CE807 – Resit-Assignment - Practical Text Analytics and Report

School of Computer Science and Electronic Engineering - University of Essex

Electronic Submission and check submission date and time on FASER (https://faser.essex.ac.uk/)

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 your own language only. Do not copy and paste and rephrase the 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 unaware that, e.g., copying entire paragraphs without clearly identifying them as quotes, etc. is a form of plagiarism, etc. Thus, please check back with your scientific writing module before you submit it! Please see the university's academic integrity policy.

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

In 2023, ~25% of the submitted reports were plagiarised. There were also multiple cases of software code plagiarism. This number is too high and shall be 0% in 2024!

MOTIVATION: Identifying sentiment is paramount to the public and companies for a better user experience. If there is a negative sentiment about a product, it might make less business.

OBJECTIVE: The objective of this assignment is to get practical experience in designing, implementing, running, and scientifically evaluating your classifier. You will train, validate, and evaluate the provided dataset. The assignment's goal is not just to build models but also to justify your modelling choices. During training, you could use other datasets of your choice with justification.

Task Description: Text classification is one of the standard tasks in text analytics. The assignment aims to classify a given text's **sentiment on Automotive review on ecommerce** using the dataset provided. Note you must use the dataset provided. You are allowed to use other datasets with explanation.

Dataset: You will be provided a dataset containing three files: train, valid, and test CSV. You will use a train file to train your model, a validation set to find the best parameters for the model, and a test file to produce model output. Remember, you only must use the test file once and can't use it to find parameters. For the train and validation set, you will be provided a "sentiment" label, and not for the test set. Using your models, you must generate a " sentiment" label for the test set. There are different train/validation/test datasets for all students.

Dataset: https://drive.google.com/file/d/1XL_KSJ5KOyyzRpWP1AdOGN0my7s8C6Ko/view?usp=sharing

Model: You will select two different models/methods such that one is **Generative classifier**, and the other is a **Discriminative classifier**. The model selection should be such that it provides the best performance on the given task and dataset. The better the performance of the model, the higher the marks. Note you don't have the ground truth label for the test set; that is, you must find the best model based on the validation set.

Code: In Lab 9/10, you are provided with a sample code and its structure. You must follow the same structure for coding purposes. Your code will be automatically executed, so you must follow the instructions to be marked. You are allowed to use codes from the lab or web with proper citations and an understanding of the code. You need to modify the code to fit the structure provided. You must make sure that your code is runnable in one click; any special package requirement needs to be automatically installed.

Output: You have to save the model's output in the "test.csv" by adding/modifying "out_label_model_gen" and "out_label_model_dis" columns in the existing columns from test.csv. Note that you shouldn't change any other column name in the "test.csv" file. You must use the same column names; otherwise, the automatic code will not work and will not be marked. You will also save models over the web (like GDrive) so that they can be automatically accessed, and your code should automatically download these models and the required files to run the model.

Report & Presentation: You must submit the report in the ACL format with the same structure as provided. The link to the report format will be provided on the Moodle page. You must present a video presentation of what you have done. The slide, your face, and audio must be present in the video presentation. You must use Zoom to record the presentation. A sample demo/presentation will be shown in the Lecture/Lab 9/10.

_____ TASKS _____

Task 1: Model Selection (20%)

Whenever you develop a new classifier or other text analytics software, you must select the best possible model considering the resource and show that your system outperforms state of the art on certain conditions. You must select two different models/methods: one **Generative** and another **Discriminative** classifier. You need to briefly summarize the selected model and critically discuss and justify the model selection. You also need to justify the pre-processing and text representation both theoretically and practically. Finding a way to combine both selected models will have extra credit.

25% weightage is for selecting models with proper reasoning beyond those taught in Lectures/Labs. Just selecting a model is insufficient; justification is the key.

