300459-M-6
Social Media and Web Analytics
Tilburg University AY: 2021/2022
PROF: Lachlan Deer / TA: Gijs van Bussel Exam

Mock Exam

Instructions to candidates

- You must write your name and student ID number on all answer pages.
- The exam is graded out of 100 points. Points for each of the three parts are clearly labeled at the beginning of each question.
- Please answer each part starting on a new page.
- Remember, your goal is to communicate. Full credit will be given only to the correct solution which is described clearly. Convoluted and obtuse descriptions might receive low marks, even when they are correct. Also, aim for concise solutions, as it will save you time and also help you conceptualize the key idea of the problem.
- No calculators, mobile phones or other electronic devices are permitted. If any of these unauthorised objects are found they will be confiscated and you may face penalties.
- Dictionaries, textbooks, written or recorded materials of any form are not permitted. Please have any unauthorised material securely out of view and reach. Failure to do so will result in confiscation and penalties at the discretion of the Examination Board. If you are unsure about whether any of your materials are unauthorised please raise your hand we will check them.
- Any communication between students during the exam, no matter whether is is about the exam or not is unacceptable and will result in harsh penalties. Attempting to view other students answers will be treated in a similar manner.
- The exam is 8 pages in total (including this page). Make sure you have all the pages.

Part A: True/False/Uncertain [20 points]

Answer each question as TRUE, FALSE or UNCERTAIN **and** write a short justification of your answer. Each question in this section is worth **2** points.

1. Post release release sentiment is a crucial factor in determining consumer demand for sequel and franchise movies.
2. Managers should provide long replies to negative reviews in order to manage their online reputation.
3. When applied to a dataset of social media posts, a Topic Model returns whether each post is classified as positive, negative or neutral.
4. Image Related Utility and Instrinsic Utility both influence people's tendency to post on social media, but which effect dominates depends on how many followers a user has.
5. Consider the simple difference in difference regression specification:

$$y_{it} = \beta_0 + \beta_1 \text{After}_t + \beta_2 \text{Treatment Group}_i + \beta_3 \text{After}_t \times \text{Treatment Group}_i + \varepsilon_{it}$$

   where i denotes individuals and t denotes time. After is a indicator/dummy variable that takes the value 1 after the experiment been implemented and is zero otherwise, and Treatment Group is a indicator/dummy variable that takes the value 1 for individuals in the treatment group and the value 0 for individuals in the control group.

   Assuming that the necessary assumptions for causal interpretation are satisfied, $\beta_1 + \beta_3$ is the estimate of the average casual effect of the treatment on the outcome variable.
6. The introduction of advertising disclosure regulations has lead to a decrease in the number of likes of posts by influencers.
7. The Louvain method for community detection within a network constructs communities using an initial list of influential users provided by the analyst writing the code.
8. Humorous ads are more likely to go viral.

9. Consider the following query sent to the Twitter API using R:

```
tweets <-
    search_tweets(
        "(@willsmith OR @chrisrock) #theslap",
        n = 1e10,
        retryonratelimit = TRUE,
        geocode = lookup_coords("usa"),
        lang = "en"
        )
```

   The data returned contains all tweets that mention Will Smith or Chris Rock or '#theslap' over the previous 14 days written in English in the United States.

10. Retweets by influential users increase TV show viewing because they bring in new followers to a show's social media accounts, which indirectly increases viewing.

Part B: Short Answer Questions [40 points]

Answer all questions below. The maximum points available for each question are indicated with [X points].

Vaccines Perceptions & Social Media

Suppose you are working for Pfizer in their Customer Experience Team in late 2020 / early 2021 as the roll out of their COVID-19 vaccine accelerates. Your team has been assigned the task of understanding consumer perceptions regarding their vaccine. Your first task is to build up this understanding using posts that mention 'Pfizer' or 'covid vaccine' from Twitter.

The Data Science team at Pfizer has provided you with 6000 tweets to begin developing the analysis. Each row of the raw data contains three columns: (1) an identifier for each tweet, 'tweet_id', (2) the tweet text, 'tweet_text' and (3) a label of whether the tweet has been classified as positive, negative or neutral by an expert, 'label'.

As a starting point you decide to focus only on tweets that mention 'Pfizer', and filtering your original data leaves you 1,324 tweets to analyse.

1. [2 points] Explain why opinions about Pfizer's COVID-19 vaccine expressed on Twitter could serve as a valid approximation for the wider public's perception of the vaccine. What limitations might you encounter using only Twitter data as a means to understand consumer perceptions?

2. [1 points] Why might you have decided to only focus on tweets explicitly mentioning 'Pfizer'?

The first component of your analysis will be understanding consumer sentiment in the tweets. The team lead has decided to examine whether the VADER lexicon can be used to measure sentiment in unlabelled tweets that may arrive in the future from the Data Science Team.

3. [3 points] What is the VADER sentiment lexicon?

