NUS-ISS

MASTER OF TECHNOLOGY IN ENTERPRISE BUSINESS ANALYTICS / ARTIFICIAL INTELLIGENCE SYSTEMS

Graduate Certificate
Sample Examination Paper

Subject: *Practical Language Processing*

SECTION A

Question 1 *(Total: 25 Marks)*

WeShare, a social media mobile app, offers people a platform to self-publish their content (typically articles) online over a wide variety of topics. The app supports two types of user accounts: public accounts (which can publish articles), and normal accounts (which can read the published articles). When reading an article, a normal account user can take actions like "*share*", "*like*", "*save*" by clicking on the corresponding buttons provided at the end of the article. He/she can also "*follow*" (subscribe to) the public accounts if he/she likes to read the articles from these public account users (the authors). Each day the number of articles uploaded by the public account users can range from 1,000 to 3,000.

To help the normal account users discover articles that they may be interested in, the app also recommends articles to these users through a '*Recommended*' page. However, one problem that ***WeShare*** faces is that user-generated content can be of very different levels of quality - self-published articles come with diverse content, different writing styles, disparate levels of semantic coherence, as well as various layout designs including text and multimedia elements like pictures and videos.

WeShare would like to build a system that can automatically assess the quality of self-published articles so that only articles of high quality are recommended to the normal account users. An internal project team has been formed to work on the task. Since this is a new task, they found that there was no readily available data containing examples of high quality or low quality articles.

To start off, the project team first engaged a magazine editor to read a sample collection of 50 randomly selected articles, and separating them into two groups: *high quality* and *low quality*. These articles are randomly selected from the articles uploaded by ***WeShare*** public account users on one randomly selected date. From this collection, the editor identified 8 articles to be of high quality.

The team then interviewed the editor for criteria of his judgement. According to the editor, besides checking whether the article has a neat layout, there are mainly two factors for consideration when evaluating the articles:

- **Writing**: The vocabulary, syntax and grammatical elements of the content and title. A high quality article should display good use of language and clear writing style.
- **Semantics**: Logical coherence of the content. A good article should make sense logically and provide useful information.

The editor also commented that based on his observation, high quality articles have a high tendency turning out to be more popular with viewers and receive more interactions like "*share*" and "*like*".

Since this task is a classification problem, the team decided to use supervised learning approach in this project. There was no large labeled data set available, therefore their first task was to construct a dataset containing articles labeled as '*high'* quality and '*low'* quality. The team produced a data construction plan as shown in Table 1 below.

Target size	Around 20,000 articles
Article selection method	Randomly selected from the most recent articles, e.g. those uploaded in the past two months.
Annotators	interns (50 year-1 college students on holiday)
Labels	'high', 'low'
Approach	- Brief the interns of the evaluation guidelines (summarized from the editor's interview) - Divide the dataset and assign each intern with 400 articles to be labeled as 'high' or 'low'

Table 1. Data Construction Plan

Considering the two factors identified by the editor in article quality evaluation, one option is to use separate models in the system taking care of different factors. However, impressed by the success of pre-trained transformer models (e.g. BERT, GPT, etc.) in transfer learning, especially their ability to capture both syntactic and semantic information of language in their vector representation of text data, the team decided to combine '*Writing'* and '*Semantics'* to be one factor '*Language'*, to be handled by one single model, **Language Model**.

The role of the **Language Model** is to assess the writing and semantic quality of the textual content of a given article, including whether the sentences and paragraphs in the article form a logical and coherent flow. Since BERT is well-known for its capability of providing deep, contextual bidirectional representation of text, the team decides to use BERT-base as a feature extractor to encode an article into a vector of fixed length. The input to BERT-base is the text of the article as a sequence of tokens. As longer sequences are disproportionately expensive to process, the sequences are truncated to 512 tokens at maximum. The final hidden vector from the model corresponding to the first input token ([CLS]) is then taken as the aggregate representation of the article. The encoded vectors of articles, together with their labels ('*high'* or '*low'*), are used to train a single layer feed-forward neural network (FFNN) to perform the article quality classification. Figure 1 illustrates the design of the Language Model.

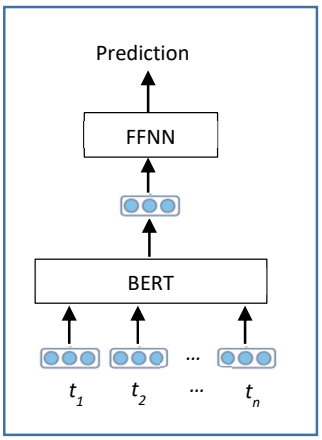

Figure 1. Language Model

To test the Language Model's ability to assess the logical coherence of the articles, the project team has conducted a '**Disruption Test**'. They took 1000 articles which were classified as '*high'* by the model, and randomly shuffled the order of the sentences in each article to disrupt its logic flow. These articles were then passed through the Language Model to get predicted labels. If the model was able to capture logical coherence effectively, the disruption of the sentences would be detected by the model and classify the article as 'low'. However, the team found that only 5.4% of the disrupted articles were labeled as '*low'* by the model.

