Marcos Alécio Spalenza

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

Vitória, ES

2023

Marcos Alécio Spalenza

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

Tese de Doutorado submetida ao Programa de Pós-Graduação em Informática da Universidade Federal do Espírito Santo, como requisito parcial para obtenção do Grau de Doutor em Ciência da Computação.

Universidade Federal do Espírito Santo – UFES

Centro Tecnológico

Programa de Pós-Graduação em Informática

Orientador: Prof. Dr. Elias de Oliveira

Coorientador: Profª. Dra. Claudine Badue

Vitória, ES

2023

Aos meus pais,
Marcos e Sirlene

Agradecimentos

Agradeço a Deus.

Agradeço aos meus pais, Marcos e Sirlene, por todo carinho e atenção.

Agradeço ao meu irmão, Murilo, pela companhia incondicional.

Agradeço a cada um dos amigos e amigas que fiz durante todo o período de pós-graduação, por cada um dos momentos compartilhados. Sem dúvida a presença e o apoio de cada um foi essencial para finalização deste trabalho.

Minha gratidão ao professor Elias pela disponibilidade, confiança no meu trabalho e pelo apreço dado a cada resultado obtido durante todo o período do mestrado e doutorado.

Agradeço à professora Claudine pelas importantes contribuições e por cada momento de incentivo, garantindo sempre caminhos e soluções aos problemas.

Agradeço também a todos os demais professores que tornaram esta tese possível, seja durante as aulas, por meio de sugestões ou colaborações em experimentos e artigos.

Agradeço a todos os membros do LCAD e do PPGI, por terem me acolhido, pelo suporte em todas as demandas e por todos esses anos de aprendizado.

Agradeço a Fundação de Amparo à Pesquisa e Inovação do Espírito Santo (FAPES) pelo incentivo a pesquisa e desenvolvimento científico (processo 80136451).

Agradeço a *NVIDIA Corporation* pela doação de uma *NVIDIA TITAN V* através do *NVIDIA Academic Hardware Grant Program*, contribuindo para o desenvolvimento desta e de várias pesquisas do laboratório.

Além disso, gostaria de expressar aqui minha gratidão a todos os profissionais que trabalharam dioturnamente contra a Covid-19, buscando reduzir o impacto e as consequências sobre a população.

"It's not the destination, it's the journey. And if you guys can understand that, then what you'll see happen is you won't accomplish your dreams, your dreams won't come true; something greater will."

(Kobe Bryant)

Publicações

Foram desenvolvidas as seguintes publicações no período do Doutorado que, em maior ou menor grau, apresentam certa relação com o tema deste trabalho.

Publicação de trabalhos em anais de congressos:

SPALENZA, M. A.; LUSQUINO-FILHO, L. A. D.; LIMA, P. M. V.; FRANÇA, F. M. G.; OLIVEIRA E. *LCAD-UFES at FakeDeS 2021: Fake News Detection Using Named Entity Recognition and Part-of-Speech Sequences. In: Proceedings of the Iberian Languages Evaluation Forum.* Málaga, Spain: CEUR-WS, 2021. (IberLEF - SEPLN 2021, v. 37), p. 646-654.

SPALENZA, M. A.; OLIVEIRA E.; LUSQUINO-FILHO, L. A. D.; LIMA, P. M. V.; FRANÇA, F. M. G. *Using NER + ML to Automatically Detect Fake News. In: Proceedings of the 20th International Conference on Intelligent Systems Design and Applications.* Online Event: Springer International Publishing, 2020. (ISDA 2020, v. 20), p. 1176-1187.

SPALENZA, M. A.; PIROVANI, J. P. C.; OLIVEIRA, E. *Structures Discovering for Optimizing External Clustering Validation Metrics. In: Proceedings of the 19th International Conference on Intelligent Systems Design and Applications.* Auburn (WA), USA: Springer International Publishing, 2019. (ISDA 2019, v. 19), p. 150-161.

SPALENZA, M. A.; NOGUEIRA, M. A.; ANDRADE, L. B.; OLIVEIRA E. *Uma Ferramenta para Mineração de Dados Educacionais: Extração de Informação em Ambientes Virtuais de Aprendizagem In: Computer on the Beach.* Florianópolis (SC), Brasil: Universidade do Vale do Itajaí - UNIVALI, 2018. v. 9, p. 741-750.

Participação em trabalhos publicados em anais de congressos:

OLIVEIRA, E.; SPALENZA, M. A. ; PIROVANI, J. P. C. *rAVA: A Robot for Virtual Support of Learning.* In: Proceedings of the 20th International Conference on Intelligent Systems Design and Applications. Online Event: Springer International Publishing, 2020. (ISDA 2020, v. 20), p. 1238-1247.

SILVA, W.; SPALENZA, M. A.; BOURGUET, J. R.; OLIVEIRA, E. *Towards a Tailored Hybrid Recommendation-based System for Computerized Adaptive Testing through Clustering and IRT. In: Proceedings of the International Conference on Computer Supported Education.* Libon, Portugal: SCITEPRESS, 2020. (CSEDU 2020, v. 12), p. 260-268.

SILVA, W.; SPALENZA, M. A.; BOURGUET, J. R.; OLIVEIRA, E. *Recommendation Filtering à la carte for Intelligent Tutoring Systems. In: Proceedings of the International Workshop on Algorithmic Bias in Search and Recommendation.* Online Event: Springer International Publishing, 2020. (BIAS 2020, v. 1), p. 58-65.

PIROVANI, J. P. C.; ALVES, J. SPALENZA, M. A.; SILVA, W.; COLOMBO, C. S.; OLIVEIRA, E. *Adapting NER (CRF+LG) for Many Textual Genres. In: Proceedings of the Iberian Languages Evaluation Forum.* Bilbao, Spain: CEUR-WS, 2019. (IberLEF - SEPLN 2019, v. 35), p. 421-433.

PIROVANI, J. P. C.; SPALENZA, M. A. ; OLIVEIRA, E. *Geração Automática de Questões a Partir do Reconhecimento de Entidades Nomeadas em Textos Didáticos In: XXVIII Simpósio Brasileiro de Informática na Educação.* Recife (PE), Brasil: Sociedade Brasileira de Computação, 2017. (SBIE 2017, v. 28), p. 1147-1156.

Resumo

O processo de avaliação é uma etapa básica que compõe a verificação de aprendizagem e garante o andamento do ensino conforme o currículo previsto. Dentro da avaliação de aprendizagem, as questões discursivas são comumente utilizadas para desenvolver o pensamento crítico e as habilidades de escrita. Com maior quantidade de estudantes, o professor precisa adaptar seus métodos de ensino, sem tornar a avaliação um fator limitante. Alinhado a isso, existe em sua totalidade uma grande quantidade de material, mesmo que a produção individual do estudante seja pequena. Apesar da quantidade, o professor precisa analisar em detalhes cada uma das respostas dos estudantes para identificar *gaps* na aprendizagem. Deste modo, a adoção de métodos de suporte educacional busca a melhoria da capacidade analítica desse professor, impactando diretamente o acompanhamento do aluno. Neste trabalho apresentamos um modelo de *Active Learning* para classificação de documentos educacionais, em especial a avaliação de respostas discursivas curtas. Para isso, combinamos métodos de clusterização e classificação com enriquecimento textual de forma gramatical, morfológica, semântica, sintática, estatística e sequencial para identificação dos padrões de respostas. Enquanto o sistema detecta os padrões textuais que se aproximam do modelo de correção do professor, este tem menor esforço de correção e suporte para seu modelo avaliativo. Para teste desse modelo, utilizamos um total de 65875 respostas em 255 questões da literatura, alcançando em média *accuracy* de 79% e *F1* ponderado de 78% em relação aos avaliadores humanos.

Palavras-chaves: Avaliação Automática de Questões Discursivas. *Active Learning*, Processamento de Linguagem Natural. Sistemas de Apoio ao Tutor. Classificação de Texto.

Abstract

The evaluation is a basic step that composes the learning assessment and guarantees progress through the planned curriculum. On learning assessment, the application of open-ended questions contributes to developing critical thinking and writing skills. The tutor must adapt your teaching methods at a large scale, avoiding assessment overload. Even though small documents, the answer collection produces a large corpus. Meanwhile, the tutor should analyze details inside these answers to identify potential learning gaps. Therefore, the adoption of educational supporting methods aims to improve the teacher's analytical capacity and, consequently, intensify monitoring of the student's learning. Over these studies, we present an *Active Learning* model for educational document classification, specifically for Short Answer Grading problems. For this purpose, we combine clustering and classification models for pattern recognition with grammatical, morphological, semantic, syntactic, statistical, and sequential analyses for textual enrichment. The system aims to approximate the textual patterns to the tutors' evaluation criteria, reducing their effort and supporting the learning assessment. We apply our method to 65875 students' answers among 255 questions found in Short Answers Graders' literature. According to the human graders, our proposal achieves an average accuracy of 79% and a weighted F1 of 78%.

Keywords: Automatic Short Answer Grader. Active Learning. Natural Language Processing. Tutor Support Systems. Text Categorization.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

CAA	*Computer-Assisted Assessment* (Avaliação Assistida por Computadores)
NLP	*Natural Language Processing* (Processamento de Linguagem Natural)
ML	*Machine Learning* (Aprendizado de Máquina)
DL	*Deep Learning* (Aprendizado Profundo)
AVA	*Ambiente Virtual de Aprendizagem*
SAG	*Short Answer Grader* (Avaliador de Respostas Discursivas Curtas)
SAS	*Short Answer Scoring* (Avaliação de Respostas Discursivas Curtas)
EDM	*Educational Data Mining* (Mineração de Dados Educacionais)
TF	*Term Frequency*
IDF	*Inverse Document Frequency*
LSA	*Latent Semantic Analysis*
LDA	*Latent Dirichlet Allocation*
CBoW	*Continuous Bag-of-Words*
EaD	Ensino a Distância
MOOCs	*Massive Open Online Course* (Curso Online Aberto e Massivo)
HTML	*HyperText Markup Language*
POS-Tags	*Part-of-Speech Tags*
NER	*Named Entity Recognition*
CHS	*Calinski-Harabasz Score*
DBS	*Davies-Bouldin Score*
SS	*Silhouette Score*
SSE	*Sum of Squared Errors*
CVS	*Coefficient of Variation: Cluster Sizes*

HS *Homogeneity*

CS *Completness*

CA *Clustering Accuracy*

KNN ou KNRG *K-Nearest Neighboors* ou *K-Nearest Neighboors Regression*

DTR ou DTRG *Decision Tree* ou *Decision Tree Regression*

SVM *Support Vector Machine*

GBC *Gradient Boosting*

RDF *Random Forest*

WSD, WSRG ou WiSARD *Wilkes, Stonham and Aleksander Recognition Device*

ACC *Accuracy*

PRE *Precision*

REC *Recall*

F1 (m, w) *F-Score (macro or weighted)*

MAE *Mean Absolute Error*

MSE *Mean Squared Error*

RMSE *Root Mean Squared Error*

LNREG *Linear Regression*

LSSR ou Lasso *Least Absolute Shrinkage and Selection Operator*

STI Sistema Tutor Inteligente

IAT *Intelligent Assessment Technologies*

ROUGE *Recall Oriented Understudy for Gisting Evaluation*

CNN *Convolutional Neural Networks*

LSTM *Long Short-Term Memory*

GG-NAP *Neural Approximate Parsing with Generative Grading*

BERT *Bidirectional Encoder Representations from Transformers*

ELMo *Embeddings from Language Models*

GPT *Generative Pre-trained Transformer*

Sumário

1 Introdução

A Classificação de Documentos é uma abrangente área de estudo que tem característica multidisciplinar. Entre larga escala de documentos, essa área é parte fundamental para redução de esforço de avaliação e aumento exponencial da capacidade analítica humana. Esses estudos compreendem formas de rotular documentos pela caracterização de seu conteúdo de forma objetiva ou subjetiva. A subjetividade da classificação é algo que torna ainda mais complexa a análise, pois os documentos individualmente podem não representar o contexto geral ou garantir a semântica exata dos termos que o compõem. Por isso, a categorização envolve um processo mais complexo do que a simples identificação de termos que estão diretamente associados a uma classe.

No contexto educacional, a avaliação de questões discursivas é um tópico da *Computer-Assisted Asssessment* (CAA) (BOGARÍN; CEREZO; ROMERO, 2018). Os estudos em CAA combinam a profundidade textual encontrada nas técnicas de Processamento de Linguagem Natural (NLP) com o vínculo entre o conteúdo e as categorias inerentes à Classificação de Documentos. Nessa linha de estudo encontramos como principal característica a identificação do conteúdo textual relevante para atribuição de notas e produção de feedbacks. Em outra perspectiva, temos o alinhamento da expectativa de nota do professor com seu padrão avaliativo. Então, a nota ideal envolve o reconhecimento do conteúdo relevante e a forma que o professor atribui notas para ele.

A avaliação é uma etapa básica do ensino, inclusive empregada para garantia da eficiência da aprendizagem. Por meio do método avaliativo o professor mensura a proficiência dos alunos no conteúdo ministrado. A proficiência envolve avaliar o raciocínio segundo a capacidade de resolver problemas, tomar decisões e realizar inferências sobre o assunto (CASIRAGHI; ALMEIDA, 2017). O papel da avaliação, portanto, é diagnosticar, apreciar e verificar o aprendizado dos alunos para que o professor atue no processo de formação de modo a consolidar seu método de ensino (OLIVEIRA; SANTOS, 2005). Portanto, por meio das avaliações, o professor observa o progresso dos alunos ao assimilar os conteúdos. Com a aplicação dessas avaliações, o professor pode observar a integração dos materiais pedagógicos em sala de aula, para reformulação e aperfeiçoamento das dinâmicas de ensino. Assim, é com o acompanhamento da disciplina e o apoio ao educando que as atividades estabelecem meios de reformular e controlar o processo de ensino-aprendizagem (BARREIRA; BOAVIDA; ARAÚJO, 2006).

Por meio da avaliação ocorre a identificação dos *gaps* de aprendizado. Esses *gaps* e as ações para contorná-los tornam a estrutura curricular personalizada, com a turma assimilando os temas de acordo com os objetivos da disciplina (BIGGS, 1998). Portanto,

os instrumentos e materiais avaliativos são uma forma de consolidar a aprendizagem, indicando quantitativamente e qualitativamente se os alunos assimilaram os principais temas da disciplina. Com o aumento do número de alunos em salas de aula por professor e a popularização dos *Massive Open Online Courses* (MOOCs), o apoio computacional se fez ainda mais necessário. Ao mesmo tempo em que, com a adoção da tecnologia, ampliamos o acesso aos conteúdos e a verificação da aprendizagem, tornando possível a redução da sobrecarga de trabalho dos professores (DUNLAP, 2005).

Com a mediação tecnológica, se consolidou a aplicação das atividades em maior escala. Desse modo, os Ambientes Virtuais de Aprendizagem (AVA) (MAQUINÉ, 2020) se tornaram plataformas de suporte para aulas em turmas presenciais e a distância (RAES et al., 2020). Com isso, o professor ganha também suporte na criação, na avaliação, na recomendação e na visualização de dados educacionais impactando diretamente o acompanhamento do currículo do aluno (PAIVA et al., 2012). Assim, com as ferramentas de apoio, o tutor verifica a aptidão dos estudantes, de forma individual ou coletiva, para melhorar a experiência da disciplina.

A avaliação é caracterizada pela união de uma série de aspectos que caracterizam o desenvolvimento do aluno. Podemos citar como um desses, a habilidade de escrita e comunicação sobre o tema. Alinhadas a essa habilidade, as questões discursivas indicam como os estudantes produzem um discurso em torno de um tópico. Entretanto, as questões discursivas incluem uma série de técnicas de escrita, partindo de respostas de preenchimento com poucas palavras até a produção de longas redações. Neste trabalho enfatizamos o suporte ao professor na avaliação de respostas discursivas curtas. Ainda assim, encontramos *datasets* de questões discursivas curtas com diferentes modelos de resposta e avaliação. Definimos como respostas curtas conjuntos textuais de até três sentenças, compostos por até cem palavras. Para caracterizar as respostas discursivas curtas, é apresentada na Tabela 1 a atividade de exemplo *PTASAG-46*.

Na Tabela 1, é apresentada uma questão com seu enunciado e uma amostra com 15 respostas, seguidas pelo identificador do estudante que realizou cada submissão. Essa questão será utilizada para ilustrar as questões discursivas curtas. As respostas foram enviadas para a questão *PTASAG-46*, cujo enunciado é “Quais são as diferenças entre veias e artérias?”. Como as amostras destacam, temos nessa questão um conjunto de respostas de diferentes tamanhos, formas e desenvolvimento segundo o tópico. Portanto, como o próprio exemplo indica, a liberdade de escrita e diversidade de conteúdos é uma característica comum dos *datasets*. Isso ocorre pois há livre produção textual por parte dos estudantes nas questões abertas. Assim, a escrita aberta é expressa pelas diferentes formas que os estudantes se referem a um ou mais modelos de resposta.

Mesmo com diversas linhas de resposta, a avaliação deve considerá-las durante a representação de conteúdo e o mapeamento de conceitos. Partindo por critérios objetivos, o

método avaliativo deve contemplar todas as perspectivas dos participantes. Nesse contexto, uma forma consistente de resposta contém total ou parcialmente as principais componentes avaliativas. No caso, dentro do nosso exemplo, as respostas deveriam contemplar *a condução do sangue entre tecidos e o coração para limpeza de toxinas*. Em destaque, segue o texto completo criado pelo professor:

> *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.*

Fica evidente no texto elaborado pelo professor que a essência da resposta deve, pelo menos, tangenciar fatores como a *troca do oxigênio* e os *ciclos de transferência do sangue entre coração e corpo*. Mesmo que as respostas não sejam iguais à proposta

Tabela 1 – Exemplo de respostas curtas com amostras da atividade 46 do *dataset PTASAG*.

PTASAG	
Atividade 46	
Quais são as diferenças entre veias e artérias?	
#	Resposta
4	veias são meio que canos que levam o sangue aos orgãos enquanto as artérias são os canos que levam o sangue ao coração
8	Veias: sao mais finas. Artéria: são mais grossas
14	As veias levam o sangue sem oxigênio e as artérias levam o com oxigênio
31	As artérias tem a função de eliminar o sangue sujo enquanto as veias fazem o contrário
49	As veias carregam o gás carbônico. As artérias carregam o oxigênio
61	Artéria passa ar e na veia passa sangue
74	As artérias bombeam o sangue para todo o corpo e as veias levam o sangue de volta ao coração
144	Veias passam sangue pouco oxigenado e artérias passam o sangue mais oxigenado.
154	Veia transporta gás carbônico e artéria o gás oxigênio.
161	as artérias circulam o ar e as veias circulam o sangue.
211	A diferença é a valvula
296	Veias: é um vaso sanguíneo que transporta o sangue em direção ao coração. Artérias: Vasos sanguíneos que carregam sangue do coração para todas as partes do corpo.
451	Veias: sangue venoso. Artéria: sangue arterial.
520	A veia é fina já a artéria é grossa pois faz as trocas e leva sangue para o coração
570	Artérias são mais grossas e são maiores, possuem maior fluxo de sangue e passa gases como O2 e CO2, já as veias são mais finas tem menor fluxo, mas também passa O2 e CO2

do professor, atribuímos a nota máxima quando são vinculados corretamente os *ciclos arteriais e venosos*. Isso ocorre pois, na aplicação de questões discursivas curtas, existe uma expectativa de convergência entre respostas. Essa convergência torna as respostas complementares aos enunciados que proporcionam uma ou poucas trilhas de resposta. Portanto a convergência não é necessariamente a simetria textual entre palavras, como restrição textual nas respostas abertas. Na verdade, essa convergência designa tendência para respostas elaboradas de forma impessoal e direcionadas ao tema. As respostas impessoais são caracterizadas pela ausência de marcas pessoais, sendo diretas quanto à expectativa com relação ao que está descrito no enunciado. Na sua construção, o estudante deve seguir uma das trilhas do tópico abordado, caso contrário não atende essa expectativa de resposta.

Nesse ponto é importante a definição de quatro níveis diferentes interpretativos das respostas. São eles: os *modelos de resposta*, a *expectativa de resposta*, as *respostas candidatas* e, por fim, os *padrões de resposta*.

Com o alinhamento das respostas com o tópico, podemos inferir que existem modelos que associam o tema abordado com o contexto da resposta. Então, chamamos de *modelos de resposta* a convergência entre o que está na resposta com o tema. Descrevemos então *modelo de respostas* como a tendência das respostas elaboradas pelos estudantes apresentarem entre si certo índice de similaridade.

Com a existência de modelos objetivos que vinculam respostas correspondentes, podemos afirmar também que a avaliação segue uma *expectativa de resposta* do avaliador. Por meio desse índice de similaridade, afirmamos que existe um nível de atingimento de resposta que permite a gradação em diferentes níveis de nota. Portanto, o atendimento da *expectativa de resposta* denota quais são as características relevantes que levam respostas divergentes para diferentes avaliações. Portanto, é a *expectativa de resposta* que compõe o alinhamento entre o texto dos estudantes, o tema e as notas atribuídas. Assim, respostas que seguem um mesmo padrão, ou modelo, devem receber notas equivalentes caso nenhum fator externo esteja envolvido no processo avaliativo. Podem ser citados como exemplo de um fator externo de avaliação os casos de plágio (**??**). Então, isso garante que, nas devidas proporções, os estudantes que seguem uma mesma linha de resposta recebam avaliações similares.

A partir da *expectativa de resposta* com a definição dos conteúdos que diferenciam e determinam as faixas de nota, podem ser determinadas as *respostas candidatas*. As *respostas candidatas* são aquelas produzidas pelo próprio professor como uma definição da *expectativa de resposta* de acordo com o que está disposto no enunciado da atividade. Portanto, são essas que caracterizam a avaliação em um exemplo textual, tal qual o que foi produzido pelos estudantes. Elas indicam o que compõe uma resposta completa para a questão na visão do avaliador, como um referencial. O critério avaliativo e a *resposta*

candidata não são determinísticos e, quando existe algum nível de subjetividade textual, destacam divergências entre os avaliadores especialistas (PADÓ; PADÓ, 2021).

Nesses três níveis, há formação implícita de um modelo, em que é envolvida diretamente a construção textual do aluno e do professor dentro do critério avaliativo. Entretanto, com sistemas de apoio ao professor adiciona-se mais um nível analítico. Sob a perspectiva do sistema há os *padrões de resposta*, que indicam o reconhecimento contextual e linguístico das respostas. Nesse nível, o sistema realiza a análise da questão de forma horizontal. Essa etapa, identifica características relevantes do critério segundo a *expectativa de resposta*. Ao se aproximar do que é a expectativa do avaliador, o sistema aprende vínculos entre notas e as partículas textuais que compõe os *modelos de resposta*. Assim, os *padrões de resposta* são critérios objetivos do sistema, compostos pela correspondência entre respostas e notas.

Com essas quatro perspectivas sobre o conjunto de respostas, designamos a forma que cada entidade participante do processo avaliativo observa a diversidade textual. Portanto, na visão dos estudantes, temos a convergência das respostas na formação de *modelos*. Na análise do professor, os modelos têm viés avaliativo, e esse viés é representado por meio da *expectativa de resposta*. Por outro lado, quando disponíveis, as *respostas candidatas* atuam para reduzir o nível de abstração das expectativas do avaliador, apresentando-a via exemplos. E, por fim, na ótica do sistema, a equivalência entre as estruturas textuais indicam *padrões*, formados por grupos de resposta com as mesmas características.

O processo avaliativo é por definição algo complexo, principalmente pela relação entre os documentos textuais e a nota atribuída. O processo inclui analisar todo contexto detalhadamente para atribuição de notas, produção de *feedbacks* e revisão dos conteúdos. Neste ainda enfatizamos a diversidade textual. A diversidade textual inclui as diferentes formas de linguagem adotadas para produção das respostas. Isso inclui as variações consideradas no atingimento do critério avaliativo. Assim a automação do processo de análise textual, mesmo que parcial, inclui compreender as tendências de resposta de acordo com o conteúdo abordado.

Pelo caráter multidisciplinar, as respostas curtas são analisadas segundo detalhes da produção textual, como sua completude e seu direcionamento ao tema. Assim, o aluno é avaliado pela coerência da resposta, pela capacidade de sumarização e pela aplicação da linguagem. Aliado a isso, o papel do sistema é dado pela sua capacidade de reconhecer padrões de escrita, com a análise de estruturas de linguagem na identificação do que é correspondente ao conteúdo ministrado. Para isso, são necessários altos níveis de interpretação textual. O objetivo geral dos sistemas de CAA é reduzir o esforço avaliativo do professor durante esses procedimentos avaliativos.

O reconhecimento das estruturas que formam a linguagem escrita é fundamental para a descoberta dos *padrões de resposta*. Cada resposta, como um documento textual,

Tabela 2 – Componentes observadas nas respostas de cada *dataset*.

Dataset	Características	Palavras	Caracteres
SEMEVAL2013 Beetle	98	10	50
SEMEVAL2013 SciEntsBank	110	13	64
Kaggle ASAP-SAS	2932	43	236
Powergrading	178	4	20
UK Open University	418	10	55
University of North Texas	140	20	107
Kaggle PTASAG	906	13	72
Projeto Feira Literária	123	8	40
VestUFES	1391	92	536

é composta por uma série de características. Cada característica é extraída de acordo com um aspecto dessa estrutura, seja ela gramatical, morfológica, semântica, sintática, estatística ou sequencial (KUMAR et al., 2019). Portanto, o reconhecimento de padrões passa pela identificação das características mais relevantes nas respostas que compõem a avaliação de uma atividade. A quantidade de características em média extraídas por atividade nos *datasets* utilizados neste trabalho é apresentada na Tabela 2.

Como é destacado na Tabela 2, mesmo sobre um tipo específico de questão, temos variações importantes no modelo de resposta. Apesar de todos os conjuntos se enquadrarem nas especificações de respostas já citadas, são evidentes as diferenças entre *datasets*. Enquanto majoritariamente temos conjuntos com respostas bem concisas (menores que 20 palavras), temos *VestUFES* e *Kaggle ASAP-SAS* mais descritivos e, possivelmente, com enunciados mais abstratos. Diante do escopo delimitado para as respostas, sobressaem alguns desafios na compreensão computacional da linguagem e dos métodos avaliativos. Desse modo, é fundamental a produção de modelos avaliativos computacionais que demonstrem fluência na análise da linguagem e conteúdo. Portanto, o sistema é composto por modelos avaliativos e deve se aproximar ao máximo do critério do professor ao realizar inferências.

