**CE807 – Assignment 1 - Interim Practical Text Analytics and Report**

School of Computer Science and Electronic Engineering - University of Essex

# Assignment Due at 11:59:59am on 06/06/2022

**Electronic Submission**

URL: https://www1.essex.ac.uk/e-learning/tools/faser2/

*Please also see your student handbook for rules regarding the late submission of assignments*

## On Plagiarism

The work you submit must be your own. Any material you use, whether it is from textbooks, classmates, the web or any other source must be acknowledged in your work.

All submissions are fairly and transparently checked for plagiarism. Please make sure that you provide frequent citations. But also make sure that each sentence written is originally yours, i.e. the material is read, understood and the report is written using your own words and own language only. Do not copy and paste and rephrase copied text.

There are many different forms of what is considered plagiarism. For example, based on the feedback from the SAO officer, many students were not aware that, e.g. copying entire paragraphs without clearly identifying them as quote etc. is a form of plagiarism etc. Thus, please check back with your scientific writing module, before you submit!

Further note that also plainly reusing software code or merely slightly adapting existing software code and submitting as one's own fulfils the matter of plagiarism. Cite any code that you reuse, too.

In 2019, 20% of the submitted reports where plagiarised. There were also multiple cases of software code plagiarism. This number is too high and shall be 0% in 2020!

**MOTIVATION:** The task of text classification is classical for text analytics ranging from binary classifications in spam detection to topical classification in scientific libraries, online retailing and others.

**OBJECTIVE:** The objective of this assignment is to explore the landscape of text classification. First, basic methods for text classification shall be identified and discussed. Subsequently, different research areas of text classification shall be identified. Finally, selected papers in the identified fields shall be discussed.

**SUBMISSION, ASSESSMENT AND RULES**

- This assignment counts towards 25% of the overall mark for CE807.
- The assignment is to be done individually or in pairs. If you work in pairs, you each need to submit the same report (please include information about which two reports should be treated as a pair, otherwise you risk that your mark will be zero). Both members of a pair will get the same mark unless there is reason to do otherwise.
- Be sure to put your registration number(s) as a comment at the top of all files. Furthermore, the assessment is blind, i.e. **do not put your name on any document or provide personally identifiable information**.
- The assignment must be submitted in a single zipped archive containing the following exactly three subfolders:

| | |
|---|---|
| CE807/Assignment1/ | All files |
| CE807/Assignment1/Task1 | List of scientific papers found and considered for the review of basic approaches for text classification. |
| CE807/Assignment1/Task2 | List of scientific papers found and considered for identifying different fields of text classification and review of advanced methods and approaches for text classification. |
| CE807/Assignment1/Task3 | The scientific report produced for Task 3. It is mandatory to include the source files (Word/LaTeX) and a PDF. |

# Task Overview: Survey on Text Classification

**MOTIVATION:** The task of text classification is classical for text analytics ranging from binary classifications in spam detection to topical classification in scientific libraries, online retailing and others.

**OBJECTIVE:** The objective of this assignment is to explore the landscape of text classification. First, basic methods for text classification shall be identified and discussed. Subsequently, different research areas of text classification shall be identified. Finally, selected papers in the identified fields shall be discussed.

Examples of recent research papers in different fields of text classification are provided below.

Getting started

Read the recent papers on text classification by

- L. Galke, F. Mai, A. Schelten, D. Brunsch, and A. Scherp: Using Titles vs. Full-text as Source for Automated Semantic Document Annotation, Knowledge Capture (KCAP); Austin, TX, USA, 2017.

- F. Mai, L. Galke, and A. Scherp: Using Deep Learning For Title-Based Semantic Subject Indexing To Reach Competitive Performance to Full-Text, Joint Conference on Digital Libraries (JCDL); Fort Worth, TX, USA, 2018.

- Liang Yao, Chengsheng Mao, Yuan Luo: Graph Convolutional Networks for Text Classification. AAAI 2019: 7370-7377

- Jake Snell, Kevin Swersky, Richard S. Zemel: Prototypical Networks for Few-shot Learning. NIPS 2017: 4077-4087

Explore further papers, both covering basic machine learning methods as well as advanced methods for text classification.

## _____ TASKS _____

Your tasks will be as follows:

In **Task 1**, you will **read the literature** to produce a concise critical discussion of the basic methods in text classification. This will serve as foundation for Task 2 and be written up in a report in Task 3

In **Task 2**, you will **identify specific fields and advanced methods of text classification**. You will **read selected literature in the identified fields** to produce a concise critical discussion of classification. This will be written up in a report in Task 3.

In **Task 3**, you will **write a scientific report** about the papers identified and read for Tasks 1 and 2. You should not only intellectually summarise the literature but also reflect on the advantages and disadvantages of the different approaches, which alternatives one could consider, lessons learned etc.

**We expect this report to be between two to three pages using ACM style plus references, which do not count to the page limit.**

Organisation and content of the report shall follow a scientific paper

As template for your report, please use the ACM style for conferences, preferably using LaTeX. There is even an Overleaf version of it available.

You can find the template here: https://www.acm.org/publications/proceedings-template

**Important Note: It is mandatory to use the ACM style for formatting the results for reasons of comparability of the different reports being submitted.**

## **MARKING BREAKDOWN** (out of 100%)

**Task 1. Basic text classification methods literature research (20%)**

- Appropriate coverage and contextualisation: up to 10%
- Critical discussion: up to 10%

**Task 2. Identification of specific fields and exploration of advanced methods for text classification research (30%) - Task 1 required**

- Identification of relevant fields for text classification: up to 10%
- Appropriate coverage and contextualisation: up to 10%
- Critical discussion: up to 10%

**Task 3. Report (50%) – Tasks 1 and 2 required**

- Contextualisation and discussion of the state of the art: up to 10%
- Discussion of advantages and disadvantages: up to 10%
- Discussion of alternative approaches for text classification: up to 10%
- Lessons learned: up to 10%
- Material submitted as requested in ACM style etc: up to 10%