Task 2: Design, implementation of classifiers (30%)

For this task, you will develop your own classifier models using two selected models. Doing this will involve one or more following steps:

- Identifying the approaches and/or features you want to extract.
- Pre-process the text based on the dataset and model requirement.
- Developing code to extract these features from the text (and weight them if you want to use more than simply binary features).
- Train a classifier that uses these features.
- Experimentally provide justification of different pre-processing and text representation.
- Use validation set and/or parameter search to find optimal model setting.
- Evaluate the performance of the classifier using a scientifically sound methodology.
- Initially, start with a subset of the large datasets. Subsequently, scale the classifier to include more and more data until you use the entire dataset or very large parts.
- You must use Google Colab for coding and submit the downloaded ".py" file. You must provide Google Colab link in the report and make sure anyone can access it.
-

25% weightage is for systematically training models on different parameters to select the best models and for properly documented codes.

Task 3: Analysis and Discussion (20%)

The goal of the task is to explain and analysis the model's performance. You need to justify and compare your model's performance using different pre-processing and text representation and compare it with the state-of-the-art (SoTA) method. Justification must contain both theoretical and practical explanations. You will also select at least 5 diverse interesting examples from the validation set, provide their ground truth/true label, and compare it with both models' output. You must compare why a model is correct or wrong. The comparison must be between your models also. You can also look at the model's confidence and comment on it.

25% weightage is for systematically comparing both selected models' performance at a granular level and connecting example output with a theoretical or experimental explanation.

Task 4: Summary (30%)

Finally, you will write a report documenting what you did and your findings. You should explain why you decided on the algorithms and features you used and how this compares to state-of-the-art. You should discuss the performance of your approach and reflect on what you have learned. Just saying you learned new model, coding etc is not sufficient. Your report must be **4-5 pages** excluding references in the ACL format.

You will also prepare a presentation and use zoom to record the presentation for **12-15min** (not more than 15min). You must use the university's zoom, save your presentation on the university cloud, and provide a shareable link to the recorded video in the report.

Video Presentation carries 25% weightage. So, if something is missing in your video, for example, if your presentation video lacks your voice, face or slide, you will lose 25% marks. It is your responsibility that your video works as required.

SUBMISSION, ASSESSMENT, AND RULES

- The assignment is to be done individually.
- Must follow the provided code and report structure.
- **Coding Practice**
  - You must use Python for the coding. You can use Google Colab for your coding and then download ".py" files.
  - Your code should be self-contained and run in one click.
  - **Data Path**: Your code should assume the data is in './data/'.
  - **Model Path**: Your train code should save in './model/student_id/Model_gen/' or './model/student_id/Model_dis/' and during testing fetch model from the same directory based on the options.
  - **Training**: Read data from the data path and train model and save it in the respective model directory.
  - **Model Save:** After training the model you need to save the model and other required files and save it over the web (like GDrive) so that it can be automatically accessed.
  - **Testing**: Read data from the data path and save the output in the respective model directory.
  - **Model output**: Your output file based on the test file will be named "test.csv" and you will add/modify "out_label_model_gen" and "out_label_model_dis" column in the existing columns from test.csv. These outputs will be generated from your trained models.

- **Seed**: Initialize all seeds to your student_id i.e. torch.seed(student_id), np.random.seed(student_id), etc This will ensure that all codes are reproducible and each student have different set of model initialization and data split.
- **Library**: For neural network/deep learning, you must use PyTorch as taught in the labs. If you use any other library like Keras, Tesorflow etc, you will lose some marks. Other than deep learning library, you could use any other library.
- **You code and data will be evaluated automatically, so if you don't follow these instructions, you will not get marks because it would give error.**
- **Code Documentation:** Make sure your code is properly documented and it prints the desired outputs using the "print" statement whenever required. You are allowed to use existing codes with appropriate citations.