Sabendo dos desafios de uma aplicação em CAA, propõe-se uma aplicação de *Machine Learning* (ML) para o reconhecimento de padrões textuais encontrados nas questões discursivas. Nela, utilizamos técnicas de *Active Learning* para compor o processo de anotação do especialista, mesclando com ciclos de *Unsupervised* e *Supervised Learning*. *Unsupervised Learning* é aplicado para identificação da distribuição espacial e amostragem. Enquanto isso, *Supervised Learning* é aplicado para construção de avaliadores alinhados ao reconhecimento de padrões textuais. Assim, o sistema chamado *p*Nota, apresenta uma combinação de técnicas de pré-processamento, clusterização, amostragem, classificação e produção de *feedbacks* para criar um modelo avaliativo de referência dado o critério do professor.

A técnica aplicada na atribuição de notas deve seguir rigorosamente as caracterísi-

cas da avaliação realizada pelo professor. Para isso, o modelo de avaliação de respostas discursivas curtas, ou *Short Answer Graders* (SAG), tem três papéis principais. O primeiro papel é a identificação dos padrões de resposta. O segundo é reproduzir o critério avaliativo do professor por meio de sua técnica de avaliação. Por fim, o terceiro compõe a descrição do método, com a criação de *feedbacks* para todos os participantes (ARTER; CHAPPUIS, 2006; SPALENZA et al., 2016b).

1.1 Problema

Na literatura dos SAGs, encontramos uma série de problemas que caracteriza a evolução da pesquisa durante os anos. Em geral, a característica principal dos problemas é a relação entre conteúdo textual e a atribuição de nota, nem sempre de forma objetiva. Com isso, os avaliadores devem aprender o método do professor como especialista, identificando padrões de resposta que levam a cada uma das categorias de nota. Assim, além de compreender a linguagem, os modelos precisam identificar o vínculo de cada resposta com o tema da questão. Nessa linha, a expectativa é a criação de modelos SAG cada vez mais similares ao formato de avaliação do professor. Portanto, com forte aderência aos desafios de NLP, os SAGs buscam estabelecer critérios na atribuição de nota similares aos do professor (PADÓ; PADÓ, 2021). Mesmo sendo um trabalho realizado há décadas (BURROWS; GUREVYCH; STEIN, 2015), a literatura dos modelos SAG descreve uma série de problemas em aberto, pouco estudados até o momento.

Nos primeiros sistemas, a modelagem de questões discursivas era um trabalho realizado com o texto na forma bruta (PÉREZ-MARÍN; PASCUAL-NIETO; RODRÍGUEZ, 2009). Neles, a atribuição de notas é dada pela compatibilidade exata entre texto e chave de resposta. São modelos sem flexibilidade e resolvem apenas problemas simples. Essa linha de trabalho falhou por inúmeras vezes na padronização e na identificação de sinônimos (LEFFA, 2003). Problemas similares aconteceram com as técnicas que usavam apenas as métricas de resposta, como a quantidade de palavras ou frases que cada resposta contém. Mas foram essas linhas de trabalho que identificaram *gaps* importantes na análise de conhecimento contextual e na formulação individual das respostas. Por conta disso, há nos algoritmos SAG atuais estudos mais profundos na construção linguística (FILIGHERA; STEUER; RENSING, 2020).

Posteriormente, os sistemas ganharam um pouco de flexibilidade textual com extração de informação e expressões regulares. Porém, mesmo esses sistemas falham para identificar a paridade dos termos dentro da linguagem e na adaptação em diferentes contextos. Atualmente a área já avalia o texto de forma mais robusta, combinando NLP e ML. Nos últimos cinco anos também começaram a figurar alguns estudos comparativos, adotando técnicas de *Deep Learning* (DL) (BONTHU; SREE; KRISHNA-PRASAD, 2021).

Porém, mesmo com técnicas mais robustas, até o momento existe essa preocupação com a profundidade do aprendizado, em tradução literal para "*depth of learning*" (BURROWS; GUREVYCH; STEIN, 2015). A profundidade indica a capacidade de recuperar a informação do texto e atuar na atribuição de notas segundo o contexto. Portanto essa é a característica dos modelos SAG que define a produção de modelos complexos de correção, interpretando computacionalmente o conteúdo de respostas curtas em textos de escrita livre.

Diretamente associada à interpretação textual, existe a busca de convergência entre as respostas. Nesse aspecto há o problema para extração do viés em respostas factuais com múltiplos contextos. A objetividade das respostas é fundamental para contextualização, atrelada a compreensão do algoritmo para fatos corretos e incorretos. Assim, as questões devem ter convergência em algum nível da interpretação textual. Por conta disso, existe ainda maior complexidade para lidar com questões que resultam em respostas opinativas, individuais ou subjetivas (BAILEY; MEURERS, 2008). Nesse aspecto, é esperado que o sistema, de forma independente do conteúdo do professor, lide com a liberdade de escrita do estudante e analise a convergência entre as respostas na tentativa de recuperar padrões compatíveis (SAHA et al., 2018; LUI; NG; CHEUNG, 2022).

Em geral, para além do reconhecimento de padrões de resposta, ainda existe o alinhamento entre o conteúdo das respostas e o critério avaliativo. De forma geral, esse fator é um reflexo das referências utilizadas na criação do modelo avaliativo (KRITHIKA; NARAYANAN, 2015). Para compreender o papel do especialista, o sistema deve seguir o professor em seu padrão de avaliação, tentando replicá-lo (JORDAN, 2012; FUNAYAMA et al., 2020). Nessa situação, é complexo ao sistema se adaptar para respostas que têm o alinhamento correto mas que recebem uma distinção nas notas. Assim, apesar da escolha de palavras alinhadas com um determinado modelo de resposta, é essencial que o sistema forme vínculos entre padrões de avaliação e de respostas para criação de modelos avaliativos complexos (HIGGINS et al., 2014). É fundamental portanto, a extração do critério avaliativo do professor, para além de um sistema tradicional de reconhecimento de padrões.

Ainda sobre a expressão da nota atribuída a cada resposta, existe também a preocupação com classificações incorretas. Em todas as formas, é determinante que a atribuição de notas seja coerente e justificável, em especial na adoção dos SAGs (FUNAYAMA et al., 2020). Assim, é recorrente o uso de formas que remontam as componentes de resposta que levam a uma correção, seja via regras de associação, expressões regulares ou extração de características textuais (CHAKRABORTY; ROY; CHOUDHURY, 2017; KUMAR et al., 2019). Nessa linha, é essencial que os sistemas compreendam o conteúdo sem avaliações tendenciosas (AZAD et al., 2020), realizando uma análise ampla do conteúdo anotado.

Aliado a isso, para a garantia da isonomia no processo avaliativo, devemos identificar

quando tais incoerências ocorrerem. Existem níveis de erro tanto durante a aplicação do modelo de avaliação quanto na anotação de respostas, algo comum até entre dois especialistas (ARTSTEIN; POESIO, 2008; PADÓ; PADÓ, 2021). Mas é essencial minorar a diferença cada vez mais entre o modelo do especialista e o modelo de avaliação automática (CONDOR, 2020). Para além da necessidade de justificativa a cada nota atribuída, ainda é possível ressaltar a necessidade de reconhecer *outliers* para que, equivocadamente, este não se torne influente no método avaliativo (DING et al., 2020). Nessa dinâmica, ressalta-se a importância em isolar comportamentos anômalos do método avaliativo para que não influencie o comportamento geral do modelo automático. Sabendo disso, um critério avaliativo robusto deve ponderar quais componentes textuais formam o modelo de avaliação. A aquisição desse modelo é feita por meio da identificação do formato avaliativo do professor em uma série de respostas.

Ainda é necessário adicionar a tal problema, o desbalanceamento entre os níveis de nota e a baixa amostragem (DZIKOVSKA; NIELSEN; BREW, 2012; LUI; NG; CHEUNG, 2022). Assim, geralmente, as primeiras aplicações de uma atividade produzem bases de dados com pequenas quantidades de amostras e com grande diferença dentro dos grupos de nota. Por conta disso, uma série de trabalhos faz uso de descritores do padrão avaliativo, como as expressões regulares, regras ou quadros de *rubrics*, para interpretar a forma como professor constrói seu padrão de avaliação (BUTCHER; JORDAN, 2010; MOHLER; BUNESCU; MIHALCEA, 2011; RAMACHANDRAN; FOLTZ, 2015; CONDOR; LITSTER; PARDOS, 2021). Porém, isso contrapõe a proposta de reduzir o esforço avaliativo do professor, se for considerada a necessidade de produção de qualquer conteúdo extra sobre seu critério (ZESCH; HEILMAN; CAHILL, 2015; HORBACH; PINKAL, 2018). É importante, portanto, a melhoria das técnicas de aprendizado, amplificando a aquisição de conhecimento pelos exemplos. Por meio dessa melhoria prioriza-se o esforço já aplicado na anotação para remontar o critério avaliativo.

Um contraponto ao aprendizado por exemplos é lidar com a amplitude da linguagem durante a atribuição de notas. Isso acontece porque os padrões desconhecidos podem conter *outliers* que recebem um modelo próprio de avaliação (FILIGHERA; STEUER; RENSING, 2020). Isso significa que, dado nível de subjetividade da questão e a diversidade textual, o modelo deve aprender formas de avaliar padrões não mapeados. Assim, a seleção de amostras é fundamental para conectar a diversidade textual aos níveis de nota. Isso torna necessárias aos sistemas a análise da distribuição de amostras e a anotação guiada das respostas, conectando as notas ao tópico abordado nos documentos textuais (MARVANIYA et al., 2018).

Como consequência da representação textual das notas, na esfera avaliativa, há um problema na criação de modelos complexos entre termos e classes (RAMACHANDRAN; FOLTZ, 2015). Sendo a nota de vínculo interpretativo, em aspectos gerais, pode-se dizer

que existe apenas a aproximação ou correlação entre as formas de avaliação. A correlação indica que os especialistas concordam e seguem as mesmas diretrizes na atribuição das notas (ARTSTEIN; POESIO, 2008). Por outro lado, isso significa também que cada especialista pode ter análises individuais do conjunto de respostas. Então, é o objetivo de um avaliador, em especial nos modelo SAG, compreender e incorporar detalhes sutis do texto em seu critério de avaliação (HORBACH; PINKAL, 2018; CONDOR; LITSTER; PARDOS, 2021). Portanto, a relação termo-classe deve ser dinâmica e extrair o modelo que melhor atenda às expectativas do professor, em especial com o fato de ela ser passível de revisões (SPALENZA et al., 2016b).

Quando se olha o conjunto de respostas, é fundamental partir da sua formação por aspectos estruturais. Por consequência, além da análise detalhada da forma de escrita, é fundamental uma extensa capacidade analítica do conteúdo (SAHA et al., 2018). Isso significa que, além do nível textual é desejável que a análise seja feita em vários níveis, incluindo verificação morfológica, semântica e sintática de cada resposta (SAKAGUCHI; HEILMAN; MADNANI, 2015; RIORDAN; FLOR; PUGH, 2019; SAHU; BHOWMICK, 2020). Desse modo, a aquisição de informação do texto deve maximizar o conhecimento na formação dos modelos, para compreensão do tema e da estrutura de escrita. Somadas a isso, algumas propostas vão além e ainda exploram a conexão semântica entre respostas, questões e domínios (DZIKOVSKA et al., 2013; SAHA et al., 2019).

A confiabilidade dos SAGs também é constituída pela soma dos fatores aqui elencados. Superficialmente é possível associar esse problema à divergência de notas entre avaliadores. Porém, em um aspecto amplo, a confiabilidade do sistema inclui desde o reconhecimento do critério avaliativo até a criação de justificativas de nota por meio de modelos descritivos de *feedback* (KUMAR et al., 2019). O papel dos modelos de *feedback* vai além de descrever o que o sistema observou na avaliação. Ele declara a todos os participantes a atribuição de notas de acordo com a composição das respostas (MARVANIYA et al., 2018; BERNIUS; KRUSCHE; BRUEGGE, 2022). Portanto, a confiabilidade do sistema passa por todos os níveis citados de representação do conhecimento.

Por fim, ainda existem complicações para encontrar *datasets* públicos na literatura (BURROWS; GUREVYCH; STEIN, 2015). Os poucos disponíveis têm métricas específicas de comparação, apresentando problemas ao estabelecer comparações com a literatura e aumentando a dificuldade de replicar o que foi realizado nos demais estudos. Desse modo, em SAG uma base de dados adequada deve simular o processo avaliativo do professor, dando visibilidade a comparações e resultados encontrados na literatura.

1.2 Proposta

Conforme os relatos encontrados na literatura dos SAGs, é apresentado neste trabalho um sistema que propõe uma análise das estruturas textuais para produção dos modelos avaliativos complexos mencionados. Portanto, há uma proposta que atende várias deficiências dos SAGs. Cada um desses problemas foi detalhadamente descrito anteriormente, na Seção 1.1. Portanto, a ideia é propor, desenvolver e analisar um método de reconhecimento do critério do professor por meio da avaliação das estruturas textuais, estabelecendo relações entre as respostas e suas respectivas notas.

Para atender as demandas encontradas nos trabalhos em SAG são utilizadas técnicas clássicas de *Educational Data Mining* (EDM) (ROMERO et al., 2010). Apesar de o método ter fundamento em modelos linguísticos complexos e comportar questões em diversas linguagens, a avaliação conta com as principais bases de dados em *inglês* e *português* da literatura. Nesses *datasets* são observados três tipos de avaliações: notas ordinais, notas discretas e notas contínuas, ao qual demandam de tratamentos estatísticos diferentes (MORETTIN; BUSSAB, 2010). Portanto, neste trabalho, são estudadas estruturas para identificação das principais respostas do conjunto, reconhecimento do método avaliativo do professor (especialista) e elaboração de *feedbacks*.

Para identificação das principais respostas é apresentado um modelo de *Active Learning*. Em *Active Learning* o especialista ativamente passa o conhecimento para o algoritmo de classificação (SILVA; RIBEIRO, 2007; MILLER; LINDER; MEBANE, 2020). O algoritmo, por sua vez, utiliza as informações passadas para criar um modelo que replica o especialista na tarefa. Nesse caso, o professor ensina ao sistema seu método avaliativo, e, por meio da atribuição de notas, é formado um modelo que realiza as avaliações das demais respostas (ROMERO et al., 2010). Cada uma das respostas enviadas para atividade é considerada uma amostra para o sistema. Entre todas as amostras, é fundamental que o sistema aprenda em detalhes as diferentes características encontradas nas respostas, identificando sua representatividade. Para essa seleção o sistema combina técnicas de otimização e clusterização (EVERITT et al., 2011; SPALENZA; PIROVANI; OLIVEIRA, 2019). As respostas selecionadas são denominadas treinamento, pois serão utilizadas para produção dos modelos, enquanto as demais formam o conjunto de teste. O conjunto de teste determina o desempenho do sistema como avaliador.

No reconhecimento do método avaliativo do professor, modelos são criados visando à atribuição de notas para respostas discursivas. Essa etapa de classificação deve se aproximar ao máximo da tarefa realizada pelo professor, analisando a similaridade entre as respostas. Quanto menor a diferença entre a nota dada pelo sistema e a nota atribuída pelo professor, melhor será o modelo criado. Consequentemente, os melhores modelos representam com coerência a diversidade de notas e respostas, apresentando menor índice de erros. Na gradação das notas, quanto maior a discrepância entre as notas mais críticos

são os erros. Os dados selecionados para treino do classificador ditam o conhecimento da gradação de notas distribuídas por ele. Assim, o classificador recebe as características de cada resposta e a sua respectiva avaliação e as compara com as amostras de teste, sem notas determinadas. Portanto, o modelo de classificação, tomado aqui como avaliador, produz as notas complementares para o conjunto de dados de teste.

Por fim, a elaboração de *feedbacks* é fundamental para o suporte ao professor, com a descrição do método avaliativo e a sumarização dos resultados. Em sala de aula, os *feedbacks* são um material que detalha a avaliação para professores e alunos, de forma a sanar dúvidas e evidenciar *gaps* no aprendizado. Por outro lado, na perspectiva da interação do professor com o sistema, os *feedbacks* caracterizam a decisão e descrevem o modelo textual e a equivalência entre respostas (BERNIUS; KRUSCHE; BRUEGGE, 2022). Portanto, em aspectos gerais, todos os ciclos do sistema atuam para reduzir o esforço de correção do tutor, apresentar resultados de alto nível com o modelo avaliativo e gerar materiais complementares explicativos sobre os níveis de nota.

1.3 Objetivos

O objetivo deste trabalho é reduzir gradativamente o esforço de correção do professor por meio de um modelo de avaliação de respostas discursivas curtas. Para isso, é essencial o aprendizado do critério avaliativo do professor por meio de exemplos, reduzindo o esforço necessário em mais aplicações do mesmo conjunto de respostas. Assim, esperamos criar modelos avaliativos que compreendam as estruturas de linguagem, o alinhamento ao tópico de cada resposta e a forma com a qual é feita a atribuição de notas pelo especialista. Dessa forma, com o *p*Nota, esperamos que o professor esteja apto para gerenciar o seu método avaliativo em um tempo menor para se concentrar na verificação de aprendizagem do aluno.

Temos, dessa forma, como objetivo principal aprender o comportamento avaliativo que vincula uma nota para determinadas estruturas textuais que compõem as respostas dos estudantes. Para isso, esperamos reproduzir por meio do *p*Nota padrões avaliativos complexos, tornando-se um avaliador correspondente ao professor.

Para isso, estudamos as diferentes formas de construção de respostas discursivas curtas, a identificação de componentes relevantes e a representação do conhecimento em questões discursivas curtas. Nessa linha, para atingir o objetivo geral descrevem-se os seguintes objetivos específicos:

- Organizar os *datasets* públicos da literatura para estabelecer uma comparação com resultados obtidos em estudos correlatos (BURROWS; GUREVYCH; STEIN, 2015);

- Estudar o impacto das técnicas de Processamento de Linguagem Natural e Recuperação da Informação para a identificação da relação termo-classe de forma gramatical, morfológica, semântica, sintática, estatística ou sequencial (GALHARDI; BRANCHER, 2018; KUMAR et al., 2019; SAHU; BHOWMICK, 2020);
- Interpretar minuciosamente as respostas e o alinhamento do conteúdo, observando a frequência de ocorrência e co-ocorrência de termos segundo sua relevância (JORDAN, 2012; SAHA et al., 2018; DING et al., 2020);
- Elaborar e ajustar a avaliação de forma eficiente, assimilando o critério estabelecido pelo professor (ZESCH; HEILMAN; CAHILL, 2015; CONDOR, 2020; LUI; NG; CHEUNG, 2022);
- Criar modelos avaliativos robustos, associando as categorias de nota aos padrões textuais (BUTCHER; JORDAN, 2010; HEILMAN; MADNANI, 2015; BURROWS; GUREVYCH; STEIN, 2015);
- Identificar estruturas textuais para cada categoria de nota, removendo *outliers* e controlando da consistência da classificação (DING et al., 2020; FILIGHERA; STEUER; RENSING, 2020);
- Apresentar avaliações adequadas ao formato de correção do professor (HIGGINS et al., 2014; FUNAYAMA et al., 2020; PADÓ; PADÓ, 2021);
- Gerar *feedbacks* que colaborem com o processo avaliativo, como o quadro de *rubrics*, de forma a contribuir com a discussão de resultados e a representação do critério de correção (MIZUMOTO et al., 2019; SÜZEN et al., 2020; BERNIUS; KRUSCHE; BRUEGGE, 2022).

1.4 Estrutura do Trabalho

A seguir são apresentados os conteúdos desta tese. A proposta é discutida em detalhes em cinco capítulos. Para além da Introdução, o trabalho é composto dos seguintes capítulos:

- **Capítulo 2 - Revisão de Literatura:** apresenta uma breve revisão da literatura sobre métodos de análise e avaliação de respostas discursivas curtas.
- **Capítulo 3 - Método:** define a estrutura do sistema *p*Nota e as formas utilizadas para efetuar de maneira abrangente a análise de respostas discursivas curtas.
- **Capítulo 4 - Experimentos e Resultados:** descreve por meio de nove *datasets* as diferentes formas de apoio avaliativo, a modelagem da relação termo-nota e a formação de *feedbacks* utilizados pelo sistema.

- **Capítulo 5 - Conclusão:** discute as contribuições deste trabalho, conclusões extraídas dos resultados obtidos e as perspectivas de trabalhos futuros.

2 Revisão da Literatura

A Classificação de Documentos, tradicional área de ML, pode ser subdividida segundo sua motivação ou o conteúdo do conjunto de documentos. Ela envolve treinar algoritmos de classificação com exemplos rotulados para replicar métodos de identificação de conteúdo conforme o especialista (BAEZA-YATES; RIBEIRO-NETO, 2011). Cada conjunto de documentos pode ser chamado também como *dataset*, base de dados ou *corpus*. A coleção destes, porém, é denominada *corpora*. Portanto, para além da origem e do conteúdo dos documentos, o algoritmo deve se adaptar à especialização na triagem dos documentos de acordo com suas características.

O especialista realiza uma leitura dos documentos e identifica informações específicas que justificam a categoria atribuída. Para replicar tal tarefa, por meio da análise do conteúdo, o sistema deve identificar características que estão diretamente relacionadas à classe que será atribuída. Dependendo da característica dos documentos, o conteúdo relevante pode incluir a identificação de poucas palavras-chave até a formação de modelos linguísticos complexos (JURAFSKY; MARTIN, 2009). Isso acontece também com os SAGs, aplicando análises complexas da relação textual para atribuição de notas (PAIVA et al., 2012; YANG et al., 2021).

Desse modo, a atribuição de notas torna os SAGs uma complexa tarefa de classificação de documentos. Para um modelo SAG é essencial a adaptação do algoritmo de acordo com o método de classificação utilizado pelo especialista. A subjetividade do critério de avaliação deve ser levada em consideração apesar do conteúdo textual (PADÓ; PADÓ, 2021). A combinação entre o reconhecimento do modelo avaliativo e do modelo textual deve atender às expectativas do professor na avaliação (CONDOR, 2020). Enquanto em parte das atividades as notas podem ser fortemente correlacionadas com a ocorrência dos termos, em outras o critério pode ter alto nível de subjetividade (AZAD et al., 2020). Então, para a construção de um SAG, são aspectos determinantes a análise contextual das respostas e a compreensão do formato avaliativo do professor (MOHLER; BUNESCU; MIHALCEA, 2011).

2.1 *Computer-Assisted Assessment*

A sala de aula é um ambiente rico em conteúdo. As informações produzidas são descritores para o desempenho dos estudantes em sala. A análise desses dados é essencial para acompanhar o aprendizado dos alunos, verificar a necessidade de reforço do conteúdo e monitorar o cumprimento do curricular (**??**). Tradicionalmente essa dinâmica faz parte dos métodos de ensino-aprendizagem empregados pelos professores, porém, superam a sua

capacidade analítica (MADERO, 2019). Por conta disso, ganharam maior notoriedade e espaço prático os sistemas de EDM, amplificando a análise dos materiais produzidos em sala (SIEMENS; BAKER, 2012; ROMERO et al., 2010).

Em EDM, a aquisição de conhecimento aplicado a dados educacionais visa o apoio e acompanhamento do ensino (**??**). A consequência da mineração de dados nesse cenário é uma expressiva redução da carga horária do professor voltada para o acompanhamento coletivo e individual (**??**). Essa redução ocorre com o professor passando para papéis de monitoramento e auditoria dos resultados.

Portanto, via mineração de dados, são possíveis a análise de todo o material produzido pelos alunos, a criação de *feedbacks* individuais e a aplicação de reforço para determinados grupos de estudantes. Os SAGs, uma das áreas de estudo em EDM, estão diretamente associados a essas três características (BURROWS; GUREVYCH; STEIN, 2015). Os SAGs são responsáveis pela avaliação em massa de respostas textuais curtas, replicando o critério avaliativo do professor. Os SAGs fazem parte de um grupo de técnicas computacionais para apoio aos métodos avaliativos, conhecidos pelos estudos em CAA (PÉREZ-MARÍN; PASCUAL-NIETO; RODRÍGUEZ, 2009).

A verificação do aprendizado nas questões discursivas curtas contribuem para identificar se os estudantes assimilaram ou não o conteúdo ministrado em sala (OLIVEIRA; CIARELLI; OLIVEIRA, 2013). Além disso, a aplicação desse tipo de atividade é base para prática da escrita, busca de informações e sumarização de conteúdo. Desta forma, tais atividades realizam uma função importante para todos os níveis de ensino, principalmente durante o desenvolvimento da escrita (JOHNSTONE; ASHBAUGH; WARFIELD, 2002).

Com a alta carga-horária do professor, existe baixo índice de aplicação desse tipo de questão, mesmo diante de sua relevância (BILGIN; ROWE; CLARK, 2017). Assim, considerando o tempo em sala juntamente com os esforços do planejamento, a revisão e a análise das atividades acabam sendo tratadas como secundárias. O apoio computacional nessa tarefa reduz o tempo que é necessário para avaliação do conteúdo fora de sala de aula, com o professor participando parcialmente da atribuição de notas (MING, 2005). Nesse processo, os sistemas reproduzem o critério avaliativo do professor, com este garantindo a coerência da avaliação, fazendo o sistema seguir fielmente seu critério. Adicionalmente, as aplicações de ML nesse cenário, produzem modelos que descrevem a atribuição de notas, formando *feedbacks* que podem ser aplicados diretamente em sala de aula (BUTCHER; JORDAN, 2010; BERNIUS; KRUSCHE; BRUEGGE, 2022).

No entanto, para interpretação computacional, as questões devem ser elaboradas com objetividade (BAILEY; MEURERS, 2008). Nesse aspecto, dentro de um tema, deve ser possível identificar alinhamento entre as respostas, definindo se cada uma segue ou não o que compõe o critério avaliativo. Assim, as questões discursivas (BEZERRA, 2008) envolvem a liberdade de escrita dos estudantes na formulação das respostas. Na Figura 1

são caracterizadas as formas de atividades segundo seu modelo de resposta (**??**).

