Spring Semester/ 2021

SWE 485

Project (15 marks)

Part I: Given Twitter Dataset (extracted from Kaggle)

* Create a data frame and call it Temp\_Sent and load all data from the given Twitter Dataset (all 1.4 million twits)
* Make shuffling (changing the orders of the rows) of Temp\_Sent (use the following statement

Temp\_Sent = Temp\_Sent.sample(frac = 1)

* Create a data frame and call it Sent and any copy 100 k twits from Temp\_Sent
* Create a function to Make **text cleansing** of the text column
* Use **CountVectorizer()** and **MultinomialNB()** to create sentiment prediction model and do the required test
* Use **TfidfVectorizer ()** and **MultinomialNB()** to create sentiment prediction model and do the required test
* Use **confusion\_matrix** to evaluate both models

Part 2: PR Dataset (found in black board)

PR is a dataset about the evaluation of some models of smart phones

* In the **product** column, show only the product general **name**: for example: Samsung, Hawawi, iPhone and remove other text
* Show in a **bar chart**, the **average** score of each model
* What is the **best** model in each **country**