**PROJECT TITLE: NLP SENTIMENT ANALYSIS USING JUPYTER NOTEBOOK (PYTHON)**

Time: May 1st – May 8th, 2024

Tool: Jupyter Notebook, MS Word, Google Colab

Client: Healthcare Insurance Firms (Medibank, Medicare, Bupa, and HFB)

Data: provided by clients (in CSV format)

**Background:**

In June 2021, nearly 14 million Australians or approximately 54.3% of the population, had some form of private health insurance which was an increase of around 1.4% since June 2020. This is the first annual increase in the proportion of Australians with health insurance since 2015. Australian consumers paid almost $25.7 billion in private health insurance premiums in 2020–21. Meanwhile the Australian government spent $37.6 billion on medical services and benefits (MBS), comprised largely of Medicare and Private Health Insurance rebate expenses in 2020-21.

Unfortunately, there appears to be a division of opinion on the healthcare system in Australia. It has been postulated that these concerns have been shared widely on various social media platforms. The important of these concerns and the availability of data is a rare opportunity for data analytics specialists like yourself to apply AI and NLP techniques to generate actionable insights that can inform marketing strategy as well as public policy on the topic of Health Insurance.

Process:

Step 1: Data Import and Pre-assessment

* Importing provided datasets as dataframes
* Checking null values and dropping unnecessary columns

Step 2: Pre-processing

* Lowercase Transformation
* Removing multi-space and characters followed by @ sign
* Removing URLs
* Removing Punctuations
* Removing Duplicates
* Removing Digits
* Stemming
* Lemmatization
* Comparing stemmed tweets and lemmatized tweets

Step 3: Stop-word and Common words Removal

* Stop-word removal
* Common words and rare words removal
* Frequency Analysis

Step 4: Text Feature Extraction

* Bigrams
* Bag of Words
* Term Frequency – Inverse Document Frequency (TF – IDF)

Step 5: Temporal Analysis

* By Date
* By Time

Step 6: Sentiment over Time

* Sentiment on Overall Data
* Sentiment on Year
* Sentiment on Month

Step 7: Topic Modelling

* Generating Topic Models
* Interactive Topic Analysers

**Analytics Techniques and Insights**

The pre-processing steps have been conducted before moving to the main part of the analysis. First, it is important to drop unnecessary rows and columns, for example, removing rows with null values. Since this is a text analysis, there is no need to impute rows containing blanks. This will efficiently reduce the data size. Several pre-processing steps are also taken to clean the data upon applying techniques such as lowercase transformation, removal of multiple spaces and user id, site links, punctuation, duplication, and digits.

Next, stemming and lemmatization approaches are used to compare whether the original tweets, stemmed tweets, and lemmatized tweets are better for further analysis. Results show that even though lemmatization algorithms worked better than stemming in this case, it also changed some of the words in the tweets to other forms which may cause some confusion afterwards. Hence, the original tweets are kept.

Common words are removed from the tweets using the NLTK library. They are deemed irrelevant for NLP purposes because they occur frequently in the language, namely English. Thus, removing them is needed as a pre-processing step.

At this stage, frequency analysis is conducted. The most common words as well as the least common words are listed out. Visualizing the word cloud gives an overview of the word frequencies and infers some of the topics that customers are concerned about.

Based on the word cloud, some insights can be extracted are:

* Most concern is about private health insurance, showing by words such as 'private', and 'health'.
* Australian would also like to inquire more about health insurance services (membership, cost, cover policy). Words such as 'member', 'cover', 'need', 'send us', 'claim', and 'dm'.
* Customers are also paying attention to insurance and health services like Amp (Amplar), Medicare.

Please kindly refer to the Google Collab file for details of the pre-processing steps.

**Text Feature Extraction**

Unlike numerical datasets, text (NL) datasets contain words as their main input. Therefore, feature extractors are applied to transform the texts into numerical values so that modelling techniques can be deployed to make analysis and predictions. In this case, N-grams, TF-IDF, and Bag of words will be conducted consecutively to see the differences.

**Bigrams**

Bigrams, which are two words coming together in the corpus, the document is divided into a set of two co-occurring words. The shift is one-step forward.

The NLTK ngrams and word\_tokenizer libraries are used for this feature extraction.

The bigram function is then applied to tweets (after cleaning) and ends up with a tuple of words and their frequencies. Next, it is visualized as a bar chart displaying the top 20 bigrams and their corresponding frequency.

Through the bar chart, it suggests that there is a significant focus among Australians on private health matters and inquiries into related insurance. Additionally, the visual representation highlights a trend where customers reach out to Medibank seeking assistance or information regarding their health insurance. Furthermore, there is notable interest in social aspects associated with Medibank, such as CEO Craig Drummond and the Melbourne Marathon (Melbourne Marathon, 2018).