4. [2 points] Why might the VADER sentiment lexicon be more appropriate than alternative lexicon such as NRC or Bing?

5. [4 points] The data are stored in an R session as 'tweets'. Write the code that would

compute the compound score and use that to decide whether each tweets is positive, negative or neutral. You can assume all necessary R libraries have been loaded.

To assess whether the VADER lexicon's classification of tweets does a good job segmenting tweets into positive, negative and neutral tweets, you decide to examine a confusion matrix using the predicted classification from VADER and the true labels from experts provided to you in the data.

6. [2 points] Explain what a confusion matrix is. Why does it help you assess how well VADER performs on tweets about Pfizer's vaccine?

The output from R yields the following confusion matrix:

		Truth		
		Negative	Neutral	Positive
Prediction	Negative	55	191	46
	Neutral	26	285	94
	Positive	32	366	229

7. [2 points] Explain the meaning of the number "366" in the confusion matrix.

Based on the confusion matrix above, the accuracy of VADER in predicting the labeled sentiments is computed to be 0.43.

8. [2 points] Would you recommend using VADER "off the shelf" to track consumer perceptions of Pfizer's vaccine? Justify your answer.

In the second phase of your analysis the team decide to develop a Topic Model to develop an understanding of what consumers are talking about when posting about the vaccine.

9. [3 points] Provide an intuitive explanation of how a Topic Model works.

10. [2 points] Is a Topic Model a descriptive, causal or predictive analytics tool? Explain your answer.

11. [4 points] What steps do you need to take to go from the data set provided by the Data Science Team to a data structure that can be used by a Topic Modeling package. Briefly explain each of these steps, but you do not need to write any code.

After some long hours your team settles on a topic model with seven (7) topics. The individual words that are most associated with each topic can be found in the table below:

	Top 5 terms				
Topic 1	grateful	pfizer	save	purchase	millions
Topic 2	pfizer	moderna	johnson	market	value
Topic 3	effect	nurse	arm	patient	#vaccinated
Topic 4	purchase	shot	expect	arrive	batch
Topic 5	vaccinate	approve	trial	clinical	south africa
Topic 6	pfizer	astrazeneca	clot	mrna	experimental
Topic 7	vaccine	adverse	reaction	severe	prevent

12. [2 points] Propose human readable labels for each of the seven topics.

13. [2 points] Which topics (if any) provide an indication of the public perception of your vaccine? Explain your answer.

Suppose the team can now use the Topic Model above to classify any new tweet about Pfizer's vaccine into one of the seven topics above in real time.

14. [3 points] Provide a brief explanation suitable for a social media marketing manager who has no quantitative training on how they can use this model as new data arrives to understand the public's perception of the vaccine.

15. [3 points] If you were tasked with developing a marketing strategy that utilizes this model to improve product image, what strategy might you propose? Explain the strategy and why you think this could work.

16. [3 points] Could your strategy in (15) backfire? Why might this be the case?

Part C: Extended Answer Question [40 points]

Answer **one** of the following questions. Students who answer more than one question will receive the grade for only one of their answers, where the answer graded will be selected randomly by the graders.

To earn full credit your answer must touch on the following points:

- Identify the business/marketing problem, and why it is relevant.
- Marketing concepts and statistical terminology must be defined when used.
- Mention explicitly any hypothesis being tested.
- Explain clearly any data used, the empirical model and/or equations used. Necessary assumptions to interpret a model in a particular way must also be made explicit and be explained.
- Identify how any empirical model answers the hypotheses posed.
- Summarize the main points in a conclusion.

Option 1:

You currently work for Unilever in their Marketing Insights team with a focus on fast moving consumer goods (FMCG). Your team has been tasked with increasing the volume of online sharing of their ads with the goal of increasing the volume of shares about Ben & Jerry's ice-cream ads in the Netherlands by 10 percent on traditional social media sites. As part of the project your team can create a new video ad, and shift around when certain images / scenes appear in ads that will air on YouTube. Design an empirical strategy to test whether the new ad(s) improve sharing of the ad, and whether the lift in volume of the "best performing" ad meets their target. In your analysis be sure to articulate what content / order of content you might change and why.

Option 2:

While working for the Democratic National Committee (The US Democratic Party's governing body) as a Marketing Analyst and Strategist, the Head of Marketing Strategy claimed in a meeting you attended that "regression estimates of advertising effectiveness from Facebook ads using observational data deliver accurate measurement of causal effects". Evaluating this statement by either (a) explaining what the academic literature knows about measuring ad-

vertising effectiveness on social media sites, or (b) developing an empirical strategy that you could implement to answer the question.

HINT: If you choose (a) be sure to reference any cited literature and explain their context, data, methods and results. The reference need not be fully correct, but needs to be identifiable to a grader who knows the literature reasonably well. If you choose (b) clearly explain the data requirements, modeling strategy and how to interpret the results of your proposed analysis.