UNIVERSITY
OF LONDON

BSc in Computer Science

Module: CM3060 Natural Language Processing

Coursework description

Coursework assignment: comparative text classification using statistical and embedding-based models

Contents

I. Introduction

This coursework requires you to develop a text classifier and apply it to a specific domain or challenge, e.g. fake news detection, sentiment analysis, spam detection, document tagging, etc., using both statistical and embedding-based language models. You will need to identify a suitable problem area with an associated data set. Additionally, your work must be situated within existing literature, with proper citations and references to high-quality sources. The core technical exercise will involve comparing the effectiveness of a traditional statistical model and a modern deep learning model in addressing the chosen problem.

1. Domain-specific area

The first step of the coursework is to identify and describe the problem or challenge. This is an area of industry or science where text classification methods can contribute. Include relevant literature to support the significance of the chosen area.

2. Objectives

Outline the goals of exploring both statistical and embedding-based models to understand their effectiveness and applicability in text classification tasks. State any contribution which the results may make to the challenge addressed, supported by relevant literature.

3. Dataset Description

The next step is to identify a suitable dataset which is representative of the challenge and will require attention to all the steps outlined in this assignment. Provide a description of the dataset, its size, data types, the way the data were acquired. Clearly state the source of the dataset. Large technology companies, such as Microsoft, Google and Amazon, provide a wide variety of datasets.
Example: 'Fake and real news' dataset available from the Kaggle official website.

4. Evaluation methodology

Describe the metrics (e.g., Accuracy, Precision, Recall, F1-Score) for assessing model performance and discuss how to apply these metrics to compare the two methodologies.

II. Implementation

This part of the coursework is the implementation of the project. It includes preprocessing the data, building and testing your classifier and obtaining results. The project is expected to be developed using the Python language and Jupyter notebook. Provide well-commented Python code, accompanied by a document describing the following steps:

5. Data Preprocessing

Convert/store the dataset locally and preprocess the data. Describe the text representation (e.g., bag of words, word embedding, etc.) and any preprocessing steps you have applied and why they were needed (e.g., tokenization, lemmatization). Address differences in data preparation for statistical models (e.g., frequency tables) versus embedding models (e.g., word vectors).

6. Baseline performance

Describe and justify the baseline against which you are going to compare the performance of your chosen approach. This can be an already published baseline (e.g. cited in the literature) or the results of a basic algorithm that you implement yourself. The baseline should represent a meaningful benchmark for comparison.

7. Comparative Classification approach

Implement both a traditional statistical model and a modern deep learning model. Build a classifier using the appropriate Python library. Detail the architecture, training, and optimization processes for each, emphasizing their strengths and weaknesses in the context of the chosen dataset. Clearly compare the performance of the two models.

8. Programming style
Ensure all code is clear and well-commented. Documentation should include detailed explanations of the rationale behind model choices, parameter settings, and any specific libraries or tools used.

III. Conclusions

9. Performance Analysis & Comparative Discussion
Present and analyze the results for both models. Use visualizations to compare performance across different classes and discuss any significant findings. Critically evaluate the advantages and disadvantages of statistical and embedding-based models based on the results. Discuss scenarios where one might be preferred over the other and hypothesize reasons for observed performance disparities.

10. Project Summary and Reflections
Reflect on the learning experience, the practicality of each model type, and their potential applications in real-world scenarios. Describe its contributions to the problem area and discuss the extent to which your solution is transferable to other domain-specific areas. Suggest improvements and future research directions.

Rubric: marks are shown in parentheses.

I. Introduction

1. Introduction to the domain-specific area (200-500 words)

[0] Missing or incorrect.
[2] Briefly discussed.
[3] Adequately discussed.
[4] Domain-specific area clearly stated, informative description, fully referenced work.
[5] Exceptional work which includes the above but goes beyond what would be expected from a student at this level.

2. Objectives of the project (200-500 words)

[0] Missing or incorrect.
[2] Briefly described.
[3] Adequately described.
[4] Objectives clearly stated with sufficient details and potential contributions.
[5] Exceptional work which includes the above but goes beyond what would be expected from a student at this level.

3. Description of the selected dataset (200-500 words)

[0] Missing or incorrect.
[2] Briefly described.
[3] Adequately described.

[4] Described in sufficient details, including origin, size, structure, data types.

[5] Exceptional work which includes the above but goes beyond what would be expected from a student at this level.

4. Evaluation methodology (200-500 words)

[0] Missing or incorrect.

[2] Briefly described.

[3] Adequately described.

[4] Methods and metrics clearly described with convincing rationale.

[5] Exceptional work which includes the above but goes beyond what would be expected from a student at this level.

II. Implementation

5. Data Preprocessing

[0] Missing or incorrect.

[2] Briefly described.

[3] Working code fragments with some preprocessing steps.

[4] All appropriate preprocessing steps undertaken, clearly described with convincing rationale.

[5] Exceptional work which includes the above but goes beyond what would be expected from a student at this level.

6. Baseline performance

[0] Missing or incorrect.

[2] Briefly described.

[3] Adequately described, some justification provided.

[4] Baseline appropriately chosen, clearly described/implemented with convincing rationale and preliminary comparison with advanced models.

[5] Exceptional work which includes the above but goes beyond what would be expected from a student at this level.

7. Comparative Classification methodology

[0] Missing or incorrect.

[2] Briefly described.

[3] Working solution with unconfirmed results.

[4] Working solution with confirmed results generated and presented appropriately, including comparative insights.

[5] Exceptional work which includes the above but goes beyond what would be expected from a student at this level.

8. Programming style

[0] Non-meaningful names in code, no comments, use of 'magic numbers'.

[2] A minimal attempt at readability.

[3] The source code is readable with some comments.

[4] The source code is of high quality and follows general coding convention.

[5] Exceptional work which includes the above but goes beyond what would be expected from a student at this level.

III. Concludions

9. Performance Analysis & Comparative Discussion (200-500 words)

[0] No evaluation or incorrect evaluation provided.
[2] Basic description of the results, minimal comparison between the two models.
[3] Results discussed but not convincingly evaluated.
[4] Results convincingly evaluated with clear quantitative and qualitative comparison.
[5] Exceptional work which includes the above but goes beyond what would be expected from a student at this level.

10. Project Summary and Reflections (200 to 400 words)

[0] Missing or incorrect.
[2] Briefly described.
[3] Some conclusions without fully evaluating the comparative effectiveness of the classifiers and the results.
[4] Detailed project evaluation including a comparative analysis (classifier, pre-processing, results, reproducibility).
[5] Exceptional work which includes the above but goes beyond what would be expected from a student at this level.

[END OF COURSEWORK ASSIGNMENT]