BUS5CA Customer Analytics and Social Media
Semester 2 2020

Assignment 1
Social Media Analysis for Understanding Customer Preferences and Sentiments

Release Date: 4th August 2020
Due Date: 24th August 2020 @ 9:00am
Assignment Type: **Individual**
Weight: 30%
Format of Submission: A report (electronic form) and electronic submissions of project files (SAS project files and R scripts) in the LMS site.

Learning Objective:
The learning objective of Assignment 1 is to further develop your understanding and skills on social media analytics via performing analysis on two case studies:

1. **Case Study A**: you will work as a social marketing analyst in a consulting company to uncover the impacts of online advertising and communication with customers. The aim of the study is to educate the marketing teams of their clients (in diverse industries) to market their products and/or services on social media to maximise customers' involvement (positive interest and sharing). The company is interested in finding out the relationship between the keywords, shares, sentiments and whether there is a relationship in different topic categories such as entertainment, technology, business, etc. that are of interest to different clients in various industries.

2. **Case Study B**: you will be a data scientist working for a hotel review firm to develop a sentiment analytics engine for Twitter, which is used to predict consumers' review sentiments. The aim is to develop both dictionary-based and machine learning-based sentiment analytics scripts using a number of R libraries and SAS Sentiment Analysis Studio (covered in the workshop activities on Week 3 and Week 4). You are required to use the developed engine to predict hotel reviewers' sentiments and benchmark various algorithms and analytics tools.

Case Study A (12%)

Leveraging the power of content and social media marketing can help elevate the audience and customer base in a dramatic way. However, using social media for marketing without any previous experience or insight could be challenging. It is vital for a marketing team to understand social media marketing fundamentals. If a company publishes exciting, high-quality content and builds an online audience of quality followers, they can share it with their own follower audience on Twitter, Facebook, LinkedIn, Google+, their own blogs and many other social media platforms. This sharing and discussing of content opens up new entry points for search engines like Google to find it in a keyword search. Those entry points could grow to hundreds or thousands or more potential ways for people to find a company, product or service online. Finding and understanding the online influencers in the market who have quality audiences and are likely to be interested in the product, service or business could make a huge positive impact.

The consulting company collected information on articles that were shared by people on social media. The dataset contains approximately 39650 articles and a large number (with the total of 31) of features were extracted from the HTML code of the article, including the title and the content of each article. (The description of the dataset is provided as an appendix.) Some of the features depend on characteristics of the service used, which could be analysed based on the meta-data provided: articles have the meta-data, such as keywords, data channel type and the total number of shares (on Facebook, Twitter, Google+, LinkedIn, Pinterest), etc. The data channel categories are: ‘Lifestyle’, ‘Business’, ‘Entertainment’, ‘Social Media’, ‘Technology’, and ‘World’. In addition, several natural language processing features were also extracted.

Task Requirements

As a data analytics team member for the consultancy firm, you are required to carry out a number of data analytics tasks for the consulting company using the data collected. You are given access to a sample of the data where some of the variables have been removed as they are not considered important for the analysis of this assignment.

The company is interested in identifying for **each data channel**:

- Investigate the impact of the article properties on sharing;
- Use the SAS Enterprise Miner for *text analysis* to identify key features in the articles and analyse their contribution towards low and high sharing.

To achieve the above, you need to carry out the following data analytics tasks:

1. **Task 1: Explore the impact of article properties (5%)**

   Explore the data and investigate what properties of the article correlate with the high number of shares of the article on social media.

   - Open the dataset ‘online_news_popularity.xlsx’ using Microsoft Excel.
   - Explore the dataset to understand and manage four channels from the six types of data channels (lifestyle, entertainment, socmed, tech) and the associating data. In

each data channel column, the value of 1 represents that the data in the row is of the corresponding data channel.

