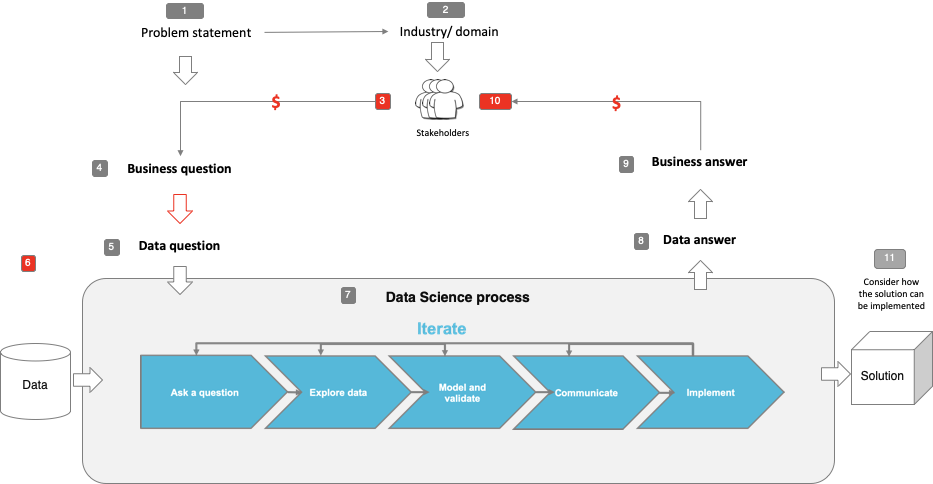
**Capstone Project**

**Document Skeleton**

# Process overview

The following diagram shows the overall end-to-end process for defining, designing and delivering the Capstone project.



Note: The following are the candidate sections of the document. They are presented here for guidance. Questions in each section could be used as possible aspects to cover. Some questions may not be applied to each project. On the other hand, additional information may be needed.

# Problem statement

* What is the problem or the opportunity that the project is investigating?
* High volumes of tweets are created every second, how do we leverage Natural Language Process algorithms to identify sentiment of tweet content to understand whether the users’ emotions
* Why is this problem valuable to address?
* Keep the business stay ahead of tweets related with business in near-real time by generating insights that might be possible to find manually.
* What is the current state (e.g. unsatisfied customers, lost revenue)?
  + Vast amounts of tweets under company’s twitter page.
  + Human have partial judgement when review company tweets.
* What is the desired state?

To use machine analysis tweets to help business aware of the sentiment of certain incoming tweets

* Has this problem been addressed by other research projects? What were the outcomes?

Yes, but other research projects used other datasets, their results are around 70% to 85%

# Industry/ domain

* What is the industry/ domain?

Social platform, shared information platform, business creating value by using information

* What is the current state of this industry? (e.g. challenges from startups)

Twitter is still a growing company, currently, vast amount of tweets are needed to be consumed in a short- timer period to influence decision-makers

* What is the overall industry value-chain?

Twitter is actually a collection of different activities that share relatedness to some extent,

* What are the key concepts in the industry?

The key concept in the industry is the information flow, the improved information flow can help the company identify and exploit new opportunities and reduce external threats.

* Is the project relevant to other industries?

Yes, using sentiment analysis on tweets content actually could strongly help business account holders to understand how their target customers’ opinions about the business’s products.

# Stakeholders

* Who are the stakeholders? (be as specific as possible)

The stakeholders will be the twitter users, twitter business account company, customers, information users, and social content analyst.

* Why do they care about this problem?

Because tweets’ users’ opinions are critical to the business account holders, they need to know their targeted existing customers, potential customers and public about their products

* What are the stakeholders’ expectations?

Business account holders and twitter company would like to identify and understand tweet posters’ sentiment in real time so that they can avoid the negative effects of bad reviews, also want to take advantages of the positive reviews to catch the market trend.

# Business question

* What is the main business question that needs to be answered?

