Agreement on Research Project/Master Thesis between:

* 1. Student

**Name: Chen Wang**

**Student number: s2230496**

**E-mail address: blcuwc@gmail.com**

* 1. Supervisor(s)

**Name: Suzan Verberne**

**Role: Primary supervisor**

**E-mail address:** [**s.verberne@liacs.leidenuniv.nl**](mailto:s.verberne@liacs.leidenuniv.nl)

**Name: Anne Dirkson**

**Role: Secondary supervisor**

**E-mail address:** [**a.r.dirkson@liacs.leidenuniv.nl**](mailto:a.r.dirkson@liacs.leidenuniv.nl)

* 1. Research project

**Project summary:**

**Sentiment analysis on medical social media could help people to understand the impact of the medical condition on the patient’s overall wellbeing. The eDiseases dataset is a dataset that contains more than 3500 posts related to three different diseases, annotated from well-known MedHelp community. By classifying these sentences into different classes, different text-representation methods and classifiers can be experimented and compared. XLNet is one of the state-of-art neural networks that used for pre-training language model, which has been verified in many NLP tasks. This project aims at through applying XLNet and different classifiers to eDiseases dataset, comparing their performance to existing experiment results, to find out the influence of pre-training language model in analyzing medical social media, and how different classifiers perform in medical social media.**

**Working title: Analyzing sentiment of medical social media with state-of-art pre-trained language model**

**Date of project start: 02-09-2019**

**Deadline for handing in the report: 29-11-2019**

**Initial project plan / global planning:**

High level work breakdown:

1. Crawl the posts from MedHelp, re-produce eDiseases Dataset

2. Read NLP and medical related literatures, be familiar with PyTorch

3. Pre-process and augment original dataset

4. Combine XLNet and different classifiers to classify eDiseases Dataset

5. Adjust experiment settings and start writing report

6. Analyze experiment results and finish report

The global planning is as follows:

|  |  |
| --- | --- |
| Sep 2019 | Crawl the dataset from MedHelp, read related literatures, and be familiar with PyTorch library |
| Oct 2019 | Pre-process original dataset, apply some text augmentation methods to original dataset, and run XLNet in PyTorch successfully |
| Nov 2019 | Combine implemented XLNet from PyTorch and different classifiers to classify eDiseases Dataset, start writing theoretical part of the report, adjust experiment settings |
| Dec 2019 | Compare and analyze experiment results under different text representation methods and different classifiers, finish report |
| Other agreements: |  |

Signatures

**Date: 10-9-2019**

**Location: Leiden**

|  |  |  |  |
| --- | --- | --- | --- |
| **Student**  **Chen Wang** |  | **Supervisor**  **Anne Dirkson** |  |
|  |  |  |  |
|  |  |  |  |