COMP6246 Machine Learning Technologies Coursework

Stuart E. Middleton, sem03@soton.ac.uk

Updated: 14th October 2021

Deliverables and deadlines

Deliverable	Deadline	Feedback	Marking Scheme	Weight
Final report	Week 12 Fri 4pm (timetable week 15)	Week 16 (timetable week 19)	Top scoring final reports will characterize the use case problem and data, and connect these characteristics to attributes of the 5 chosen algorithm designs. They will justify algorithm design choices in the context of other algorithm options available (from course text or wider literature). They will provide a critical review of the strengths and weaknesses of the chosen designs, and rank them with clearly explained justifications.	50%

Introduction

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

Background: 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 investigation needs to be carried out to debunk it and decide whether the shared multimedia represents real information. As there is 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 imposes a strict code of faithfulness to reality and objectivity.

The task is to design/build algorithm(s) 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 to a current event
- Digitally manipulated multimedia
- Synthetic multimedia, such as artworks or snapshots presented as real imagery

You will evaluate a set of possible machine learning algorithm designs to classify posts within the MediaEval 2015 "verifying multimedia use" dataset. You will critically analyse the use case (task and

data) and identify 5 possible algorithms designs. Each algorithm design will include a choice of pre-processing, feature selection, dimensionality reduction technique(s) and machine learning algorithm. In addition you will write a final report. This will explain your use case analysis, justifying your 5 algorithm design choices and critically review them. This critical review will identify for each algorithm design 3 strengths and 3 weaknesses, compare all 5 algorithm designs against each other, and then rank them in order of suitability to the use case problem (with justifications for the ranking).

You are not expected to implement any of the 5 algorithm designs, but you are expected to perform analysis of the dataset (which will probably need code) for the purpose of providing evidence to underpin algorithm design choices. The strengths and weaknesses analysis of algorithm designs should be based on a critical analysis of the theoretical properties of these designs and evidence from your data analysis. It should not be based on a comparative evaluation of 5 full implementations of your 5 algorithm designs.

Task dataset

The dataset and ground truth labels are provided on the module wiki page (https://secure.ecs.soton.ac.uk/notes/comp3222/coursework/assignment-comp3222-comp6246-mediaeval2015-dataset.zip). Do not use any other dataset for this assignment (e.g. datasets shared via MediaEval website will 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.

Final report

The final report 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) **Algorithm design** >> Describe 5 possible algorithm designs, each including pre-processing, feature selection, dimensionality reduction and a machine learning algorithm. Outline all choices made when selecting these designs from the many designs possible, justifying why they were considered good in the context of the wider options available in the literature and your analysis of data characteristics.
3) **Evaluation** >> Describe for each algorithm design 3 strengths and 3 weaknesses, then critically compare all 5 algorithm designs against each other using these strengths and weaknesses. Rank your algorithm designs in order of suitability to the task, and include justifications for this ranking.
4) **Conclusion** >> Summarize your findings, and suggest some areas for future improvement and lessons learnt.
5) **References** >> List of reference papers cited in 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. This may involve writing software or using tools for data analysis. You are not expected to implement any of your 5 algorithms, you need only critically analyse the use case data and problem, and identify 5 possible algorithms designs with enough evidence to justify your choices.

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

FAQ

How can I evaluate my 5 designs without actually coding them?

You are not expected to implement the 5 designs you select for your final report. Instead you are expected to critique a range of possible machine learning approaches based on characteristics reported in the literature, select with justification the best 5 for your task data and problem, and do a full strengths and weaknesses analysis of these 5 to rank them. You are not expected to code the 5 designs and report F1 scores for each. The MSc coursework is evaluating your ability to analyse data and critically appraise possible machine learning approaches for a concrete problem, not your coding ability for a specific approach.

Am I allowed to write code to analyse the data?

Yes. You can write simple code to help analyse and visualize the data in addition to using existing tools (e.g. MS Excel). You should not submit any code you write. You should instead add graphs & histograms you produce using your code and tools into your final report to help you show the important data characteristics.

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 do extracting to 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).

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'.

Do I need to submit software used for data visualization and characterisation?

No. You can use any software you like (e.g. some Python scripts) to segment and analyse the dataset and provide you with evidence to justify your algorithm design decisions. You should include evidence of data characterisation work in your final report, such as graphs & histograms to underpin your design choices and problem/data characterisation. However, the code to generate these graphs & histograms should not be submitted.

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

The report need to be the student's own work unless mentioned otherwise.

You are allowed to use ideas and strategies reported in academic papers, as long as you 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/