School of Electronic Engineering and Computer Science
Queen Mary University of London

MSc PROJECT DEFINITION
D2022-23

This project definition must be undertaken in consultation with your supervisor. The feasibility of the project should have been assessed and the project aims should be clearly defined.
Submission of this document implies that you have discussed the specification with your supervisor.

Project Title: Optimisation of Good News Classification using Large Language Models

Supervisor: Dr Bruno Ordozgoiti

Student name: Hannah Melkemaryam Claus

Student e-mail: h.m.claus@se22.qmul.ac.uk

PROJECT AIMS:
State the design, development or research challenge (problem) that the project aims to solve.

The main goal is to create a machine learning (ML) classifier which discriminates between relatively positive and relatively negative news. The motivation behind this goal is to provide users with international uplifting news which potentially improves the user's mental health and supports progress in society.

At the moment, trained journalists go through news providers manually to select uplifting news. However, this can be simplified and automated by utilising a neural network. This way, more positive news can be made accessible in less time and without linguistic limitations regarding different languages. Nevertheless, the first step is to create such a classifier for international news written in English. This can then be extended to be applied in other languages as well.

PROJECT OBJECTIVES:
List a series of objectives you need to achieve in order to fulfil the aims of your project.

The main objective is to create a framework for a classifier which discriminates between positive and negative news. In order to receive the best possible outcomes, there will be two types of models:

- optimised baseline model
- optimised Large Language Models (LLMs)

These three types of models will be trained and tested on three types of classifications of news articles:

- headlines
- content
- headlines + content

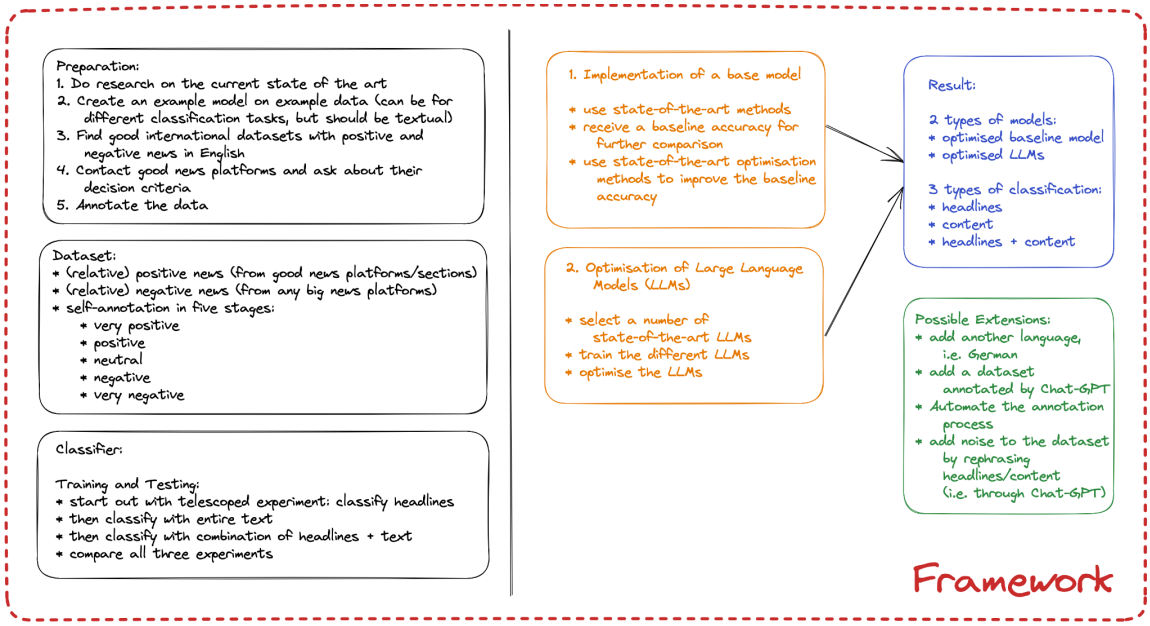
Hence, at the end, there will be a minimum of six different classifiers that can be compared to each other and to the current state-of-the-art.

