Assignment 2 – CSC3066 Deep Learning (Fake News Detection)

04 March 2024

Note: This assignment amounts to 30% of the module mark.

Deadline: 23:59 Friday, 12th April 2024

Summary: This assignment entails the implementation of various architectures of Artificial Neural Networks (ANN) and word embedding models aimed at solving a text classification problem. The primary tasks include:

- Obtaining suitable representations of text data through word embedding models.
- Implementing and assessing different ANN models utilizing the Keras library.
- Presenting, interpreting, and explaining the outcomes derived from various ANN models.
- Contrasting different ANN models and identifying the most effective solution for the provided problem statement.

Problem Specification: You, as a data scientist in your company, have been tasked with addressing the following challenge: a customer seeks the ability to analyse Twitter posts to determine their veracity. Historically, this analysis has relied on a large team of human fact checkers, which is no longer sustainable. Consequently, the company seeks automated solutions. Specifically, they desire a system capable of automatically classifying each tweet as either True (containing genuine information) or Fake (containing false information). Importantly, the customer wishes to minimize instances where fake messages are inaccurately classify as genuine.

They are keen to explore the potential of machine learning to meet this objective but lack the expertise to implement a solution themselves. To support this effort, they have provided a dataset comprising 2,500 tweets, each manually assigned a label: 1 for true and 0 for false. The dataset has been divided into training (2,000 tweets, train.csv) and testing (500 tweets, test.csv) sets.

Your role is to investigate the problem, analyse the data, and propose viable solutions to the customer. Upon thorough exploration, you are expected to provide insights and recommendations regarding the utilization of machine learning techniques to effectively address the task at hand.

Task One: Implementation of ANN architectures (40 marks)

In this task, your objective is to implement and evaluate various ANN models. You will explore the following architectures:

a) Multilayer Perceptron (MLP) model: You should investigate two distinct approaches. Firstly, each record (model input) will be represented as a single vector derived from pre-trained word embedding vectors. Secondly, each record will be represented as a sequence of pre-trained word embedding vectors. For the second approach, utilize the Keras Embedding layer.
b) Convolutional Neural Network (CNN) model: In this approach, each record should be represented as a sequence of pre-trained word embedding vectors, leveraging the Keras Embedding layer.
c) Recurrent Neural Network (RNN) model: Here, each record should be represented as a sequence of pre-trained word embedding vectors, using the Keras Embedding layer.

Utilise the training set for constructing the models and the testing set for evaluation purposes. Given the nature of the project, you are encouraged to apply various techniques to enhance the performance of the models.

Deliverables: code containing the implementation of the final version of each of the four model architectures – specifically, the versions that yielded the best results in your evaluation process.

Task Two: Analysis and reporting of the results from Task One (35 marks)

For this task, your objective is to compose a comprehensive report analysing the outcomes achieved in Task One. Your report should encompass:

a) A detailed description of various techniques and settings you experimented with to enhance the performance of the models in Task One.
b) Justification for exploring each technique, explaining the rationale behind their selection.
c) A thorough discussion on the impact of these techniques on each of the models.
d) Conclusions drawn from the observed results, highlighting key insights gleaned from the experimentation process.

Feel free to incorporate graphs and tables to visually represent different results.

Deliverables: A comprehensive report discussing the results obtained in Task One.

Task Three: Final outcome of the project (25 marks)

In this task, your objective is to compose a reflective report for the customer, assessing the success of the project. The report should cover the following key points:

a) Explain which model was selected as the best candidate for the task. Provide detailed justification of your choice.
b) Provide an assessment of the project's success. Address any challenges or issues encountered during the project delivery process. Discuss the limitations of the developed model, if any, and their implications for future use.
c) Discuss strategies for enhancing the performance of the ANN model in this specific task.
d) In a case where the model is not ready to be used as a fully automated solution for fake content detection, propose strategies for using the model as a supportive tool for manual checkers.

Deliverables: A report discussing the points listed above.

Assessment Criteria:

Task One: The evaluation of this task will be based on the correctness of the implemented algorithms (four different architectures of ANN models) and their experimental settings. You can use existing libraries like Keras, Sklearn and others for this task. Use comments and markdown cells to explain your code's functionality.

Task Two: The evaluation of this task will be based on your creativity and well-reasoned analysis of the results obtained in your experimental analysis. In the report, the results obtained in task one should be presented in a clear and coherent manner. Each set of results should be followed by a comprehensive discussion. Use your own words to explain and interpret the results. Motivate each of the techniques you applied to try and improve the performance of your models. You should demonstrate in-depth knowledge and understanding of different techniques applied in your experiments.

Task Three: The evaluation of this task will be based on (points a & b) well-reasoned cumulative analysis of the results obtained in the project and the rationality behind the applied selection criteria; (points c & d) your creativity.

Guidelines and Submission:

Submission including your code (1 Jupyter Notebook file for Task One) and report (1 pdf file for Tasks Two and Three) must be submitted on Canvas by 23:59, Friday 12th April 2024. For Task 1 only provide one implementation of each model, the one that gave you the best results after tuning. The report should be no longer than four A4 pages (Arial 11 font). Please compress the Jupyter Notebook file and the PDF file into one Zip file (call it your student number) before submission.

Note that your report should not contain details of the code implementation. Use code comments and markdown cells in Jupyter Notebook for code explanation. The report should have information (student name, student number, date, and assignment number) clearly written on the first page.

It's noted that your submission is your own work, and you are aware of the University policies regarding plagiarism and collusion. Late submissions will be treated according to standard university penalties.

Good Luck!