implementing methods to clean it, you will need to have a better idea about what the data is about. Wrangling needs to be done in specific manners, based on some criteria which could demarcate and divide the data accordingly.

2) Structuring

Raw data is given to you in a haphazard manner, in most cases, there will not be any structure to it. This needs to be rectified, and the data needs to be restructured in a manner that better suits the analytical method used. Based on the criteria identified in the first step, the data will need to be separated for ease of use. One column may become two, or rows may be split, whatever needs to be done for better analysis.

3) Cleaning

All datasets are sure to have some outliers, which can skew the results of the analysis. These will have to be cleaned, for the best results. In this step, the data is cleaned thoroughly for high-quality analysis. Null values will have to be changed, and the formatting will be standardized to make the data of higher quality.

4) Enriching

After cleaning, it will have to be enriched, it is done in the fourth step. This means that you will have to take stock of what is in the data and strategy whether you will have to augment it using some additional data in order to make it better. You should also brainstorm about whether you can derive any new data from the existing clean data set that you have.

5) Validating

Validation rules refer to some repetitive programming steps which are used to verify the consistency, quality and the security of the data you have. For example, you will have to ascertain whether the fields in the data set are accurate via a check across the data or see whether the attributes are normally distributed.

6) Publishing

The prepared wrangled data is published so that it can be used further down the line – that is its purpose after all. If needed, you will also have to document the steps which were taken, or logic used to wrangle the said data.

**Report finding from exploratory data analysis**

In data mining, Exploratory Data Analysis (EDA) is an approach to analysing datasets to summarize their main characteristics, often with visual methods. EDA is used for seeing what the data can tell us before the modelling task. It is not easy to look at a column of numbers or a whole spreadsheet and determine important characteristics of the data. It may be tedious, boring, or overwhelming to derive insights by looking at plain numbers. Exploratory data analysis technique has been devised as an aid in this situation.

For our project, we analysis about a fake news in general news. Fake news is written and published usually with the intent to mislead to damage an agency, entity, or person or gain financially or politically, often using sensationalist, dishonest, or outright fabricated headlines to increase readership. Main our project to detect the number of fake news after the data will be clean or wrangling. We have 2 data which is Data from link and Data scraping.

A screenshot of a cell phone

Description automatically generated

Figure 6: Bar Chart of Real News before Data Wrangling

A screenshot of a cell phone

Description automatically generated

Figure 7: Bar Chart of Real News after Data Wrangling

This bar chart shows before the data wrangling and after data wrangling, in this bar chart of real news have 176 datasets and 7 columns. We used date as data to know the number of real news. From both bar chart, the highest number of real news is 14 December 2017 which is 23 and the lowest number of real news is 2 which is in 6 November 2017, 12 December 2017, 24 December 2017, and 30 December 2017. Next, at the bar chart of before data wrangling, shows have a date which is not have number of real news, for example 8 November 2017. However, after data will wrangling, the bar charts show, the date does not have number of real news have been deleted. From both bar chart shows, the bias in bar chart before wrangling is more than the bias in bar chart after data wrangling.

A screenshot of a cell phone

Description automatically generated

Figure 8: Bar Chart of Fake News before Data Wrangling

A screenshot of a cell phone

Description automatically generated

Figure 9: Bar Chart of Fake News After Data Wrangling

For both bar chart, they indicate the number of fake news, where one is before data wrangling and after data wrangling. In both of bar chart have include 105 datasets and 7 columns. For fake news, we also use the date as a data to know the number of fake news. From bar chart before data wrangling show have 4 same highest number of fake news is 9. After the data will be wrangling, the bar chart shows only have 1 highest is 9 at 12/12/2017. After data will be clean or wrangling, it shows the date from 28/11/2017 until 10/12/2017 have been delete and bias will be decrease compared to bar chart of fake news before data wrangling.

A picture containing screenshot, drawing

Description automatically generated

Figure 10: The Label Percentage

From this label show percentage overall about the number of fake news and the number of real news. We can see label of orange as the percentage number of real news while label of blue as a the percentage number of fake news. Both of labels shows same percentage which is 50 %.

In conclusion, Exploratory Data Analysis is a crucial step to take before diving into machine learning or statistical modelling because it provides the context needed to develop an appropriate model for the problem at hand and to correctly interpret its results. EDA is valuable to the data scientist to make certain that the results they produce are valid, correctly interpreted, and applicable to the desired business contexts.