- Copy the separate datasets for each channel to different Excel sheets (sort and filter by each data channel to separate in Microsoft Excel or apply proper R code in R Studio).
- In each data channel, identify the articles with a high number of shares (with the threshold of top 15% in the dataset).
- Investigate the following properties and explain how they could have affected the high number of shares. *You should provide explanations to support your argument.*
    - Number of links
    - Number of images
    - Number of videos
    - Number of keywords in the meta data
    - Was the article published on the weekend

(Hint: To do this, you can create plots in R with proper measures between the corresponding columns and the number of shares. You may want to include a fitted line to your plots to investigate the correlation for continuous variables.)

2. **Task 2: Use SAS Enterprise Miner for keyword analysis (7%)**

- Use the SAS Enterprise Miner to extract the keywords from the title in each data channel. (Hint: To do this, you can refer to the workshop activities in Week 2 and Week3; by setting 'Title' column as the only 'Text' role in the variable setting.)
- What are the highly used (top 5) topics in each category? Use the SAS Result window to explain your answers.
  (Hint: 'Topic' column will need to be set as the only 'Text' role.)
- Are there common topics which span across the six data channels and relate to a high number of shares and a low number of shares? Use the whole dataset in the SAS Enterprise Miner to identify the relationship. *You should provide explanations to support your argument.*
  (Hint: Use the whole dataset to identify the articles with the high number of shares and the low number of shares – by using appropriate thresholds with the top 15% and the bottom 15% in the dataset. Separate the dataset using Excel based on this before the analysis and use these two datasets to analyse the common topics in each of them. In this question, please use 'Title' column as the only 'Text' role for topic modelling.)

You are required to:

a) Prepare a report for the Case Study A with all the analytics results to the above two key tasks. (You can use an appendix for any additional screenshots, figures and tables, which you feel are important for the report). The report should be named as: *<student_id>Assignment1A_Report.doc*

b) Save the R script after Task 1 above as: *<student_id>Assignment1A.r*

c) Save the SAS project for Task 2 above as <student_id>Assignment1_Task1.spk. You may zip the SPKs files if you have multiple of them.

(The detailed procedures for exporting a model package SPK file can be found in Assignment 1 Additional Technical Support file.)

d) The SAS project file should be named as: *<student_id>Assignment1_SAS1.zip*

Case Study B (18%)

Sentiment analysis is the technique aiming to gauge the attitudes of customers in relation to topics, products and services of interests. It is a pivotal technology for providing insights to enhance the business bottom line in campaign tracking, customer-centric marketing strategy and brand awareness. Sentiment analytics approaches are used to produce sentiment categories such as 'positive', 'negative' and 'neutral'. More specific human emotions are also the topic of interest. There are two major streams of methods to develop sentiment analytics engine: the dictionary-based and machine learning-based approaches. In this assignment, you are required to perform sentiment analytics based on both approaches.

Task Requirements

As a data scientist, you are required to perform a number of data analytics tasks. You are tasked to develop both dictionary-based and machine-learning sentiment analytics engines using R programming language and apply it to predict the sentiments of hotel review tweets from a sample of data. You are also required to use the SAS Sentiment Analysis Studio to compare the results.

To achieve the above, you need to carry out the following data analytics tasks:

1. **Develop a dictionary-based sentiment analytics engine based on the R library 'syuzhet' and 'tidytext' to analyse the different emotions from hotel review tweets (8%).**
    - Analyse and aggregate the eight emotions (anger, anticipation, disgust, fear, joy, sadness, surprise and trust) from the hotel review tweets file 'hotel_tweets.csv' using the function **'get_nrc_sentiment'**. (You should combine both negative and positive tweets into one before conducting the analysis. Additionally, you are required to plot a chart to visualise these emotions using the R library 'ggplot2'.)
    - Finding the top 5 most frequent words in all the hotel reviews for each of the eight emotions (anger, anticipation, disgust, fear, joy, sadness, surprise and trust). Analyse the results.
2. **Develop a machine learning-based model using the R libraries 'tm' and 'e1071' as well as evaluate the predictive accuracies of SVM classifier (5%).**
    - Develop R scripts and import the data set 'hotel_tweets.csv' for training and testing.
    - Use the first 200 negative tweets and the first 200 positive tweets as the training dataset; and use the rest of the 63 negative tweets and 63 positive tweets as the testing dataset.
      (Hint: You may need to use as.character() function to convert a dataframe column from factors to characters.)
    - Develop a machine learning-based sentiment analytics engine and predict sentiment categories (only 'positive' and 'negative') using **'tm' and 'e1071'** with the **SVM** classifier.
    - Evaluate the testing accuracies and report the predicted results.