State how your project will be aligned with the learning outcomes of your programme of study.

The framework will be written in Python, using TensorFlow and Keras libraries, as well as other packages that were introduced during the lectures. Utilising the skills and knowledge acquired from the Natural Language Processing and Machine Learning lectures, the neural networks will be built using familiar environments, converting theory into practice.

METHODOLOGY:

Describe the various steps that you intend to follow in order for you to achieve your project aims.



The above image shows the general methodology regarding the final framework.

1. Review and analyse the current state-of-the-art methods
2. Annotate the dataset, which contains various kinds of news, to allow for a diverse dataset
3. Implement a base model
4. Optimise the base model
5. Implement Large Language Models and optimise them
6. Compare the results with different datasets
7. Create a pipeline to automate these processes and put everything together as a framework for future work

The above steps leave room for further possible extensions, if time and resources allow it:

1. add news articles in other languages than English to the dataset
2. let Chat-GPT annotate the dataset
3. automate the entire annotation process

4. add noise to the dataset by rephrasing headlines/content, for example using Chat-GPT

Justification:

- In order to use classifiers in applications, they need to undergo certain training and testing stages to ensure trustworthiness and quality of the final results
- However, these stages are time-consuming and therefore, need automated techniques, so that more time can be spent on the actual applications of said classifiers
- The framework will be created, so that in future arbitrary classifiers (e.g. for NLP) can undergo the same automated processes as the developed classifier

PROJECT MILESTONES

Indicate what measurable/tangible components you will produce as part of this project. This may take the form of deliverable document(s) or developmental milestones such as a working piece of software/hardware.

The final component will be the framework, which would be available open-source. When breaking the various parts of the framework down, it is easily separated into the dataset and the classifier. The dataset contains self-annotated data, which will be made publicly-available.
The classifier component entails the two developed types of classifiers, which were all trained and tested with three types of data, hence, at least six different classifiers.

This framework can then be used to successfully classify a news article as positive or negative.

REQUIRED KNOWLEDGE/ SKILLS/TOOLS/RESOURCES:

Indicate as far as possible the skills that are required for you to undertake this project. Also include any software, hardware or other tools or resources that you believe you will need.

GitHub Repository: https://github.com/melkemaryam/m_thesis

Programming:

- Python

    - TensorFlow/Keras
    - Numpy
    - Pandas
    - NLTK

- Data:
    - Kaggle

- Software:
    - Visual Studio Code/Sublime Text
    - Jupyter Notebooks
    - Google Colab

Resources:

- Datasets
- Laptop/Computer
- GPUs/TPUs

TIMEPLAN

This can be a GANTT chart submitted with this document or a list of tasks, milestones and deliverables with timings.

Weeks left	Week of	Task	Sumission
	April	- Review and analyse the current state-of-the-art - Collect datasets	
	May	- Annotate datasets - Create an example model - Create base model - Optimise base model	
	June	- Optimise LLMs - Compare results - Start writing the first draft of the dissertation	

		paper	
7	03.07. - 09.07.	- Finalise the first draft of the dissertation paper	
6	10.07. - 16.07.	- Start writing the reflective essay	Draft Dissertation Paper
5	17.07. - 23.07.	- Finalise the setup of the framework (open-source) - Add full code documentation	
4	24.07. - 30.07.	- Get feedback and revise the texts	
3	31.07. - 06.08.	- Polish the models and the data	
2	07.08. - 13.08.	- Get feedback and revise the texts	
1	14.08. - 20.08.	- Record the video	
0	21.08. - 28.08.	- Final read through	- Dissertation Paper - Reflective Essay - Video
0	28.08. - 15.09.	- Viva preparation	Viva