Text mining student discussion forum data: common pitfalls and how to avoid them

Elaine Farrow[1], Johanna Moore[1] and Dragan Gašević[2]

[1]School of Informatics University of Edinburgh, UK
[2]Faculty of Education, Monash University, Australia

The widespread use of online discussion forums in educational settings provides a rich source of data for researchers interested in how collaboration and interaction can foster effective learning. Natural language processing and machine learning techniques allow discussion forum texts to be analysed in an automated, efficient way. Here, we present our findings related to the robustness and generalisability of automated text classification methods in common use (Farrow et al., 2019). We closely examined one published state-of-the-art model, comparing different approaches to (a) managing unbalanced classes in the data, and (b) selecting a suitable data set to use for evaluation. By demonstrating how commonly-used data preprocessing practices can lead to over-optimistic results, we contributed to the development of the field so that the results of automated content analysis can be used with confidence.

We ran a replication study focusing on one specific data set and classifier type, allowing us to critically examine some of the common pitfalls associated with typical data preparation practices that are in widespread use. We recreated a state-of-the-art predictive model (Kovanović et al., 2016) using the original data and methodology, then compared different approaches to dealing with the unbalanced classes in the outcome variable. Building on these results, we also explored the effect of splitting the data by course offering instead of using a random split.

In the prior work, the full data set was first processed to redress the class imbalance, before splitting it into training and test sets. When we instead applied the class rebalancing only to the training data, our results were lower on every outcome metric. We conclude that the prior results were affected by data contamination between the training and test sets, leading to an over-estimation of that model's predictive power. Further, we found that rebalancing classes across the whole training data set before tuning the model decreased the final model's performance, compared to training on unbalanced data; whereas moving the class rebalancing step inside the cross-validation loop improved the results. This is consistent with prior work on parameter tuning with small data sets (Kuncheva and Rodríguez, 2018).

Finally, using a session-based data split (training on the earlier course offerings and evaluating on the final one) led to much lower results on every metric than when using a stratified random split. These results are consistent with recent work on replication in MOOCs (Gardner et al., 2018). One explanation is that the data points are not independent: the forum messages form a natural sequence and share commonalities, such as vocabulary used. Taking several messages from a discussion thread to use for training and then using another message from the same thread for testing is thus likely to give biased results.

We conclude with two recommendations for the field: perform model tuning inside the cross-validation loop in order to build classifiers that will generalise better to future data; and evaluate models using data from the most recent run of a course, rather than using a stratified random sample.

References

Farrow, E., Moore, J., and Gasevic, D. (2019). Analysing discussion forum data: a replication study avoiding data contamination. In *LAK '19*, Tempe, AZ. ACM.

Gardner, J., Brooks, C., Andres, J. M., and Baker, R. (2018). Replicating MOOC predictive models at scale. In *L@S '18*, pages 1–10, New York, New York, USA. ACM Press.

Kovanović, V., Joksimović, S., Waters, Z., Gašević, D., Kitto, K., Hatala, M., and Siemens, G. (2016). Towards Automated Content Analysis of Discussion Transcripts: A Cognitive Presence Case. In *LAK '16*, pages 15–24.

Kuncheva, L. I. and Rodríguez, J. J. (2018). On feature selection protocols for very low-sample-size data. *Pattern Recognition*, 81:660–673.