How can we use machine to identify sentiments of customer users.

* What is the business value of answering this question? (quantify value and make necessary assumptions)

Machine can help to organize massive amounts of tweets into information in real-time for example it helps business to be aware of negative reviews about an important product launch before it gets worse, or it can help business account holders to analysis positive comments so they can understand customers’ desired needs.

* What is the required accuracy? What are the implications of false positives or false negatives?

The accuracy is the higher the better. The false positives are the negative tweets have been identified as positive tweets, the false negatives are the positive tweets that have been identified as negative tweets.

# Data question

* What is the data question that needs to be answered?

How the machine can help to analyse vast amount of tweets in real-time

* What is the data required to answer the question?

Real twitter content with the original tweets posted by variety users

# Data

* Where was the data sourced?

The data sourced from Kaggle opendatasets, for saving computing power, I only select 3000 tweets from the whole datasets to be used in models

* What is the volume and attributes of the data?

The data have 3000 rows and have columns including target, ids, date, flag, user and tweets.

* How reliable is the data?

Over 90 percent, the data is published around 5 years, a lot of users download the data, 350 comments about data and 1500 kaggle users upvote it, therefore it should be reliable.

* What is the quality of the raw data?

The data is unstructured data, the post date is around May 2016, the data is outdated, id number of the tweet is 2087. Overall, the data is able to be used in the models

* How was this data generated?

The original data contains 1.6 million tweets, are extracted using twitter API by the open data publisher. The tweets are labelled as 0 is equal to negative 2 represents neutral, 4 shows positive. The data is suitable for detecting sentiment.

* Is this data available on an ongoing basis?

No, the data is one-time extraction through API

# Data science process

## Data analysis

* What data pipeline was to wrangle the raw data?

After read csv file, I calculate the length of the each tweets to see whether there is a significant difference in length of the data.

* What are the highlights of the Exploratory Data Analysis (EDA)?

The length of the positive tweets is similar with the negative tweets.

There are two outliers shows that positive tweets have more than 150 length text.

People with positive emotions tends to post around 40-words text.

* Is the pipeline reusable? (for example, to process future data?)

Yes, it’s reusable, we extract more data from twitter API, even streaming it.

* What are the intermediary data structures used (if any)?

The notebook used SpaCy English natural language process library to help identify vocabulary, syntax and entities to generate text features.

## Modelling

* What are the main features used?

The main features are count vectors, length, if-idf vectors, ngram level if-idf, characters level ti-idf, character counts, word counts, word density, punctuation count, title word count, uppercase word count.

* Did you find any interesting interactions between features?

There is no significantly difference in the positive or negative tweets.

* Is there a subset of features that would get a significant portion of your final performance? Which features?

On final performance, count vectors feature actually tends to have higher scores. I believe it get a significant portion of the final performance.

* How did you select features?

I only have one feature at the first time, after analysing the length, I add 13 features to make sure the text change to numbers is ready to be fitted in the models.

* What feature engineering techniques are used?

I used count vector, word-leve ti-idf, ngram level tf-idf, characters level tf-idf to turn tweets to numeric value

* What are the models used?

I used Logistic Regression, Naïve Bayes, Support Vector Machine, Random Forest, Gradient boosting models.

