School of Computing
Project Initiation Document

William Green

Domain analysis of feature implementations between Classic and Deep NLP models.

PJE40

1. Basic details

Student name:	William Green
Draft project title:	Domain analysis of feature implementations between Classic and Deep NLP models.
Course:	BSc Computer Science - C0056S
Project supervisor:	Dr. Alaa Mohasseb
Client organisation:	N/A
Client contact name:	N/A

2. Degree suitability

Please describe how your project satisfies the criteria for your current course. For example, if you are a Software Engineering student, please explain why your project is suitable for a Software Engineering degree.

In each section, please write your text below ours in regular (non-italic) font.

The proposed project will be suitable for a BSc Hons Computer Science degree because it will include current areas of interest and research, I have chosen to base my project on Natural Language Processing which will identify different model implementations. For this project, I will be including classical models, machine learning and programming skills acquired throughout my time at University.

When choosing my project, I sought to find a challenging but manageable idea which includes opportunities to learn various aspects of Computer Science Theory, specifically how I can view mathematical models and representations of current Computer Science Theories. I am taking on a new area of interest, one that has not been formally taught to me and am taking full advantage of the resources offered dedicated to my degree. My degree teaches me how to process information and create solutions based on know-how and intuition with an efficient result; this project will test the past 2 years of my undergrad and present new philosophies of data-science.

Throughout my project, I will be mainly programming in Python, in which I will be implementing (Convolutional) Neural Networks to train a machine learning model to apply transformations on a dataset. “Computer science is the study of algorithmic processes and computational machines”, I aim to look at several algorithmic implementations and features to best decide the efficiency between classic models and a newer if not new implementation of how I can compute information.

3. Outline of the project environment and problem to be solved

*For engineering **projects without a client**:*	*For projects **with a client**:*	*For theoretical or **study** projects:*
What is the problem that you will investigate? *Why is it worth working on?*	*Who is the client?* *What do they do?* *What is their problem?* *Why does it need to be solved?*	*Who is the intended readership/audience?* *What is the contextual significance of this topic?* *What are the research questions you are seeking to answer?*

My project focuses on the efficiency of different NLP models and how they interact with their respected datasets; there are many domains within NLP which have their subsequent areas of interest, I will be looking for variable changes as to how classical approaches differ to deep-learning implementations of the same domain. In this instance, can the use of deep learning affect the performance results of text classification using ensemble learning.

Research suggests there are inconsistencies in newer implementations as buzzwords are being attached to different problems, it seems this idea is novel as it is based upon current research which has not applied these adjacently applied these techniques. It is worth the time and effort as it is new and may have potential to give valuable insight when designing or merging algorithmic approaches.

4. Project aim and objectives

What is the overall aim of the project?
What are the objectives that will lead to you meeting that aim?

This project seeks to outline any performance differences between classical NLP models and the use of deep learning to existing approaches; ideally, this project aims to profile and focus on any form of performance gains by applying CNNs to classical models such as the bag-of-words concept.

I will achieve this by creating three to five different NLP models with their deep learning counterpart, this will allow me to have enough variation in my results to ensure there is no overlap and that the results are accurate.

5. Project deliverables

For an engineering project, what information system artefacts will be developed? What documents will be produced? This always includes your project report but could also include supporting documentation for your client such as requirement and design specifications, test strategies, user guides, that are useful outside of the project report.

For a study project, are there anticipated outcomes besides the report, for example datasets or recommendations to external bodies?

This project will have several direct and related deliverables with a detailed final report, my report will include all, and any python source code developed throughout this dissertation, this will most likely be handled by screengrabs and version control software in-order to display sequential updates in the form of commits. I intend to produce and deliver three to five deep learning models each possessing different characteristics.

Title	Type	Description
Dissertation Report	PDF/ Report	A concise overview of the project.
Source Code	.py files	Python code written during report.
Machine Learning Networks	.py files and mathematical models	Evidence to support my report.
Results	Graphs and Tables	All outcomes of this project.

6. Project constraints

What constraints are there on your solution to the problem? For example, you could not test a medical system on real patients.

This project will likely come with constraints common to most projects.

Constraint	Description
Sample Size	The chosen data set may not enough entries to produce valuable results.
Lack of Reliable Data	My research and or results may not produce reliable/ consistent data.
Lack of Prior Research	As this idea is seemingly novel, there may not be enough supporting material to use as guidance for improvement.
Lack of Technical Ability	Oneself may not be able to produce the required program.
Self-Reported Data	The quality of the results may be questionable.
Time	We can only devote so much time to a dissertation, it may not be enough to achieve a level of satisfaction or desired outcome.

7. Project approach

How will you go about doing your project? What background research do you need to do? For an engineering project, how will you establish your requirements? For a study project - can you refine your larger research area into research questions that you can meaningfully answer? What skills do you require and how are you going to acquire those that you do not already have? What methodologies are you going to use?

