COMP3222 Machine Learning Technologies Coursework 2023/24

Deliverables and deadlines

Deliverable	Deadline	Feedback	Weight
Machine Learning pipelines implementation 1. Source_code.ipynb 2. Requirements.txt The deliverable is used to provide evidence of practical work.	Semester Week 12 - 12 January 2024, Fri @ 4pm	Semester Week 16	15%
Final report to be submitted as Report.pdf Top scoring final reports will characterize the use case problem and data, and connect these characteristics to attributes of the chosen pre-processing, feature selection, dimensionality reduction for **two** machine learning algorithms. They will provide a rigorous evaluation of the implementation and justify chosen designs in the context of other algorithm options available (from course text or wider literature).	Semester Week 12 - 12 January 2024, Fri @ 4pm	Semester Week 16	35%

Introduction

This assignment is about analysing use cases, designing machine learning pipelines with different algorithms and evaluating the resulting design/implementation. The use case is based on the MediaEval 2015 "verifying multimedia use" task.

The MediaEval 2015 "verifying multimedia use" task aims to test automatic ways to classify viral social media content propagating fake images or presenting real images in a false context. After a high impact event has taken place, a lot of controversial information goes viral on social media and an investigation needs to be carried out to debunk it and decide whether the shared multimedia represents real information. As there is a lack of publicly accessible tools for assessing the veracity of user-generated content, the task intends to aid news professionals, such as journalists, to verify their sources and fulfil the goals of journalism that impose a strict code of faithfulness to reality and objectivity.
The task is to design/build pipelines to classify social media posts within the MediaEval 2015 "verifying multimedia use" challenge dataset as 'real' or 'fake'.

Definition of fake posts:

- Reposting of real multimedia, such as real photos from the past re-posted as being associated with a current event
- Digitally manipulated multimedia
- Synthetic multimedia, such as artworks or snapshots presented as real imagery

You will accomplish the task through the following steps:

1. Design and implement **two** pipelines using two different machine learning algorithms to classify posts within the MediaEval 2015 "verifying multimedia use" dataset.
2. Analyse the use case (task and data) and design a data pipeline with pre-processing steps, feature selection methods and dimensionality reduction techniques.
3. Train your **two** algorithms using the MediaEval 2015 "verifying multimedia use" training dataset provided. You may use the algorithm implementations from the libraries learned in the labs (Scikit, TensorFlow, Keras)
4. Evaluate your pipelines by classifying posts in the MediaEval 2015 "verifying multimedia use" test dataset and scoring the answers against the test set ground truth labels.

In addition, you will write a final report. This will explain your use case analysis, justifying your choice of pre-processing step(s), feature selection method(s), dimensionality reduction technique(s) and the **two** machine learning algorithms. You will report your results using the MediaEval 2015 "verifying multimedia use" evaluation metrics.

The F1 scores of algorithm implementations will not affect the final assignment marks (unless they are incorrectly calculated - see section 'machine learning algorithm implementation'). As a guide only, the highest F1 results for this case study coursework in previous years were over 0.9.

Task dataset

The dataset and ground truth labels are provided on the Blackboard. Do not use any other dataset for this assignment (e.g. datasets shared via the MediaEval website must not be used).

The MediaEval 2015 "verifying multimedia use" dataset consists of social media posts (e.g. Twitter, Facebook and blog posts) for which the social media identifiers are shared along with the post text and some additional characteristics of the post. In the original MediaEval challenge multimedia features (image, video) were provided in addition to text and metadata. However, only the text and metadata features have been provided to you to simplify the problem.

A set of ground truth labels (i.e., 'fake' or 'real') are provided in the dataset for both the training and test set. Algorithms will only train using ground truth labels in the training set. The algorithm must not use the test ground truth labels for anything other than computing the final scoring.

Machine learning pipeline implementation

The machine learning pipeline implementation comprises **15 out of the 50 marks** for the coursework. You must implement data pipelines for **two** different machine learning algorithms and conduct evaluations based on classifying the test set and using the ground truth labels to compute the F1 scores.

The code must be implemented using python and submitted as a jupyter notebook. The code must include explanatory comments. You must also include requirements.txt including any library installations required. The code must be able to run and generate the F1 scores reported in the final report.

Final report

The final report comprises **35 out of 50 marks** and MUST have the following five sections:

1. **Introduction and data analysis:** Describe the problem being addressed. Provide a detailed characterization of the task dataset in terms of format, volume, quality and bias.
2. **Pipeline design:** This section must include the entire data pipeline in detail. Describe pre-processing, feature selection and dimensionality reduction used (if any). Describe the machine learning algorithm used. Outline all choices made and justify why they were considered best in the context of the wider options available in the literature and your analysis of data characteristics.
3. **Evaluation:** Report using the F1 metric scores for different configurations of the **two** algorithm implementations you have tested.
4. **Conclusion:** Summarize your findings and suggest some areas for future improvement and lessons learnt.
5. **References:** Provide a list of reference papers cited in the report.