WeShare also plans to add to the app a new feature, **Audio Publishing**, which allows public account users to publish audio content with speech recordings, such as chapters of audio books, commentaries of stocks and markets, etc. There are generally two ways to create the speech content:

1. Users may record their own voices and publish directly. In this way, an additional process is needed to convert speech into text before the content can be evaluated. Thus an **automatic speech recognition** (**ASR**) engine needs to be used.
2. Or they may prepare the text content first and use the app to convert the text into speech. Using this approach, a **text-to-speech** (**TTS**) engine is required.

Answer the following questions based on the information provided above.

a. Review and critique the team's **data construction plan**. Can the plan ensure a good dataset with consistent labels? If not, what are the problems in the plan? How would you improve it? Justify your answer using the information provided in the case study.

(6 Marks)

b. Assuming that the '*high'* articles used in the '**Disruption Test**' are labelled correctly, the result of this test reveals the inadequacy of the existing transfer-learning approach adopted by the project team to capture the logical coherence of sentences. Identify <u>two issues</u> in the existing design of **Language Model** that are most likely the causes of this problem, and justify your answers.

(*4 Marks*)

c. For <u>each</u> issue identified in b, propose how it can be resolved to improve the model's ability to capture the logical coherence of an article, with limited GPU resources available. Describe the changes to the model's network structure if any. Your answer should be based on the information given in the case study.

(5 Marks)

d. While developing the ASR component for **Audio Publishing**, two candidate ASR systems (A and B) were evaluated to test their performances. The same set of testing speech recordings were passed to the two ASR systems, and two sample recognition results were recorded as shown in the following table. Please compare the result and explain the reasons for the difference between the two systems in terms of the major components of speech recognition systems.

	Sample 1	Sample 2
Speech Content	The consumer watchdog's statement comes as tensions rise between private insurers and doctors over the use of Integrated Shield Plans panels.	It takes about half an hour from Jurong East to Tiong Bahru.
Result from system A	The consumer watchdog's statement comes as tensions rise between private insurers and doctors over the use of integrated shield plans panels.	It takes about half an hour from rural East to Jonesboro
Result from system B	The consumer watchtowers detrimental comes as the tensions arise between private insurers at the doctors over the use of the integrated Shield plans panels	It takes about half an hour from Jurong East to Tiong Bahru

Table 2. Sample Recognition Results from the Two ASR systems

(3 Marks)

e. For auto speech content creation from text article, **WeShare** wants to try it in one of the most popular article categories among their users, 'stock and market commentaries'. Two sample snippets of such articles are shown in Figure 3.

i. What are the challenges for off-the-shelf TTS engines to work well for such commentaries? What text normalization step you need to do if you expect the TTS system may not work well for some of the text content? Support your answer with examples from Figure 3.

(2 Marks)

ii. Suppose we decide to create a text normalization dataset with the help of the interns and build a text normalization model with machine learning method. Please propose a method to achieve this goal. Kindly describe (1) the data set needs to be created, and (2) the machine learning method to be used.

(2 Marks)

[Snippet 1]
Meanwhile, the major European markets have all moved to the upside on the day. While the German DAX Index has risen by 0.4%, the French CAC 40 Index is up by 0.3% and the U.K.'s FTSE 100 Index is up by 0.2%.
In commodities trading, crude oil futures are slipping $0.16 to $64.89 a barrel after slumping $1.04 to $65.05 a barrel on Monday. Meanwhile, after tumbling $20.50 to $1,678 an ounce in the previous session, gold futures are spiking $29.90 to $1,707.90 an ounce.

[Snippet 2]
If we start thinking of the cryptocurrency as a cultural product, last week's sudden jump in Dogecoin's price makes sense. The boost came just after a meme-centric community managed to drive the share price of videogame retailer GameStop from US$20 to US$350 in mere days.
One particularly interesting aspect of the Reddit forum r/WallStreetBets – which coordinated the attack on the hedge fund that had effectively bet on GameStop's share price falling – was how many users were having fun.

Figure 3. Sample Text Snippets

f. **WeShare** also decides to use voice print as one of the ways to access the system. There are two options (A and B) of voice print solutions under consideration. The performances of them are shown in Figure 4, in which FAR (false acceptance rate) means the chance of an unauthorized user is accepted, and FRR (false reject rate) means the chance of an authorized user is rejected. If we would like to choose a system that has higher security level, which one should we choose? If we concern more about convenience of access to the system, which one should we choose? Justify your choices.

