*p*Nota: Análise das Estruturas Textuais com *Active Learning* para Avaliação de Respostas Discursivas

Marcos A. Spalenza
Orientador: Elias de Oliveira

UFES

PPGI - Programa de Pós Graduação em Informática
Universidade Federal do Espírito Santo

Sumário

- Introdução
- Problema
- Objetivo
- Método
- Experimentos
- Conclusão

Introdução

As avaliações de aprendizado são fundamentais para o ensino ao apontar o desempenho da turma com o progresso nos conteúdos. Com aplicações frequentes, as atividades permitem ao professor interagir com os alunos e com os materiais pedagógicos para reformulação e aperfeiçoamento da sua metodologia.

Desse modo, é com o acompanhamento da disciplina e o apoio ao educando que as atividades formativas permitem a reformulação do processo de ensino-aprendizagem (BARREIRA; BOAVIDA; ARAÚJO, 2006).

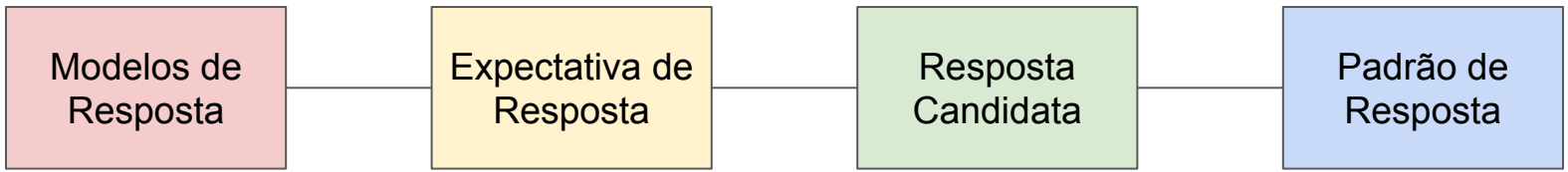
O papel da avaliação, portanto, é diagnosticar, apreciar e verificar a proficiência dos alunos para que o professor atue no processo de formação de modo a consolidar o aprendizado (OLIVEIRA; SANTOS, 2005).

Introdução

Enunciado: Quais são as diferenças entre veias e artérias?
Amostras:

#1 A diferença é a válvula
#2 As veias levam o sangue sem oxigênio e as artérias levam o com oxigênio
#3 Veia transporta gás carbônico e artéria o gás oxigênio.

Resposta candidata:
Quando o coração bombeia o sangue, ele bombeia este sangue diretamente nas artérias com grande pressão, para que as artérias possam conduzir o sangue na direção dos tecidos. As veias são responsáveis por conduzir o sangue de volta ao coração e removem as toxinas dos tecidos para que elas sejam eliminadas.



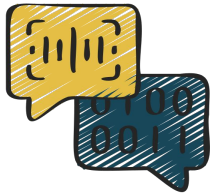
Problema

Existe o alto custo de correção de respostas discursivas atrelado ao processo de avaliação. Sendo que tais são essenciais para o desenvolvimento das técnicas de escrita e leitura em qualquer nível de ensino (HORBACH; PINKAL; 2018).

Portanto, os sistemas *Short Answers Grader* (SAG) (BURROWS; GUREVYCH; STEIN, 2015) tornaram possível um aumento na capacidade de aplicação e correção de atividades para avaliação estudantil (ZESCH; HEILMAN; CAHILL; 2015).

Problema

- Aplicação em diferentes conjuntos de dados (BURROWS; GUREVYCH; STEIN, 2015)
- Análise de escrita em diferentes níveis da linguagem (BURROWS; GUREVYCH; STEIN, 2015; KUMAR et al., 2019; SAHU; BHOWMICK, 2020)
- Alinhamento com o critério do professor (FILIGHERA; STEUER; RENSING, 2020, PADÓ; PADÓ, 2021)
- Identificação de conteúdos relevantes (BUTCHER; JORDAN, 2010; SAHA et al., 2018)
- Reconhecimento de outliers (DING et al., 2020; FUNAYAMA et al., 2020)
- Explicação para a atribuição de notas (MIZUMOTO et al., 2019; SÜZEN et al., 2020; BERNIUS; KRUSCHE; BRUEGGE, 2022)