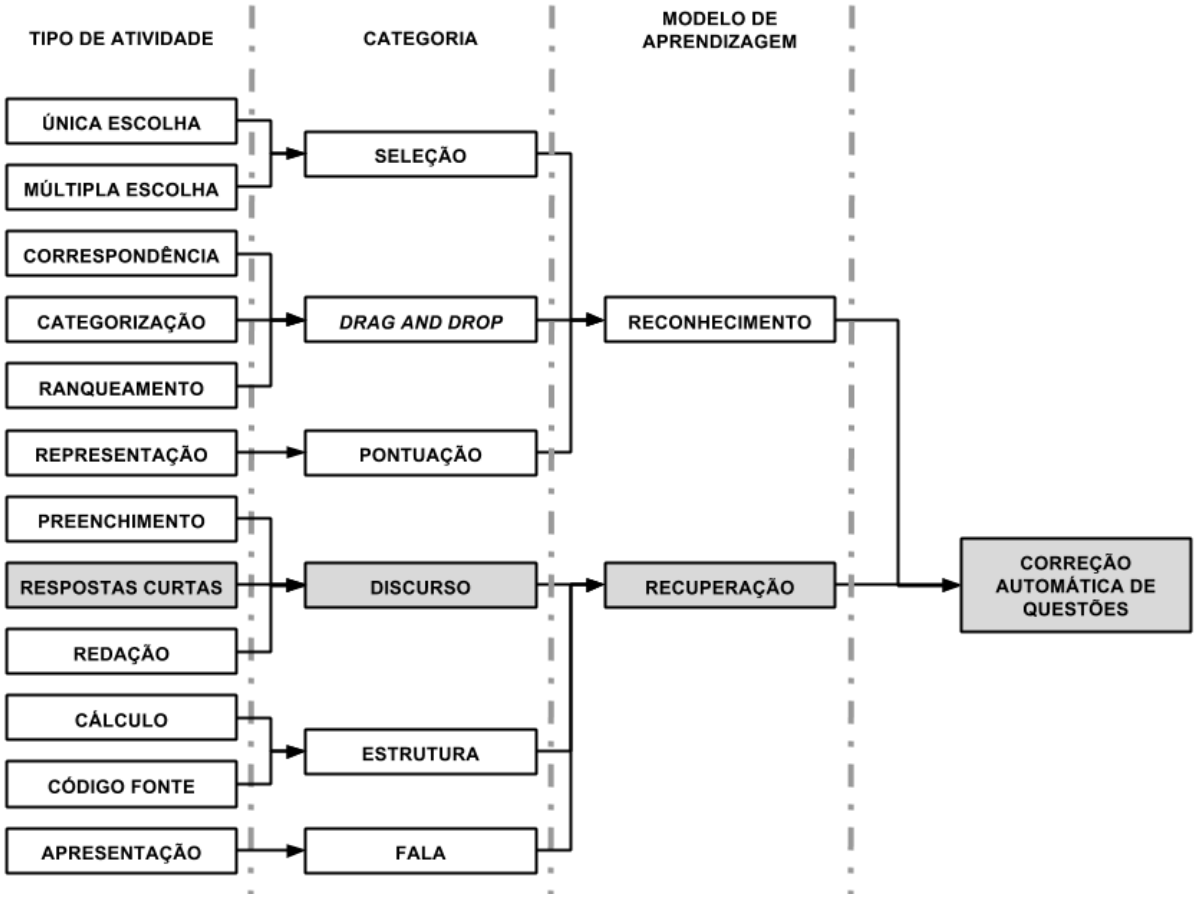


Figura 1 – A extração da informação e os tipos tradicionais de atividade aplicados no cotidiano de sala de aula.

Como apresentado na Figura 1 o professor dispõe de alguns modelos de atividades que, refletem diferentes aspectos do aprendizado. Há quesões que são abertas e com pouca ou nenhuma restrição como as redações (ALMEIDA-JÚNIOR; SPALENZA; OLIVEIRA, 2017) ou respostas fechadas para uma única palavra, guiadas pelo enunciado. As respostas discursivas encontram-se em âmbito intermediário (BAILEY; MEURERS, 2008). As respostas curtas, por sua essência, visam estabelecer a relação entre o conhecimento do aluno e o conteúdo encontrado no material didático. Na Figura 2 é demonstrado o espectro de questões trabalhados por meio das respostas discursivas curtas (**??**).

A Figura 2 posiciona as respostas curtas enquanto um nicho das questões discursivas. O ideal é que, entre os conhecimentos, a questão deve evitar abordar temas de cunho interpretativo ou temas individuais, que tangenciam experiências específicas de cada aluno (SIDDIQI; HARRISON, 2008). A representação da resposta deve ser completa, dando embasamento para a correção e evitando informações restritas ou codificadas (DING et al., 2020). O enunciado deve guiar o aluno, de forma que as respostas sejam convergentes (SÜZEN et al., 2020; FILIGHERA; STEUER; RENSING, 2020). Para isso, é fundamental que o sistema realize ao menos três etapas. A primeira etapa é o aprendizado do modelo

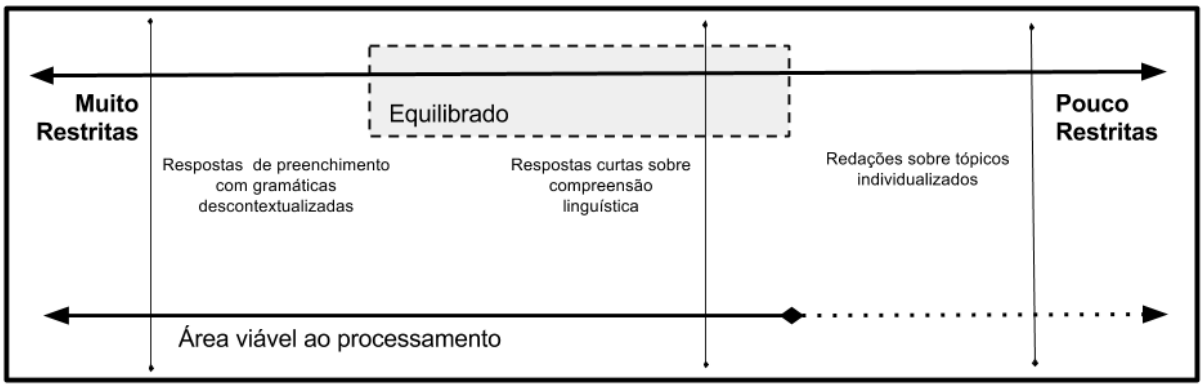


Figura 2 – Extração de informação em questões discursivas: entre respostas pequenas não-convergentes e a subjetividade das competências na avaliação de redações.

de respostas do aluno (RAMACHANDRAN; CHENG; FOLTZ, 2015). A segunda etapa é reconhecer o padrão avaliativo do professor por meio do modelo de respostas (FUNAYAMA et al., 2020). A terceira etapa é replicar o modelo avaliativo e elaborar *feedbacks* coerentes (FOWLER et al., 2021).

2.2 *Active Learning*

Em ML, o aprendizado ocorre com a formação de conhecimentos a partir da interpretação dos padrões (**??**). Esse procedimento dita a forma de aquisição de conhecimento do sistema para treinamento de modelos, buscando desempenho similar ao humano. Neste trabalho apresenta-se um método de amostragem por *Active Learning* (MILLER; LINDER; MEBANE, 2020; **??**), com anotação iterativa do professor em itens selecionados por meio das etapas de clusterização (HORBACH; PINKAL, 2018). Presentes em uma série de sistemas SAG, as técnicas de clusterização utilizam *Unsupervised Learning* para avaliação com base na similaridade das amostras (BASU; JACOBS; VANDERWENDE, 2013; ZHANG; SHAH; CHI, 2016; MARVANIYA et al., 2018). Porém, os trabalhos na literatura geralmente utilizam *Supervised Learning*, criando modelos por meio de uma série de amostras pré-avaliadas pelo especialista.

Portanto, a grande maioria dos estudos utiliza o particionamento entre treino e teste dos dados, de acordo com o determinado em cada *dataset*. Considerando cada resposta dos estudantes uma amostra, o particionamento em treino e teste reflete a divisão *a priori* do conjunto de dados em um grupo para criação do modelo e outro para avaliação (HEILMAN; MADNANI, 2015). Esse formato clássico permite ao sistema observar apenas uma parcela dos dados para reconhecimento dos padrões, realizando a inferência nas demais amostras até o momento desconhecidas (**??**). Assim, esses sistemas utilizam um conjunto de treino para extração do critério avaliativo, criando o modelo para replicar o método de atribuição de notas, pressupondo a equivalência deles.

O conjunto de amostras de treino e teste não tem necessariamente a mesma origem (SUNG et al., 2019). O uso dos sistemas SAG pelo professor compreende sua aplicação durante a disciplina. Portanto, uma atividade avaliada com um modelo SAG pode ser utilizada para uma série de aplicações, em um diferente momento e com outro grupo de estudantes. Com uma nova iteração, outros tópicos podem ser levantados em cada resposta. A tendência nesses casos é a avaliação incorreta de parte dos dados por conta da rigidez do conjunto fixado para o treinamento.

Alguns métodos também são baseados em exemplos da resposta-alvo, denominadas respostas candidatas (BANJADE; RUS; NIRAULA, 2015; ROY et al., 2016). As respostas candidatas, são amostras elaboradas pelo professor e anotadas para representar seus padrões avaliativos. Os sistemas SAG com base nesse tipo de dado buscam, em geral, a comparação direta entre as respostas e o índice de sobreposição (JIMENEZ; BECERRA; GELBUKH, 2013; KAR; CHATTERJEE; MANDAL, 2017; ZHANG; LIN; CHI, 2020). Porém, as limitações dos dados em análise são um contraponto à liberdade textual das questões discursivas curtas. Esse tipo de treinamento gera uma tendência na avaliação, com limitações na capacidade de o modelo interpretar conteúdos adversos nas respostas dos alunos (RAMACHANDRAN; FOLTZ, 2015). Além de tornar-se engessada, a resposta candidata também não garante o alinhamento para os demais documentos do *dataset*.

Nesse contexto, para contornar parcialmente as limitações, ainda existem alguns métodos que utilizam mais informações sobre a atividade na etapa de treinamento, como por exemplo o enunciado, o material de apoio e o quadro de *rubrics* (RAMACHANDRAN; CHENG; FOLTZ, 2015; WANG et al., 2019). O enunciado e o material de apoio adicionam ao sistema conhecimento externo sobre o tema. Já as respostas candidatas e o quadro de *rubrics* são materiais descritivos do modelo avaliativo do professor para todos, inclusive o próprio SAG (MIZUMOTO et al., 2019; MARVANIYA et al., 2018). Existem ainda sistemas que precisam de mais detalhes sobre a avaliação, com a confecção de regras e filtros de conteúdo (BUTCHER; JORDAN, 2010; PRIBADI et al., 2017).

Para lidar com a diversidade textual também são empregadas estratégias de *Data Augmentation*. Com *Data Augmentation* as amostras passadas como treinamento são combinadas para representar de forma mais complexa o modelo avaliativo. O uso do aumento de dados torna os sistemas tradicionais um pouco mais robustos a alterações e mudanças nos padrões básicos, reduzindo a ocorrência de classificações tendenciosas (KUMAR et al., 2019; LUN et al., 2020). Assim, a quantidade de amostras para treinamento de variações para cada modelo de resposta torna-se muito superior à quantidade dada inicialmente.

Diferentemente das técnicas citadas, o método proposto de *Active Learning* prioriza a seleção das principais amostras para otimização do esforço de anotação (**??**). A proposta combina os métodos de *clusterização* (SPALENZA; PIROVANI; OLIVEIRA, 2019) e

classificação (OLIVEIRA et al., 2014) para identificação iterativa dos diferentes tópicos abordados nas respostas. Assim, a evolução do conjunto de dados acontece durante cada uma das iterações. A clusterização, via *Unsupervised Learning*, não recebe dados anotados e extrai as respostas distintas com base no nível de similaridade (EVERITT et al., 2011). E a classificação, via *Supervised Learning*, coleta as anotações do especialista nas amostras selecionadas para treinamento do modelo (**??**). A partir daí, o modelo treinado replica a avaliação para as demais respostas.

2.3 Processamento de Linguagem Natural

Para criação de um modelo linguístico, os sistemas utilizam técnicas de NLP como estratégias de aquisição de informação. As primeiras técnicas de SAG da literatura e os primeiros sistemas propostos utilizavam descritores (GALHARDI; BRANCHER, 2018). Os descritores são características simples extraídas segundo o formato da escrita de cada documento. Em geral, são formados por características predefinidas, de acordo com a estrutura da resposta do aluno, sem levar em consideração a profundidade do conteúdo (MOHLER; MIHALCEA, 2009). Entre os descritores, os mais comuns eram a contagem de erros da linguagem, a quantidade de palavras e a frequência de certas classes gramaticais (GALHARDI et al., 2018; RIORDAN; FLOR; PUGH, 2019). Porém, tais características predefinidas não atendem a uma grande quantidade de respostas, criando modelos linguísticos com pouca aderência ao conteúdo.

Posteriormente, surgiram estruturas para maior aquisição de informação e modelagem linguística ao observar os diferentes propósitos das questões discursivas curtas e sua aplicação multidisciplinar (SAHA et al., 2018; KUMAR et al., 2019). Nesse novo cenário, os modelos linguísticos ampliaram a aderência do sistema ao tema das atividades. Por meio do conjunto de respostas, os sistemas começaram a elaborar modelos linguísticos contextuais, até o momento suficientes para encontrar associações entre as palavras (TAN et al., 2020). A partir dessas associações, os sistemas estabeleceram as primeiras conexões entre os termos de cada respostas e o critério avaliativo do professor (SAHU; BHOWMICK, 2020).

A evolução das estratégias, agora voltadas para análise do texto por completo, adicionou mais informações aos SAGs. Porém, a resposta é dada por detalhes do conjunto de respostas, sendo o todo não necessariamente relevante para avaliação. Nesse aspecto, podemos citar como adições importantes as técnicas de seleção de características, ponderação e reconhecimento de padrões (BANJADE et al., 2016). Para ponderação textual o modelo mais comum é o Term Frequency - Inverse Document Frequency (TF-IDF) (BAEZA-YATES; RIBEIRO-NETO, 2011). O TF-IDF é um método clássico que realiza a ponderação de acordo com a frequência dos termos, equilibrando a relevância de cada termo

segundo sua ocorrência nos documentos e no *dataset* (SULTAN; SALAZAR; SUMNER, 2016). Entretanto, com a evolução dos métodos neurais, ficaram em evidência as *word embeddings* (JURAFSKY; MARTIN, 2009). As *embeddings* são modelos linguísticos de grande dimensionalidade adquiridos de uma coleção de documentos (GOLDBERG; HIRST, 2017). Esses modelos analisam a conexão semântica dado o emprego conjunto dos pares de termos em *corpora* de larga escala. Assim, de forma pareada, os sistemas avaliam o nível de correspondência dos termos pelo contexto. A partir disso, os sistemas SAG avaliam a proximidade entre as respostas dos estudantes para diferentes termos, frases e contextos (SUNG; DHAMECHA; MUKHI, 2019; GHAVIDEL; ZOUAQ; DESMARAIS, 2020; GALHARDI; SOUZA; BRANCHER, 2020; HALLER et al., 2022).

Por outro lado, na seleção de características e sumarizalção de conteúdo, se destacaram as técnicas como o *Latent Semantic Analysis* (LSA) (LANDAUER; FOLTZ; LAHAM, 1998) e o *Latent Dirichlet Allocation* (LDA) (**??**). O LSA é uma das mais utilizadas na literatura (BASU; JACOBS; VANDERWENDE, 2013; SAHU; BHOWMICK, 2020). O uso dessa técnica compreende identificar relações semânticas dentro do conjunto de respostas (MOHLER; MIHALCEA, 2009). Assim, com o LSA, os sistemas reúnem o conteúdo que potencialmente contém maior significância no tema. O mesmo intuito é compartilhado pelo LDA. Esse algoritmo utiliza a análise probabilística para *ranqueamento* dos termos encontrados no texto segundo sua identificação com grupos de documentos. No âmbito dos SAGs, essa técnica de extração de tópicos, é utilizada para agrupamento pelas referências encontradas em cada grupo de nota (BASU; JACOBS; VANDERWENDE, 2013; **??**).

Entretanto, nesse nível, os modelos linguísticos criados pela frequência dos termos de cada resposta dos estudantes ainda não refletem uma análise complexa tal qual a do especialista. Portanto, na literatura existem estudos que propõem maior extração de informação textual, ainda que em textos curtos, para formação de componentes linguísticos mais robustos (SAHA et al., 2018; ZESCH; HORBACH, 2018). Assim, foram realizadas análises da estrutura textual segundo suas camadas de construção, sejam elas sintática, semântica, léxica, morfológica ou gramatical (RAMACHANDRAN; CHENG; FOLTZ, 2015; ROY et al., 2016). Ainda nessa linha, alguns estudos também remontam o conteúdo das respostas sob a perspectiva sequencial da construção textual (KUMAR; CHAKRABARTI; ROY, 2017). Essas sequências subdividem cada resposta em pequenos trechos que contêm de um a n termos para aplicar na análise de equivalência e sobreposição entre respostas (JIMENEZ; BECERRA; GELBUKH, 2013; SAKAGUCHI; HEILMAN; MADNANI, 2015; SULTAN; SALAZAR; SUMNER, 2016).

Os modelos de DL também contribuíram nessa linha. Utilizando técnicas de *Continuous Bag-of-Words* (CBoW) ou *skip-gram* e suas derivações, estes construíram representações de padrões mais complexos da vizinhança dos *tokens* (**??**). Essas técnicas buscam a identificação contextual em segmentos do conteúdo com a atribuição de pesos

ponderando sua relevância. Assim, as redes neurais substituíram boa parte das estratégias de quantificação de equivalência e a aplicação das métricas de sobreposição (HALLER et al., 2022). Através das redes, foram incorporadas melhores formas de identificar a ocorrência dos segmentos de termos. Com tais segmentos, são aplicadas enriquecimentos estruturais e semânticos na análise documental (CAMUS; FILIGHERA, 2020). No entanto, uma dificuldade encontrada na construção desses SAG são os *datasets* com poucas amostras para treinamento (BONTHU; SREE; KRISHNA-PRASAD, 2021). Com o avanço dos s SAG, a combinação dos aspectos de NLP que investigam a forma e a construção de cada sentença deve contemplar também tais representações de vizinhança das respostas (RIORDAN; FLOR; PUGH, 2019; KUMAR et al., 2019).

2.4 Avaliadores de Questões Discursivas Curtas

Os sistemas SAG, para uma análise documental complexa, são compostos por um conjunto de métodos que incluem a criação do modelo linguístico, a organização do conhecimento e a identificação de características relevantes. Apesar disso, uma parte fundamental dos sistemas SAG são os classificadores de alta qualidade (FUNAYAMA et al., 2020). Portanto, são os classificadores que destacam o conhecimento adquirido nas etapas anteriores e o aprendizado do modelo avaliativo (MOHLER; BUNESCU; MIHALCEA, 2011).

O propósito do classificador é compreender, replicar e descrever o modelo do professor (especialista) (YANG et al., 2021). Assim, é função do sistema identificar características relevantes para assimilar a forma que o professor avalia cada resposta enviada pelos estudantes (JORDAN, 2012; MAO et al., 2018). Em geral, os avaliadores automáticos são divididos segundo quatro diferentes técnicas: por mapeamento de conceitos, extração de informação, análise de *corpus*, algoritmos de ML (BURROWS; GUREVYCH; STEIN, 2015).

O método de mapeamento de conceitos consiste em um processo de detecção de determinado conteúdo nas respostas produzidas pelos estudantes. O reconhecimento de conteúdo, portanto, é realizado com análise de alinhamento entre termos de respostas (JIMENEZ; BECERRA; GELBUKH, 2013; ZHANG; LIN; CHI, 2020). Nesse método avaliativo, a principal característica é a existência dos conceitos nas respostas de maior grau de nota (KAR; CHATTERJEE; MANDAL, 2017; CHAKRABORTY; ROY; CHOUDHURY, 2017). Porém, mesmo com a construção automática de padrões através da amostragem, não é garantida a consistência dos modelos produzidos (AZAD et al., 2020). Desse modo, tendo como principal fator a compatibilidade entre respostas, o mapeamento de conceitos tende a ser muito dependente do objetivo da questão e do conteúdo do conjunto de respostas (FILIGHERA; STEUER; RENSING, 2020).

Já a extração de informação, apresenta sistemas caracterizados por um primeiro contato com estratégias de identificação factual e contextual. Nesses modelos, existe uma evolução dos métodos para análise do conteúdo, sendo compostos por técnicas de reconhecimento de padrões e séries de expressões regulares (RAMACHANDRAN; CHENG; FOLTZ, 2015; BUTCHER; JORDAN, 2010). Assim, sistemas SAG com base na extração de informação apresentam uma coleção de padrões em análise para o alinhamento entre a resposta e a expectativa do professor (TAN et al., 2020). Então, o modelo de avaliação utilizado se aproxima da leitura do professor. Porém, até aqui, essas técnicas atendem apenas aos modelos predefinidos.

Os métodos baseados em *corpus* distinguem-se pelo uso da análise estatística com base na frequência dos termos em cada conjunto de dados (KUMAR et al., 2019). Nesse método, os sistemas utilizam a linguagem para validação do alinhamento entre respostas, interpretação de variações de uso e caracterização do conteúdo (ZIAI; OTT; MEURERS, 2012; MENINI et al., 2019). Para além dos termos, essas técnicas aplicam adição de informação para maior diversidade semântica, tornando modelos mais flexíveis para análise do vocabulário do material (FOWLER et al., 2021).

Apesar da consistência dos modelos anteriores, em um âmbito geral existem limitações para aplicação de cada uma das técnicas em diferentes *datasets* (RIORDAN; FLOR; PUGH, 2019; DING et al., 2020). Em geral, as representações do critério avaliativo por parte do especialista por si não retratam bem o conhecimento para sua reprodução pela sistema (FILIGHERA; STEUER; RENSING, 2020). Em contraste aos modelos superficiais, as técnicas de ML foram incorporadas na análise textual para criação de modelos mais robustos, com fundamentação estatística (GALHARDI et al., 2018). Assim, esses modelos visam compreender o conteúdo dos documentos, pelas diferentes componentes textuais, para realizar o reconhecimento de padrões (SÜZEN et al., 2020). Distintamente das técnicas baseadas em regras e expressões regulares, os modelos de ML são capazes de se adaptar a diferentes temas e modelos de resposta (ZHANG; SHAH; CHI, 2016; SAHA et al., 2019; CAMUS; FILIGHERA, 2020). Assim, a capacidade de adaptação desses modelos permite a associação de padrões não convergentes. Como consequência, mesmo com amostras divergentes, existe a formação de critérios mais complexos para atribuição da nota.

Em geral, um objetivo dos sistemas SAG, descrito pela literatura, é mesclar os métodos e suas dinâmicas de aprendizado para evolução do modelo avaliativo (BURROWS; GUREVYCH; STEIN, 2015; ZESCH; HORBACH, 2018). Dessa maneira, é essencial a construção de modelos que reproduzam com alta qualidade a atribuição de notas realizada pelo especialista (JORDAN, 2012). Apesar das dificuldades e dos detalhes subjetivos da avaliação (ROY; RAJKUMAR; NARAHARI, 2018), o intuito é que o desenvolvimento do SAG compreenda a relação entre diferentes características de avaliação e a capacidade de atender diferentes domínios (SUNG et al., 2019; SAHA et al., 2019). Dessa forma,

esperamos o desenvolvimento de sistemas mais robustos, que compreendam o vínculo entre o critério avaliativo e a escrita dos estudantes.

3 Método

Neste trabalho, apresentamos o *p*Nota, um SAG que aplica *Active Learning* para análise da relação entre o conteúdo das respostas dos estudantes e o método avaliativo do professor. Acompanhando o desenvolvimento recente da literatura dos sistemas SAG (BURROWS; GUREVYCH; STEIN, 2015; BONTHU; SREE; KRISHNA-PRASAD, 2021; HALLER et al., 2022), identificamos pontos sensíveis e problemas descritos nesses estudos. Utilizamos como base os fundamentos de análise documental e modelagem do método avaliativo do tutor para a criação de uma proposta SAG. Com esse direcionamento, é possível verificar os principais métodos para análise das componentes textuais para elaborar um conjunto robusto de informações sobre cada resposta. O conhecimento das respostas é vinculado ao critério de avaliação do professor. Com isso, espera-se construir modelos que maximizem os resultados na atribuição de notas, aproximando-se do formato de correções do especialista. A estrutura do *p*Nota é particionada em módulos responsáveis por diferentes etapas do processo, como apresentado na Figura 3.

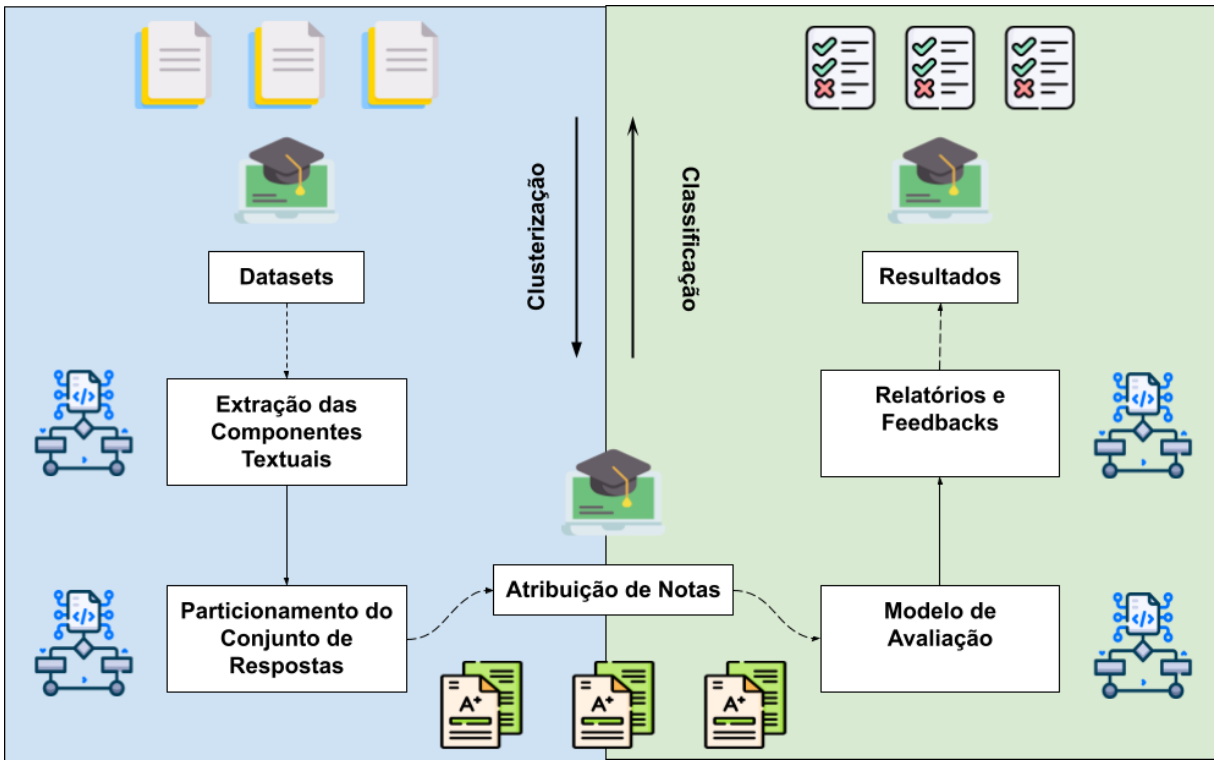


Figura 3 – Esquema do *p*Nota dividido em seus quatro módulos.

