e-Learning with *Kaggle in Class*: Adapting the ALTA Shared Task 2013 to a Class Project

Karin Verspoor[1,2] **and Jeremy Nicholson**[2]
[1]National ICT Australia
[2]Department of Computing and Information Systems
The University of Melbourne
Melbourne VIC 3010 Australia
karin.verspoor@nicta.com.au, nj@unimelb.edu.au

Abstract

The 2013 ALTA Shared Task was utilised as a class project for a subject taught at The University of Melbourne in the second semester of 2013. This paper reviews the experience of using an on-line, *Kaggle in Class*-based shared task for class work. Adoption of the shared task enables a *blended learning* paradigm that engages students in problem-based learning in a shared and open context.

1 Introduction

As in recent years, the Australasian Language Technology Association sponsored a shared task in 2013 to stimulate interest in language technology tasks among university students (Molla, 2013). This year's task was primarily organised by Diego Molla of Macquarie University and addressed the restoration of normal case (capitalisation) and punctuation to a noisy text input not conforming to conventional use of case and punctuation. As described in (Molla, 2013) the task is framed as a simplification of the general task explored by (Baldwin and Joseph, 2009).

Because the task is specifically aimed at university students with programming skills, and as it can be approached as a classification task, it is appropriate to consider as a project for a university subject that addresses machine learning algorithms. Furthermore, it provides an opportunity to make use of *blended learning* (Garrison and Kanuka, 2004) in the classroom; that is, integrating face-to-face learning with on-line asynchronous learning opportunities. Enrichment of the traditional classroom learning experience with on-line activities has been suggested to have positive benefits for student learning, in addition to student satisfaction and retention.

The task was therefore selected as a project for approximately 115 students registered for The University of Melbourne's *Knowledge Technologies* subject, a subject in the Department of Computing and Information Systems for which the stated objective is to "learn algorithms and data structures for extracting, retrieving and storing explicit knowledge from various data sources, and methods for data mining and machine learning with complex data." The students were given the option to register for the shared task formally through the *Kaggle In Class* system.

2 Organisation

To adapt the shared task to the classroom context, the project was split into several stages.

2.1 Data pre-processing and task familiarisation

The students were introduced to the task through the ALTA shared task data, without an explicit reference to the shared task itself. They were given the context of the task in a project specification, provided with the training data and asked to write scripts to manipulate the data in various ways. In one subtask, they were asked to map the provided ALTA Shared Task data format to the ARFF (Attribute-Relation File Format) format which is used in several machine learning frameworks. This was intended to get the students comfortable with regular expressions for simple data transformations, and to enable them to produce files appropriate for use in the next stage of the project.

A second subtask required the students to write a program for producing training data for the task from natural language data. That is, to write a program that given normally cased and punctuated text, would produce the appropriate lower-cased and unpunctuated, but appropriately tokenised and labelled structured output for the task. The hope with this subtask was that students would realise that they could in principle produce very large quantities of their own training data for their eventual shared task solutions, by downloading text sources and stripping case and appropriate punctuation.

In a third subtask, post-graduate students (primar-

Karin Verspoor and Jeremy Nicholson. 2013. e-Learning with Kaggle in Class: Adapting the ALTA Shared Task 2013 to a Class Project. In *Proceedings of Australasian Language Technology Association Workshop*, pages 142−145.

ily Master by coursework students) were additionally asked to explore some preliminary rule-based methods for solving the classification tasks. This step was optional for undergraduate students, however they were encouraged to attempt an initial solution. The purpose of this subtask was to encourage the students to begin thinking about the features that might be relevant to addressing the problem, and to give them some sense of the difficulty of a rule-based approach to the task.

2.2 Machine learning-based problem exploration

The students were introduced to various machine learning algorithms in class, as well as approaches to feature engineering and system evaluation. They were pointed specifically towards the WEKA machine learning toolkit, which provides good implementations of many algorithms (Hall et al., 2009), although they were allowed to use any machine learning implementation they were familiar with. The students were instructed to construct and experiment with different features that would be helpful for solving the two classification tasks (i.e. features that may be useful indicators of the appropriate punctuation and capitalisation of a word), and to explore different classification algorithms on the shared task data. The distributed ALTA Shared Task data was divided into pre-defined training and development subsets, with the test set provided as held-out (blind), unlabelled data.

