Comparing Natural Language Processing Methods for Text Classification of Small Educational Data

Tanner Phillips, Asmalina Saleh, Krista D. Glazewski, Cindy E. Hmelo-Silver
Indiana University
tanphil@indiana.edu, asmsaleh@indiana.edu, glaze@indiana.edu, chmelosi@indiana.edu,

Seung Lee, Bradford Mott, James C. Lester
North Carolina State University
sylee@ncsu.edu, bwmott@ncsu.edu, lester@ncsu.edu

ABSTRACT: Over the past decade, new natural language processing techniques have been developed that show promise when applied to educational data. Although these methods can be effective, little work has been done to measure the comparative strength of these methods when applied to small data sets. This poster presents an analysis of student chat from a collaborative game-based learning environment. Natural language processing techniques were used to attempt to match human coding of 2877 student chat messages. Findings showed that simple feature engineering methods such as latent semantic analysis outperformed neural networks, which may suggest that it is not appropriate to apply neural networks to the small data sets often found in educational settings.

Keywords: Computer-supported collaborative learning, conversational agents, small data, text classification.

1 INTRODUCTION

Computer supported-collaborative learning (CSCL) environments can offer spaces for students to participate in socially supported education. However, CSCL environments require continued monitoring by instructors and often take place in brick-and-mortar classrooms (Kapur & Kinzer, 2007). One way to address the need for teacher guidance is through conversational agents as a method for augmenting teacher interaction. Conversational agents can monitor student performance in CSCL and offer guidance when students get stuck or go off-task, alerting teachers when human intervention might be needed. However, no consensus has been reached in the learning analytics community as to what method best accomplishes the task of understanding student utterances—the first step in creating a conversational agent. There is high variability in the methods utilized to understand student utterances. They include the utilization of hard-coded or pre-trained linguistic models (Jung & Wise, 2020; Kovanović et al., 2018; Pennebaker et al., 2007), statistical dimensionality reduction techniques (Kovanović et al., 2018; Stone et al., 2019; Vytasek et al., 2019), and neural network models (Fiacco et al., 2019; Stone et al., 2019). In most studies, only one type of model is considered, making comparisons between these models difficult. In this study, we address this issue by comparing a variety of methods for understanding student utterances.

2 METHODS

The student utterances analyzed in this study were gathered from a collaborative game-based learning environment designed to teach students to 14 about environmental science (N=45). In the game, students visit an island in the Philippines where they discover that the tilapia in the local fish farm are sick. Over eight sessions, students visit collected information from characters and objects. Working in groups of four, students use an in-game whiteboard to share their findings with fellow students, engage in group inquiry, and generate hypotheses. They repeat this cycle of collecting data and brainstorming three times before determining why the fish are sick. Throughout the experience, students communicated with each other using an in-game chat feature.

2.1 Data Collection and Utterance Coding

Students' utterances were coded based on the accountable talk and problem-based learning frameworks (Resnick et al., 2018; Saleh et al., 2020). There was a total of eight codes based on student talk (*Agreement*, *Rebuttal*, *Descriptions*, *Hedges*, *Relational*, *Regulation*, and *Questions*). Utterances that could not be coded under these categories were coded as *Other*. The chat data was collected from 45 students between the ages of 12-13. Students worked in 11 groups of four, with one group of five. The students generated a total of 2877 unique chat utterances. We performed standard data cleaning such as removing capitalization and punctuation. However, we did not use certain common NLP practices, such as removing stop words. This was because many utterances contained only a single word (e.g., "no," "hi," or "okay") and removing stop words would have deleted over 20% of the data.

2.2 Classification Methods

The results of ten different NLP models are presented in this paper. These models used one of four methods for feature engineering (LDA, LSA, Pre-trained ELMo word embeddings, untrained word embeddings) and one of three different classification methods (Random Forest (RF), Multinomial Regression (MR), or Long-Short Term (LSTM) Neural Networks). For brevity, several classification methods that performed poorly are not included in this paper including Support Vector Machines and other pre-trained word embeddings such as Word2Vec, GloVe, and BERT. A baseline of 31% was used to measure model accuracy, as this was the size of the largest coding category.