Desse modo, tal qual ilustrado na Figura 3, o sistema é composto por quatro módulos. Além destes, outros três processos dependem do estado do documento na avaliação. O primeiro módulo é o de *Extração das Componentes Textuais*, que realiza os ciclos de coleta de dados, verificação textual, extração da informação e organização do conhecimento.

Nesse módulo o sistema analisa cada resposta com aplicação de tratamentos textuais para padronização e aquisição de conhecimento. O resultado dessa etapa é um conjunto de vetores de documentos com múltiplos níveis de análise da linguagem empregada.

O segundo módulo é composto pelas técnicas de *Particionamento do Conjunto de Respostas*. Nesse núcleo são empregados métodos de otimização em *clustering* para uma seleção representativa do conteúdo textual. A representação criada pelas amostras é o que define o aprendizado do sistema. Por conta disso, as amostras são escolhidas pela sua distribuição espacial (**??**BAEZA-YATES; RIBEIRO-NETO, 2011), buscando incluir todos os tópicos abordados no tema da questão. Aqui, aplicam-se técnicas de otimização selecionando o agrupamento com melhor desempenho nos índices qualitativos.

A próxima etapa recebe as atividades particionadas em *clusters* e com a atribuição de notas nas amostras selecionadas. Neste terceiro módulo, com a construção dos *Modelos Avaliativos*, é realizada a calibração dos classificadores e a atribuição das demais notas. A calibração dos algoritmos busca refinar o critério avaliativo para compreender qual é a relação entre os termos e a nota resultante. Ao fim dessa etapa, as notas geradas colaborativamente entre professor e sistema são encaminhadas aos alunos.

Quando os resultados estão prontos, o sistema atua na etapa de *Relatórios e Feedbacks*. Nesse ponto, todas as notas já foram atribuídas e é possível atuar na transparência do modelo avaliativo. Assim, com o conjunto de informações utilizadas durante os processos, são produzidos relatórios e *feedbacks* que descrevem as notas atribuídas e os resultados do *dataset*. Por fim, os relatórios proporcionam acesso ao formato da amostragem, distribuição de notas, análise de desempenho e descrição dos padrões de resposta.

Antes da execução do sistema, existe a aplicação em sala de aula. Por meio dos AVA, o *p*Nota busca acompanhar a evolução das salas de aula digitais, com o *Ensino a Distância* (EAD) e a disseminação dos MOOCs (MOHAPATRA; MOHANTY, 2017). Para isso, utiliza-se um *framework* de coleta para transferência e controle das atividades da sala virtual para processamento externo (SPALENZA et al., 2018). Portanto são responsabilidades da aplicação a coleta das atividades no ambiente virtual, a transferência para um servidor de processamento e o envio de resultados para o professor. Na Figura 4 é apresentado o funcionamento do método de coleta de dados em diferentes plataformas de ensino.

Na Figura 4 é apresentada a forma empregada na extração de informação dos AVA. Com a configuração, o módulo acessa cada cliente e transfere as atividades para o servidor de processamento do *p*Nota. Então, o *p*Nota solicita via AVA as requisições de avaliação e, após os resultados, também envia as notas para a plataforma. Adicionalmente, os *feedbacks* gerados também são encaminhados individualmente a cada aluno. Na plataforma, após a apresentação das notas, o professor também pode realizar os ajustes necessários caso o resultado não esteja totalmente alinhado ao seu critério.

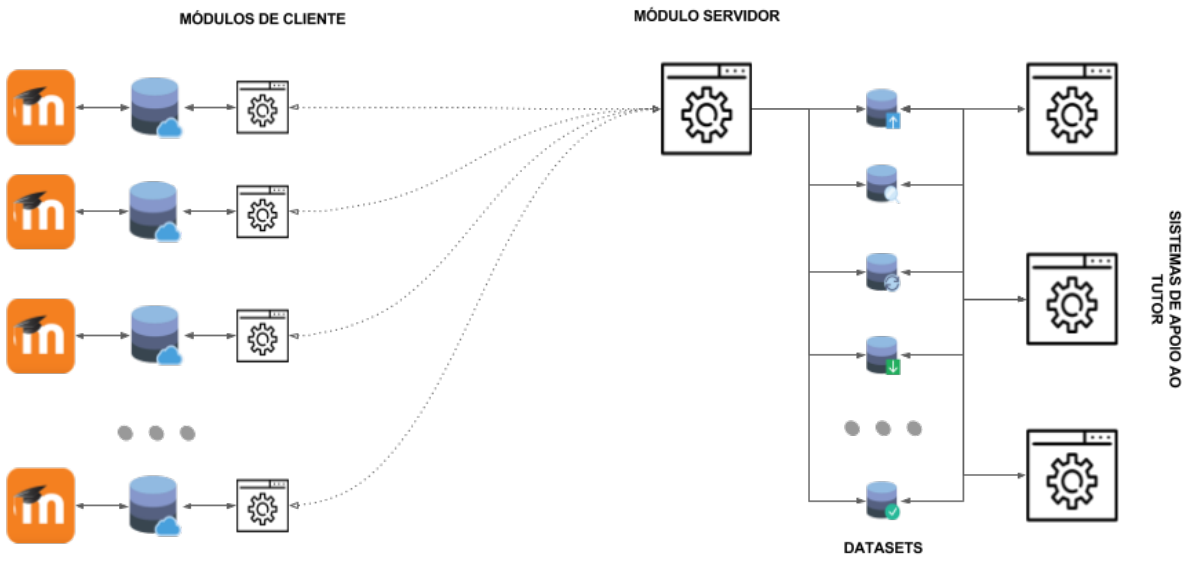


Figura 4 – *Framework* utilizado para transferência de dados, interligando plataformas AVA e o servidor do *p*Nota.

O professor, controla via plataforma o processamento, sendo aberto para determinar a finalização da atividade diretamente nela. O professor é livre para realizar alterações de qualquer nota mesmo que ainda esteja em análise pelo sistema. No controle do processo avaliativo, o professor fica responsável por monitorar a atribuição de notas e ajustar os resultados propostos pelo sistema.

3.1 Extração das Componentes Textuais

A primeira etapa, *Extração das Componentes Textuais*, realiza o carregamento e a análise do conteúdo textual. Inicialmente é realizada a leitura dos dados, carregando o conjunto de respostas que forma a atividade. É fundamental para extração que o arquivo seja recebido da forma como foi escrito pelo aluno na plataforma. Por conta disso, na sequência é realizada uma série de pré-processamentos que efetuam a limpeza destes documentos, com padronização, segmentação, filtragem, transformação e vetorização. O resultado após essa série de processos é a informação extraída, no formato de vetores com as componentes textuais de cada documento. Na Figura 5 são apresentados os processos que compõem essa etapa.

Como é mostrado na Figura 5, realizamos todo o tratamento do texto nessa etapa, com coleta das informações. A padronização é composta pela coleta do conteúdo da resposta, remoção de dados não-informativos e aumento da equivalência entre as ocorrências dos termos. A segmentação efetua o particionamento das respostas em segmentos (*tokens*) para verificação de cada termo. Com as frases particionadas, a filtragem seleciona as palavras com maior potencial de ter vínculo com o conteúdo. A transformação realiza

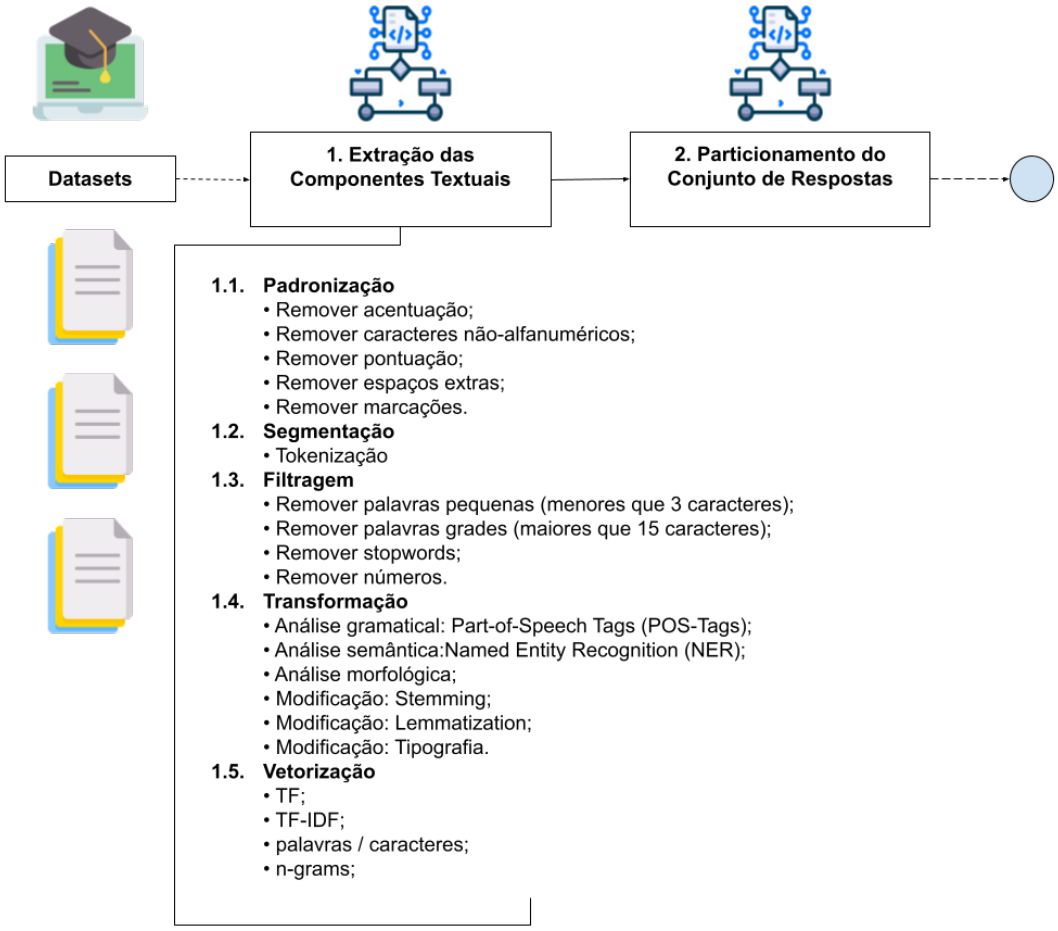


Figura 5 – Detalhe do módulo de *Extração das Componentes Textuais* no esquema do *p*Nota.

análise do conteúdo para aquisição de informação em outros níveis de linguagem. Por fim, com documentos em padrões textuais realiza-se a vetorização, gerando vetores com cada *token* ou série de *tokens* representando o conteúdo enviado pelos alunos.

As respostas, que são recebidas em formato bruto, partem então para pré-processamentos e análises de conteúdo durante essa etapa. Aqui é fundamental a adição dos níveis de compreensão linguística, para amplificar a capacidade de aprendizado dos algoritmos nas próximas etapas. Assim, é pela forma que o texto é passado no formato vetorial (BAEZA-YATES; RIBEIRO-NETO, 2011) que são identificadas respostas compatíveis até o passo da *Avaliação*.

3.1.1 Padronização

Após o documento ser enviado pelo aluno, o conteúdo do(s) arquivo(s) está em estado bruto. No estado bruto, o conteúdo não segue padrões, em especial de codificação, espaçamento, acentuação e pontuação. É necessário portanto que o formato bruto chegue ao nível de escrita que o aluno enviou ao professor. Além disso, é fundamental remover

conteúdos não interpretáveis, como caracteres não alfanuméricos e *tags* (marcações). Portanto, essa etapa é composta pelos seguintes processos:

- Remover acentuação;
- Remover caracteres não-alfanuméricos;
- Remover pontuação;
- Remover espaços extras;
- Remover marcações.

Após cada um dos passos, o texto do aluno está normalizado, tornando possível sua manipulação em nível de conteúdo. Pode-se inferir que a navegação no conteúdo só é possível após a remoção dos ruídos. São usados como exemplos as tags *HTML* e a acentuação. Por um lado, os sinais da linguagem, como acentuação, são fundamentais para leitura e pronúncia dos termos. Mas estes são irrelevantes para identificação dos termos pelo sistema. Por outro lado, o inverso ocorre com marcações de arquivos textuais. São estruturas para leitura do sistema, mas não fazem parte do conteúdo produzido pelo estudante. Em ambos os casos, não existe qualquer relação desses dados com a semântica das respostas, e consequentemente, eles são removidos durante o processo.

Os ruídos são muito comuns nos textos produzidos na internet, causados pela transferência de arquivos em repositórios externos, *crawlers* ou *web services* (HAN; PEI; KAMBER, 2011). Portanto, é a remoção de dados que aproxima a interpretação computacional proposta do envio do estudante na plataforma.

3.1.2 Segmentação

Com os documentos passíveis de interpretação, e bem próximos ao que foi enviado pelo aluno, começa-se uma análise detalhada de seu conteúdo. Este é iniciado com o particionamento de cada texto em segmentos menores. A segmentação divide em menores componentes de resposta, seja por caracteres, frases ou parágrafos. Cada particionamento, no entanto, é apenas uma forma de entrada para os procedimentos realizados na sequência. Enquanto parte dos processos faz uso do texto em formato de segmentos de palavras, outros fazem análise contextual, comumente aplicada em formato de segmentos de frase.

É mais comum o formato de segmentos de palavras, denominados *tokens*. Em todos os casos os segmentos são extraídos com base em uma *heurística*, que delimita cada segmento. A *heurística* mais comum para *tokenização* é a divisão pelo espaçamento, eliminando os espaços em branco e considerando as palavras. Porém, esses métodos simples são sujeitos a muitas falhas. Nesses casos, são melhores os métodos construídos

especificamente para a linguagem, considerando formas específicas de pontuação e divisões textuais. A *tokenização*, então, é o método que transforma o conteúdo em uma lista de palavras.

A sequência de palavras permite que os próximos níveis trabalhem a perspectiva de cada *token* dessa lista ou de sua vizinhança. Mesmo assim, é muito comum que, durante o processo, o documento seja manipulado de diferentes formas, inclusive passando várias vezes pela transformação de texto em lista de *tokens* e vice-versa. Nesse formato, os *tokens* permitem que haja navegação pelos termos adjacentes tal qual a análise dos termos de forma independente.

3.1.3 Filtragem

A filtragem de conteúdo é uma componente muito importante desse processo. Apesar de ser uma etapa que causa perda na informação inicial dos *sets* de resposta, a proposta é identificar *features* que adicionam pouco ou nenhum dado relevante. É esperado, que a inerente perda de informação cause melhoria na consistência e na equivalência entre os documentos. Guiada pelo sistema, a limpeza representa itens que têm baixa correlação com o tema, não sendo componentes do núcleo das respostas. Assim, podemos incluir os seguintes componentes como parte desse processo:

- Remover palavras pequenas (menores que três caracteres);
- Remover palavras grandes (maiores que 15 caracteres);
- Remover *stopwords*;
- Remover números.

Com a filtragem, busca-se uma avaliação fortemente ligada ao tema e o emprego contextualizado. É necessário remover os demais termos, que seriam de baixa significância, pouca capacidade de interpretação e menor relação com o que conteúdo. Esse é o caso das *stopwords*. As *stopwords* são palavras que são empregadas na linguagem como conectivos e não estão conectadas com o conteúdo passado. Elas são extremamente importantes para a nossa interpretação e reconhecimento de contexto, mas não adicionam informação quando empregadas. Assim, a lista de *stopwords* é composta por palavras fundamentais para a linguagem durante a comunicação, mas sem potencial para a análise do contexto.

No caso do tamanho das palavras ainda há uma situação adicional para além da aderência ao tema. A filtragem garante que possíveis ruídos que escaparam dos processos anteriores sejam removidos. Casos específicos como caracteres isolados, links e problemas de segmentação na origem podem gerar ruídos ainda nessa parte. Inclusive, sem uma análise matemática complexa, a verificação numérica também entra em boa parte desses

casos e pode ser incluída na filtragem. Mas é importante reconhecer os impactos quando os filtros são aplicados. O filtro numérico, por exemplo, afeta diretamente a capacidade de análise de conjuntos de respostas compostas por valores ou datas.

Uma dificuldade, entretanto, é quantificar qual é o nível de filtragem desejável, balanceando a aquisição da informação. O ideal é que todos os processos de filtragem não causem impacto nos núcleos da resposta, que contêm os termos essenciais e fortemente vinculados ao tema. Nessa linha, em sua maioria, os casos de filtragem de algumas palavras específicas não impactam a forma e mantêm os termos com aderência ao tema.

Nessa etapa, os filtros de conteúdos são métodos de redução de ruído, responsáveis por discernir quais termos podem ser extraídos de cada item de resposta. Os ruídos podem causar a queda no desempenho do algoritmo da mesma forma ou pior do que a perda de informação causada na filtragem. Assim, o ruído em meio ao texto pode ser um grande problema para o sistema durante a interpretação do conhecimento. Com isso, esperamos que a filtragem auxilie os processos subsequentes com a capacidade interpretativa e relacional entre as respostas na formação do *Modelo Avaliativo*.

3.1.4 Transformação

As análises de conteúdo são realizadas assim que os níveis anteriores prepararam o texto. Na transformação, a linguagem é analisada em níveis linguísticos. Neste processo, são interpretados alguns detalhes da construção textual para extração de *features* via técnicas de NLP. Essas técnicas observam, entre as funções de cada palavra no texto, aspectos desde sua formação até sua função dentro da frase. Os diferentes níveis analisados nessa etapa são apresentados a seguir:

- Modificação: Tipografia;
- Modificação: *Stemming*;
- Modificação: *Lemmatization*;
- Análise Gramatical: *Part-of-Speech Tags* (POS-Tags);
- Análise Semântica:*Named Entity Recognition* (NER);
- Análise Morfológica;

Cada uma das técnicas da lista aplica uma diferente transformação no texto. A primeira, bem simples, realiza a conversão do texto para uma tipografia comum, seja ela com letras maiúsculas ou minúsculas. Por outro lado, *stemming* realiza conversão mais complexa, recuperando a raiz da palavra na construção da linguagem. Com *stemming*, as palavras são convertidas para um núcleo comum, removendo as flexões, os prefixos e

os sufixos. Por outro lado, um método equivalente é realizado com *lemmatization*. Nesse outro, as palavras são convertidas para o *lemma*, a palavra base, na forma com a qual é encontrada nos dicionários (BAEZA-YATES; RIBEIRO-NETO, 2011).

Os métodos analíticos compõem ainda outras três formas mais robustas de identificação linguística. A técnica de *POS-Tags* aplica a extração da função gramatical de cada palavra segundo seu emprego na frase. Em âmbito gramatical, identifica-se qual é o papel de cada palavra no contexto, entre verbos, adjetivos, pronomes, totalizando 17 categorias (MARNEFFE et al., 2021).

Em nível semântico, aplica-se *NER*, classificando o tipo de entidade nomeada de cada um dos *tokens*. Com o NER, os nomes encontrados no texto são caracterizados pela classe que eles representam (PIROVANI et al., 2019). Entre as categorias reconhecidas há *pessoa* (PER), *local* (LOC), *organização* (ORG) e *diversos* (MISC). Isso permite que o sistema reconheça de forma simétrica diferentes menções dentro do conjunto de respostas para as principais categorias de entidades.

Por fim, o analisador morfológico identifica características da construção de cada palavra. Pela análise morfológica, as palavras são representadas pela sua flexão. Entre as flexões classificadas por cada termo estão as nominais (como gênero, número e definição) e verbais (pessoa, modo, tempo, voz). Além disso, esse mesmo processo também é responsável por algumas classificações léxicas de pronomes, adjetivos e advérbios (MARNEFFE et al., 2021).

Cada uma dessas transformações é realizada para ampliar o conhecimento de cada *token*. As análises mais complexas da linguagem e as categorizações dos termos permitem que as respostas sejam interpretadas de forma mais profunda. Essa profundidade é necessária para que, além do nível textual, a simetria das respostas seja levada em consideração. Desse modo, nesse ponto, a linguagem torna-se mais próxima da compreensão do algoritmo do que da forma original, aplicada na escrita.

A resultante desses processos é uma forte análise das componentes textuais de cada documento, buscando a identificação e a compreensão das estruturas textuais (SPALENZA et al., 2020). Com maior profundidade textual, espera-se tornar o sistema mais flexível para lidar com texto livre (DING et al., 2020). Assim, apesar das nuances da linguagem, o sistema consegue reconhecer e lidar com padrões que estão na composição de cada sentença (FILIGHERA; STEUER; RENSING, 2020). Então, a compreensão de diferentes níveis linguísticos é fundamental para a construção do SAG (SAHU; BHOWMICK, 2020).

3.1.5 Vetorização

A vetorização, como última etapa do pré-processamento, é responsável por extrair o modelo numérico de cada documento, permitindo mensurar a diferença ou a equivalência

para os demais da coleção. Dessa forma, os documentos são representados por vetores numéricos segundo seu padrão de características. Cada uma das *features* é analisada conforme sua frequência de ocorrência em cada documento do *dataset*. A representação vetorial numérica de cada documento pela frequência é denominada *Term Frequency* (TF). Sendo a coleção de documentos $D = \{d_0, d_1, d_2, \dots d_i\}$ e as *features* encontradas nos documentos $F = \{f_0, f_1, f_2, \dots f_j\}$. Portanto, para cada documento d na coleção D, conta-se a frequência de cada *feature* f_j do vocabulário F. Assim, a forma vetorial do documento de índice i é dada por d_i, sendo o vetor composto pela frequência n de cada *feature* no documento $n_{i,j}$. Então, cada documento é representado em D por sua forma vetorial $d_i = [\ n_{i,0}, n_{i,1}, n_{i,2} \dots n_{i,j}]$, usando TF.

Dadas as diferenças entre a frequência de cada termo em cada documento, é aplicada a ponderação para equilibrar a relação de frequência. A ponderação é denominada *Inverse Document Frequency* (IDF). O *Term Frequency-Inverse Document Frequency* (TF-IDF) estabelece a relação de que termos que ocorrem em muitos documentos têm menor relevância (BAEZA-YATES; RIBEIRO-NETO, 2011). A ponderação ocorre conforme a Equação 3.1.

$$TF - IDF = d_{i,j} * \log \frac{n_D}{n_{d_j}} \tag{3.1}$$

O IDF é uma ponderação na frequência de cada *feature* no vetor $d_{i,j}$, segundo o total de documentos n_{d_j} que contém f_j em relação ao total de documentos da coleção D. Essa ponderação reduz a diferença numérica entre uma *feature* encontrada em todos os documentos para as *features* que estão apenas em grupos específicos de documentos. Assim, o uso dessa forma potencialmente delimita melhores características (*features*) para a construção de modelos avaliativos. A aplicação desse modelo está associada à capacidade de identificação de características com alta correlação a grupos específicos de nota.

No método de vetorização, durante a contagem de frequência de cada *feature*, é priorizada a análise de vizinhança entre os termos. Preservando o aspecto textual, os *n-grams* estabelecem a frequência conjunta dos termos dentro de sequências. Em vez de cada documento ser representado por um vetor simples da frequência de cada *feature*, essa frequência é calculada segundo a sequências dos n termos. Sendo assim, aplicamos valores de n entre 1 a 5-*grams*. São utilizadas simultaneamente as sequências de um a cinco termos, de forma a capturar o comportamento das estruturas textuais na formação dos documentos. Dessa forma, os padrões identificados em *n-grams* (SPALENZA et al., 2020) visam à associação entre *features* fortemente correlacionadas nos vetores, compondo o modelo avaliativo.

3.2 Particionamento do Conjunto de Respostas

No formato vetorial, temos uma representação numérica dos documentos. Por meio dessa representação, podem ser comparadas as estruturas textuais que cada item de resposta contém. A comparabilidade permite identificar o comportamento das respostas definindo padrões. O conjunto desses padrões forma a distribuição das respostas no espaço vetorial. É possível assim, mensurar a proximidade entre as respostas, enquanto amostras, para formação de *clusters* para análise. Aplicando *Unsupervised Learning*, é realizado o *Particionamento do Conjunto de Respostas* para identificar estruturas textuais similares e contextualizar as diferentes formas de resposta. Essa forma de extração de conhecimento por clusterização é apresentada na Figura 6.