I will conduct preliminary research of a wide-scale approach, using existing research to gain knowledge on the individual concepts and interconnectivity between them, I plan to use this research to support my hypothesis and adapt any future work.

I will then conduct specifically targeted research to find possible gaps in existing research to support why my idea may be a viable option to fill potential gaps.

Once I have ironed out the theoretical concepts I will be using, I will start to look at language specifics, in this case Python Libraries and or Frameworks I can use to aid the development of this project's program. I am aware of libraries such as: NumPy, SciKit-Learn, NLTK and a higher-level abstraction service WordNet API for python.

I will approach my project problem with a logical methodology, in this case it is most effective to start with the Top-Down approach as it implies the problem can be compartmentalized into smaller modules and tackled separately until there is a full hierarchical module. Starting with a base idea, I can have a high-level approach and engineer lower-level aspects as I get deeper into the problem set.

8. Literature review plan

What are the starting points for your research? (e.g. specific books or papers in journals, existing reports or documents, online resources, existing systems)

For my literature review, I will start by looking at the combined words of Andrea Ferrario and Mara Naegelin for their paper titled "The Art of Natural Language Processing: Classical, Modern and Contemporary Approaches to Text Document Classification" as it looks at preprocessing of a dataset and classical bag-of models compares with NNs for text classification. In addition, I will be reading "Combining Machine Learning and Natural Language Processing to Assess Literary Text Comprehension" by Balyan and McCarthy due to its nature of applying machine learning to an ensemble transformer. These are starting points for my research.

9. Facilities and resources

What computing/IT facilities will you use/require?
What other facilities/resources will you use/require?
Are there constraints on their availability? If funds are required to acquire them, have these been allocated? Will they be available in time?
For example, you might need a specialist lab or equipment at the university, which might be in use in teaching and by other project students. Your own computer and free software, or software you already have, do not normally need to be mentioned.

For this project, I will not require the use of university facilities or any external resources, I do not plan for this project to incur costs.

10. Log of risks

What risks will you encounter when doing your project? What backup plans do you have if identified things go wrong?
What is your plan for reviewing risks? Remember that risk probabilities, and hence priorities, will change over the course of the project, so this section should be maintained. Use a table like below.

Description	Impact	Likelihood	Mitigation	First indicator
COVID-19 outbreak means I cannot get into a lab for usability testing	*Severe*	*Likely*	*Get in while I can, prioritise lab tasks in time.* *Make an alternate test plan that does not need the lab.*	*University informs that lab closure is likely*

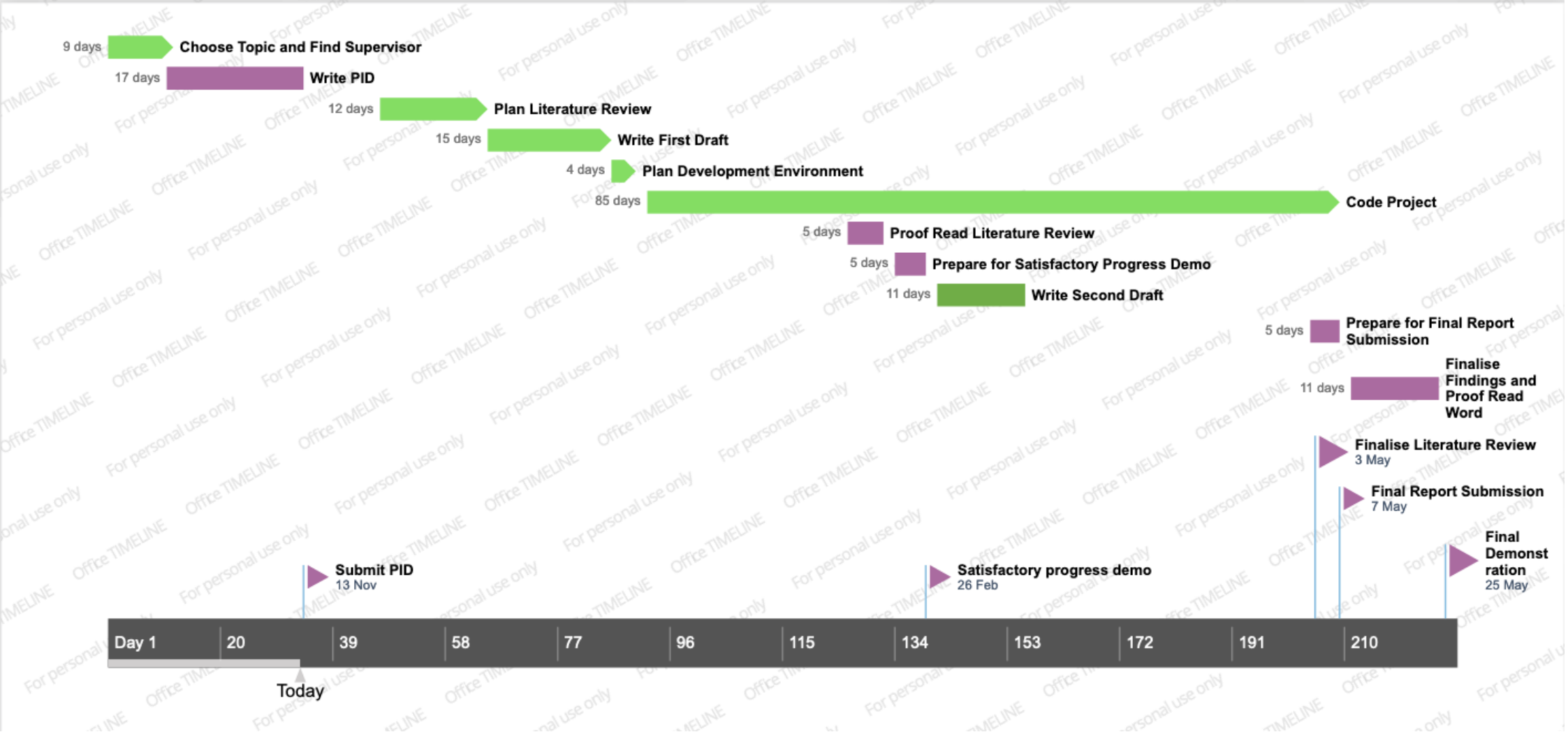
Description	**Impact**	**Likelihood**	**Mitigation**	**First indicator**
Time Issues	Severe	Likely	Create a detailed plan for time management allowing for unforeseen circumstances.	Falling outside of the created plan.
Learning Curve	Severe	Likely	Prepare research and give enough time to learn new material.	Struggling to conceptualize new material and implement theoretical aspects.