3. **Develop a statistical model using SAS Sentiment Analysis studio and evaluate the accuracies (5%).**
   - Use the data folder: ‘hotel_tweets’ which contain ‘negative’ and ‘positive’ tweets for training and testing.
   - Build a statistical model using SAS Sentiment Analysis (either simple or advanced), you may need to change configurations in the advanced model to obtain the best training accuracy and keep a record of how you manage to improve the accuracy.
     (Hint: Refer to the SAS Sentiment Analysis Studio tutorial.)
   - Evaluate and compare the testing accuracies for different models and report the results.
   - Compare this result with the previous predictive results using R and discuss (Note: the hotel tweets used in this task is the same tweets as in Task 2).

You are required to:

a) Prepare a report for Case Study B with all the analytics results to the above three key tasks. (You can use an appendix for any additional screenshots which you feel are important for the report). The report should be named as:
<student_id>Assignment1B_Report.doc

b) Save the R script after Task 2 above as: *<student_id>Assignment1B.r*

c) Save the SAS Sentiment Studio project as: *<student_id>Assignment1_SAS2.zip*

(The detailed saving procedures can be found in Assignment 1 Additional Technical Support file.)

Important: You should submit all the reports, R scripts and SAS Sentiment Studio project via the LMS Assignment1 submission link.

Report Guidelines

1. The report should consist of a table of contents, an introduction, and logically organised sections/topics (such as ‘case study A’, ‘case study B’), a conclusion and a list ofreferences where necessary.
2. Choose a fitting sequence of sections/topics for the body of the report. Two sections for the two case studies are essential, you may add other sub-sections deemed relevant.
3. You should include diagrams, tables and charts from the analytics solutions to effectively present your results. (Consider using Alt + Print Screen to capture screenshots if needed.)
4. Page limit: For each case study, five (5) pages for the main report writing but not more than ten (10) pages including appendices.
5. Reports should be written in Microsoft Word (font size 11) and submitted as a Word file.
6. Final submission will comprise **six separate files**:
    a. <student_id>Assignment1A_Report.doc (should not be zipped);
    b. <student_id>Assignment1B_Report.doc (should not be zipped);
    c. <student_id>Assignment1A.r;
    d. <student_id>Assignment1B.r;
    e. <student_id>Assignment1_SAS1.zip;
    f. <student_id>Assignment1_SAS2.zip.

Marking Rubrics

A grade will be awarded to each of the tasks and then an overall mark determined for the entire assessment. The rubric below gives you an idea of what you must achieve to earn a certain 'grade'.

As a general rule, to meet a 'C', you must first satisfy the requirements of a 'D'. And for an 'A', you must first satisfy the requirements of a 'B', which must of course first meet the requirements of a 'C' and so on.

The marking rubric for this assignment is given below.