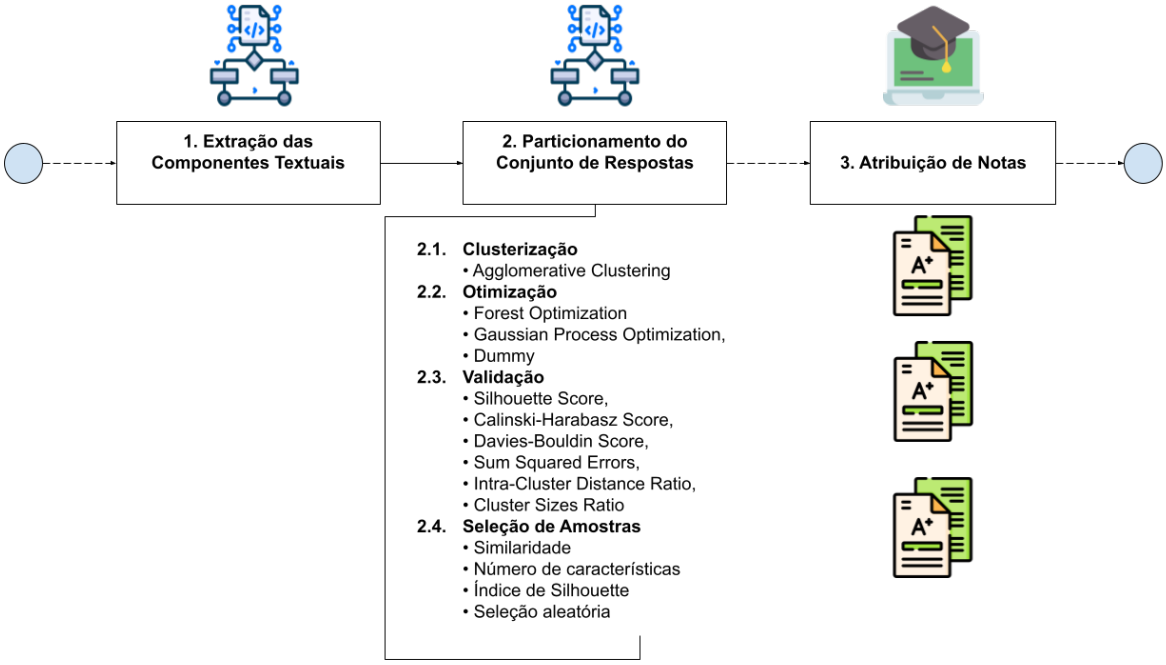


Figura 6 – Módulo de *Particionamento do Conjunto de Respostas* no esquema do *p*Nota.

Na Figura 6 é apresentado um primeiro passo do método de *Active Learning* empregado neste trabalho. São identificados diferentes componentes de resposta pela distribuição espacial dos vetores para anotação de amostras selecionadas do conjunto pelo especialista. A partir dessa forma de amostragem, são coletadas as notas, de acordo com o conteúdo que forma cada resposta ou grupo de respostas. Consequentemente, o processo de anotação requisitado pelo sistema vincula o conteúdo destas respostas com os padrões de avaliação do especialista. Diferentemente da maioria dos sistemas, que realizam amostragem aleatória, o *p*Nota analisa as instâncias que compõem cada *cluster* formado.

3.2.1 Clusterização

É realizado o particionamento das respostas utilizando técnicas de clusterização. Esse processo é responsável por agrupar respostas em grupos por similaridade. O algoritmo de *clusterização* utilizado é o *Agglomerative Clustering* (SPALENZA; PIROVANI; OLIVEIRA, 2019), um método hierárquico de agrupamento por proximidade. O *Agglomerative* compreende formar *clusters* agrupando item a item até que um limiar de proximidade seja alcançado dado um k número de *clusters* (EVERITT et al., 2011). Os grupos formados, ou *clusters*, indicam algum nível de compatibilidade entre as estruturas que formam as respostas. Assim, a análise entre a equivalência e a divergência das respostas permite a contextualização da avaliação do especialista.

Para isso, precisamos que os *clusters* sejam bons descritores do contexto, captando bem esse aspecto de equivalências e divergências textuais. A forma que foi adotada para definir o equilíbrio é a otimização via *elbow method* (EVERITT et al., 2011). Esse método compreende testar uma sequência de parâmetros da clusterização para identificar a melhor combinação de *clusters* formados segundo uma métrica de qualidade. Em geral, a métrica de qualidade é diretamente relacionada ao propósito de uso dos *clusters*.

Entre as métricas estudadas estão *Calinski-Harabasz Score* (CHS) (CALIŃSKI; J., 1974), *Davies-Bouldin Score* (DBS) (DAVIES; BOULDIN, 1979), *Silhouette Score* (SS) (ROUSSEEUW, 1987), *Sum of Squared Errors* (SSE) (MAIMON; ROKACH, 2005) e o *Coeficiente de Variação* do tamanho do cluster (CVS). Essas métricas são denominadas *Índices de Validação Interna* e avaliam os agrupamentos sem considerar a classe de cada amostra.

Cada índice é uma heurística utilizada para mensurar, sob diferentes perspectivas, a qualidade dos *clusters* gerados em relação a outras formas de agrupamento de um mesmo *dataset*. CHS mensura a razão entre a dispersão dos itens intra-*cluster* e a dispersão extra-*cluster*. DBS é o índice que estabelece a relação entre a média de similaridade entre as amostras do *cluster* para a média de similaridade entre-*clusters*. SS é a média entre as distâncias das amostras pertencentes a um *cluster* em relação às amostras do *cluster* mais próximo. SSE é uma métrica que avalia o erro de cada amostra que compõe um *cluster* em relação ao seu centroide. O centroide é o ponto médio dos itens que constituem cada *cluster*. Portanto, o centroide é uma instância representante da dispersão dos itens no *cluster*, porém é um ponto artificial e não necessariamente uma amostra que o compõe. Por fim, CVS avalia o equilíbrio entre o número de amostras agrupadas em cada *cluster*, considerando a diferença entre o maior grupo e o menor grupo formados.

Para a avaliação de respostas abertas, consideramos que o ideal são as análises que balanceiam os itens de cada *cluster* em relação aos *cluters* adjacentes. Por isso, *clusters* com padrões muito específicos não devem formar bons descritores para a distribuição, mas

sim para uma ou poucas amostras que compõe o grupo. Assim, para a proposta de *Active Learning*, os grupos equilibrados têm maior potencial para aquisição de informação. Por outro lado, também é fundamental reconhecer a proximidade intra e inter-*cluster*. Assim, foram combinados CVS e SS para avaliação amostral. Aplicamos CVS na formação dos agrupamentos enquanto o SS é observado durante a seleção amostral.

O processo de otimização, para teste dos parâmetros de clusterização com o *elbow method* visa reduzir o intervalo de busca enquanto maximiza os resultados do índice. Foram avaliados três métodos *Forest Optimization*, *Gaussian Process Optimization* e *Dummy*. Os resultados obtidos com os dois primeiros foram equivalentes nesse contexto, sendo escolhido o *Gaussian Process* para a aplicação (SPALENZA; PIROVANI; OLIVEIRA, 2019). Esse método analisa cada teste pela distribuição dos valores da métrica de qualidade como uma *gaussiana*, buscando pontos de máxima da função. A resultante é dada pelo melhor valor encontrado. O parâmetro sob controle é o k, número de *clusters*. O intervalo de k é definido por valores de 2 até $2 * \sqrt{n}$, sendo n o número de amostras do *dataset* (HAN; PEI; KAMBER, 2011). Simultaneamente, para cada combinação de k são testadas 20 métricas de distância.

- braycurtis
- canberra
- chebyshev
- correlation
- cosine
- dice
- euclidean
- hamming
- haversine
- jaccard
- kulsinski
- mahalanobis
- manhattan
- matching
- minkowski
- rogerstanimoto
- russellrao
- sokalmichener
- sokalsneath
- yule

O agrupamento selecionado é utilizado para amostragem em um percentual do conjunto de respostas disponíveis. Ainda avaliamos de forma qualitativa essa seleção segundo três índices *Homogeneity* (HS), *Completness* (CS) e *Clustering Accuracy* (CA). Em uma ótica diferente da formação dos *clusters*, com os índices qualitativos é mensurado o impacto de cada resultado da clusterização pela distribuição das classes. Esses são chamados *Índices de Validação Externa*.

CA é o índice que avalia o desempenho da clusterização enquanto classificador por voto majoritário. Nesse cenário, cada *cluster* é representado pela sua principal classe, mostrando a coesão dos grupos para sua representação de classe. Esse índice também estabelece um *baseline* de desempenho de classificação. Essa métrica é simétrica a ACC, descrita na Equação 3.3 da Seção 3.3.1. HS é o índice que mensura se os *clusters* são formados apenas por uma classe (**??**). CS por outro lado, avalia se todos os itens de uma classe estão presentes em um mesmo *cluster* (**??**). Ambos são métricas que avaliam a entropia (H) dos *clusters* dada a anotação das amostras, apresentadas na Equação 3.2.

$$Homogeneity = 1 - \frac{H(y_c|\hat{y}_c)}{H(y_c)} \tag{3.2}$$

$$Completness = 1 - \frac{H(y_c|\hat{y}_c)}{H(\hat{y}_c)}$$

A Equação 3.2 apresenta as métricas HS e CS como referências para a concentração das classes reais (y_c) dada a distribuição dos *clusters* ($\hat{y}_c$). Assim, identifica-se o comportamento das classes de nota na distribuição pela entropia. Tais métricas permitem a identificação *a posteriori* dos resultados mais coesos de clusterização. A concentração de classe por *cluster*, permite uma melhor amostragem sendo possível amplificar os resultados obtidos na etapa de classificação.

3.2.2 Seleção de Amostras

Com a formação dos *clusters*, identificam-se as principais respostas de cada agrupamento para inferência do modelo avaliativo do professor. A amostragem é realizada com a coleta de um percentual dos itens mais representativos que compõem o *dataset*. Essa coleta analisa padrões de documentos de cada *cluster*, a fim de compreender como é dada a avaliação do especialista para os diferentes padrões de resposta. As amostras são selecionadas conforme critérios específicos, descrevendo componentes do *cluster*.

A nossa amostragem segue alguns critérios. Os critérios definem a sequência de seleção para atingir o percentual escolhido para anotação. No primeiro grupo de amostras são selecionados os pares de amostras que apresentam maior e menor similaridade de cada *cluster*. O segundo grupo é composto por amostras com mais e menos características. O terceiro grupo é formado de duas formas: pelo coeficiente *silhouette* (ROUSSEEUW, 1987) de cada amostra ou pela seleção aleatória.

Nesse último grupo a análise de dispersão calcula o coeficiente de *silhouette* da amostra. Tal qual o SS, esse índice determina a razão entre a distância da amostra para os demais itens do grupo em relação aos itens do *cluster* mais próximo. Dessa forma, por meio desse método é incrementado o número de amostras por dispersão até alcançar o percentual de amostragem selecionado. Uma outra opção de seleção é a escolha de amostras pelo balanceamento do tamanho dos *clusters*. Nesse caso, o método determina que um item seja aleatoriamente selecionado, ponderado de acordo com a quantidade de itens que compõem cada grupo. Descartando as duas opções anteriores, as amostras são selecionadas aleatoriamente entre todo o *dataset*.

Terminando esse procedimento de seleção, as amostras são enviadas para atribuição de notas pelo professor. Na plataforma de correção que o professor adotou, ele realiza a atribuição de notas para cada item sugerido pelo sistema após a amostragem. Finalizado esse

processo, com o conjunto de respostas representativas e suas respectivas notas atribuídas pelo professor, começa-se a análise de padrões para inferência das notas para as demais respostas.

3.3 Modelo Avaliativo

O passo posterior à atribuição de notas, etapa com participação do professor, é a criação dos modelos computacionais. O *Modelo Avaliativo*, é a etapa do *p*Nota que desenvolve o SAG para replicar a forma de avaliação do professor. Assim, após a atribuição de notas, são criados os padrões que vinculam o texto e a avaliação. O sistema, portanto, deve se aproximar da forma como o professor gera avaliações. A plataforma utilizada para reconhecimento de padrões de nota é apresentada na Figura 7.

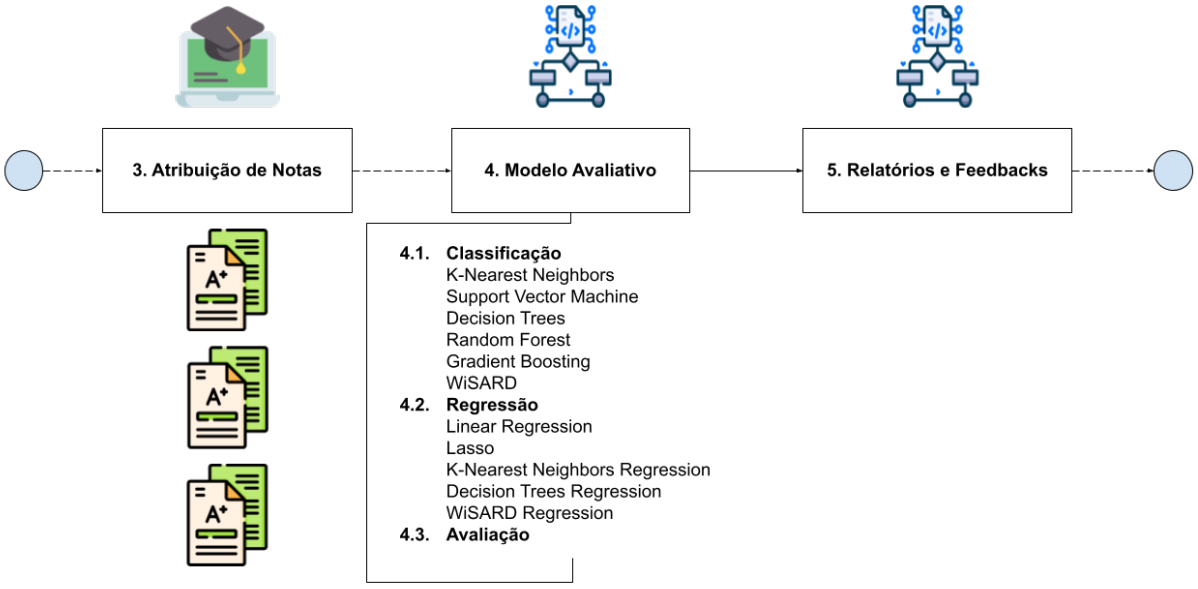


Figura 7 – Etapa de construção do *Modelo Avaliativo* no esquema do *p*Nota.

Ao receber todas as notas para as amostras, o processo começa a construção dos modelos, tal qual ilustrado na Figura 7. A criação de um modelo SAG complexo compreende identificar detalhes correspondentes entre as respostas. O *p*Nota analisa a correspondência entre notas e *features* textuais para analisar equivalência. É possível considerar que padrões equivalentes de uma mesma nota têm alta probabilidade de serem relacionados ao que o professor considerou para avaliação. Cada atividade tem especificamente uma forma de avaliação e um padrão avaliativo alinhado com a prática do professor. Portanto, a identificação do critério avaliativo não é trivial. Por isso, a técnica de *Active Learning* proposta neste estudo é fundamental para compreender o modelo e a forma que o professor trabalha durante a avaliação.

O *Modelo Avaliativo* gerado pelo *p*Nota é responsável por atender a expectativa do

professor. Sua função é vincular o padrão avaliativo com o padrão textual (*features*) do conjunto de resposta. Assim, a função da técnica de *Active Learning* proposta é treinar classificadores contextuais, transformando os métodos tradicionais de ML em avaliadores especializados. Por meio do conhecimento de uma série de níveis de linguagem e da otimização da seleção de amostras busca-se reduzir os problemas que são incorretamente atribuídos apenas à técnica aplicada na atribuição de notas. Assim, espera-se que os *Modelos Avaliativos* lidem com a variação linguística, a individualidade dos *outliers* e o desbalanceamento dos níveis de nota.

3.3.1 Classificação

O processo de classificação é utilizado em dois tipos distintos de avaliação: com notas ordinais e discretas. As notas ordinais permitem estabelecer ordem de escala numérica enquanto as discretas são textuais e requerem interpretação. No entanto, aos classificadores empregados, trabalha-se com a relação de diferença entre os níveis para aprendizado de convergência e divergência. Por um lado, a convergência indica equivalência entre os padrões de avaliação e texto encontrados nas amostras. Por outro lado, a divergência indica os padrões incompatíveis de texto por nota e entre níveis de nota, degradando sua influência avaliativa. A essência das técnicas empregadas o conhecimento do que compõe um nível de nota (equivalência) e o que é informação irrelevante ou auxiliar (divergência). Para estudar esses aspectos são aplicadas cinco diferentes formas de reconhecimento de padrões por meio dos algoritmos: *K-Nearest Neighboors*, *Decision Tree*, *Support Vector Machine*, *Gradient Boosting*, *Random Forest* e *WiSARD*.

O *K-Nearest Neighboors* (KNN) é o algoritmo de classificação pela análise da vizinhança amostral. No KNN, cada amostra é categorizada pela distribuição local dos seus k vizinhos. A atribuição do rótulo é por voto majoritário, atribuindo o mesmo valor à amostra não anotada. Diferentemente deste, o algoritmo *Decision Tree* (DTR) estabelece a equivalência entre amostras, sob uma perspectiva das características que as compõem. O DTR associa os grupos anotados com a mesma classe pelos limiar das características, gerando regras de decisão. As regras, elaboradas automaticamente, delimitam as principais *features* segundo os valores de tendência de classe. O processo de classificação, então, acontece com a comparação de cada um dos itens dentre a cadeia de decisões na estrutura de árvore.

Outro tradicional algoritmo, o *Support Vector Machine* (SVM) estabelece uma forma distinta de observar os dados. Os dados, em grupos por categoria, formam um *kernel*. O *kernel*, diferentemente do DTR, cria modelo espacial que delimita a diferença entre categorias. Então, cada amostra, é identificada segundo sua posição em relação ao limiar de características dado o modelo representante da classe. De forma similar é aplicado o algoritmo *Wilkes, Stonham and Aleksander Recognition Device* (WSD) (ALEKSANDER;

THOMAS; BOWDEN, 1984; LIMA-FILHO et al., 2020), conhecido como *WiSARD*[1]. O algoritmo produz um modelo binário com o registro de padrões de características. Cada padrão é reconhecido em análise sequencial de um intervalo de bits predefinido. O modelo binário criado é comparado com as respostas não avaliadas, categorizando-as pela similaridade entre padrões. Especificamente para esse algoritmo, a conversão dos vetores TF-IDF em seu formato binário foi dada com 1 *bit* por característica, de acordo com a esparsidade observada em dados textuais (MANNING; SCHUTZE, 1999). Assim, dado como pré-requisito de sua aplicação, o padrão submetido é dado pela existência (valor 1) ou não (valor 0) de cada característica na resposta.

Adicionalmente, dois modelos de *ensemble* foram aplicados. Os *ensembles* são técnicas que combinam vários classificadores mais simples para determinar áreas de decisão mais robustas. Os classificadores simples são denominados *weak learners*, em busca de detalhes na avaliação entre termos e classes. Nesse aspecto, *Random Forest* (RDF) é um algoritmo que combina o método tradicional *Decision Tree* com *subsets* de amostras. Desta forma, o RDF combina análises parciais do conjunto de dados para definir regras de decisão mais complexas sobre a distribuição de amostras. De modo similar, o *Gradient Boosting* (GBC) combina uma série de *Regression Trees* para otimização diferencial da função de perda (*loss*). Nessa linha, o GBC observa o gradiente da função de perda com *Logistic Regression*. Nesse aspecto, com uma série de amostragens, a técnica procura minorar o erro de classificação obtido com a calibração do modelo segundo uma sequência de *subsets*.

A combinação com modelos tradicionais e técnicas de *ensemble* visa potencializar a capacidade analítica do método. Com diferentes formatos de dados, a proposta deste trabalho testa diferentes modelos procurando o que melhor se adapta ao padrão avaliativo do professor. Nesse aspecto, o método de classificação é escolhido de acordo com a similaridade entre o modelo automático com o critério do professor (PADÓ; PADÓ, 2021). Para avaliar esse aspecto utiliza-se o coeficiente *kappa* quadrático (COHEN, 1960). As amostras são separadas em dois grupos acordo com o *Stratified K-Fold*, para mensurar a capacidade de cada algoritmo na categorização das amostras. Dentro do próprio conjunto utilizado para treinamento dos modelos SAG, é possível avaliar a paridade dos resultados com o avaliador humano (ARTSTEIN; POESIO, 2008).

Na sequência, a qualidade de cada um é avaliada com quatro métricas. A *Accuracy* (ACC), ou acurácia, mensura a equivalência percentual entre as avaliações. A *Precision* (PRE), ou precisão, estabelece a razão entre a atribuição correta de rótulos e a quantidade de atribuições incoerentes da mesma categoria. De forma similar, a *Recall* (REC), ou revocação, estabelece a razão entre a atribuição correta de rótulos e os itens de determinada categoria que foram classificados de forma incorreta. Por fim, F1 é o balanceamento entre PRE e REC, observando simultaneamente os erros de e para cada classe. A Equação

[1] wisardpkg - https://github.com/IAZero/wisardpkg

3.3 apresenta a fórmula de cada uma das métricas citadas para avaliação qualitativa dos algoritmos de classificação testados.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.3}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall}$$

Na Equação 3.3, é possível observar as fórmulas para mensurar a qualidade dos classificadores. Nelas T refere-se aos casos verdadeiros e F aos falsos. Da mesma forma, P refere-se aos casos positivos e N aos negativos (MANNING; RAGHAVAN; SCHUTZE, 2008). Porém, tradicionalmente a avaliação tem nuances que se extendem para além das marcações entre certo (ou verdadeiro) e errado (ou falso). Assim, as métricas são balanceadas conforme o número de classes, como determinado na Equação 3.4.

$$Accuracy = \frac{1}{n_{\text{amostras}}} \sum_{i=0}^{n_{\text{amostras}}-1} 1(\hat{y}_i = y_i) \tag{3.4}$$

$$Precision_{macro} = \frac{1}{|n_{classes}|} \sum_{c \in n_{classes}} Precision(y_c, \hat{y}_c)$$

$$Recall_{macro} = \frac{1}{|n_{classes}|} \sum_{c \in n_{classes}} Recall(y_c, \hat{y}_c)$$

$$F1_{macro} = \frac{1}{|n_{classes}|} \sum_{c \in n_{classes}} F1_{\beta}(y_c, \hat{y}_c)$$

$$F1_{ponderado} = \frac{1}{\sum_{c \in n_{classes}} |y_c|} \sum_{c \in n_{classes}} |y_c| \, F1(y_c, \hat{y}_c)$$

A Equação 3.4 mostra as métricas qualitativas aplicadas em avaliações com múltiplos níveis de nota, sendo y o valor de nota atribuído para cada amostra ou grupo de amostras (c) (MANNING; RAGHAVAN; SCHUTZE, 2008). Por definição, os SAGs são majoritariamente criados para avaliações com mais de uma classe de nota. É usada para mensurar o desempenho a média (macro) da atribuição de notas. Porém, pelo já esperado desbalanceamento entre notas, também avalia-se o F1 ponderado pela quantidade de amostras por classe. Comparando estatisticamente os desempenhos, via *kappa*, a expectativa

é selecionar o que tem notas mais adequadas ao modelo avaliativo do professor. Mas, o melhor modelo é o que efetivamente apresenta maior ganho de qualidade nessas métricas quando comparado com a avaliação final do professor.

3.3.2 Regressão

Outra forma de atribuição de nota é a não-categórica. Nesses casos, são chamadas de notas contínuas, pois apresentam um intervalo de notas possíveis mas sem níveis específicos. São aplicados os métodos de regressão, estimando a nota pela similaridade entre respostas. Nesse formato ainda se enquadram as noções de *equivalência* e *divergência* entre as *features* das respostas na avaliação. Os cinco métodos de regressão aplicados são *Regressão Linear*, *Lasso*, *K-Nearest Neighboors*, *Decision Tree* e *WiSARD*.

A Regressão Linear (LNRG) é um algoritmo que avalia a tendência linear das amostras segundo sua distribuição. Essa tendência linear busca, no espaço n-dimensional das características, definir os coeficientes do hiperplano que minimizam o resíduo entre as amostras. É importante para o algoritmo determinar uma função de tendência dos dados. Minimizar o erro pelos coeficientes da função reflete na simplificação do conjunto de dados. Contudo é determinante que o modelo não apresente *overfitting* e um baixo desempenho com o viés dos dados de treinamento. Por outro lado, como espera-se do algoritmo, a aquisição de informação deve extrair um modelo que minimamente descreva os dados conhecidos, evitando a ocorrência de *underfitting*. Assim, o modelo simplificado deve ser direcionado ao desempenho linear e não apenas à associação forte com o conjunto de treinamento. Também é utilizada uma variante do LNRG tradicional, denominada *Least Absolute Shrinkage and Selection Operator - Lasso* (LSSR), que utiliza a normalização dos dados com a função $L1$, reduzindo a complexidade do modelo de dados e prevenindo o *overfitting*.

Os demais três modelos, são similares aos modelos utilizados na classificação. O *K-Nearest Neighboors* (KNRG), assim como o algoritmo de classificação, observa a distribuição dos dados e define o valor resultante de acordo com a vizinhança. Assim, o resultado de cada amostra de valor desconhecido é a interpolação entre os valores das K amostras mais próximas conhecidas. De forma semelhante, *Decision Tree* (DTRG) observa características semelhantes entre amostras e, por equivalência, divide em subgrupos. A subdivisão dos itens na árvore e o particionamento em subgrupos delimita regiões específicas com resultantes correspondentes por aproximação. Dessa maneira, após o particionamento das regiões amostrais em zonas de decisão, o valor dado para todas as amostras ali categorizadas é a média conhecida do subgrupo de treinamento. De forma similar funciona a WiSARD (WSRG), organizando registradores com as notas das respostas similares atribuindo o valor médio do registrador para respostas de padrão equivalente.

Para seleção do regressor mais adequado utiliza-se a correlação de *Pearson*. Pela

correlação mensura-se a compatibilidade dos avaliadores como pares, do sistema e do professor. Nessa visão, maiores índices de correlação indicam distribuições equivalentes de distribuição de notas (MORETTIN; BUSSAB, 2010). Isso implica modelos avaliativos mais equivalentes ao método avaliativo do professor. Já a avaliação é dada pelo resíduo entre as duas notas, considerando como ideal o modelo que apresenta uma série de notas próximas do que foi atribuído pelo professor. Assim, para mensurar a diferença entre a expectativa do professor e a nota resultante do sistema são utilizadas três métricas: o *Mean Absolute Error* (MAE), o *Mean Squared Error* (MSE) e o *Root Mean Squared Error* (RMSE)

O MAE, erro médio absoluto, mensura a resíduo absoluto entre a nota predita e a nota dada pelo professor. Em outras palavras, o MAE avalia as diferenças em módulo entre os valores obtidos, segundo o alinhamento de cada predição com a expectativa do professor. Enquanto isso, MSE ou erro médio quadrático, é uma medida do resíduo entre os valores com penalização dos erros absolutos. Assim, por meio do MSE erros maiores têm maior impacto no sistema quando comparados com erros de menor grau. Por fim, o RMSE ou raiz do erro médio quadrático, é a raiz quadrada do valor obtido no MSE, normalizando o erro obtido nessa métrica em relação à avaliação do professor. A Equação 3.5 apresenta a fórmula de cada uma das métricas utilizadas para avaliação dos métodos de regressão citados.

$$MAE = \sum_{i=0}^{D} |y_i - p| \tag{3.5}$$

$$MSE = \sum_{i=0}^{D} (y_i - p)^2$$

$$RMSE = \sqrt{\sum_{i=0}^{D} (y_i - p)^2}$$

Na Equação 3.5 são apresentadas as fórmulas de avaliar o erro do modelo criado conforme a expectativa de nota. Assim, em cada fórmula as amostra i da coleção são comparadas com as notas atribuídas do avaliador humano (professor) y e pelo sistema p. O melhor modelo avaliativo é o que apresenta menor nível de erro em relação à atribuição de notas do professor. Apesar de serem comuns os erros entre modelos computacionais e a expectativa do especialista, é crucial para um bom avaliador automático a proximidade entre os modelos. Nos sistemas SAG, foram observados durante a correção entre professores até 0,66 pontos de divergência em notas de zero a cinco pontos (MOHLER; BUNESCU; MIHALCEA, 2011). Em uma escala de zero a dez pontos, representaria 1,32 pontos de divergência entre avaliadores humanos. Assim, é esperado que o sistema minimize os erros em relação ao professor, reduzindo a divergência para a nota do professor.