**Trigrams**

Like Bigram, Trigram is a technique of dividing documents into three co-occurring words. Since containing more words, the information in each text is richer.

Accordingly, the bar chart above shows some differences with the Bi-gram chart in terms of the most common texts. For trigrams, it is noticeable that the most common terms are about inquiries of health insurance details. Social topics such as Melbourne Marathon Festival and CEO Craig Drummond are also mentioned, and with a higher rank compared to that of bigrams. Moreover, there is positive feedback from customers about the Medibank program that could not be seen in the bigram chart. Instead, the bigram bar chart shows more concerns about insurance policy with texts such as health system, Gough Whitlam, and health funds.

**Bag of words**

This feature extraction mechanism will separate words in a text and describe the occurrence of words within a document. In this technique, CountVectorizer library is used, and the results will end up with a term-document matrix containing tweet id, word id (id of the word in the bag of words model dictionary), and word counts.

**Term frequency - Inverse Document Frequency (TF-IDF)**

To evaluate the importance of a word in a document of a collection, a numerical statistic called Term frequency – Inverse Document Frequency is proposed. This technique has two components: Term Frequency (TF) and Inverse Document Frequency (IDF).

In this case, the measurement of TF shows how often a word appears in the tweet. Term – Frequency (TF) is calculated by: 𝑇𝐹= 𝑁𝑢𝑚𝑏𝑒𝑟 𝑜𝑓 𝑡𝑖𝑚𝑒𝑠 𝑡ℎ𝑒 𝑤𝑜𝑟𝑑 𝑎𝑝𝑝𝑒𝑎𝑟𝑠 𝑖𝑛 𝑡ℎ𝑒 𝑡𝑤𝑒𝑒𝑡𝑇𝑜𝑡𝑎𝑙 𝑛𝑢𝑚𝑏𝑒𝑟 𝑜𝑓 𝑤𝑜𝑟𝑑𝑠 𝑖𝑛 𝑡ℎ𝑒 𝑡𝑤𝑒𝑒𝑡

Inverse Document Frequency (IDF), on the other hand, measures how unique a word is across the entire collection of tweets. IDF is measured as follows: 𝐼𝐷𝐹= log (𝑇𝑜𝑡𝑎𝑙 𝑛𝑢𝑚𝑏𝑒𝑟 𝑜𝑓 𝑡𝑤𝑒𝑒𝑡𝑠 𝑁𝑢𝑚𝑏𝑒𝑟 𝑜𝑓 𝑡𝑤𝑒𝑒𝑡𝑠 𝑐𝑜𝑛𝑡𝑎𝑖𝑛𝑖𝑛𝑔 𝑡ℎ𝑒 𝑤𝑜𝑟𝑑)

Hence, the TF-IDF is calculated by multiplying the TF and IDF.

**Temporal Analysis**

**Tweet by Date**

Temporal Analysis is an important part of this report since the tweets correspond to their date and time shared on social media. To observe the change of tweet counts by date, a visualization with x axis as date and y axis as tweet counts is generated.

Understanding the tweet counts over date draws an overall picture of customer engagement on social media and how they cope with specific events that lead to the high and/or low volume of tweets about Medibank.

There are periods of time that Medibank gathered a high level of social attention shown by the number of tweets by date. For instance, during September 2017, there was an exceptional record of tweet counts (over 160 a day). From 2019 to the middle of 2020, it also observed periods of considerable tweet counts, followed by the middle of 2021 before setting a lower intensity of media attention.

Events that were possibly associated with these phenomena could be when “The ACCC alleged Medibank made false, misleading or deceptive representations and engaged in unconscionable conduct in relation to its failure to notify Medibank’s, and its subsidiary ahm’s, members of its

decision to limit benefits for in-hospital pathology and radiology services, despite representing across several of its communication and marketing materials that it would” (Australian Competition and Consumer Commission, 2018). The COVID-19 pandemic occurred from 2019 up until 2021 involved in global health crisis.

**Tweet by Time**

Another temporal analysis is taken by separating the given time into Morning time, Afternoon time, Evening time, and Nighttime to see how customers engage with Medibank during hours of a day.

Assuming Morning time starts from 6AM to 1PM, Afternoon starts from 1PM to 6PM, Evening starts from 6PM to midnight, and Nighttime from midnight to 6AM, the tweet counts are distributed into 4 different bins.

This bar chart shows that most people tweeted about Medibank at nighttime, followed by morning and evening time. The least number of tweets was corresponding to noon time.

Though nighttime sees the most tweets about Medibank, mornings and evenings are active too. Capitalize on this by scheduling social media posts during these high-engagement times to maximize reach.