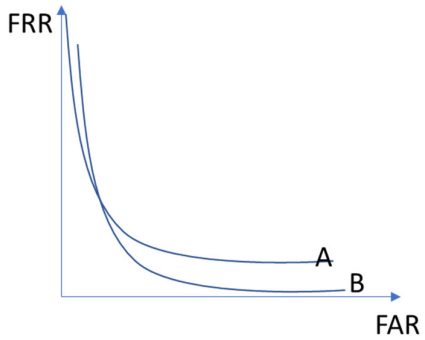

Figure 4. Performance of Two Solutions

(3 Marks)

Question 2 *(Total: 20 Marks)*

With the increasing demand for public opinion data monitoring and analysis, financial investment research and risk control logic has gradually become the research focus while integrating with the in-depth mining on public opinion data.

Market Intelligence Platform is a product that analyzes the sentiment, opinions and attitudes of public opinion data from different dimensions based on natural language processing technology for news, announcements, and other public opinion data. Subscribers can easily view the major events, opinions and polarities extracted and summarised from millions of recent news articles by AI, with respect to any financial entities they are interested in. It also provides timely and accurate quantitative factors for investment, risk control, margin financing and risk assessment services.

Figure 5 demonstrates the highlevel framwork of the platform.

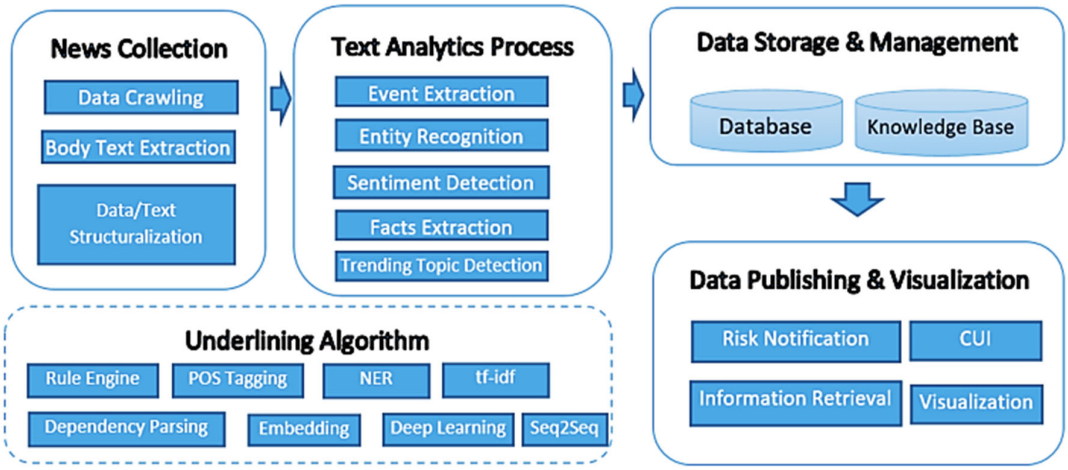

Figure 5. Platform framework

Figure 6 shows one example of articles crawled from *Yahoo Finance*, which is one of the major channels of financial news. After one months' data crawling, the collected corpus contains about 2 million recent articles from *Yahoo Finance*.


CAIRO (Reuters) - Shell Egypt has entered into an agreement to sell upstream assets in Egypt's Western Desert to Cheiron Petroleum Corporation and Cairn Energy Plc for up to $926 million, the company said on Tuesday.

The deal includes Shell Egypt's stake in 13 onshore concessions and its share in the Badr El-Din Petroleum Company (BAPETCO). It is subject to government and regulatory approvals and is expected to be completed in the second half of 2021, the statement said.

It sets a base price of $646 million, with additional payments of up to $280 million between 2021 and 2024, subject to the oil price and exploration results.

Figure 6. Example of News Article

You are currently leading a team to design and develop such a market intelligence platform in order to monitor the public opinion and identify the risk of financial market by analyzing big data related to particular companies or industries. With the information provided above, answer the following questions.

a. In order to achieve the business goal of analysing and monitoring the recent public opinions with respect to a particular company, the first real word problem is to identify the subset of articles reporting on this company from the 2 million articles collection. Your teammates designed a filtering system with the keyword matching strategy, which was to select the articles based on predefined company names as keywords. By observing its performance on the production dataset, this naïve strategy was not satisfying due to two major issues. Firstly there are always out-of-vocabulary financial entities beyond the coverage of any keyword list. Around 35% percentage of the important financial articles failed to match any keywords. It makes the business owner less confident to trust the platform since it probably overlooks many important market news articles. Moreover, for example in Figure 6, multiple company names (*i.e., "Reuters", "Shell Egypt, "Western Desert" and so on*) could be matched against the article based on this naive strategy. However, *"Shell Egypt"* should be the only primary financial entity being described. In this case please decide which underlining algorithms should be used here to recognize the financial entities and determine how to rank the detected entities by importance.