3.4 Relatórios e *Feedbacks*

Com as notas atribuídas pelo sistema, ocorre a criação de relatórios e *feedbacks*. Eles contribuem para descrição do modelo de Inteligência Artificial aplicado na atribuição de notas. Em especial, os *feedbacks* devem destacar quais são as *features* relevantes levadas em conta na criação do modelo avaliativo. Nessa mesma linha, os relatórios são a forma de descrever os processos realizados para todos os participantes. Esta etapa, aplicada na descrição dos processos internos do *p*Nota, é ilustrada na Figura 8.

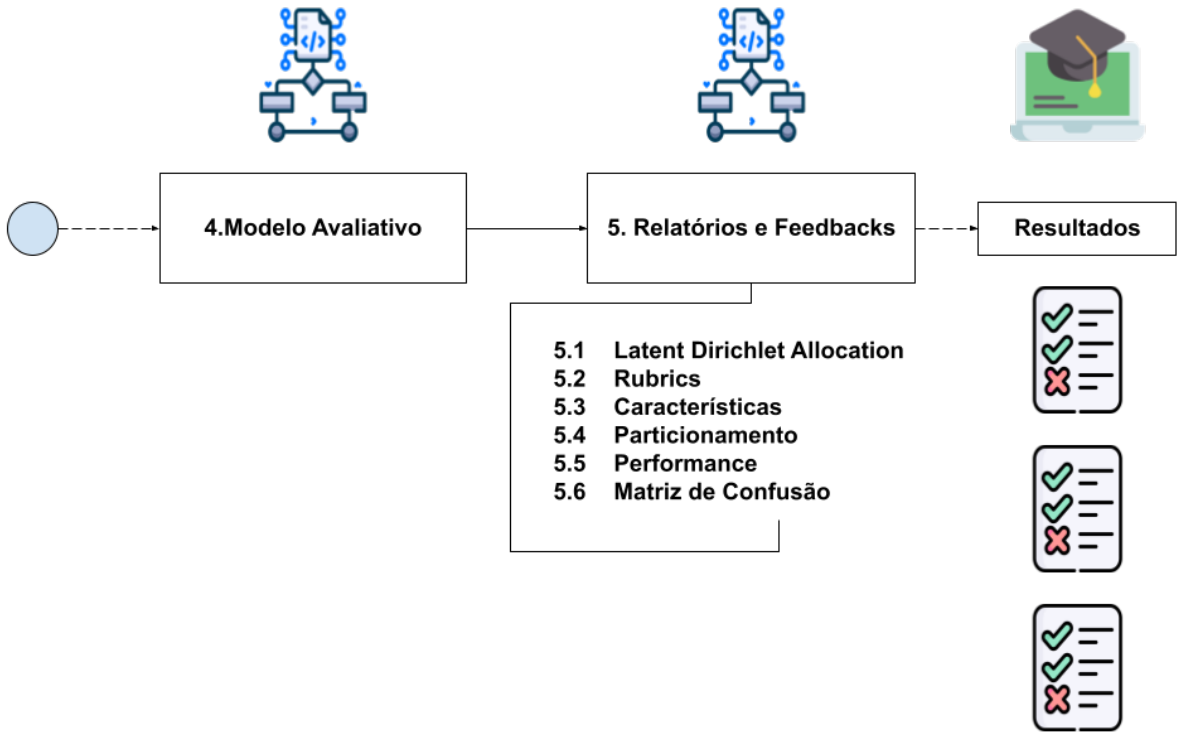


Figura 8 – Módulo de *Relatórios e Feedbacks* no esquema do *p*Nota.

Conforme a etapa apresentada na Figura 8, os relatórios precedem o envio dos resultados para o AVA, de forma a descrever o *Modelo Avaliativo*. Com fundamento na relação entre termos e notas, a caracterização das respostas é determinante para conectar usuários com os métodos aplicados pelo *p*Nota. Assim, os relatórios e os *feedbacks* devem descrever em detalhes a forma de avaliação aplicada para instruir os usuários.

Os procedimentos de descrição do *p*Nota incluem, desde características superficiais, conectadas às calibrações e às etapas do sistema, até características mais profundas, que contextualizam a interpretação do sistema sobre a atividade. É possível citar como exemplos do primeiro os *clusters* formados, amostras selecionadas para anotação ou características

mais frequentes. No segundo, por outro lado, são apresentados em linguagem natural os resultados e a proximidade entre as notas.

3.4.1 Relatório dos Processos

Os relatórios buscam passar por cada etapa do *p*Nota para mostrar como foram as interações com o professor e os resultados obtidos. Um desses é o relatório de esforço de anotação e treinamento do algoritmo. Retomando o exemplo do Capítulo 4.1, na Tabela 3 é apresentado tal relatório para a atividade 46 do *dataset PTASAG*.

Na Tabela 3 é apresentado um exemplo de relatório utilizado para explicar o que foi realizado em um dos processos. O mesmo informa qual foi o particionamento de amostras utilizado e qual foi o esforço de correção do professor. Porém, o nível descritivo deve ser maior quando caracterizamos aspectos avaliativos, tornando cada vez mais transparente a avaliação. Durante a elaboração destes, identifica-se uma certa dificuldade de interpretação das métricas categóricas em relação ao observado com métricas contínuas. Por conta disso, são estabelecidos três níveis de desempenho para as métricas percentuais: *Avançado*, *Adequado* e *Insuficiente* (NASCIMENTO; KAUARK; MOURA, 2020).

- Intervalo de 75% - 100%: Nível Avançado;
- Intervalo de 35% - 75%: Nível Adequado;
- Intervalo de 0% - 35%: Nível Insuficiente.

Os níveis são similares aos que o professor utiliza para determinar o conteúdo assimilado pelos alunos durante a avaliação (NASCIMENTO; KAUARK; MOURA, 2020). Em nível Insuficiente a relação entre as notas finais divulgadas pelo professor e as notas do sistema apresenta índices abaixo do esperado. Em nível Adequado as notas apresentam alinhamento com as que foram atribuídas pelo professor. E em nível Avançado, os resultados do modelo avaliativo foram próximos aos divulgados pelo especialista, identificando bem seu método avaliativo. Assim, foi necessário trazer para a realidade em sala de aula os resultados da classificação, tal qual já é a realidade quando é aplicado o nível de erro entre avaliadores (ALMEIDA-JÚNIOR; SPALENZA; OLIVEIRA, 2017). Na Figura 9 é apresentado o desempenho de categorização conforme esses três níveis.

Tabela 3 – Particionamento das amostras em treino e teste na atividade exemplo *PTASAG Atividade 46*.

Dataset			Amostras
PTASAG : Atividade 46			655
Treino (Un.)	Treino (%)	Teste (Un.)	Teste (%)
524	80.0	131	20.0

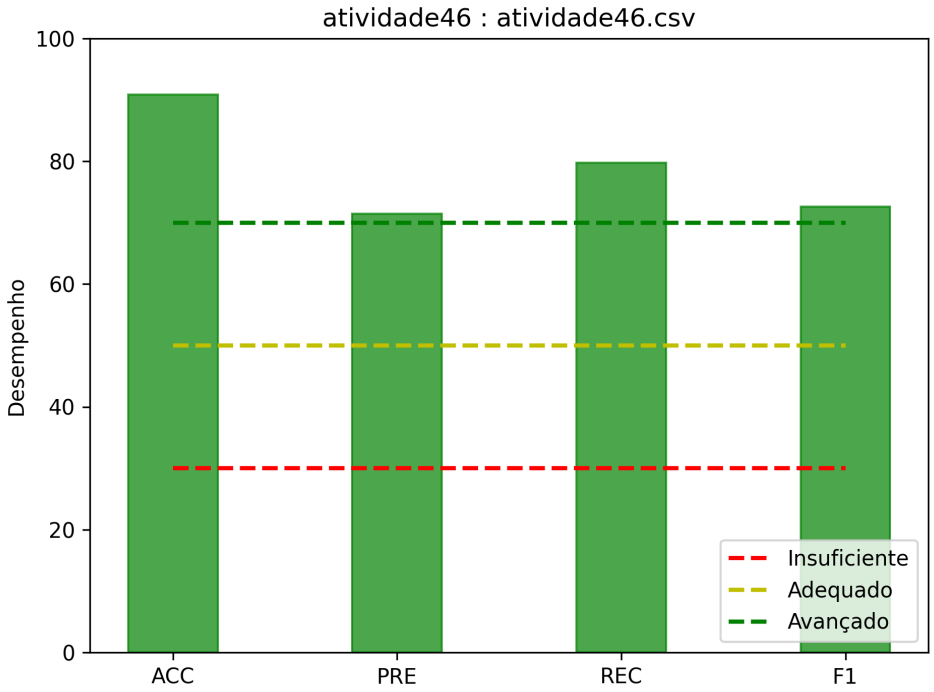


Figura 9 – Resultados de desempenho do exemplo *PTASAG Atividade 46*.

Na Figura 9, há um resultado de alto desempenho de classificação usando o classificador RDF. Com os níveis e as cores, os resultados e as métricas de avaliação tornam-se um pouco mais simples para interpretação e análise dos professores, sendo assim, uma ferramenta útil para comparação e uso na sua rotina de validação do sistema.

3.4.2 *Feedbacks* Contextuais

Os *feedbacks* contextuais são métodos que descrevem como foi o comportamento da avaliação segundo o conteúdo. Esses métodos aplicam-se diretamente à realidade da disciplina e visa caracterizar o vínculo textual de cada nível de nota. O objetivo é levar para as salas de aula um material que apoia os estudantes na compreensão da disciplina, dando suporte ao método do professor.

O primeiro modelo realiza a aplicação de cores nas respostas, identificando quais são as palavras mais correlacionadas com cada nota. Essa técnica é realizada com otimização por Algoritmo Genético (SPALENZA et al., 2016b) ou com Lime[2]. O Lime é uma ferramenta de visualização que descreve o processo de classificação de acordo com os padrões do conteúdo (RIBEIRO; SINGH; GUESTRIN, 2016). Em ambos, a ideia é atribuir coloração por nota e mostrar os termos mais correlacionados com cada uma das classes. Assim, associa-se a menção de cada termo com a nota recebida pela resposta e seu alinhamento, definindo *status* negativo ou positivo. Na Figura 10, identificam-se termos que ampliam a relação da resposta com a nota 3 atribuída.

[2] Lime - https://github.com/marcotcr/lime

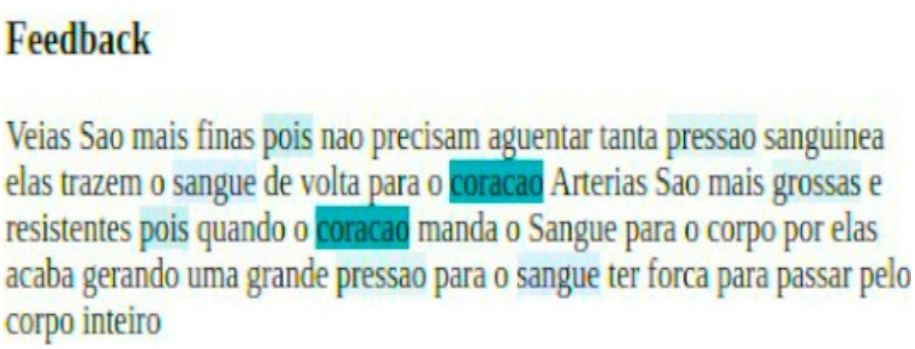


Figura 10 – Destaques nos principais termos da resposta do estudante #1995 do *PTASAG Atividade 46*.

Como é exposto na Figura 10, os termos *coração*, *sangue* e *pressão* são destaques dessa resposta. Porém, os termos poderiam ser encontrados nas demais respostas. Por isso, foi identificado que o contexto apresenta 63% de correlação com demais menções que receberam nota 3. Um segundo modelo, também em níveis textuais, extrai o conteúdo chave, de acordo com os tópicos mencionados por nível de nota. Nesse nível aplica-se LDA (**??**) no reconhecimento da composição de nota, ou seja os termos que em tese são essenciais para receber cada uma das classes. O LDA é uma técnica que aplica estatística descritiva para equivalência parcial dos dados. Nesse caso, a identificação da compatibilidade vetorial entre os textos que compõem um determinado nível de nota (SAHU; BHOWMICK, 2020). Portanto, os dois são complementares. Enquanto o primeiro observa cada resposta pela perspectiva da avaliação, o segundo realiza o oposto.

Apesar do acompanhamento da dinâmica do sistema no processo avaliativo, é complexo ao sistema identificar padrões coerentes de resposta. Para isso, utiliza-se o quadro de *rubrics* para representar o modelo avaliativo elaborado pelo sistema em conjunto com o professor. O quadro de *rubrics* é um modelo de caracterização do processo avaliativo conforme o modelo de resposta esperado para cada nota. Após o processo avaliativo, este torna-se um descritor, determinando na perspectiva dos estudantes quais foram as principais características elencadas para cada nota. Na Tabela 4 há um exemplo do quadro de *rubrics* para a nota 3 da *Atividade 46*.

Na Tabela 4 foram coletados os tópicos mais relevantes em cada uma das categorias para destacar os principais termos das respostas. No exemplo a resposta nota 3 fica evidente. Esta deve citar a relação entre o *corpo* e o *coração*, com a *pressão arterial* levando o sangue, e retornando pelas *veias*. Então, a função do *rubrics* é definir os padrões de resposta do sistema e, adicionalmente, criar relatórios que expliquem o formato avaliativo em apoio ao professor.

Tabela 4 – Tabela de *rubrics* para as duas notas encontradas na atividade exemplo e as respostas mais alinhadas com as palavras selecionadas pelo LDA.

PTASAG : Atividade 46

Nota: 3	
Tópicos: arterias coracao corpo levam pressao rico sangue veias	
#	Exemplos
19	Veias *levam* o *sangue* para o *coracao* e as *arterias levam* o *sangue* do *coracao* As *veias* sao mais finas e as *arterias* sao grossas e resistentes
78	As *veias levam sangue* ate o *coracao* elas nao aguentam muita *pressao* As *arterias levam* o *sangue* do *coracao* para o resto do *corpo* pois aguentam maior *pressao* e sao maiores
242	As *veias levam sangue* do *corpo* para o *coracao* onde ele possa ser bombeado novamente para o *corpo* As *arterias* saem do *coracao* tornando se cada vez mais finas esses vasos *levam* o *sangue* nutrientes e oxigenio do *coracao* para os tecidos
328	As *arterias* transportam o *sangue* que sai do *coracao* inicialmente *rico* em O2 as diversas partes do *corpo* as *veias* recolhem esse *sangue rico* em CO2 e *levam* de volta para o *coracao* As *arterias* possuem paredes mais grossas e as *veias* possuem valvulas que impedem o *sangue* de voltar
372	As *veias levam sangue* do *corpo* ao *coracao* e as *arterias* do *coracao* ao *corpo* Alem disso as *arterias* sao mais grossas que as *veias* para suportar *pressao*
444	As *veias* realizam o transporte do *sangue* venoso *rico* em CO2 no sentido do *corpo* para o *coracao* Ja as *arterias* carregam o *sangue rico* em O2 do *coracao* para o *corpo* Alem disso as *arterias* sao mais grossas que as *veias* pois tem que aguentar a *pressao* exercido pelos batimentos cardiacos que bombam o *sangue*
456	As *arterias* sao vasos sanguineos responsaveis por conduzir o *sangue* para fora do *coracao* carregando *sangue* artereal *rico* em O2 As *veias* sao responsaveis por conduzir o *sangue* proveniente dos tecidos para o *coracao* onde e *rico* em CO2 e carrega *sangue* venoso
479	As *veias levam* o *sangue* dos tecidos para o *coracao* e possuem baixa *pressao* ja as *arterias* possuem alta *pressao* e *levam sangue* do *coracao* para o *corpo*

4 Experimentos e Resultados

O *p*Nota é um sistema desenvolvido como laboratório para integrar melhorias na análise e avaliação de propostas de forma multidisciplinar, como discutido anteriormente. Para interagir com os usuários, o *p*Nota se apoia em plataformas AVA na Universidade Federal do Espírito Santo (UFES). Nesse período, durante o desenvolvimento foram adicionadas diferentes técnicas para desenvolver a compreensão linguística e avaliativa do modelo SAG proposto. Os testes envolveram um conjunto de atividades com mais de 126 questões e 2805 respostas (SPALENZA et al., 2016b). Nesses testes foram analisadas formas distintas de escrita e em diferentes níveis de instrução, do Ensino Básico à Pós-Graduação, abordando tópicos distintos como Computação, Arquivologia, Ciências, Filosofia, Economia e Medicina. Entretanto, para avaliar o sistema é exposta aqui uma série de experimentos com *datasets* encontrados na literatura.

Este capítulo apresenta os experimentos das duas etapas avaliativas. A primeira avalia a parte fundamental da técnica de *Active Learning* do sistema *p*Nota, com o uso da *clusterização* para a identificação dos principais itens de resposta em cada base de dados. Na sequência, a segunda apresenta os métodos de classificação, a qualidade do aprendizado do sistema na predição de notas e a sua adequação ao modelo esperado pelo tutor.

4.1 *Datasets*

De acordo com a literatura, foram selecionadas nove bases de dados, em português e inglês. Cada base de dados foi utilizada conforme sua disponibilidade e suas características. Alguns *datasets* estão em linguagens que seria possível o processamento mas é necessário maior conhecimento da linguagem para adaptação ao *p*Nota, como o *JapaneseSAS* (japonês) (ISHIOKA; KAMEDA, 2017), *Cairo University Dataset* (árabe) (GOMAA; FAHMY, 2019), *Corpus of Reading comprehension Exercises in German (CREG)* (alemão) (ZIAI; OTT; MEURERS, 2012), *Science Answer Assessment (ScAA)* (hindi) (AGARWAL; GUPTA; BAGHEL, 2020) e o *Chinese Educational Short Answers (CESA)* (chinês) (DING et al., 2020). Outros *datasets* foram descartados pela indisponibilidade, como o *Critical Reasoning for College Readiness (CR4CR) Assessment* (CONDOR; LITSTER; PARDOS, 2021), o *Cordillera Corpus* (ZHANG; LIN; CHI, 2020), entre outros inúmeros privados. Nesse caso enquadra-se grande quantidade de artigos, com coleta de dados local e sem acesso público. As nove bases de dados da literatura acessíveis foram organizadas por formato das notas atribuídas: ordinais, discretas e contínua (MORETTIN; BUSSAB, 2010).

Em bases de dados com notas *ordinais* o método avaliativo do tutor é dado de forma textual e categórica. A representação do rótulo não estabelece escalas para o sistema,

não sendo possível mensurar a diferenças na escala *a priori*. O modelo formado deve compreender as estruturas textuais de forma simbólica, caracterizando a essência de cada nível. Portanto, o classificador deve ser robusto para aprender a relevância das respostas pela equivalência de palavras-chave. Basicamente, é fundamental para o classificador produzir um modelo com as informações essenciais para a resposta receber tal categoria e reproduzir o modelo.

Por outro lado, outra situação acontece com bases de dados avaliadas com notas contínuas. As notas *contínuas* não apresentam níveis, mas sim intervalos numéricos. As respostas recebem notas de acordo com o intervalo avaliativo. Apesar de numérico, o fato de a variável não definir uma categoria que represente a divergência entre respostas dificulta o aprendizado do modelo avaliativo. Ao sistema, isso torna subjetiva a expectativa de resposta. Assim, esse tipo de atividade é avaliado por interpolação. Nesse caso, o sistema realiza uma regressão de acordo com os pontos conhecidos, gerando a nota pela referência ao grau de similaridade para as demais respostas.

Por fim, a avaliação *discreta* numérica é a mais comum. Esse modelo favorece também os sistemas computacionais na criação da representação de resposta por categoria de nota à medida que a categoria induz a equivalência de todas as respostas associadas. Assim, o sistema consegue mensurar equivalência e divergências pelos indícios de proximidade entre respostas avaliadas já conhecidas para além da mesma categoria. O desafio do sistema com esse tipo de nota é criar um bom modelo de classificação que aprenda essa relação dupla. Para além da categoria das respostas, o sistema passa a ter que interpretar as informações fundamentais de cada classe e a escala de divergência para as demais categorias. Na Tabela 5 cada *dataset* é apresentado em detalhes, incluindo o número de questões, o total de respostas, o modelo avaliativo aplicado e a linguagem.

Tabela 5 – Bases de dados e suas principais características.

Dataset	Questões	Respostas	Modelo Avaliativo	Linguagem
SEMEVAL2013 Beetle	47	4380	ordinal	Inglês
SEMEVAL2013 SciEntsBank	143	5509	ordinal	Inglês
Kaggle ASAP-SAS	10	17043	discreto	Inglês
Powergrading	10	6980	discreto	Inglês
UK Open University	20	23790	discreto	Inglês
University of North Texas	87	2610	contínuo	Inglês
Kaggle PTASAG	15	7473	discreto	Português
Projeto Feira Literária	10	700	discreto	Português
VestUFES	5	460	contínuo	Português

Na Tabela 5 são descritos os nove *datasets* utilizados nos experimentos deste capítulo. Por meio das características apresentadas, é possível identificar uma regularidade interna de cada *dataset*. Entretanto, entre os *datasets* existe uma variação muito grande com questões de 30 até mais de 1800 respostas. É importante destacar ainda que, a definição

de respostas curtas e as características textuais das respostas de cada um desses *datasets* foi descrita no Capítulo , em especial na Tabela 2. No total, esse *corpora* contém 347 questões e 68.945 respostas. Cada base de dados e sua descrição completa são apresentadas a seguir:

4.1.1 *Dataset Beetle* do *SEMEVAL'2013 : Task 7 (Inglês)*

Beetle (DZIKOVSKA; NIELSEN; BREW, 2012) é um dos *datasets* utilizados durante o *International Workshop on Semantic Evaluation - SEMEVAL'2013*. O *SEMEVAL* seleciona anualmente uma série de desafios em análise semântica e apresenta no formato de competição. O *corpus Beetle* foi selecionado para a *Task 7: The Joint Student Response Analysis and 8th Recognizing Textual Entailment Challenge* (DZIKOVSKA et al., 2013). Portanto, a competição consistia em duas propostas. A primeira é a análise e avaliação das respostas obtidas e a segunda o reconhecimento da relação textual entre as respostas coletadas e a expectativa de resposta do professor.

Esse *dataset* consiste em uma coleção de interações entre estudantes e o sistema *Beetle II*. Beetle II é um Sistema Tutor Inteligente (STI) para aprendizado de conhecimentos básicos em Eletricidade e Eletrônica do Ensino Médio. Os alunos foram acompanhados durante três a cinco horas para preparar materiais, construir e observar circuitos no simulador e interagir com o STI. Esse sistema apresenta as questões aos alunos, avalia as respostas e envia *feedbacks* via *chat*. Na construção desse *dataset* foram acompanhados 73 estudantes voluntários da *Southeastern University* dos Estados Unidos.

Foram aplicadas questões categorizadas em dois tipos: factuais e explicativas. As questões factuais requerem que o aluno nomeie diretamente determinados objetos ou propriedades. Enquanto isso, as questões explicativas demandam que o aluno desenvolva a resposta em uma ou duas frases. Para a formação do *dataset* foram adicionadas apenas as atividades do segundo tipo, pois representam maior complexidade para sistemas computacionais. No total foram selecionadas 47 questões com 4380 respostas. A avaliação foi feita conforme o domínio demonstrado sobre o assunto em cinco categorias: *correct*, *partially-correct-incomplete*, *contradictory*, *irrelevant* e *non-domain*. Durante a anotação o coeficiente *Kappa* obtido foi de 69% de concordância.

4.1.2 *Dataset SciEntsBank* do *SEMEVAL'2013 : Task 7 (Inglês)*

O *corpus Science Entailments Bank (SciEntsBank)* (DZIKOVSKA; NIELSEN; BREW, 2012) é um dos *datasets* utilizados durante o *International Workshop on Semantic Evaluation (SEMEVAL'2013)* (DZIKOVSKA et al., 2013), com foco na avaliação de sistemas conforme a sua capacidade de análise e exploração semântica da linguagem. É uma base de dados formada por questões da disciplina de ciências. Na avaliação 16 assuntos

distintos são abordados entre ciências físicas, ciências da terra, ciências da vida, ciências do espaço, pensamento científico e tecnologia.

As questões são parte da *Berkeley Lawrence Hall of Science Assessing Science Knowledge (ASK)* com avaliações padronizadas de acordo com o material de apoio *Full Option Science System (FOSS)*. Participaram estudantes dos Estados Unidos de terceira a sexta série, coletando em torno de 16 mil respostas. Porém, entre as questões de preenchimento, objetivas e discursivas, foram utilizadas apenas as discursivas, que requisitavam explicações dos alunos segundo o tema. As respostas foram graduadas em cinco notas ordinais: *correct*, *partially-correct-incomplete*, *contradictory*, *irrelevant* e *non-domain*. O *SciEntsBank*, então, consiste em um conjunto com 143 questões selecionadas e 5509 respostas. No processo de avaliação foi observado o coeficiente *Kappa* com 72,8% de concordância.