**Data Variables**



Figure 11: The Variables of the Data

The data we have has 5 different variables which are title, text, subject, date, and label. The title variable is self-explanatory as it is for the news title while the text variable is for the description of the news. The subject is to classify whether it is a Political News, US News, and others. The date taken for each data is the date the news was released on while the label is to classify the news as fake or real.

The variables used to predict the outcome of the label, whether it is fake or real, are the title and the text. This is possible by using the TfidfVectorizer. The TfidfVectorizer transforms text to feature vectors that can be used as input to estimator or in other words, it is used to convert a collection of raw documents into a matrix of 2 different pieces which are the Term Frequency (TF) and IDF (Inverse Document Frequency).

The TF is the number of times a word appears in a single document. When the TF obtains a high value, this would mean that the term appears more often that others which means the document is a good match when the term is part of the search teams. In this project, we declared the stop words in English for the TF. Stop words are words that are frequent but do not hold much meaning or thematic component within them. Stop words will only take up space in and take up valuable processing time thus they are needed to be filtered out before processing the natural language data.

Next, the IDF on the other hand is used as a measure of calculating how significant a word is in an entire corpus. This is done by calculating how many times a word appears on a set of documents. For this project we used a document frequency of 0.7. When both IDF and TF or the TfidfVectorizer is applied, the PassiveAggressiveClassifier comes to action.

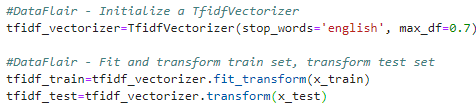


Figure 11: Sample coding for TfidfVectorizer

This classifier is a type of online learning algorithms. The PassiveAggressiveClassifier only turns aggressive when a miscalculation occurs to which it responds by updating and adjusting, otherwise it will remain passive. It does not converge as its purpose is only to make updates that correct the loss, which causes very little change in the norm of the weight vector.



Figure 12: Sample coding for PassiveAggressiveClassifier

With both of the TfifdVectorizer and the PassiveAggresiveClassifier in use, the data can be classified as fake or not based on the attributes stated before.

**Analysis and Prediction of Data to Form Hypotheses**

After the data is cleaned, the data is then inserted into a word cloud to calculate its most frequent word so we can analyse the trend of the data.

|  |  |  |
| --- | --- | --- |
|  | Frequency | Word |
| 1 | 100 | trump |
| 2 | 43 | video |
| 3 | 22 | president |
| 4 | 21 | watch |
| 5 | 16 | news |
| 6 | 15 | just in |
| 7 | 12 | obama |
| 8 | 11 | fbi |
| 9 | 11 | donald |
| 10 | 11 | democrat |
| 11 | 10 | fake |

Table 1: Word frequency of the Fake News Data

|  |  |  |
| --- | --- | --- |
|  | Frequency | Word |
| 1 | 62 | tax |
| 2 | 56 | trump |
| 3 | 45 | bill |
| 4 | 25 | house |
| 5 | 21 | republican |
| 6 | 20 | senate |
| 7 | 16 | vote |
| 8 | 15 | senator |
| 9 | 12 | republicans |
| 10 | 11 | twitter |

Table 2: Word frequency of the Real News Data

From the tables above, without including the word Trump which indicates a name, the highest for the fake news data is video. The possibility that a fake news has a video linked to it is high as this would act as ‘fish bait’ for people to click on the link of the fake news. The phrase “Just In” is also common as it also acts as a fish bait to viewers. Compared to the real news data, the words most frequent are considered as high-level words that are more formal. From this, we can conclude that any data with informal words and the word ‘video’ in it are most likely fake news while the ones with formal words are most likely to be real news.

A screenshot of a cell phone

Description automatically generated

Figure 13: Fake News Data after cleaning

A screenshot of a cell phone

Description automatically generated

Figure 14: Real News Data after cleaning

After both the data were cleaned so that the timelines match and the outliers are removed. From a rough observation, we can see that the fake news is more frequent when real news are not. This might be so the fake news will get more attention and recognition compared to on days where the real news is more frequent. This could also be because some fake news are just speculations made for a certain event, which is misleading and will give people the wrong idea, thus they are released earlier than the real news which needs further investigation and fact check before releasing it.

From the trends we observed, we can see that fake news uses the term clickbait so people will be attracted to the fake news, which will gain them more view while real news uses formal words. Fake news also is most likely to be released on the days real news are not though this observation can be inaccurate as the data we have might be randomly chosen without taking this into account.

Thus, we can expect these hypotheses to help us identify the classification of the data.