(3 Marks)

b. A fine-tuned BERT model is widely applied for sentiment detection on documents according to the best practice of industry. It is requiring small collection of annotated data on the target task but offering reliable performance due to its powerful pre-trained model. However, for this platform, you are facing a different but real world problem of how to mine the sentiment **with respect to each financial entity** detected from articles. For example, given the following paragraph: *"It will enable Shell to concentrate on its offshore exploration and integrated value chain in Egypt, including seven new blocks in the Nile Delta, West Mediterranean and Red Sea," the company said."*, the entity detected should be "*Shell*", the sentiment for "*Shell*" should be *positive*. Given the practical constraint of not having labour and resources to make large scale annotation for this problem, please create your model by using or adapting BERT to detect the sentiment with respect to the entity, given the paragraph where the entity is recognized from.

(3 Marks)

c. To fulfil the business requirement of mining and demonstrating recent key events, the **Event Extraction** module is designed here to extract major elements of events from news articles. For example, taking the article from Figure 6 as the input, the expected extraction result is summarized in Table 3, having 5 major elements with extracted values. All the events extracted for the companies being monitored, will be further rendered and presented to subscribers. Before selecting the sequence labelling models,

in the real world situation, the most important thing will always be preparing data annotation with high quality. After carrying out the manual annotation, Table 4 shows the number of training instances annotated for individual event types.

Major Entity	Action	Time	Amount	Objective Entity
Shell Egypt	sell	second half of 2021	926 million	Cheiron Petroleum Corporation
				Cair Energy Plc

Table 3. Extraction Results

Major Entity	Action	Time	Amount	Objective Entity
820	354	280	77	427

Table 4. Number of Training Instances

A critical alert has been raised by your teammates because they realize a practical issue, which is the number of training instances are not balanced over different types of events. Your teammate suggests to balance the training instances for better labelling performance. Please make your judgement on this suggestion and justify your opinion.

(4 Marks)

d. To further fulfil the business requirement of mining and demonstrating recent key events, after resolving the issues of the training data preparation (addressed in question c), it is time for the team to build the models. Knowing from the textbooks that sequence labelling problem can be addressed by CRF, LSTM and BERT models, your teammates propose to combine the three models (as Figure 7) together, expecting it to have a better performance than individual models. However, in the real world, there are always constrains caused by hardware, resources and time value, which will make the textbook methods difficult to apply.

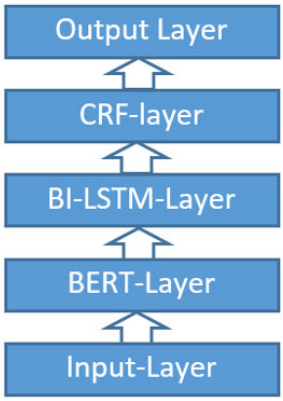

Figure 7. Model Architecture

i. Considering the situation that there are about 2 million raw articles collected (as described in question a), 1000 news articles have been annotated (as described in question c) , and 2 GPUs available for training, in actual practice, please

determine which layer(s) in Figure 7 should hold trainable parameters and which layers should be locked during training phase?

(2 Marks)

ii. Please diagnose this model (*i.e.* expected performance, efficiency) and determine the way to improve this model, given the constraints and the below Figure 8.

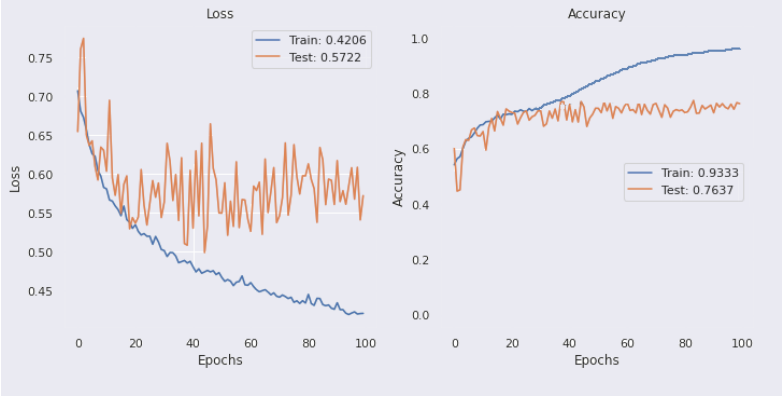

Figure 8. Plots of the evaluation

(4 Marks)

e. Subscribers can hardly read all the news articles in detail, but they benefit from reading abstracts as many as they can. To further fulfil the business requirement of providing short abstracts to the subscribers, events extracted from question c & d should be further transformed to summarize the original article. Your teammates proposed to implement the template-based method to generate the text based on the structured data from Table3. Please judge the feasibility of this method, evaluate its pros and cons, and design an enhanced way of doing it.

(4 Marks)