De-identification of Student Writing in MOOCs

Langdon Holmes and Scott Crossley

AI-ALOE – EAB 2022

Background

- Massive Open Online Courses (MOOCs) collect substantial student data.
- Personally Identifiable Information (PII) poses a significant barrier to the creation of open datasets.
- Variation in formatting conventions and text type make automatic de-identification of unstructured text difficult.
- Student names are particularly difficult to automatically identify.
- Large, pre-trained language models have improved performance in virtually all natural language processing tasks
- Large, pre-trained language models have successfully used to de-identify medical data (Murugadoss et al. 2021)
- Small neural networks have shown some promise in the educational domain (Bosch et al. 2020)

PII in MOOCs

- Currently there is little regulation or oversight of PII in MOOC data.
- FERPA generally not considered to be applicable.
- "Control" and "Transparency" have been promoted as values alongside "Privacy" (Young, 2015).

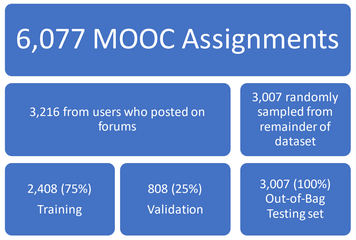
Our Purpose

- Implement state of the art natural language processing tools to identify names in student writing
- Compare performance characteristics between this approach and human annotations

1) What is the classification accuracy of a fine-tuned language model applied to MOOC data?
2) How does the performance of this approach compare to human annotations of student names?

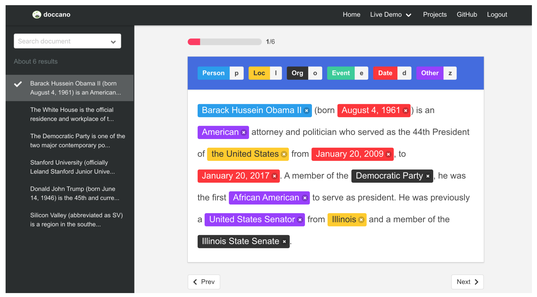
Methods

Corpus



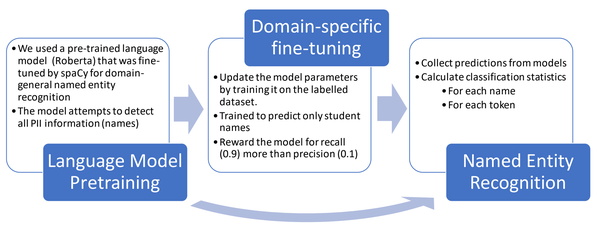
- Texts were parsed from PDF file uploads.
- Assignment submission is 100% of course grade.
- Submissions were peer reviewed.
- Online course hosted on Coursera.

Creating Gold Labels: Human Annotations



Two human raters labeled all 6,077 submissions for potential student names.

NLP Model Development

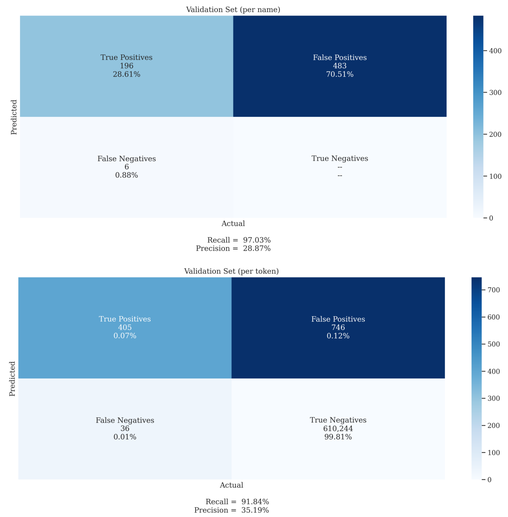


Performance evaluation

1. Results were evaluated per-name and per-token.
2. False positives and false negatives were characterized.

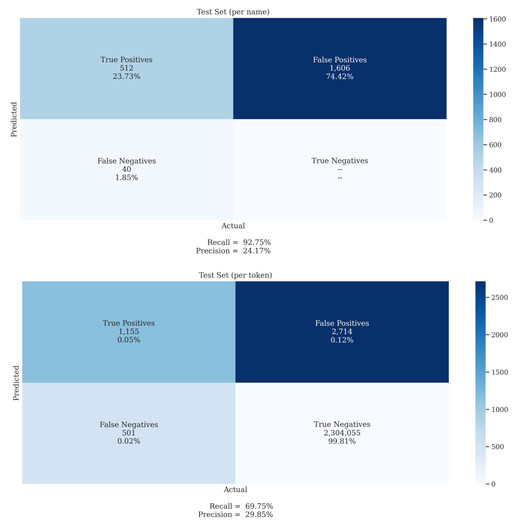
Results

Validation Set



A manual, post-hoc review of the false positives in this set showed approximately 70 of the false positives (per name) were full (first + last) names that may have referred to students.

Test Set



Of 40 false negative names, 21 were potential student names. 11 of these were full (first + last) names.

Of 501 false negative token labels, approximately 40 were alphabetic tokens belonged to potential student names (including the 21 names missed entirely).

Discussion

- Deep-learning based model performed well on the validation and out-of-bag test set.
- The fine-tuned model complemented the pre-trained model by adapting to patterns specific in the data, such as the tendency for names to appear in headers, footers, and underneath the document's title.
- False positives were mostly names: authors, lecturers, historical figures, and students missed by human annotators.
- The two models collectively outperformed human raters in terms of identifying potential PII in student submissions.
- However, the ensemble still failed to detect several complete first and last names due to inconsistent formatting.

Conclusion

Perfect recall cannot be expected from these systems.

One approach to de-identification is to obfuscate rather than remove PII. This protects student identities by scrambling any potentially leaked signals.

The deep learning approach shows substantial enough promise that it may soon be implemented unsupervised, although domain-specific labelled datasets will still be needed.

References

Young, E. M. (2015). Educational Privacy in the Online Classroom: FERPA, MOOCS, and the Big Data Conundrum. *Harvard Journal of Law & Technology*, *28*(2).

Murugadoss, K., Rajasekharan, A., Malin, B., Agarwal, V., Bade, S., Anderson, J. R., Ross, J. L., Faubion, W. A., Halamka, J. D., Soundararajan, V., & Ardhanari, S. (2021). Building a best-in-class automated de-identification tool for electronic health records through ensemble learning. *Patterns*, *2*(6).

Bosch, N., Crues, R. W., & Shaik, N. (2020). "Hello, [REDACTED]": Protecting Student Privacy in Analyses of Online Discussion Forums. *Proceedings of The 13th International Conference on Educational Data Mining*, 11.