- **Report:** You will be provided overleaf and pdf for the report structure. You can add section/sub-section(s) but can't delete any section/sub-section. Your report must be **4-5 pages** excluding references.
  - No need to include code snipped in the report.
  - You can appendix to add some extra examples.
- **Presentation:** You will record a presentation using Zoom. Your presentation should follow the same order as the report and must not exceed 15min. Both your slide and face should be visible in the recording. It is your responsibility that your slide, face, and audio are present in the video. If anything is missing, you will be given zero marks.
- **Links:** It is your responsibility that Model link and Zoom presentation links are working in the report. If it is not working, those parts will not be marked.
- **Report Template**: Report must be written in the ACL Template. It has Latex and Word version; you can use anyone but submit **a PDF file only**. It is **mandatory to use the ACL style** for formatting the results for reasons of comparability of the different reports being submitted**.**
- **Submission**: The assignment must be submitted following files having same name
  - **report.pdf** → Your report file.
  - **presentation.pdf** → Your presentation slide
  - **code.py** → Your code as a single file
  - **test.csv** → Your output for both model

Extenuating circumstances or Late of Submission

For any reason, you are late in your assignment submission. Please don't contact me. There is university wise guideline for this, you should follow that. In general, you should read your student handbook for more details.

- Making an extenuating circumstance claim
    - https://www.essex.ac.uk/student/exams-and-coursework/extenuating-circumstances
- Late submission of coursework
    - https://www.essex.ac.uk/student/exams-and-coursework/late-submission-of-coursework

If you are in any queries or need some advice, please contact csee-assessment@essex.ac.uk or the school Office team for further advice.

Again, please don't contact me regarding EC.

Plagiarism Policy

In case you are reported for suspected academic offences. You will get information from the school office on how to proceed next. Suspected academic offences don't mean you will have a penalty. A team will be involved and decide the next course of action in which I am not involved. **Please don't contact me regarding suspected academic offences.** At every step, the school office will keep you informed. In the meantime, you could look at the University's guidance and procedures about academic offences and understand beforehand what will happen at the meeting. If you think you need some support or guidance before the meeting, you get in touch with SU Advice, who provide free, expert, and independent advice to all students.

The wait time or the process could be stressful. Please get in touch with the University's student wellbeing service, which provides confidential advice and assistance.

Re-marking Policy

It's important to note that being dissatisfied with your mark is not grounds for requesting a remark. You can find the grounds on which you can appeal a mark in section 3 of the marking policy. You can review the marking policy at https://www.essex.ac.uk/student/exams-and-coursework/assessment-and-marking-policies

In minor cases, some feedback might not match the exact submission. For example, in "Training of classifier using selected models," feedback says, "no use of validation set," but you have used a validation set. This makes a case for a relook. You must provide all such feedback, and only those will be relooked. However, that might have a minor or no impact on overall marks because marking is done by looking at everything together, and wrong feedback was selected. Please mail me for these cases with the email subject "**CE807-24-SU: Assignment (1234567)**", where 1234567 is your 7/8 digits student id.

MARKING BREAKDOWN (out of 100%)

Task 1: Model Selection (20%)

- Summary of 2 selected Models **(up to 10%)**
- Critical discussion and justification of model selection **(up to 10%)**
- 25% weightage is for selecting models with proper reasoning beyond those taught in Lectures/Labs. Just selecting a model is insufficient; justification is the key.

Task 2: Design and implementation of classifiers (35%) - Previous required

- Training & Testing of classifier using selected models **(up to 25%)**
- High-quality code including comments and printing required measures etc **(up to 10%)**
- 25% weightage is for systematically training models on different parameters to select the best models and for properly documented code.

Task 3: Analysis and Discussion (25%) - Previous required

- Providing Justification of Model performance and comparing with SoTA **(up to 10%)**
- Example Selection and it's explanation, and other analysis **(up to 15%)**
- 25% weightage is for systematically comparing both selected models' performance at a granular level and connecting example output with a theoretical or experimental explanation.

Task 4: Summary (20%) - Previous required

- Lessons Learned **(up to 10%)**
- Material submitted in appropriate format **(up to 10%)**
- 25% weightage is for video presentation.

Other Points

- **Wrong Model Selection**
    - If you select a wrong model, say instead of 1 generative and 1 discriminative classifier you selected 2 generative classifiers. In that case you will get only 50% maximum marks
- **No report submission**
    - Submission without a report will get you 50% maximum marks
- **No Workable Code submission**
    - Submission without a code will get you 50% maximum marks.