2.3 Kaggle In Class

Along with the project specification for the second, machine learning-based problem exploration stage, the students were introduced to the ALTA Shared Task platform in the *Kaggle in Class* site and provided with invitation links associated to their student IDs. Submission of the results to the Kaggle In Class site for the shared task was entirely optional, although the students were encouraged to participate. Submitting to Kaggle required the students to work out how to map their system results (from WEKA, or whatever toolkit they selected), back into the ALTA Shared Task format.

Participation in the official shared task gave students access to a baseline solution, and allowed them to receive an immediate evaluation of their system results on the held-out test data. It provided an opportunity for immediate feedback on the effectiveness of their solutions; through the Leaderboard the students could see concretely how well their solutions were performing relative to other students.

2.4 Report writing

The students were asked to write a report describing their approach, summarising their exploration of the features and algorithms on the task, and providing observations and critical analysis of their results. The objective of the report was to demonstrate their understanding of the task, methods, and results and to highlight creativity in their solution. Marks were primarily based on the student's critical analysis of their results, rather than the overall score of their solution.

2.5 Peer review

Using an on-line peer review system, TurnItIn's PeerMark, that is integrated into The University of Melbourne's on-line Learning Management System, each student provided feedback on two other students' reports. This enabled Contributing Student Pedagogy (CSP) (Hamer et al., 2008), a participatory learning strategy in which students are encouraged to contribute to the learning of others and to value the contributions of others.

The students were specifically asked to address three points:

1. A summary of the author's work; the approach to the task and the analysis in the report.

2. What they felt that the author had done well, and for what reasons. For example, novel use of features, interesting methodology, or insightful discussion.

3. What they felt were the weak points of the submission, including suggestions of avenues for further research.

The quality of the student peer review reports was quite high; students largely provided thoughtful feedback and critical assessment of their peers' work.

3 Results

The students generally appeared to find the task quite challenging. For most students it was their first exposure to hands-on application of machine learning algorithms to solve a problem, as well as their first exposure to text classification. Lectures covered algorithms and evaluation strategies in detail, and several pointers were provided about good features to experiment with, such as token "shape" and character or token n-grams. However, many student solutions applied WEKA in a narrow range of configurations and with a limited set of features. Some students—typically those with prior exposure to natural language processing through another subject in the department—made use of linguistic features such as part of speech tags, and some used gazetteers of English names or common words specifically to help with the capitalisation task. A few students used machine learning frameworks other than WEKA.

A number of students did submit their results to the main Kaggle ALTA Shared Task site, and some even included those results in their project reports. It was observed that several of the students' submissions to the Kaggle site displayed identical performance. Further investigation revealed that their scores matched exactly the performance of the baseline model provided along with the ALTA Shared Task data upon registration to Kaggle In Class. This suggests that these students likely made test submissions using the baseline model, rather than submitting results based on their own systems or solutions.

4 Discussion

4.1 Interaction with Kaggle

Since the ALTA Shared Task was run using the Kaggle in Class framework (`https://inclass.kaggle.com/c/alta-2013-challenge`), students were encouraged to submit results directly to the on-line system. This required generating individual invitation links to join the shared task site for each student. While this was easily generated through the Kaggle system by the organiser of the task, it was also important to associate Kaggle logins with individual student IDs, so that Kaggle submissions from our student cohort could be identified. For a class with over one hundred students, this created a logistical hassle for managing login-student ID associations, and the distribution of the invitation links to the individual students.

4.2 Timing considerations

An important factor in the decision to utilise the ALTA Shared Task as a class project was whether the timing would fit in with the overall timeline for the subject. The dates generally aligned well; the shared task was announced in mid-July, while the semester began at the end of July.

The final submission date for the official ALTA Shared Task was set at 04 October. That date fell during the non-teaching week of the semester (semester break) and did not allow adequate time for a second project during the second half of the semester. Therefore it was decided to set the deadline for the class project ahead of the final ALTA Shared Task deadline, on 20 September. In the end, as the students found the assignment challenging, the deadline was extended to 27 September (compressing the second project somewhat) to give them more time to make adequate progress.

Since the deadline for the class project was ahead of the shared task deadline, the students were told that they could continue to attempt to improve their results after submission of the project report if they were enjoying participating in the shared task. Reviewing the time stamps on the Kaggle Leaderboard, most students did not continue working on the project after the submission deadline. Three students did at least take the time to submit results on the "final" ALTA Shared Task data (on a sister Kaggle site, `https://inclass.kaggle.com/c/alta-2013-challenge-final`) on 06 October; two did not do particularly well (obtaining scores of 0.3 and 0.08, respectively), while one student obtained 0.65, second to the winning system score of 0.74. This second-place result was consistent with the leader board results for the original shared task; i.e. that student also placed second to the winning system on the original data. Interestingly, one of the students who submitted results on the final data hadn't participated in the original shared task leader board at all.