Sequencial

Sintático

Semântico

Léxico

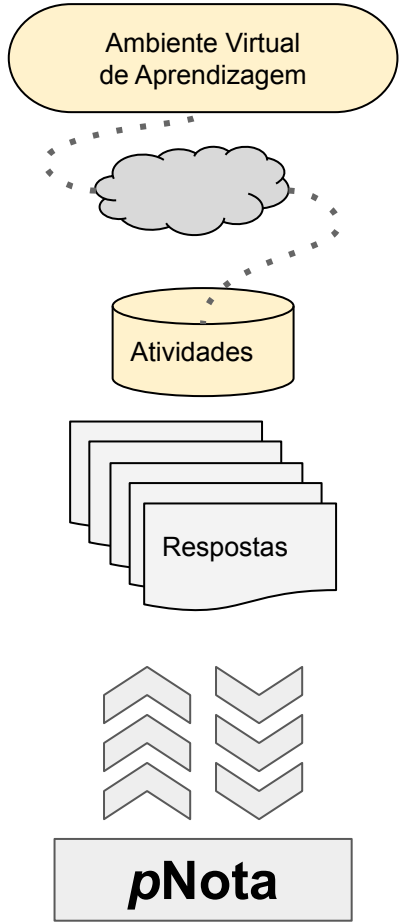
Morfológico

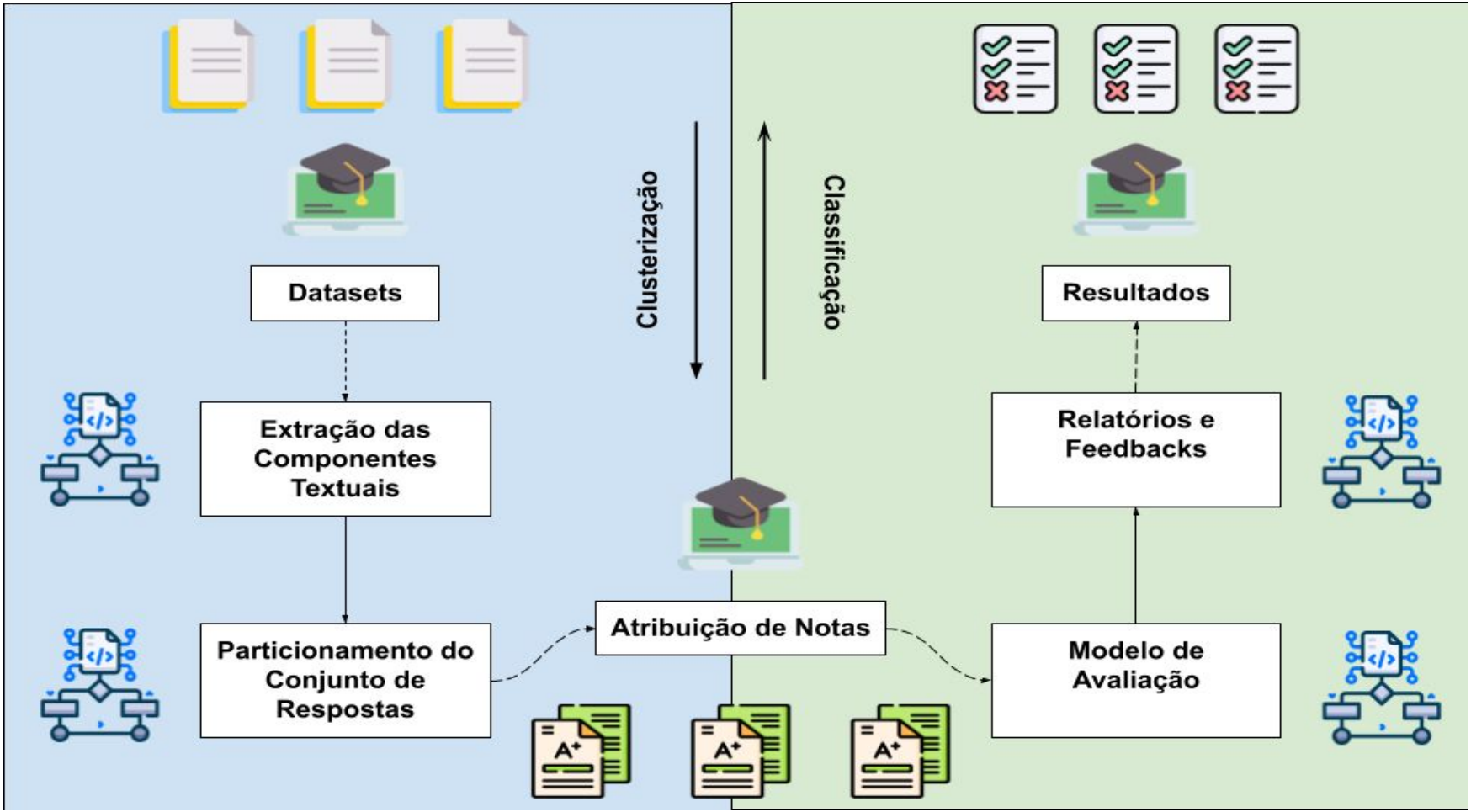
Objetivo

Aprender o comportamento avaliativo que vincula uma nota para determinadas estruturas textuais que compõem as respostas dos estudantes. Para isso, esperamos reproduzir padrões avaliativos complexos, tornando-se um avaliador correspondente ao professor, buscando a redução do esforço de correção do mesmo.

Método

- Estrutura textual nos principais níveis linguísticos citados na literatura
- Active Learning com ciclos de clusterização e classificação
- Processo de amostragem dos documentos por representatividade intracluster ou intercluster
- Identificação de padrões entre texto e nota
- Classificação / regressão para atribuição de notas
- Elaboração de feedbacks e relatórios

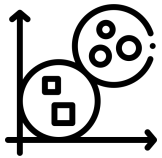



Datasets
Extração das Componentes Textuais
Particionamento do Conjunto de Respostas
Clusterização
Atribuição de Notas
Classificação
Modelo de Avaliação
Relatórios e Feedbacks
Resultados

Método

- Padronização, Segmentação, Filtragem, Transformação e Vetorização
- Agglomerative Clustering + Otimização (índices de validação interna)
- 20 métricas de distância
- 6 algoritmos de classificação:
  K-Nearest Neighbors, Decision Tree, Support Vector Machine, Gradient Boosting, Random Forest e WiSARD
- 5 algoritmos de regressão:
  Regressão Linear, Lasso, K-Nearest Neighbors, Decision Tree e WiSARD






Experimentos

Datasets	**Palavras**	**Caracteres**	**Linguagem**	**Tema**
SEMEVAL2013 Beetle	10	50	Inglês	Ciências (BeetleII)
SEMEVAL2013 SciEntBank	13	13	Inglês	Ciências (ASK)
UK Open University	10	55	Inglês	Introdução a Ciências
Kaggle ASAP-SAS	43	236	Inglês	Ciências, Inglês, Artes
Powergrading	4	20	Inglês	História (USCIS)
University of North Texas	20	107	Inglês	Estrutura de Dados
Kaggle PTASAG	13	72	Português	Biologia
Projeto Feira Literária	8	40	Português	Quimica
VestUFES	92	536	Português	Português (Vestibular)