The report PDF document should be between 5 and 10 pages long. The 10-page limit is not a target to aim for, and shorter reports that present information concisely are better - find your perfect balance. Use tables and figures to show data cleanly and highlight key information clearly such as the main findings. Use any document style (e.g. reference style) as long as it's clear and easy to read. Reports over 10 pages long may incur a 5-mark penalty for demonstrating a poor ability to summarize key information.

You need to explain both your design and the design choices, alternatives considered, and justifications for each choice in the context of the data and problem characteristics. You should describe in the final report how you iteratively developed your pipeline, what the evaluation results were at different design iterations and how you used these evaluation results to revise and refine your design choices.

The marking scheme shows you how marks are allocated to each section.

FAQs

Can I use external task-specific data?

No. Use only training and test data from the assignment ZIP file. MediaEval image content feature data (for example) is not provided in the ZIP file, so should not be used. Twitter profile pages and users home pages are not provided in the ZIP file, so should not be used. This is intended to simplify the assignment and allow easier comparison of how you extract most from the text-based features provided.

Can I use external generic data?

Yes. You can use static external resources such as NLTK stopwords, POS tagging, NER, lists of first names, lists of respected news organizations, sentiment word lists etc. These can generate additional useful features from the dataset which might be useful. Static resources should not be tailored to the test set as this would be cheating (e.g. no lists of usernames in testset who are fakers).

Can I edit the test data to improve my results?

No. All posts in the test data must be used for the final evaluation when calculating the final F1 scores. Your algorithm can filter the training data as part of its pre-processing if that is useful. You can pre-process test and training data as you wish to clean or add features, but ultimately you need to produce a classification for every entry in the testset. You must not use the test set for training in any way.

What's the humour label?

Humour label should be treated as a Fake label. The assignment is to create a binary classifier, so treat a Humour as a fake label when calculating F1 scores. You are allowed to use Humour labels to gain an advantage during training, if you want to, for example segmenting the training data to allow discovery of better discriminating features.

How should I define TP, FP, TN, FN?

The task is to classify 'fake' as defined by MediaEval. The binary classifier thus labels data as 'fake' (positive) or 'real' (negative). So a TP is a correct 'fake' classification. A FP is an incorrect 'fake' classification when its 'real'. A TN is a correct 'real' classification. A FN is an incorrect 'real' classification when its 'fake'.

What software can I use for the implementation?

The code must be implemented using Python and submitted as a jupyter notebook. You can make use of lab sessions and materials to help you with the coursework. You can also use any python libraries such as numpy, sklearn, nltk, etc. that help. You may not use any third-party open-source algorithm implementation that was specifically designed to work on a MediaEval challenge (your work must be your own).

The task dataset posts contain text in multiple languages, do I need to translate them?

It is up to each student to analyse the dataset, and decide for themselves what to do with non-English posts. You can detect them (there are PyPI Python libs to detect languages), translate them, ignore them or just allow non-English phrases as features. There is no right answer. You need to analyse the problem and data yourself, then decide what algorithm design to use. Use the methodology taught in the lectures to analyse the problem and data space and match the characteristics to the algorithms available.

Plagiarism

Both the machine learning algorithm implementation and the final report need to be the student's own work unless mentioned otherwise.

For the machine learning algorithm, you can reuse example code from the course textbook or lab materials, but anything else needs to be clearly acknowledged in the report and when submitting. You are not allowed to copy the open-source algorithm releases from those who have previously participated in MediaEval competitions. You are expected to implement your own approach.

You are allowed to use ideas and strategies reported in academic papers, as long as you implement these strategies yourselves and acknowledge the papers in your report. In case of doubt, feel free to ask! This is important as any violations, deliberate or otherwise, will be automatically reported to the Academic Integrity Officer.

Late submissions

Late submissions will be penalised according to the standard rules.

References

[Boididou 2014] Boididou, C., Papadopoulos, S., Kompatsiaris, Y., Schifferes, S., Newman, N. Challenges of computational verification in social multimedia. In Proceedings of the companion publication of the 23rd international conference on World wide web companion (WWW Companion '14), pp. 743-748

[Boididou 2016] Boididou, C. Papadopoulos, S. Middleton, S.E. Dang Nguyen, D.T. Riegler, M. Petlund, A. Kompatsiaris, Y. The VMU Participation @ Verifying Multimedia Use 2016. In Proceedings of the MediaEval 2016 Workshop, Hilversum, The Netherlands, October 20-21, 2016.

[Conotter 2014] Conotter, V., Dang-Nguyen, D.-T., Riegler, M., Boato, G., Larson, M. A Crowdsourced Data Set of Edited Images Online. In Proceedings of the 2014 International ACM Workshop on Crowdsourcing for Multimedia (CrowdMM '14). ACM, New York, NY, USA, 49-52

[MediaEval 2015 proceedings] http://ceur-ws.org/Vol-1436/

[MediaEval 2015] http://www.multimediaeval.org/mediaeval2015/verifyingmultimediause/

[MediaEval 2016 proceedings] http://ceur-ws.org/Vol-1739/