4.3 Set-up of the Shared task

Due to the separation of the ALTA Shared Task into a development competition and a "final" competition, with the final data not being released until well after the class project deadline, it proved difficult for the students enrolled in the subject to submit results to the final test. As indicated above, only three students did so while there were about 50 students who made at least one submission to the original Kaggle ALTA Shared Task site.

4.4 The Leaderboard

The students who participated in the on-line competition were not systematically compared to the students who did not participate on-line; significant variations in how students set up their training and testing scenarios for their final reports would have made this very difficult. In contrast, the availability of the on-line framework and the Leaderboard provided a consistent testing scenario for comparing student performance on the task: the relative performance of different systems over the same held-out data were immediately available upon submission. While we did not systematically cross-reference Kaggle results with student reports, our general impression was that students who showed creativity in their feature engineering did appear to achieve higher results on the leader board for the shared task. Participating in the on-line task seemed to spur experimentation. While we cannot know how many configurations the students who did not participate on-line explored, most students participating on-line submitted multiple runs. This suggests that they were experimenting with various configurations to obtain better results. One student made 14 entries, and indeed the winning system submitted 13 sets of results.

4.5 Emphasis

The focus of the shared task is competitive; entrants aim to achieve the best possible results on the task. In contrast, the aim of the class project was to provide the students with an opportunity to apply newly acquired knowledge of machine learning and feature engineering, and to demonstrate understanding of that application through critical exploration of the problem and different approaches to solving it. A student who scored high on the leader board was not guaranteed to have a good mark for the project; as indicated above, the mark was based on the report. Conversely, a student could achieve a good mark for the project without creating a high-performing solution to the task, for instance by exploring and explaining the performance of a broad range of features that may not have proven particularly effective for solving the task. However, given the above observation that participation in the on-line shared task seemed to result in substantial experimentation, and the context of comparative, immediate feedback, it seems likely that students who actively participated on-line would have been thinking relatively more creatively about their approach. In turn, the objectives of the project would have been met, and their marks would likely have reflected this creativity.

5 Conclusions

Nearly one-half of the students in a subject taught at The University of Melbourne who were given the (completely voluntary) opportunity to participate in the *Kaggle in class* on-line component for the ALTA Shared Task elected to sign up and participate in the open competition. While the emphasis of the students' assignment was on problem exploration rather than system performance, it appeared, based on an informal and unsystematic review of the assignments, that students who performed well on the on-line task also had made a significant effort to explore creative strategies for solving the task.

Use of the ALTA Shared Task as a class project was generally successful despite some differences in objectives. Participation in the on-line experience afforded by the ALTA Shared Task seemed to enhance overall student learning.

Acknowledgements

Dr. Verspoor's teaching and research time is supported by National ICT Australia (NICTA). NICTA is funded by the Australian Government through the Department of Communications and the Australian Research Council through the ICT Centre of Excellence Program. Mr. Nicholson is supported by The University of Melbourne Department of Computing and Information Systems as Head Tutor of the Knowledge Technologies subject, among several other subjects.

References

Timothy Baldwin and Manuel Paul Anil Kumar Joseph. 2009. Restoring punctuation and casing in english text. In *AI 09 Proceedings of the 22nd Australasian Joint Conference on Advances in Artificial Intelligence*, pages 547–556.

D. Randy Garrison and Heather Kanuka. 2004. Blended learning: Uncovering its transformative potential in higher education. *The Internet and Higher Education*, 7(2):95–105.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The WEKA data mining software: An update. *SIGKDD Explorations*, 11(1):10–18.

John Hamer, Quintin Cutts, Jana Jackova, Andrew Luxton-Reilly, Robert McCartney, Helen Purchase, Charles Riedesel, Mara Saeli, Kate Sanders, and Judithe Sheard. 2008. Contributing student pedagogy. *ACM SIGCSE Bulletin*, 40(4):194–212.

Diego Molla. 2013. Overview of the 2013 alta shared task. In *Proceedings of the Australasian Language Technology Association Workshop 2013*, pages 132–136, Brisbane, Australia, December.