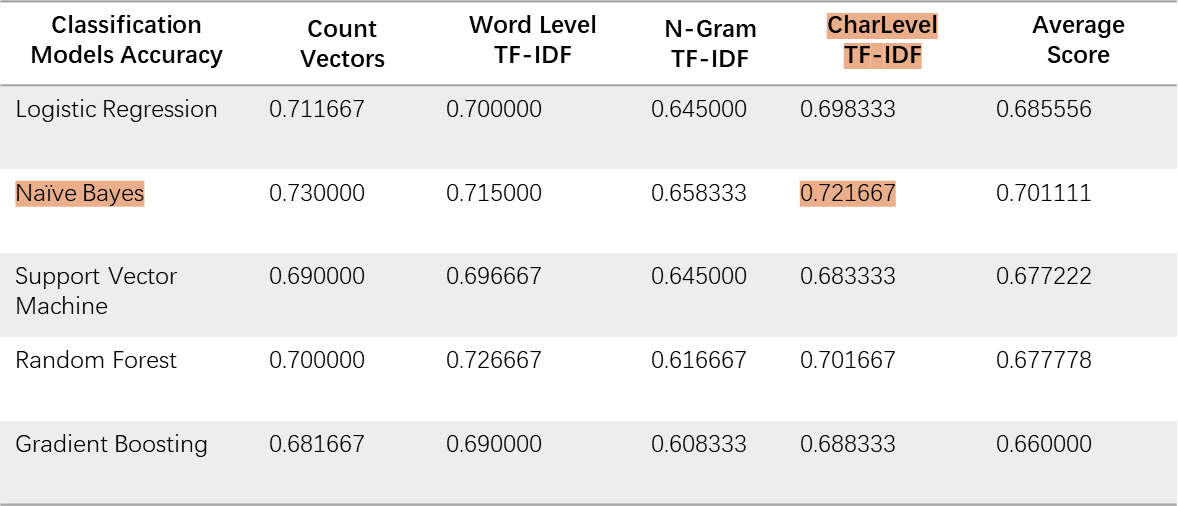
* How long does it take to train your model?

If I use original dataset which including 1.6 million tweets, it took 11 hours even I hosted on google colab. After I processed, I only select 3000 tweets, I only need 5 minutes to train my model.

* What are the tools used? (Cloud platform, for example)

I used colab notebook to host on the cloud for running code for me

* What are the model performance metrics?



* Which model was selected?

Naïve bayes model has been selected

## Outcomes

* What are the main findings and conclusions of the data science process?

The model accuracy scores are having similar scores, Naïve Bayes have the highest score maybe because Naïve Bayes model is good at probability rate. Support Vector Machine doesn’t perform very well compared with others overall.

## Implementation

* What are the considerations for implementing the model in production?

This model can implement on the business intelligence platform about the potential customers’ sentiment to aware customers’ interests and needs about company products.

# Data answer

* Was the data question answered satisfactorily?

Yes, the question is answered, I used density chart and swarm strip to check the balance level of the datasets. Also used word cloud to visualise the most popular words.

* What is the confidence level in the data answer?

High confidence level.

# Business answer

* Was the business question answered satisfactorily?

Yes, the question is answered, I successfully used sentiment analysis to process text predict sentiment of the given texts

* What is the confidence level in the business answer?

High confidence level.

# Response to stakeholders

* What are the overall messages and recommendations to the stakeholders?

Recommendations to the stakeholders are:

Using sentiment analysis existing customers and potential customers’ needs are valuable for business, it brings critical information will affect the

# End-to-end solution

* What is the overall end-to-end solution to use the model developed in the project?

The overall end-to-end solution is extract data, slicing data to 3000 tweets in total, including 1500 positive tweets and 1500 negative tweets, I did add length in the model, I also add count vectors, word vectors, word density and text-based features to get the dataset ready for models, next step is using multiple classification models including logistic, naïve bayes, support vector machine, random forest, gradient boosting to compare the accuracy scores, after this step, I used given unsee text sample to test the model

# References

* Where are the data and code used in the project? (show a simplified list of main items: notebooks, datasets, exported models)
* What are the resources used in the project? (libraries, algorithms, etc)
* The libraries include:
  + Gc, re, string, spacy, numpy, pandas, matplolib, seaborn, datetime, collections, wordcloud, nltk, classification models.
* The algorithms:
  + Count vectors, TF-IDF vectors, texted based vectors
  + Logistic Regression, Naïve Bayes, Support Vector Machine, Random Forest, Gradient boosting models