Criterion	**Pass**	**Credit**	**Distinction**	**High Distinction**
Case study one: Impact of article properties **(5 marks)**	Limited effort to structure and present information and insights.	Fair effort to structure and present information and insights.	Excellent effort to structure and present information and insights.	Exceptional effort to structure and present information and insights.
Case study one: Use SAS Enterprise Miner for keyword analysis **(7 marks)**	Limited effort to structure and present information and insights. Limited knowledge of SAS Enterprise Miner.	Fair effort to structure and present information and insights. Fair knowledge of SAS Enterprise Miner.	Excellent effort to structure and present information and insights. Excellent knowledge of SAS Enterprise Miner.	Exceptional effort to structure and present information and insights. Comprehensive knowledge of SAS Enterprise Miner.
Case study two: Develop dictionary-based sentiment analytic engine and analyse emotions **(8 marks)**	Limited effort to structure and present insights for emotions from tweets. Limited knowledge of the R programming.	Fair effort to structure and present insights for emotions from tweets. Fair knowledge of the R programming.	Excellent effort to structure and present insights for emotions from tweets. Excellent knowledge of the R programming.	Exceptional effort to structure and present insights for emotions from tweets. Comprehensive knowledge of the R programming.
Case study two: Develop machine learning-based sentiment analytic engine and evaluate predictive accuracies using R **(5 marks)**	Limited effort to structure and present information and insights. Limited knowledge of the R programming.	Fair effort to structure and present information and insights. Fair knowledge of the R programming.	Excellent effort to structure and present information and insights. Excellent knowledge of the R programming.	Exceptional effort to structure and present information and insights. Comprehensive knowledge of the R programming.
Case study two: Develop sentiment analytic engine using SAS Sentiment Analysis Studio **(5 marks)**	Limited effort to structure and present information and insights. Limited knowledge of SAS Sentiment Studio.	Fair effort to structure and present information and insights. Fair knowledge of SAS Sentiment Studio.	Excellent effort to structure and present information and insights. Excellent knowledge of SAS Sentiment Studio.	Exceptional effort to structure and present information and insights. Comprehensive knowledge of SAS Sentiment Studio.

Other information

- Standard plagiarism and collusion policy, and extension and special consideration policy of this university apply to this assignment.
- A cover sheet is NOT required. By submitting your work online, the declaration on the university's assignment cover sheet is implied and agreed to by you.

Appendix – Attribute Information

1. This section contains a description of the attributes of the dataset ‘online_news_popularity.xlsx’.
   {‘name of the column’: ‘description’}

   1. url: URL of the article (unique)
   2. title: Title of the article
   3. topic: topics related to the article
   4. content: content of the article
   5. timedelta: Days between the article publication and the dataset acquisition
   6. n_tokens_title: Number of words in the title
   7. n_tokens_content: Number of words in the content
   8. n_unique_tokens: Rate of unique words in the content
   9. n_non_stop_words: Rate of non-stop words in the content
   10. n_non_stop_unique_tokens: Rate of unique non-stop words in the content
   11. num_hrefs: Number of links
   12. num_self_hrefs: Number of links to other articles published by Mashable
   13. num_imgs: Number of images
   14. num_videos: Number of videos
   15. average_token_length: Average length of the words in the content
   16. num_keywords: Number of keywords in the metadata
   17. data_channel_is_lifestyle: Is data channel 'Lifestyle'?
   18. data_channel_is_entertainment: Is data channel 'Entertainment'?
   19. data_channel_is_bus: Is data channel 'Business'?
   20. data_channel_is_socmed: Is data channel 'Social Media'?
   21. data_channel_is_tech: Is data channel 'Tech'?
   22. data_channel_is_world: Is data channel 'World'?
   23. weekday_is_monday: Was the article published on a Monday?
   24. weekday_is_tuesday: Was the article published on a Tuesday?
   25. weekday_is_wednesday: Was the article published on a Wednesday?
   26. weekday_is_thursday: Was the article published on a Thursday?
   27. weekday_is_friday: Was the article published on a Friday?
   28. weekday_is_saturday: Was the article published on a Saturday?
   29. weekday_is_sunday: Was the article published on a Sunday?
   30. is_weekend: Was the article published on the weekend?
   31. shares: Number of shares

2. The description for the dataset ‘hotel_tweets.csv’.
   {‘name of the column’: ‘description’}
   1. Negative: negative tweets content
   2. Positive: positive tweets content

3. The description for the dataset ‘hotel_tweets’ folder.
   Same content as the dataset in (2), grouped as negative and positive