Requirements are not well defined	Severe	Unlikely	Thorough requirement elicitation	Project scope changes.
Unplanned Work relating to knowledge gaps	Moderate	Likely	Rely on current programming skills and logic overcome academic challenges	Struggling to complete or implement theoretical aspects.
dependency issues	Low - Moderate	Likely	Virtual Environments for Python	Dependencies do not work.

11. Project plan

What do you need to do to create the artefact / do the primary research and write the report? Walk through your proposed approach and break it down into tasks.
When are you planning to perform these tasks? When do you need access to other people or resources? Usually a Gantt chart is a good way of presenting the plan.
Note that plans can change over the course of the project, so this plan should be maintained.

Project overview Gantt Chart



12. Legal, ethical, professional, social issues (mandatory)

What are the legal/ethical/professional/social issues that may impose constraints on the project? How will you ensure that they will be addressed, or what steps will you take to avoid/mitigate their effects?

Whatever project work you are doing, you must consider its security implications, for the data you generate or use, or for the software artefact itself. Please describe how you are taking these into account. There is also a question about security on the ethics review form (see below)

*All students must complete the ethics review form at https://ethicsreview.port.ac.uk at this time. Has your supervisor (and the FEC representative, if required) seen and approved your ethics form? **Remember** – this is obligatory and must be completed **now**.*
The school's FEC representatives are Dr Matt Dennis and Dr Philip Scott (not Dr Carl Adams as the output of the review may say).

UNIVERSITY OF PORTSMOUTH

Certificate of Ethics Review

Project Title: Domain analysis of feature implementations between Classic and Deep NLP models.

Name: WILL GREEN **User ID:** 853829 **Application Date:** 12-Nov-2020 23:24 **ER Number:** ETHIC-2020-1528

You must download your referral certificate, print a copy and keep it as a record of this review.

The FEC representative for the School of Computing is Carl Adams

It is your responsibility to follow the University Code of Practice on Ethical Standards and any Department/School or professional guidelines in the conduct of your study including relevant guidelines regarding health and safety of researchers including the following:

- University Policy
- Safety on Geological Fieldwork

It is also your responsibility to follow University guidance on Data Protection Policy:

- General guidance for all data protection issues
- University Data Protection Policy

Which school/department do you belong to?: **SOC**
What is your primary role at the University?: **UndergraduateStudent**
What is the name of the member of staff who is responsible for supervising your project?: **Dr Alaa Mohasseb**
Is the study likely to involve human subjects (observation) or participants?: **No**
Are there risks of significant damage to physical and/or ecological environmental features?: **No**
Are there risks of significant damage to features of historical or cultural heritage (e.g. impacts of study techniques, taking of samples)?: **No**
Does the project involve animals in any way?: **No**
Could the research outputs potentially be harmful to third parties?: **No**
Could your research/artefact be adapted and be misused?: **No**
Does your project or project deliverable have any security implications?: **No**
Please read and confirm that you agree with the following statements: **Confirmed**
Please read and confirm that you agree with the following statements: **Confirmed**
Please read and confirm that you agree with the following statements: **Confirmed**

Supervisor Review

As supervisor, I will ensure that this work will be conducted in an ethical manner in line with the University Ethics Policy.

Supervisor signature: Alaa Mohasseb **Date:** 13/11/2020