4.1.3 *Dataset* do Concurso ASAP-SAS no *Kaggle (Inglês)*

A origem da base de dados *ASAP - SAS*, *Automated Student Assessment Prize - Short Answer Scoring* é uma competição proposta pela *Hewllet Foundation* na plataforma *Kaggle*[1]. A competição consistiu em três fases:

- Fase 1: Demonstração em respostas longas (redações);
- Fase 2: Demonstração em respostas curtas (discursivas);
- Fase 3: Demonstração simbólica matemática/lógica (gráficos e diagramas).

Seu objetivo era descobrir novos sistemas de apoio ao desenvolvimento de escolas e professores. Especificamente, as três fases destacam a atividade lenta e de alto custo de avaliar manualmente testes, mesmo que com padrões bem definidos. Uma consequência disso é a redução do uso de questões discursivas nas escolas, dando preferência para as questões objetivas para evitar a sobrecarga de trabalho. Isso evidencia uma gradativa redução da capacidade dos professores em incentivar o pensamento crítico e as habilidades de escrita. Portanto, os sistemas de apoio, são uma possível solução para suportar os métodos de correção, avaliação e feedback ao conteúdo textual dos alunos.

Nesse contexto, a competição apresentou dez questões multidisciplinares, de ciências a artes. Estão distribuídas 17.043 respostas de alunos entre essas atividades. Para chegar a essa quantidade, foram selecionadas por volta de 1.700 respostas entre 3.000 respostas em cada atividade. Cada resposta tem aproximadamente 50 palavras. A primeira avaliação foi dada pelo primeiro especialista como nota final e a segunda nota foi atribuída apenas para demonstrar o nível de confiança da primeira nota. A avaliação apresentada por dois especialistas demonstrou concordância de 90% no coeficiente *Kappa*.

[1] The Hewlett Foundation - Short Answer Scoring: https://www.kaggle.com/c/asap-sas

4.1.4 *Dataset Powergrading (Inglês)*

Elaborado com o *United States Citizenship Exam* (USCIS) em 2012, a base de dados *Powergrading* contém dez questões e 6980 respostas (BASU; JACOBS; VANDERWENDE, 2013). Desenvolvida originalmente para destacar a possibilidade de avaliação massiva, o *dataset* selecionou 698 respostas para cada uma das questões. As respostas foram geradas com *Amazon Mechanical Turk*, serviço remoto de análise manual de conteúdo para anotação da *Amazon*. Foi coletada por um grupo de pesquisa da *Microsoft*[2] e cada questão acompanha um modelo de resposta e as respostas recebidas para cada questão. Além disso, acompanha o *dataset* outras dez questões não avaliadas.

Foram requisitadas respostas com poucas palavras, atingindo no máximo uma ou duas sentenças. Por conta disso, os resultados são respostas muito curtas com quatro palavras em média. Em geral, por conta da convergência, vários padrões de resposta se repetem (RIORDAN et al., 2017). Com avaliações binárias, 1 para resposta correta e 0 para incorreta, cada resposta apresenta avaliações de três diferentes tutores. Apesar de alguns trabalhos assumirem um dos avaliadores como padrão, foi utilizado como modelo de avaliação a resultante da comparação entre os três. Apesar de não ter complexidade linguística, ocorreu contradição entre os avaliadores em 470 respostas. Em valores percentuais, isso representa 7% do total de respostas do conjunto de dados.

4.1.5 *Dataset* da *UK Open University (Inglês)*

A base de dados da *UK Open University* é um conjunto de questões coletadas na disciplina de introdução a ciências, denominada *S103 - Discovering Science* (JORDAN, 2012). O foco do conjunto de atividades são abordagens em questões factuais bem concisas, não excedendo 20 palavras. Os alunos receberam as atividades via ambiente virtual da *Intelligent Assessment Technologies (IAT)*, o *FreeText Author*. O *FreeText Author* foi utilizado como um método de *CAA* de modo interativo e com resultado automático analisando a resposta do aluno segundo os padrões de resposta conhecidos. O sistema permitiu uma sequência de envios e apresentava comentários da resposta como *feedback* para os alunos. Dependendo da complexidade da resposta, o tempo de retorno dos resultados varia muito entre alguns poucos minutos até mais do que um dia.

O *dataset* é de acesso privado, mas foi disponibilizado pelos autores para este estudo. Entre as suas 20 questões, esse *dataset* apresenta diferentes quantidades de respostas entre 511 e 1897. A avaliação é discreta e binária, definindo cada resposta como correta ou incorreta. Não existem notas intermediárias, representando diretamente se o aluno atendeu ou não os requisitos da resposta.

[2] Powergrading: https://www.microsoft.com/en-us/download/details.aspx?id=52397

4.1.6 *Dataset* da *University of North Texas (Inglês)*

O *dataset* da *University of North Texas - UNT* (MOHLER; BUNESCU; MIHALCEA, 2011), conhecido como *Texas dataset*[3], é uma coleção de questões discursivas extraída no curso de Ciência da Computação. Composto por 80 atividades únicas, esse conjunto contempla dez listas de exercícios com até sete questões e dois testes com dez questões cada. Foram aplicados em um ambiente virtual de aprendizagem durante a disciplina de Estrutura de Dados para 31 alunos. No total o *dataset* é composto por 2.273 respostas de alunos entre as 80 atividades.

A avaliação foi feita com cinco notas discretas, de 5 equivalente a resposta perfeita até 0 completamente incorreta. Foram avaliadas por dois avaliadores independentes, estudantes do curso de Ciência da Computação. Para os autores, o modelo seguido pelo sistema deve ser a resultante da média ente os avaliadores, em intervalo contínuo. Entre as notas atribuídas, 57,7% das respostas receberam a mesma pontuação. Enquanto isso, um nível de diferença entre as notas representou 22,9% do total de respostas. Foi constatado também que, dentre as diferenças na avaliação, o avaliador 1 atribuía notas maiores notas 76% das vezes.

4.1.7 *Dataset PTASAG* no *Kaggle (Português)*

A *PTASAG - Portuguese Automatic Short Answer Grading Data* é uma base de dados brasileira apresentada por (GALHARDI et al., 2018) e disponibilizada na plataforma *Kaggle*[4]. Foi coletada pela Universidade Federal do Pampa - Unipampa em conjunto com cinco professores de biologia do Ensino Fundamental. Foram criadas 15 atividades com base no sistema Auto-Avaliador CIR. Em biologia, os tópicos abordados foram sobre o corpo humano. Cada questão acompanha uma lista de conceitos, as respostas avaliadas e as respostas candidatas criadas pelos professores como referência. Foram criadas entre duas e quatro respostas candidatas contendo entre três e seis palavras-chave.

As atividades foram aplicadas no Ensino Fundamental para 326 estudantes de 12 a 14 anos do 8º e do 9º ano. Também foram aplicadas a 333 alunos do Ensino Médio de 14 a 17 anos. As respostas foram avaliadas por 14 estudantes de uma turma do último ano, considerando uma escala de notas de 0 a 3:

- Nota 0: Majoritariamente incorreta, fora de tópico ou sem sentido;
- Nota 1: Incorreta ou incompleta mas com trechos corretos;
- Nota 2: Correta mas com importantes trechos faltantes;

[3] Texas Dataset: https://web.eecs.umich.edu/~mihalcea/downloads.html
[4] PT-ASAG-2018: https://www.kaggle.com/lucasbgalhardi/pt-asag-2018

- Nota 3: Majoritariamente correta apresentando os principais pontos.

No total, participaram 659 estudantes com um total de 7.473 respostas. Cada uma das 15 questões apresenta entre 348 e 615 respostas. Apenas quatro questões foram examinadas por mais de um avaliador para verificar a concordância entre eles. O coefficiente *Kappa* observado foi de, em média, 53.25%.

4.1.8 *Dataset* do Projeto Feira Literária das Ciências Exatas *(Português)*

É um conjunto de dados coletados durante o Projeto Feira Literária das Ciências Exatas (NASCIMENTO; KAUARK; MOURA, 2020). As questões foram obtidas durante uma Atividade Experimental Problematizada por meio de um livro paradidático, ou seja, cujo objetivo primário não é o apoio didático. O livro escolhido foi *A Fórmula Secreta* de David Shephard.

Conforme o livro, os professores formularam dez atividades e ministraram para 70 alunos do 5º ano de Ensino Fundamental. Essas atividades ministradas descreviam situações práticas de química básica. No total, o conjunto de dados conta com dez questões, 700 respostas e suas respectivas avaliações.

4.1.9 *Dataset* do Vestibular UFES *(Português)*

A base de dados VestUFES (PISSINATI, 2014) é uma amostra das questões discursivas de Português do vestibular da UFES em 2012. A amostra selecionada contém 460 respostas divididas igualmente entre as cinco questões de língua portuguesa, também referentes a respostas dadas por 92 diferentes alunos.

Cada resposta foi avaliada por dois avaliadores. De acordo com o vestibular da universidade, os avaliadores atribuíram notas entre 0 e 2 pontos em cada questão, totalizando 10 na soma da prova. Caso houvesse divergências de mais de 1 ponto entre as correções um terceiro avaliador era acionado para reavaliar a coerência das notas. A nota das respostas do *dataset* foram redimensionadas pelo autor para o intervalo de 0 a 10 pontos. Na nova escala, as diferenças observadas entre os avaliadores foi de, em média, 1,38 ponto com desvio padrão de 1,75.

4.2 Experimentos

Os experimentos realizados têm como intuito demonstrar a qualidade do método proposto de *Active Learning*, com otimização do esforço de anotação humano. Assim, é esperada uma boa forma na amostragem tal qual o ganho de desempenho nos resultados de classificação. Com os *datasets* conhecidos da literatura, é possível comparar nossa

proposta de análise das estruturas textuais em relação aos principais trabalhos publicados. Inicialmente, os experimentos mostram o desempenho da etapa de *clusterização* com a caracterização dos *clusters* formados. Essa etapa indica a qualidade da amostragem em relação ao objetivo da atribuição de notas. Na sequência, apresentam-se comparativos com os classificadores da literatura, tentando maximizar os resultados obtidos com qualidade compatível aos demais trabalhos.

4.2.1 Resultados de *Clusterização*

Nessa primeira análise, observa-se a qualidade dos *clusters* formados. Assim, é investigado se a forma utilizada para construção dos *clusters* foi efetiva para formação das amostras. Nessa etapa é essencial que a amostragem colete toda a diversidade de notas, adquirindo conhecimento que torna possível comparar os níveis de nota. Neste identificam-se as regiões que estabelecem diferenças entre contextos, com identificação de equivalências e divergências, isolando *outliers*. Assim, essa primeira etapa é definitiva para a qualidade do aprendizado via *Active Learning*.

A formação dos *clusters* foi feita com base em uma otimização do *elbow method*, caracterizando a dispersão das amostras (SPALENZA; PIROVANI; OLIVEIRA, 2019). Os resultados obtidos, visam ter *clusters* mais homogêneos, segundo os índices de validação. O índice utilizado foi o CVS, avaliando o coeficiente de variação do tamanho dos *clusters* formados e evitando grande concentração. Os múltiplos contextos e a sobreposição entre as notas atribuídas são detalhes observados em grandes *clusters*. Na avaliação dos clusters, utilizam-se as métricas CA, HS e CS, descritas na Seção 3.2.1. Os resultados obtidos em cada um dos *datasets* é apresentado na Tabela 6.

Tabela 6 – Bases de dados e índices qualitativos de *clusterização*.

Dataset	CA	HS	CS
SEMEVAL2013 Beetle	0,5897	0,2895	0,1856
SEMEVAL2013 SciEntsBank SciEntsBank	0,6064	0,2357	0,1789
UK Open University	0,7628	0,2825	0,0667
Projeto Feira Literária	0,6414	0,3920	0,2574
Kaggle ASAP-SAS	0,5364	0,1065	0,0741
Powergrading	0,9032	0,6773	0,0760
Kaggle PTASAG	0,5208	0,2282	0,1418
University of North Texas	-	-	-
VestUFES	-	-	-

Na Tabela 6 é mostrado um alto índice de relação entre *clusters* e itens que receberam uma mesma nota. Isso implica em um CA médio de 65,15%, atingindo até 90,32% no *dataset Powergrading*. O CA, via voto majoritário, indica quantas amostras compartilham *clusters* com diferentes avaliações. Pela atribuição de notas contínuas, os *datasets University of North Texas* e *VestUFES* foram desconsiderados nessa análise.

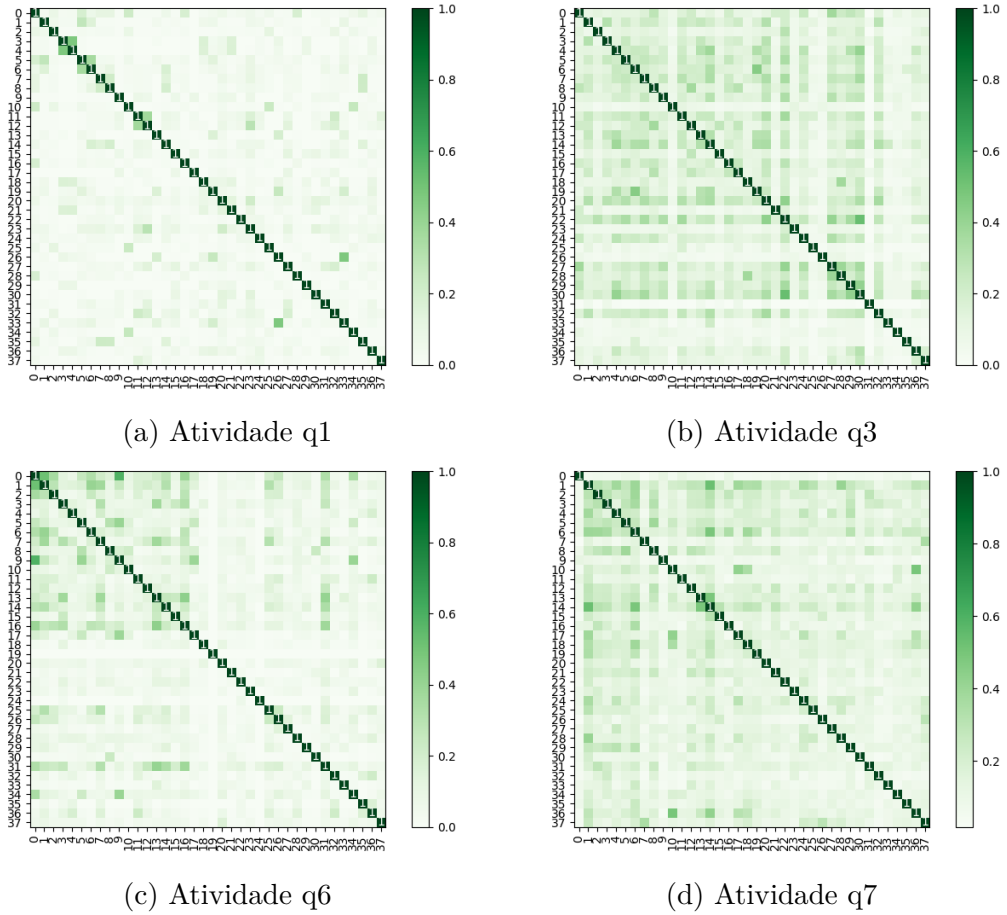
É um dos objetivos da etapa de *clusterização* que os grupos formados representem os tópicos abordados na questão. Assim, com efeito descritivo sobre os vínculos entre os *cluster* e as notas, o CA indica o nível de complexidade para o reconhecimento de padrões e avaliação. No entanto, os índices HS e CS apontam a conexão entre cada classe de nota com os *clusters* formados. Esses dois índices refletem algumas características dos problemas textuais.

O primeiro indica o nível de homogeneidade, ou seja, o percentual de *clusters* formados por uma mesma nota. Estabelecer essa separabilidade entre classes e *clusters* é uma tarefa muito complexa. Em especial na análise textual, é comum que as avaliações distintas ocorram em respostas com algum nível de sobreposição de termos. Ou seja, notas distintas podem apresentar algum nível de similaridade e formar um mesmo *cluster*. Já o segundo indica a equivalência entre os *clusters* formados e cada uma das instâncias de nota. No entanto, dada a liberdade textual, dificilmente as respostas que recebem uma mesma nota apresentam os mesmos termos como referência.

Desse modo, *datasets* com classificação binária tendem a ter alto desempenho na *clusterização*, como o *Powergrading*. Porém, com muitos níveis de nota e textos maiores, *datasets* como *ASAP-SAS* e *PTASAG*, tornam-se muito complexos para essa primeira etapa. O desempenho antes da etapa de amostragem mostra a eficiência dos classificadores para levar *clusters* com divergências para uma melhoria considerável de desempenho. O desempenho das classificações após a amostragem é apresentado em detalhes na Seção 4.2.2. Para mostrar como os *clusters* são formados por diferentes contextos, analisa-se também a similaridade entre centroides, indicando nessa etapa de *Active Learning* que a *clusterização* regionaliza as respostas segundo a interpretação dos contextos. Na Figura 11 é exposto o grau de similaridade encontrado nos *clusters* formados para o *dataset Powergrading*.

Como pode ser observado na Figura 11, os grupos de respostas formados são muito consistentes, apresentando baixa similaridade para os demais *clusters*. Há nesses casos uma divergência do item médio (*centroide*) do *cluster* em relação aos demais grupos. São denominados agrupamentos consistentes aqueles que todas as respostas têm um mesmo alinhamento, recebendo posteriormente uma mesma classe de nota. A atividade *q1*, por exemplo, apresenta similaridade média de apenas 0,0355. Mesmo a que apresenta *clusters* mais próximos, a atividade *q7* apresenta em média 0,1387 de similaridade. As outras duas em destaque, atividades *q3* e *q6*, atingem 0,1207 e 0,0730. Assim, a etapa de *clustering* têm potencial de formar zonas de equivalência e divergência com alta qualidade para as etapas de atribuição de notas. Essa alta qualidade indica *clusters* bem separados de acordo com o contexto. A Figura 12 apresenta um caso oposto com a diferença entre os *centroides* dos *clusters* para a atividade *SciEntsBank*.

Como ilustrado na Figura 12, há um resultado mais complexo de *clusterização*,



(a) Atividade q1

(b) Atividade q3

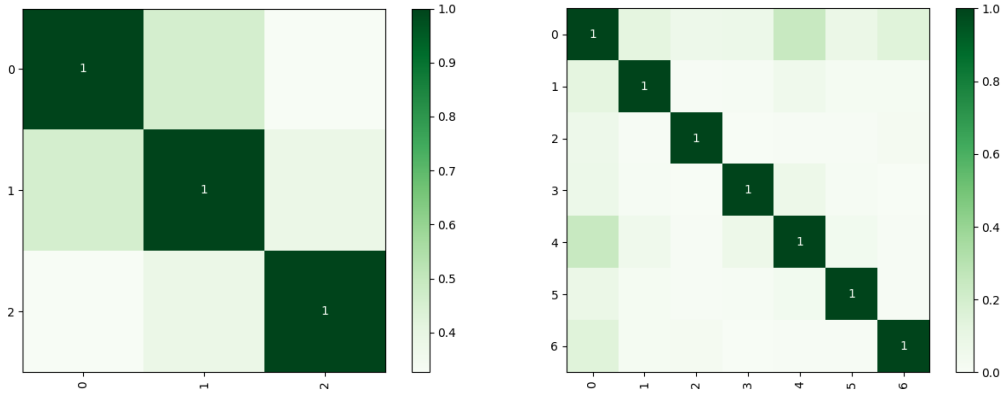
(c) Atividade q6

(d) Atividade q7

Figura 11 – Similaridade entre *centroides* para as atividades *q1*, *q3* e *q6* e *q7* em *Power-grading*.

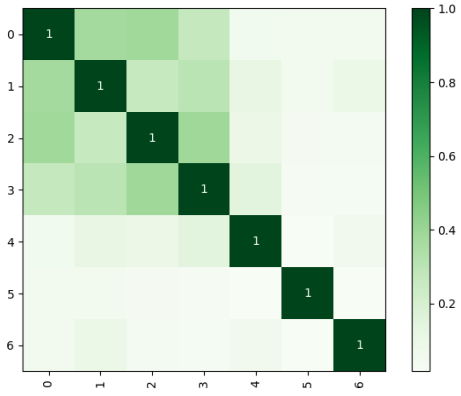
em especial, com poucas amostras. Diferentemente do que foi notado na questão anterior, existem situações distintas dentro de um mesmo *dataset*. A primeira atividade *EM16b* apresenta poucos *clusters* e similaridade média de 0,3880. A segunda atividade *EM21b* mostra uma boa separabilidade entre os dados com média de 0,0476 e máxima de 0,2442. A terceira atividade *EM27b* mostra uma concentração entre alguns *clusters*, com um núcleo de *centroides* similares e outra parte bem distante, com similaridade média de 0,1361 e máxima entre *centroides* de 0,3256. Por fim, a quarta atividade *MX16a* mostra alguns *clusters* muito próximos atingindo similaridade máxima de 0,6125 sendo a média entre *centroides* de 0,1974.

Pelo índice de similaridade parcial ou total dos clusters formados pode-se considerar que algumas atividades mostram diferentes perspectivas de uma mesma classe ou mesclando diferentes classes. Assim, a identificação de padrões entre esses grupos de alta similaridade depende de um bom reconhecimento de padrões *a posteriori*. Assim, é compreensível que os níveis de CA desses *clusters* sejam baixos. No entanto, é importante obter ganhos na etapa de classificação, elevando o nível dos resultados alcançados.

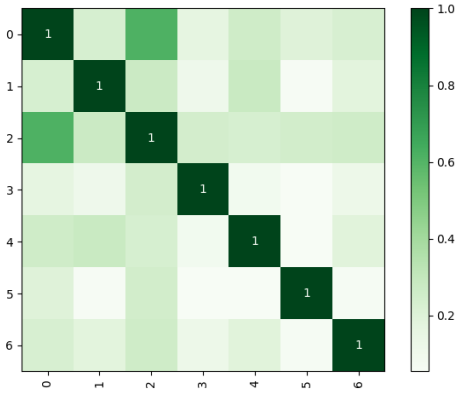


(a) Atividade EM16b

(b) Atividade EM21b



(c) Atividade EM27b



(d) Atividade MX16a

Figura 12 – Similaridade entre *centroides* para as atividades *EM16b*, *EM21b*, *EM27b* e *MX16a* em *SciEntsBank*.

4.2.2 Resultados de *Classificação*

Após os resultados obtidos na etapa de *clusterização*, é realizada a análise de classificação. Nesta, usam-se os classificadores tradicionais para avaliar o ganho em desempenho proporcionado pela etapa de *Active Learning*. Tais experimentos, diferentemente do descrito para estudo dos *clusters* formados, caracterizam-se pela simetria com os artigos da literatura e os desafios da área. Como os *datasets*, geralmente vêm particionados, foram agrupados os conjuntos de treino e teste para que a nosso método *não-supervisionado* realize os particionamentos, preservando os percentuais.

Neste estudo, cada atividade foi processada isoladamente, distinguindo assim de alguns estudos que utilizam todas as atividades durante o treinamento. Essa técnica é comum em abordagens que precisam de maior quantidade de amostras para aprendizado. No entanto, proporcionalmente, ainda é necessário garantir a coesão do *dataset* com o tema. Nesse caso, com o volume de dados ainda reduzido, os sistemas tornam-se específicos para um contexto. Além disso, em caráter multidisciplinar, a tendência é o sistema receber blocos distintos de atividades, cada qual em um contexto.

Outro fator comum encontrado na literatura é a adoção de métricas distintas na

avaliação dos modelos. Os principais exemplos incluem métricas diretamente relacionadas com os testes e o grupo de amostras. Assim, são comuns os casos que a avaliação não leve em conta os problemas de desbalanceamento e a paridade entre experimentos. Em boa parte dos casos dos SAGs é reportado apenas o coeficiente *Kappa* ou correlação de *Pearson*. O *Kappa* é considerada uma métrica robusta, que representa detalhes por categoria e a relação de avaliação (**??**). Entretanto, individualmente, a escolha de uma métrica como essa indica apenas uma visão unilateral do problema (**??**). Uma consequência da aplicação de *Active Learning* é o aprendizado gradual dos níveis de nota, conforme as iterações avaliativas e o ganho de informação. Durante cada ciclo o modelo busca se adaptar da melhor forma à diversidade de notas. De fato, com a integração humano-computador no processo avaliativo, é importante atentar-se na prevalência dos resultados de classificação, intra-classe e inter-classes de nota.

Adicionalmente, ressalta-se o modelo de amostragem observado nos trabalhos da literatura. O percentual de amostragem foi aplicado de acordo com o conjunto de dados e os demais trabalhos, variando entre 70% e 90% das respostas. O alto percentual é comum pois é possível reaproveitar o treinamento para aplicar em novas amostras. Após o primeiro treinamento, apenas algumas intervenções do professor são necessárias para correção. Desta forma, gradativamente o esforço do professor pode ser minimizado. Nesses casos, principalmente com o *p*Nota, com novos grupos de dados para cada atividade tornam-se necessárias apenas eventuais avaliações assimilar novos tópicos.

Seguindo a característica da avaliação, serão apresentados os resultados obtidos segundo a forma avaliativa adotada pelos professores. Os conjuntos de dados *Beetle* e *SciEntsBank* podem ser considerados mais complexos pela baixa quantidade de amostras para um modelo de avaliação de cinco categorias textuais (ordinais). Nestes *datasets* do evento *SEMEVAL' 2013*, foram apresentados três níveis de desafios. O primeiro nível é a avaliação de respostas não conhecidas, selecionadas aleatoriamente no conjunto de respostas (*Unseen Answers*). O segundo nível compreende a correção de respostas em questões desconhecidas, ainda em um determinado domínio (*Unseen Questions*). E, por fim, o terceiro nível está relacionado à análise de respostas em um domínio desconhecido (*Unseen Domain*). Assim como a maioria dos sistemas SAG, o desafio que se enquadra no tópico aqui abordado é apenas o primeiro (*Unseen Answers*), avaliando conjuntos de respostas dentro do tópico.

Sendo a principal característica deste *dataset* o desbalanceamento das classes (DZIKOVSKA et al., 2013), ambos os *datasets* foram anotados em níveis de nota: *correct*, *partially-correct-incomplete*, *contradictory*, *irrelevant* e *non-domain*. Evidencia-se pela complexidade, inclusive semântica, de separar as três categorias inferiores, *contradictory*, *irrelevant* e *non-domain*. Utilizando os seis classificadores descritos na Seção 3.3.1, são apresentados os resultados obtidos na Tabela 7.