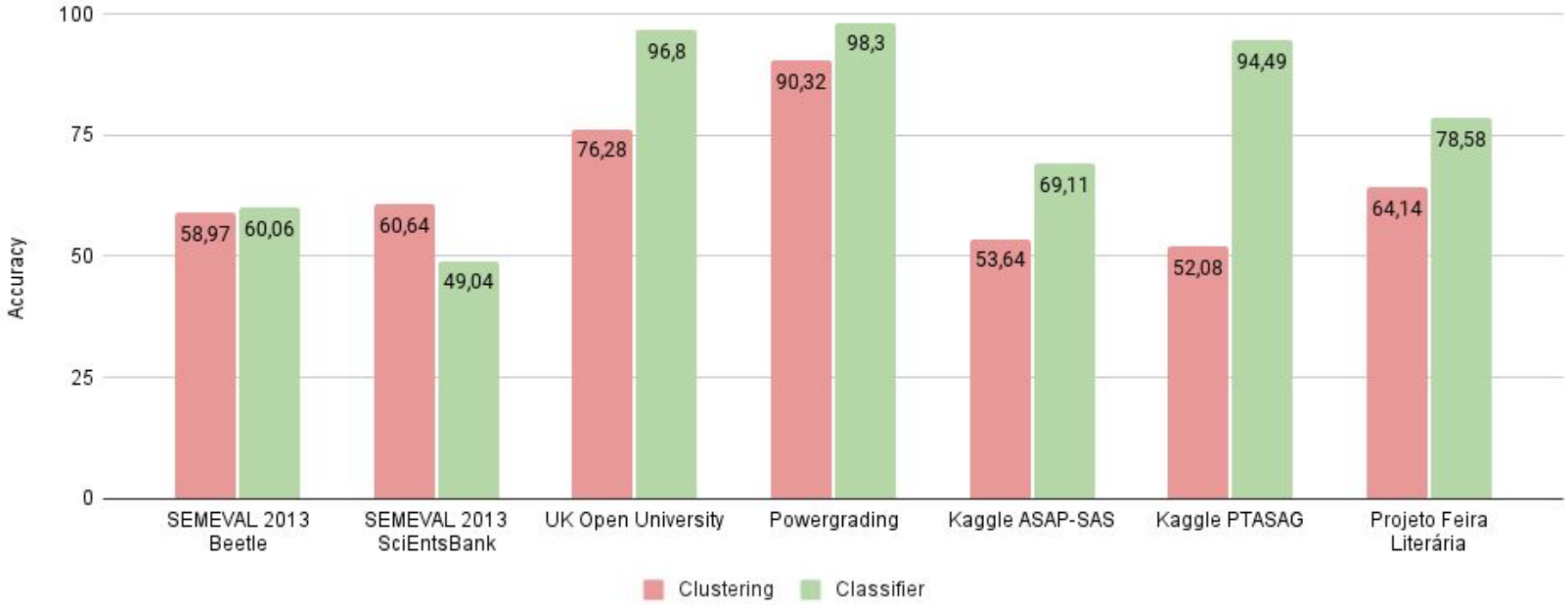
Experimentos

Datasets	Amostragem	Questões	Respostas	Modelo Avaliativo
SEMEVAL2013 Beetle	90%/10%	47	4380	Ordinal
SEMEVAL2013 SciEntBank	88%/12%	143	5509	Ordinal
UK Open University	90%/10%	20	23790	Discreto
Kaggle ASAP-SAS	77%/23%	10	17043	Discreto
Powergrading	80%/20%	10	6980	Discreto
University of North Texas	92%/8%	87	2610	Contínuo
Kaggle PTASAG	80%/20%	15	7473	Discreto
Projeto Feira Literária	75%/25%	10	700	Discreto
VestUFES	30%/70%	5	460	Contínuo

Experimentos

Datasets	**Métrica**	**pNota**	**Literatura**	
SEMEVAL2013 Beetle	F1(w)	60,06	70,91	Sahu(2020)
SEMEVAL2013 SciEntBank	F1(w)	49,04	92,50	Sahu(2020)
UK Open University	ACC	**96,80**	**96,48**	Butcher(2010)
Kaggle ASAP-SAS	RMSE	0,4527	0,2055	Steilmel(2020)
Powergrading	F1(m)	**98,30**	**97,03**	Lui(2022)
University of North Texas	RMSE	**0,619**	**0,65**	Roy(2016)
Kaggle PTASAG	ACC	**94,49**	**68,8**	Galhardi(2018)
Projeto Feira Literária	ACC	**78,58**	-	-
VestUFES	MAE	**1,77**	**1,78**	Spalenza(2017)

Experimentos



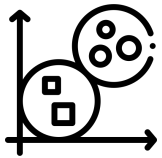
Contribuições

- Ciclo avaliativo desenvolvido em conjunto com os professores
- Análise textual em uma série de níveis linguísticos
- Otimização da seleção de hiperparâmetros
- Refinamento do modelo com ciclos avaliativos via Active Learning
- Reconhecimento de padrões durante o ciclo avaliativo
- Criação de formatos próprios para relatórios e feedbacks
- Suporte para uma série de pesquisas pedagógicas, colaborando com duas Dissertações de Mestrado do Mestrado Profissional em Química em Rede Nacional (ProfQui) do IFES

Conclusão

- Ganho de informação obtido via Active Learning
- Descartada a diferença estatística entre os classificadores
- Níveis de desempenho similares aos observados entre dois humanos
- Resultados de ACC de 79% e F1-ponderado médio de 78%
- Dentre as 255 atividades de classificação 137 foram avaliadas com nível cognitivo avançado (75% de ACC ou mais).
- Integração na rotina pedagógica do professor

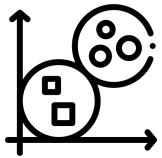





Trabalhos Futuros

- Refinamento e enriquecimento dos modelos com ganho analítico e de domínio
- Otimização dos resultados com critérios mais sofisticados de calibração
- Mensurar conhecimento da amostragem e do vínculo entre termos e classes
- Acompanhar os ciclos avaliativos do professor para vincular o sistema com a rotina desse profissional e a descrição de resultados