**Sentiment Analysis**

**Sentiment by Year**

Using the Textblob library, sentiment analysis is conducted to determine the attitude of the emotion of the users in terms of numerical values, whether it is positive, negative, or neutral.

Assigning another column to categorize the sentiment values into Positive (greater than 0), Negative (less than 0), and Neutral (equal to 0). Then, visualizing the distribution of sentiment using a bar chart to compare the general customer sentiment on Medibank.

Most of the tweets are positive and neutral. This suggests that Medibank marketing strategies and insurance policy were doing good and/or enough to reach customers. However, negative sentiment illustrates that the firm still needs to improve whether on their service quality or insurance policy.

Even though observing the customer sentiment distribution suggests a pattern of good customer satisfaction on Medibank, note that this only mentions the social media aspects, it is necessary to research whether the firm is doing their best to maintain customer satisfaction over time. Hence, an analysis of sentiment over date is conducted. The average sentiment is calculated by date and results are shown using a time series chart below.

There is a behaviour of volatility clustering over time and the high volatility tends to appear more frequently since the end of 2019. This might be associated with the health crisis during the time of pandemic. Thus, customer sentiment, represented by tweet sentiment, changed to a larger extent. This infers that Medibank was not coping well with customer’s needs, leading to low values of sentiment appearing more frequent in the recent time.

the assumption that customer satisfaction on Medibank services was dropping significantly over the years is supported. Sentiment decreased over time, except the period between 2017 and 2018. Also, the decreasing tendency was moving largely from 2019. There are three main events that could possibly be responsible for this phenomenon which were the pandemic, the case of Medibank and ACCC, and the data breach.

**Sentiment by Month**

An analysis of sentiment by Month is also provided to have a closer look at the average customer sentiment within a year time.

Sentiment by month showed an increasing pattern, however, fluctuated to a considerable extent. This suggests that customer sentiment might improve at the end rather than in the middle of the year. However, the marketing and customer service departments should try to improve some aspects to tackle with the uncertainty in customer sentiment during a year time.

**Topic Modelling**

Uncovering the main themes or topics presented in the collection of tweets will help understand the underlying patterns in tweets without labelling.

First, a corpus of tweets is created and tokenizing words of all the tweets. Then, a dictionary based on the tokenized words is generated and saved as a local file for LDA model to access.

Topic Modelling is now ready to fit the feature inputs, which is the TF-IDF from the above feature extraction steps. Top 10 topics are presented combining of keywords (6 keywords for each topic).

For example, a topic is represented as 0.024\*"health" + 0.017\*"private" + 0.013\*"not" + 0.010\*"cover" + 0.009\*"amp" + 0.008\*"time".

Based on the keywords in the topic and their weights, we can summarize it as the problem of a private health that is not covered.

Another example is 0.037\*"free" + 0.024\*"feelgoodprogram" + 0.019\*"parkrun" + 0.013\*"fitness" + 0.011\*"get" + 0.010\*"legend".

This may be assumed as the free parkrun program receiving good feedback from participants.

An interactive Topic Analyzer with the relevance metric of 0.6 is created. Based on the figure, some of the most common topics can be inferred.

* Topic 1 mainly involves in insurance policy.
* Topic 2 is about inquiries of customers on details of insurance services.
* Topic 3 discusses the payment, claim, conditions, and cover of health insurance.

**Conclusions and Recommendations**

In conclusion, this report analyses the customer tweets regarding Medibank with the deployment of several techniques including NLP, feature extraction, temporal analysis, sentiment analysis and topic modelling. The findings uncovered the insights from word frequency, customer engagement, Medibank’s social attention, sentiment scoring, and insurance policy.

Particularly, analysis of word frequency generally disclosed the overall attention of customers on Medibank. This related to various aspects of the firm including insurance policy, health cover details, and services. Sentiment analysis revealed the asymmetry between Medibank’s substantial customer base and the current downward trend in social sentiment associated with recent scandals and forensic events. Furthermore, the topic modelling shed light on customer inquiries, feedback on services, and perceptions of some programs.

Upon the gathered insights, Medibank should prioritize their marketing strategies to tackle with the turmoil in customer sentiment after their recent data breach, leading to a significant plunge in

sentiment scores. Considering proactive communication strategies as the high volume of customer inquiries signified the need for transparency in providing information and marketing campaigns. Besides, service quality should be improved to mitigate negative sentiment and enhance trust and loyalty. Additionally, monitoring of external events remains a good option to launch Medibank marketing strategies. For example, the parkrun program and marathon festival were positively contributed and received successful feedback from customers. Leveraging these events could attract more participants and new customers. Lastly, investing in data analytics will empower Medibank to extract actionable insights, identify emerging trends and drive informed decision-making to quickly respond to the Health Insurance industry.