3 RESULTS

Results show that common LDA and LSA machine-learning methods outperformed neural networks at classification (see table 1) when measured against hand coding. MR and LDA performed best overall, the differences were slight between LDA and LSA models. The pre-trained word embedding performed worse than chance, while the free embedding model performed better than chance, but not as well as LDA and LSA models.

Table 1: Results of NLP Classification Models

Feature	Classification	Accuracy	% Above Baseline	Precision	Recall	F-1 Score
LDA	RF	0.425	11.1%	0.424	0.471	0.442
LDA	**MR**	**0.435**	**12.3%**	0.422	0.488	0.440

LSA	**RF**	0.427	11.5%	0.427	0.472	**0.445**
LSA	MR	0.434	12.2%	0.422	0.488	0.440
ELMo	RF	0.234	-7.8%	0.221	0.274	0.246
ELMo	LR	0.269	-4.3%	0.235	0.269	0.278
ELMo	LSTM	0.310	-0.02%	0.309	0.318	0.320
Free Embedding	LSTM	0.380	6.8%	0.355	0.411	0.400

4 DISCUSSION & CONCLUSION

Results support several common assumption of natural language processing, while also suggesting some implications specific to educational data. First, the relatively small data size used in this model does not appear to be large enough for training of neural networks. In the context of K-12 student chat, utterances also often include informal, colloquial, and incomplete sentences, this may help to explain why the ELMo word embedding, which was mainly trained on the Wikipedia corpus, performed so poorly, as it was not familiar with the context it was being asked to analyze. This study also supports the interchangeable usage of LDA and LSA. While all models failed to perform at particularly high accuracy, this study still revealed that what is considered "state-of-the art" in computer-science may not be applicable when dealing with educational data.

REFERENCES

Dascalu, M., Stavarache, L. L., Trausan-Matu, S., Dessus, P., Bianco, M., & McNamara, D. S. (2015). *ReaderBench: An Integrated Tool Supporting both Individual and Collaborative Learning*. 436–437. https://doi.org/10.1145/2723576.2723647

Fiacco, J., Cotos, E., & Rosé, C. (2019). Towards enabling feedback on rhetorical structure with neural sequence models. *ACM International Conference Proceeding Series*, 310–319. https://doi.org/10.1145/3303772.3303808

Jung, Y., & Wise, A. F. (2020). How and how well do students reflect?: Multi-dimensional automated reflection assessment in health professions education. *ACM International Conference Proceeding Series*, 595–604. https://doi.org/10.1145/3375462.3375528

Kapur, M., & Kinzer, C. K. (2007). Examining the effect of problem type in a synchronous computer-supported collaborative learning (CSCL) environment. *Educational Technology Research and Development*, *55*(5), 439–459. https://doi.org/10.1007/s11423-007-9045-6

Kovanović, V., Joksimović, S., Mirriahi, N., Blaine, E., Gašević, D., Siemens, G., & Dawson, S. (2018). Understand students' self-Reflections through learning analytics. *ACM International Conference Proceeding Series*, 389–398. https://doi.org/10.1145/3170358.3170374

Pennebaker, J. W., Chung, C. K., Ireland, M., Gonzales, A., & Booth, R. J. (2007). The Development and Psychometric Properties of LIWC2007 The University of Texas at Austin. *Development*, *1*(2), 1–22. https://doi.org/10.1068/d010163

Saleh, A., Yuxin, C., Hmelo-Silver, C. E., Glazewski, K. D., Mott, B. W., & Lester, J. C. (2020). Coordinating scaffolds for collaborative inquiry in a game-based learning environment. Journal of Research in Science Teaching(57), 1490-1518. https://doi.org/10.1002/tea.21656

Stone, C., Quirk, A., Gardener, M., Hutt, S., Duckworth, A. L., & D'Mello, S. K. (2019). *Language as Thought*. 320–329. https://doi.org/10.1145/3303772.3303801

Vytasek, J. M., Patzak, A., & Winne, P. H. (2019). Topic development to support revision feedback. *ACM International Conference Proceeding Series*, 220–224. https://doi.org/10.1145/3303772.3303816