MSc. in Computing
Practicum Approval Form

Section 1: Student Details

Project Title:	Predicting The Legitimacy Of Reviews Using Machine Learning Algorithms
Student ID:	20210415 21263045
Student name:	Kashish Kishinchandani Aditi Bhat
Student email	kashish.kishinchandani2@mail.dcu.ie aditi.bhat2@mail.dcu.ie
Chosen major:	Data Analytics
Supervisor	Prof. Manoj Kesavulu
Date of Submission	28/01/2022

Section 2: About your Practicum

Please answer all questions below. Please pay special attention to the word counts in all cases.

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

Reviews help the consumers choose between products. However, sometimes vendors use bots to post reviews(bogus/fake) to increase the ratings and reach of the products. This undermines the product's actual reviews, thereby negatively affecting consumer satisfaction. Our research focuses on predicting the legitimacy of reviews using Machine Learning algorithms to ensure consumers are mislead by fake reviews.

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

1. Khan, Hanif, et al. "Fake review classification using supervised machine learning." *International Conference on Pattern Recognition*. Springer, Cham, 2021.

2. P. Shetgaonkar, J. T. Rodrigues, S. Aswale, V. L. K. Gonsalves, J. C. Rodrigues and A. Naik, "Fake Review Detection Using Sentiment Analysis and Deep Learning," 2021 International Conference on Technological Advancements and Innovations (ICTAI), 2021, pp. 140-145, doi: 10.1109/ICTAI53825.2021.9673375.

3. S. M. Anas and S. Kumari, "Opinion Mining based Fake Product review Monitoring and Removal System," 2021 6th International Conference on Inventive Computation Technologies (ICICT), 2021, pp. 985-988, doi: 10.1109/ICICT50816.2021.9358716.

4. I. Amin and M. Kumar Dubey, "An overview of soft computing techniques on Review Spam Detection," 2021 2nd International Conference on Intelligent Engineering and Management (ICIEM), 2021, pp. 91-96, doi: 10.1109/ICIEM51511.2021.9445280.

5. G. Shahariar, Swapnil Biswas, Faiza Omar, Faisal Shah and Samiha Hassan, "Spam Review Detection Using Deep Learning", 2019 IEEE 10th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)

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

The proposal focuses on predicting the legitimacy of the reviews from the dataset. In [1], the supervised learning technique is employed for classifying spam and genuine reviews starts by entering input review, pre-processing the review, and finally classifying it as fake and genuine. In [2][3][5], different techniques are used to detect fake reviews such as NB, SVM, LSTM, Bi-LSTM, GRU and more. Using the research done in [4], we intend to use it as a framework for our research, we will be using ML algorithm and Deep Learning to predict the legitimacy of the review and will be evaluating the results using the activation function. There have been a few gaps in the existing research where the right characteristics or all characteristics have been used which has led to the increase in the computational speed. The focus of our research primarily relies on using new and relevant characteristics to be passed to the model to increase the accuracy along with reduced computational power. Some of the relevant characteristics which will be used are number of funny feedbacks, suspicion degree and review frequency. CNN used in some places requires fine-tuning of the model's hyper parameters. In addition to this, in the feature selection we will be employing the GloVe method for word embedding.

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

- How to apply ML algorithm(s) along with activation functions on a dataset of reviews for the prediction of fake and genuine reviews?
- How to evaluate the efficiency of the method employed?
- How does the employed method differ from the previous methods?

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?**

- Programming Language: Python
- Environment: Jupyter Notebooks, Jupyter Labs, AWS/GCP

- What coding/development will you do?

- The dataset will be acquired from an already existing dataset. The dataset will then be preprocessed using NLP techniques.
- Feature engineering techniques will be employed on this dataset and machine learning algorithm will be employed to train the model.

- Lastly, the accuracy will be evaluated using activation functions.

- What data will be used for your investigations?

- The dataset will be acquired from:
  - Yelp reviews dataset - https://www.yelp.com/dataset
  - Kaggle e-commerce reviews dataset - https://www.kaggle.com/furkangozukara/turkish-product-reviews

- Is this data currently available, it not, where will it come from?

- The Kaggle datasets are readily available.
- The Yelp dataset has been requested and acquried.

- What experiments do you expect to run?

- We will perform preprocessing on the dataset using methods like tokenization, stop word removal and other NLP processes. Additionally, we will employ techniques to reduce noise in the data to avoid faulty training.
- We will use GloVe method for feature extraction. This technique allows us to find the distance between two word vectors thereby helping us measure the semantic similarity between two word vectors.
- Lastly, we will use ML algorithm to train the data and evaluate results using activation functions.

- What output do you expect to gather?

- The practicum's output is to predict the legitimacy of reviews with high training and validation accuracy and low loss.