MSc. in Computing
Practicum Approval Form

Section 1: Student Details

Project Title:	Exploring different approaches for sentiment analysis.
Student ID:	20210378 20210149
Student name:	Rohit Nair Ashwin Bharadwaj
Student email	rohit.nair2@mail.dcu.ie ashwin.sridhar3@mail.dcu.ie
Chosen major:	Data Analytics
Supervisor	Dr. Jennifer Foster
Date of Submission	06/12/2020

Section 2: About your Practicum

What is the topic of your proposed practicum? (100 words)

Proposed Topic: Comparing machine learning and lexicon-based approach for sentiment analysis.

Please provide details of the papers you have read on this topic (details of 5 papers expected).

1. R. K. Bakshi, N. Kaur, R. Kaur and G. Kaur, "Opinion mining and sentiment analysis," 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom), New Delhi, 2016, pp. 452-455.
2. M. Wongkar and A. Angdresey, "Sentiment Analysis Using Naive Bayes Algorithm Of The Data Crawler: Twitter," 2019 Fourth International Conference on Informatics and Computing (ICIC), Semarang, Indonesia, 2019, pp. 1-5, doi: 10.1109/ICIC47613.2019.8985884.
3. X. Fang, J. Zhan, Sentiment analysis using product review data. Journal of Big Data 2, 5 (2015).
4. U. Kumari, A. K. Sharma and D. Soni, "Sentiment analysis of smart phone product review using SVM classification technique," 2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS), Chennai, 2017, pp. 1469-1474, doi: 10.1109/ICECDS.2017.8389689.
5. Y. Wang, "Advanced Naïve Bayes Algorithm Design with Part-of-Speech Tagger on Sentiment Analysis," 2017 International Conference on Computer Systems, Electronics and Control (ICCSEC), Dalian, 2017, pp. 1382-1385, doi: 10.1109/ICCSEC.2017.8446798.

How does your proposal relate to existing work on this topic described in these papers? (200 words)

R. K. Bakshi et al. [1] used a mathematical approach to derive sentiment from data collected from microblogging websites. Sentiment score for each sentence is calculated by tokenisation of words from the dictionary. M. Wongkar and A. Angdresey [2] used Naïve Bayes algorithm to identify user sentiment from data fetched from twitter. Probability of each word being positive or negative is calculated and multiplied. This result is used to classify the word as positive or negative. Xing Fang

and Justin Zhan [3] determine sentiment of amazon reviews using POS tagging and feature vector classification. U. Kumari et al. [4] explore SVM classification technique to classify user sentiment for smart phone product review. Y. Wang [5] used Naïve Bayes algorithm in combination with POS tagger to improve the efficiency of the prediction model.

What are the research questions that you will attempt to answer? (200 words)

1. How different techniques compare in terms of accuracy and precision?
2. Which approach performs better for each type of data set?

How will you explore these questions? (Please address the following points. Note that three or four sentences on each will suffice.)

- What software and programming environment will you use?
    - Spyder and Jupyter Notebook will be used for building the model. Additional libraries such as NLTK, Scikit-learn and Numpy will also be used.
- What coding/development will you do?
    - Different models will be trained using various algorithms on Python. Implementation of lexicon-based approach and Naïve Bayes algorithm will be done. 3 standard data sets will be used to train and test the models.
- What data will be used for your investigations?
    - Labelled data for a given subdomain will be used to train the model and test its accuracy.
- Is this data currently available, it not, where will it come from?
    - Stanford Artificial Intelligence Laboratory contains labelled data sets of IMDB movie reviews and amazon user reviews.
    - There are 12000 positive and negative movie reviews for training as well as testing.
    - Amazon data set has approximately 120000 records and the ratings are on a scale of 5 which was acquired from Stanford Network Analysis Platform. This will have to be converted to positive or negative to train the model.
    - The twitter data set consists of 1.5 million tweets labelled with 1 and 0. 1 being positive and 0 being negative. This data set was acquired from Thinknook.com.
- What experiments do you expect to run?
    - Implement lexicon-based approach and Naïve Bayes algorithm and test how accurate the predicted sentiment is.
- What output do you expect to gather?
    - The trained model will be able to identify the sentiment of users from the data provided. Performance of the 2 models will be compared for each subdomain.
- How will the results be evaluated?
    - The results will be evaluated by calculating the accuracy percentage of the model based on the difference in predicted sentiment and actual sentiment.