UFES

Marcos A. Spalenza

Doutorando em Ciência da Computação

Laboratório de Computação de Alto Desempenho - LCAD
Programa de Pós-Graduação em Informática - PPGI
Universidade Federal do Espírito Santo - UFES

marcos.spalenza@gmail.com

Referências

- BARREIRA, C.; BOAVIDA, J.; ARAÚJO, N. Avaliação Formativa: Novas Formas de Ensinar e Aprender. Revista Portuguesa de Pedagogia, Universidade de Coimbra, v. 40, n. 3, p. 95–133, 2006.
- BERNIUS, J. P.; KRUSCHE, S.; BRUEGGE, B. Machine Learning Based Feedback on Textual Student Answers in Large Courses. Computers and Education: Artificial Intelligence, Elsevier, v. 3, n. 1, p. 100081.1–100081.16, 2022.
- BURROWS, S.; GUREVYCH, I.; STEIN, B. The Eras and Trends of Automatic Short Answer Grading. International Journal of Artificial Intelligence in Education, Springer, v. 25, n. 1, p. 60–117, 2015.
- BUTCHER, P. G.; JORDAN, S. E. A Comparison of Human and Computer Marking of Short Free-Text Student Responses. Computers & Education, Elsevier, v. 55, n. 2, p. 489–499, 2010.
- DING, Y. et al. Don't Take "nswvtnvakgxpm" for an Answer - The Surprising Vulnerability of Automatic Content Scoring Systems to Adversarial Input. In: Proceedings of the 28th International Conference on Computational Linguistics. Barcelona, Spain (Virtual Event): International Committee on Computational Linguistics, 2020. v. 28, p. 882–892.
- FILIGHERA, A.; STEUER, T.; RENSING, C. Fooling Automatic Short Answer Grading Systems. In: Proceedings of the 21st International Conference on Artificial Intelligence in Education. Ifrane, Morocco: Springer International Publishing, 2020. (AIED' 2020, v. 21), p. 177–190.
- FUNAYAMA, H. et al. Preventing critical scoring errors in short answer scoring with confidence estimation. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop. Online Event: Association for Computational Linguistics, 2020. v. 58, p. 237–243.
- HORBACH, A.; PINKAL, M. Semi-supervised clustering for short answer scoring. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). Miyazaki, Japan: European Language Resources Association (ELRA), 2018. (LREC 2018, v. 11), p. 4065–407.
- KUMAR, Y. et al. Get it Scored Using AutoSAS - An Automated System for Scoring Short Answers. In: Proceedings of the AAAI Conference on Artificial Intelligence. Honolulu (HI), USA: AAAI Press, 2019. v. 33, p. 9662–9669.1.
- MIZUMOTO, T. et al. Analytic Score Prediction and Justification Identification in Automated Short Answer Scoring. In: Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications. Florence, Italy: Association for Computational Linguistics, 2019. v. 14, p. 316–325.
- OLIVEIRA, K. L. d.; SANTOS, A. A. A. Compreensão em Leitura e Avaliação da Aprendizagem em Universitários. Psicologia: Reflexão e Crítica, SciELO, v. 18, p. 118 – 124, 04 2005.
- PADÓ, U.; PADÓ, S. Determinants of Grader Agreement: An Analysis of Multiple Short Answer Corpora. Language Resources and Evaluation, Springer, v. 55, n. 2, p. 1–30, 2021.
- SAHA, S. et al. Sentence Level or Token Level Features for Automatic Short Answer Grading?: Use Both. In: Proceedings of the 19th International Conference on Artificial Intelligence in Education. London, United Kingdom: Springer International Publishing, 2018. (AIED' 2018, v. 19), p. 503–517.
- SAHU, A.; BHOWMICK, P. K. Feature Engineering and Ensemble-Based Approach for Improving Automatic Short-Answer Grading Performance. IEEE Transactions on Learning Technologies, IEEE, v. 13, n. 1, p. 77–90, 2020.
- SÜZEN, N. et al. Automatic Short Answer Grading and Feedback Using Text Mining Methods. Procedia Computer Science, Elsevier, v. 169, n. 1, p. 726–743, 2020.
- ZESCH, T.; HEILMAN, M.; CAHILL, A. Reducing Annotation Efforts in Supervised Short Answer Scoring. In: Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications. Denver (CO), USA: Association for Computational Linguistics, 2015. v. 10, p. 124–132.