Tabela 7 – Resultados dos seis classificadores testados nos *datasets* do *SEMEVAL' 2013*.

Beetle					(5 Categorias)
	Métricas				
	ACC	PRE	REC	F1(m)	F1(w)
DTR	61,90%	41,14%	43,08%	40,82%	60,01%
GBC	**62,32%**	**41,63%**	**43,66%**	**41,25%**	**60,06%**
KNN	59,80%	35,98%	39,21%	36,38%	56,26%
RDF	60,67%	39,21%	40,65%	38,67%	58,35%
SVM	60,06%	36,86%	42,56%	38,10%	54,70%
WSD	60,76%	37,28%	40,40%	37,59%	56,95%

SciEntsBank					(5 Categorias)
	Métricas				
	ACC	PRE	REC	F1(m)	F1(w)
DTR	48,79%	38,94%	39,30%	38,01%	39,30%
GBC	**50,62%**	**40,43%**	42,50%	**39,93%**	**49,04%**
KNN	43,16%	34,88%	36,13%	33,89%	41,86%
RDF	49,49%	37,84%	**43,25%**	38,96%	45,67%
SVM	46,51%	32,89%	41,01%	35,31%	40,45%
WSD	47,43%	36,45%	40,21%	36,97%	44,43%

Na Tabela 7 é mostrado o desempenho do sistema com os seis classificadores testados. Fica evidente que a performance do sistema é superior no *Beetle* em relação ao *SciEntsBank*. No *dataset Beetle* o melhor resultado obtido foi com o GBC, com ACC médio de 62,32% e F1-ponderado de 60,06%. No entanto, a diferença foi bem pequena em relação aos demais classificadores. No *dataset SciEntsBank* o resultado foi diferente, com o GBC apresentando resultados bem superiores em relação a alguns classificadores, com ACC médio de 50,62% e F1-ponderado de 49,04%. O KNN apresentou baixo desempenho, com ACC de apenas 43,16%, devido aos vários níveis de nota entre as poucas respostas. Foram comparados os resultados obtidos no *dataset Beetle* na Figura 13.

Na Figura 13, é mostrado o desempenho do *p*Nota e dos trabalhos da literatura em ACC, F1-macro e F1-ponderado. O melhor resultado é apresentado por (SAHU; BHOWMICK, 2020), com uma estratégia que realiza a combinação de vários níveis da estrutura textual, atingindo ACC de 66,6% e F1-ponderado de 70,91%. Neste, os autores incluem *features* de similaridade semântica, de sobreposição léxica, de recuperação da informação, de similaridade de tópicos, similaridade entre *feedbacks* e de alinhamento textual. Portanto, temos uma ampla aquisição de informação, combinando os métodos tradicionais *Latent Semantic Analysis* (LSA), *word embeddings*, *Recall Oriented Understudy for Gisting Evaluation* (ROUGE), *TF-IDF*, *LDA*, dentre outros. A avaliação é dada posteriormente com *ensembles* usando *Random-Forest*.

No estudo realizado por (GALHARDI et al., 2018) encontra-se também uma

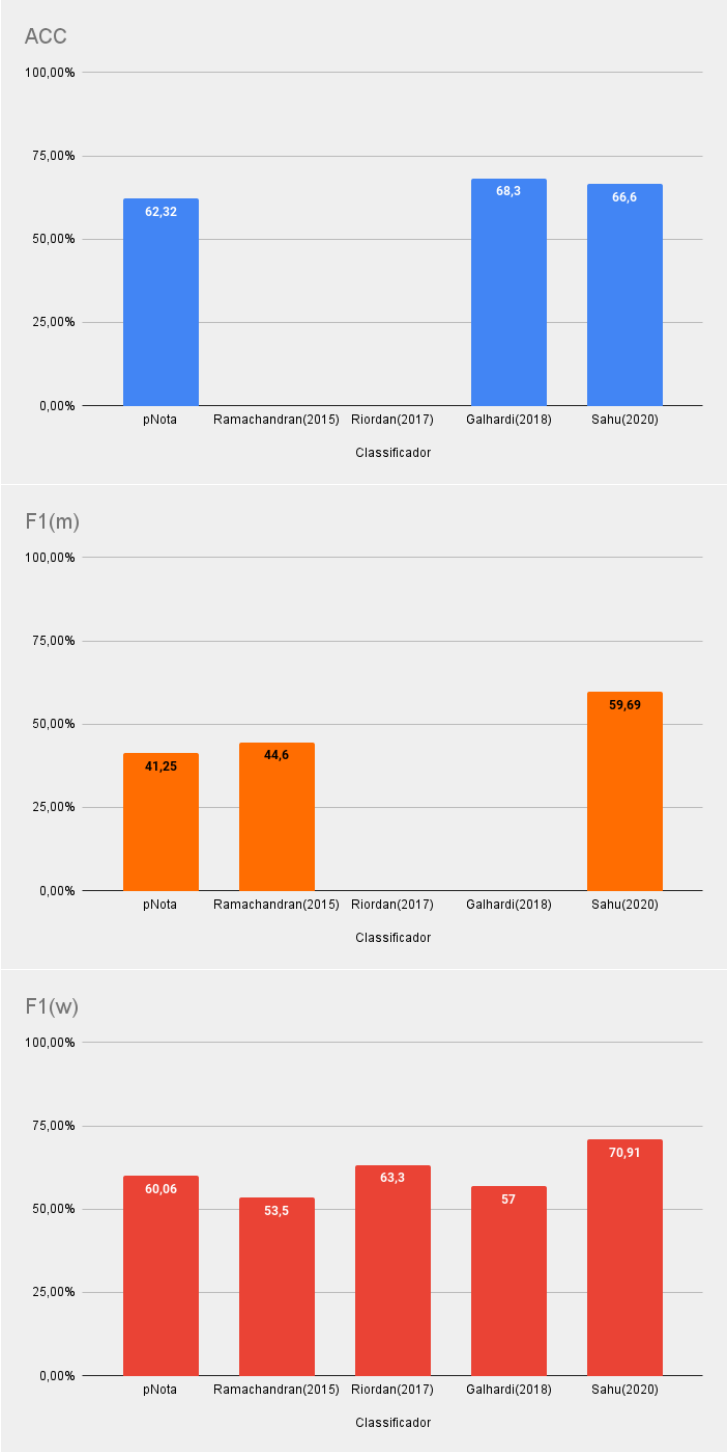


Figura 13 – Resultados obtidos no *dataset Beetle*.

combinação de *features* que incluem estatísticas textuais, como o cálculo de erros, tamanhos de resposta e contagem de palavras por sentença. Essas estruturas eram comuns nos primeiros trabalhos da área, hoje avançando com análise da linguagem (BURROWS; GUREVYCH; STEIN, 2015). Usando *Random Forests* e *Extreme Gradient Boosting* os autores alcançaram ACC de 68,3% e F1-ponderado de 57%. Um terceiro destaque vale para o trabalho de (RIORDAN et al., 2017), com F1-ponderado de 63,3%. Nesse caso,

os autores não divulgaram as demais métricas. O resultado foi alcançado combinando *word-embeddings*, *Convolutional Neural Networks* (CNN) e *Long Short-Term Memory* (LSTM). As notas são dadas no *layer* de agregação por um modelo linear. Outra técnica aplicada é a sumarização, avaliando as respostas ao mensurar a similaridade dos grafos textuais (RAMACHANDRAN; FOLTZ, 2015; RAMACHANDRAN; CHENG; FOLTZ, 2015). Essa estratégia ainda depende de combinar vários trechos da atividade, desde o enunciado até o quadro de *rubrics*. De forma geral, os resultados obtidos no *Beetle* são bem próximos entre eles. O desempenho intermediário indica o desafio de compreender as cinco categorias do *dataset*. Outra perspectiva é dada nos resultados do *SciEntsBank*, apresentados na Figura 14.

Na Figura 14 vê-se que, apesar da aplicação no mesmo desafio, temos vários trabalhos que foram aplicados apenas nesse segundo. O melhor resultado é apresentado por (SAHU; BHOWMICK, 2020), com F1-ponderado de 92,5%. Porém, esse desempenho muito superior é incomum, pois não apresenta o mesmo resultado no *Beetle*. Os autores relatam como maior ganho de desempenho as técnicas de similaridade semântica e sobreposição textual. Portanto, as atividades têm potencial de ter alto desempenho com técnicas de construção de regras e expressões regulares.

O estudo realizado por (ROY et al., 2016) apresentou ACC de 56,5% e F1-ponderado de 67,2%. Tal como mencionado, os autores estudaram técnicas para aquisição de padrões sequenciais, comparando com as respostas candidatas. O método em uma forma geral foi feito no nível de *tokens*, identificando sobreposição entre as respostas. A técnica tem desempenho muito superior em algumas classes, dado o alto F1-ponderado mas acompanhado de um ACC menor que outros trabalhos da literatura. Técnicas mais recentes foram propostas por (GALHARDI et al., 2018) e (SAHA et al., 2018). O primeiro alcançou resultados ACC média de 65,9% e F1-ponderado de 62,8% usando *Random Forests* e *Extreme Gradient Boosting*, enquanto o segundo aplicou seis níveis de análise dos termos, apostando no *Histogram of Partial Similarity*. O método de *partial similarity* aplica um score de similaridade entre a resposta candidata e cada uma das respostas. As palavras são comparadas pela similaridade de cosseno dos pares em *word embeddings*.

A aplicação feita com LSTMs realizada por (RIORDAN et al., 2017) alcançou F1-ponderado de 53,3%. Uma proposta parecida foi realizada por (GOMAA; FAHMY, 2019). Os autores realizam uma composição com vetores semânticos via *skip-thought*, comparando os vetores resultantes. Os vetores são dados por meio de pares, pelo produto e pela diferença entre uma resposta e a resposta candidata. O resultado alcançado é de F1-ponderado de 65,6% em média.

Pela baixa quantidade de amostras, no *SciEntsBank* os sistemas têm maior dificuldade de adquirir conhecimento por categoria de nota. Por isso, os sistemas que apresentaram melhores resultados foram os que utilizaram a sobreposição dos termos com

Figura 14 – Resultados obtidos no *dataset SciEntsBank*.

resposta candidata. No entanto, isso indica que tais estratégias seguem um único viés de resposta, tornando-se menos efetivas ao lidar com variações linguísticas (FILIGHERA; STEUER; RENSING, 2020).

A hipótese avaliada é que o diferencial encontrado por esses estudos contemplam a adição de informação e enriquecimento das respostas, dado o pequeno número de amostras por classe. Ao todo, o *SciEntsBank* contém em média apenas 37 respostas por

Tabela 8 – Resultados de classificação para o *dataset OpenUniversity*.

Open University					(2 Categorias)
	Métricas				
	ACC	PRE	REC	F1(m)	F1(w)
DTR	96,44%	93,66%	**93,26%**	93,21%	96,46%
GBC	**96,80%**	**96,39%**	92,80%	**93,64%**	**96,60%**
KNN	84,62%	82,09%	78,12%	77,50%	83,80%
RDF	93,62%	93,58%	88,75%	89,79%	93,37%
SVM	91,45%	91,04%	85,21%	85,41%	90,16%
WSD	90,44%	90,47%	84,89%	84,78%	89,56%

questão, enquanto o *Beetle* apresenta 84 respostas. Entre as 4.380 respostas do *Beetle*, 1.841 foram anotadas como *correct*, 1.160 como *contradictory* e 1.031 como *partially-correct-incomplete*. Por outro lado, apenas 218 foram avaliadas como *non-domain* e 130 como *irrelevant*. Considerando ainda a distribuição de classes, a situação é agravada em relação ao *SciEntsBank*. Entre as 5.509 respostas, 2241 foram dadas como *correct*, 1.437 como *partially-correct-incomplete*, 1.248 como *irrelevant* e 557 como *contradictory*. Só constam nesse *dataset* 26 respostas anotadas como *non-domain* entre as 143 questões. Notoriamente, são poucas amostras para algumas categorias que se destacam quando vê-se que, em média, o primeiro *dataset* apresenta 93 respostas por questão enquanto o segundo apresenta apenas 38 respostas. Portanto, apesar da complexidade de avaliar tal questão, os resultados são positivos, aprimorando resultados esperados conforme a distribuição de *clusters*.

Na sequência há os resultados obtidos com o *dataset Open University*. Nesse conjunto de dados a avaliação designada foi binária, com notas 0 ou 1. Assim, a avaliação foi dada apenas como corretas ou incorretas. Bem distinto dos *datasets Beetle* e *SciEntsBank*, esse conjunto contém mais de 23 mil respostas e, em média, 1190 respostas para cada questão. Isso impacta diretamente a construção de modelos de resposta, com uma variedade de padrões para uma mesma classe, sendo possível a identificação de núcleos de resposta bem consistentes segundo a simetria da classe. Os resultados apresentados na Tabela 8 refletem justamente esse aspecto.

É evidente, por meio dos resultados obtidos e apresentados na Tabela 8, que a grande quantidade de amostras melhorou consideravelmente o desempenho do sistema. Sendo assim, o *p*Nota atingiu média de 96,8% de ACC e 96,6% de F1-ponderado. A simplificação dos padrões de nota para um esquema de avaliação binário é um fator que evidencia as diferenças, com detalhes que contribuem positivamente e negativamente com cada classe. Na Figura 15 é apresentado um comparativo do desempenho do *p*Nota em relação à publicação dos autores.

Na Figura 15 é demonstrada a alta qualidade da classificação automática em relação

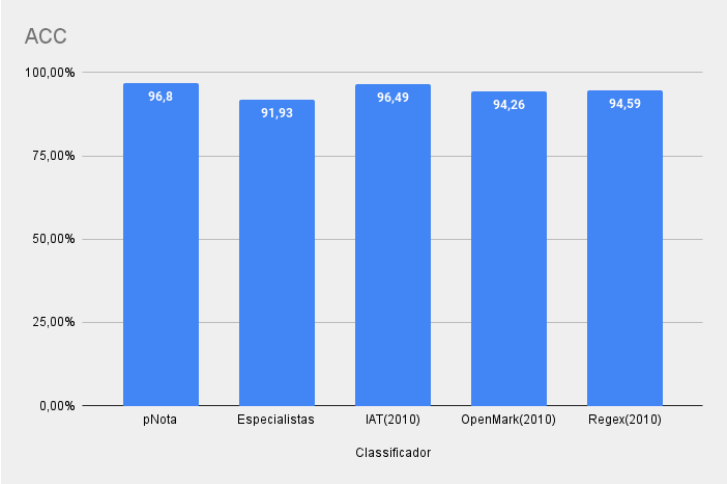


Figura 15 – Resultados obtidos no *dataset* da *Open University*.

Tabela 9 – Resultados de classificação para o *dataset Powergrading*.

Powergrading					(2 Categorias)
	Métricas				
	ACC	PRE	REC	F1(m)	F1(w)
DTR	99,79%	97,96%	**98,68%**	98,28%	99,80%
GBC	99,71%	97,46%	98,64%	97,93%	99,74%
KNN	99,79%	97,96%	**98,68%**	98,28%	99,80%
RDF	99,79%	97,96%	**98,68%**	98,28%	99,80%
SVM	**99,86%**	**99,93%**	97,50%	**98,30%**	**99,83%**
WSD	99,79%	97,96%	**98,68%**	98,28%	99,80%

ao que foi reportado entre especialistas. Os especialistas atingiram entre eles ACC média de 91,33%. Os autores (BUTCHER; JORDAN, 2010) reportaram ACC de 96,49%. Porém, os três sistemas aplicados utilizam regras e expressões regulares, o que demanda de maior esforço para início das correções. Assim, o sistema deve aguardar enquanto o professor elabora as regras de correção. Como o estudo destaca, o IAT usa o conhecimento sobre o conteúdo e as regras de associação de respostas para criação de *feedbacks* direcionados ao tema.

Outro *dataset* similar é o *Powergrading*. Esse *dataset* também conta com muitas amostras e classificação binária. Foi criado especificamente para estudos que aplicam técnicas de *clusterização* na atribuição de notas (BASU; JACOBS; VANDERWENDE, 2013). Portanto, uma hipótese é que neste caso há maior separabilidade entre as classes e baixa taxa de sobreposição, algo incomum para respostas discursivas. Os resultados obtidos pelo *p*Nota são apresentados na Tabela 9.

Como é apontado na Tabela 9, alcançamos um bom desempenho com o *p*Nota. Neste, todos os seis classificadores apresentam poucos erros, atingindo ACC de 99,86% e F1-ponderado de 99,83%. Nesse caso, o alto desempenho do SVM reforça a hipótese de baixa sobreposição de *features* entre as classes. Os resultados em relação aos demais

trabalhos da literatura são apresentados na Figura 16.

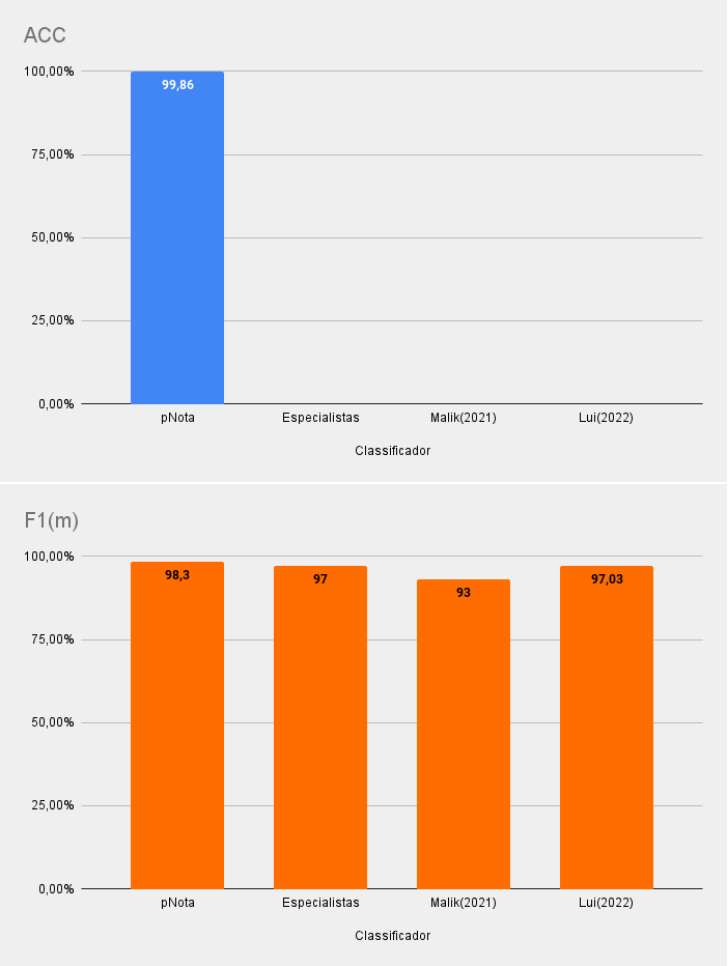


Figura 16 – Resultados dos classificadores com dados do *dataset Powergrading*.

Na Figura 16 é caracterizada pela alta qualidade obtida na avaliação entre humanos, mas replicada também pelos sistemas. Entre especialistas o F1-macro observado foi de 97%. O estudo foi realizado por (MALIK et al., 2021), com foco na formação de *feedbacks* explicativos. Os autores utilizaram DL para elaborar a *Neural Approximate Parsing with Generative Grading* (GG-NAP). Essa técnica visa decompor as respostas para identificar os trechos que compõem uma resposta correta. Entretanto o método não foi feito apenas para respostas discursivas, mas também para atividades com linguagens de programação e programação em blocos. O desempenho como avaliador alcançado é de 93% de F1-macro.

Outra proposta foi apontada por (LUI; NG; CHEUNG, 2022), usando *clusterização* e análise de divergências entre as amostras, atingindo F1-macro médio de 97,03%. O método de *clusterização* aplicado é um *multi-objective evolutionary clustering*, trabalhando os agrupamentos como uma população, buscando o refinamento para aproximar de uma resposta otimizada por *cluster*. A ideia é aumentar a qualidade dos *clusters* segundo suas menções, elegendo sua representação. Portanto, esse conjunto de dados reflete como a etapa de *clusterização* é efetiva para coleta de informações, recuperando a resposta ideal.

Entretanto, por mais que ocorram as subpartições entre *clusters* em ambas as propostas, há preocupação com as características que não encontramos no *dataset Powergrading*, como altos índices de *outliers* e a *subjetividade* na avaliação.

Seguindo a mesma plataforma avaliativa, o *PTASAG* foi descrito no trabalho de (GALHARDI; SOUZA; BRANCHER, 2020). Neste, os autores investigam o impacto das *features* na formação de um bom avaliador. Os níveis estudados foram representação em *n-grams*, representação em *word embeddings*, similaridade léxica, similaridade em *word embeddings*, similaridade em *WordNet* e estatísticas textuais, apresentando nessa ordem maior qualidade. Nessa linha, foram estudadas técnicas primárias (como o tamanho das respostas) até as técnicas consolidadas de *word embeddings* (Word2Vec, GloVe e FastText). O estudo, então, sugere que o maior ganho na avaliação desse *dataset* foi a análise das sequências textuais. O desempenho apresentado pelo *p*Nota é mostrado na Tabela 10.

Tabela 10 – Resultados de classificação para o *PTASAG*.

PTASAG					(4 Categorias)
	Métricas				
	ACC	PRE	REC	F1(m)	F1(w)
DTR	88,10%	61,39%	60,98%	60,25%	88,01%
GBC	90,29%	66,14%	66,27%	64,55%	91,07%
KNN	87,78%	65,77%	67,36%	64,44%	88,46%
RDF	**94,49%**	**69,04%**	**67,98%**	**67,32%**	**94,05%**
SVM	79,13%	61,22%	61,91%	56,55%	78,61%
WSD	86,33%	61,83%	62,71%	60,48%	87,60%

Na Tabela 10 é mostrado o desempenho da nossa proposta de SAG. Os resultados obtidos expõe o alto desempenho com RDF. Nossa técnica atingiu ACC e F1-ponderado de 94,49% e 94,05% respectivamente. Comparando com os níveis de F1-macro, é possível identificar que existe um grande desbalanceamento entre os quatro níveis de nota. Assim, o maior desafio encontrado nesse *dataset* é aprender a avaliar as categorias com amostragem desbalanceada. Os resultados em relação a outros trabalhos da literatura são apresentados na Figura 17.

Ainda com poucos trabalhos na literatura, na Figura 17 é mostrado o desempenho dos estudos realizados no *dataset PTASAG*. Nesse estudo, os autores reportaram ACC média de 68,8% e F1-ponderado de 68,09% com a técnica aqui mencionada (GALHARDI et al., 2018). Porém, o ganho obtido com as técnicas de *word-embeddings* não foi reportado nas métricas acima. Ainda assim, segundo o autor, a técnica de *n-grams* apresentou desempenho superior pelo coeficiente *kappa* (GALHARDI; SOUZA; BRANCHER, 2020). O coeficiente *kappa* é recomendado para mensurar a equivalência entre pares de avaliadores, entretanto não permite uma comparação justa com diferentes grupos de amostra.

Em um trabalho mais recente, foi estudada a calibração de oito classificadores

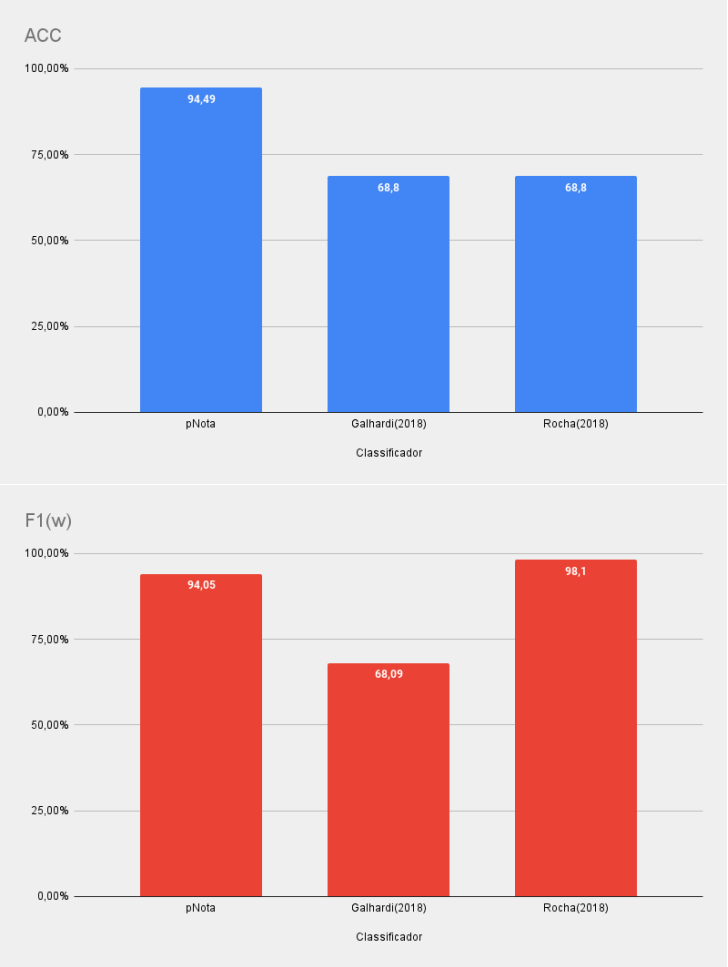


Figura 17 – Resultados dos classificadores com dados do *dataset PTASAG*.

para este *dataset* (GOMES-ROCHA et al., 2021). Dentre os testes estão os classificadores *Logistic Regression*, *K-Nearest Neighbors*, *Support Vector Machines*, *Bernoulli Naive Bayes*, *Extra Trees*, *Random Forest*, *AdaBoost* e *Multi-layer Perceptron*. *Random Forest* apresentou os melhores resultados com ACC de 68,8%, mas com ganho marginal em relação a outros classificadores. Os autores também apresentam a F1 e PRE acima de 90%, mas os resultados aparentam incompatibilidade com a ACC descrita.

Esse processo de avaliação também ocorre com um dos principais *datasets* da literatura. Com origem em uma competição, o *ASAP-SAS* foi criado com partições fixas de treino e teste, com a proposta de avaliação pelo coeficiente *Kappa* quadrático. Os resultados da classificação obtidos pelo *p*Nota são apresentados na Tabela 11.

Na Tabela 11 é exposto o desempenho de cada um dos classificadores com o *p*Nota. Um destaque é a grande diferença entre as técnicas, caracterizando a formação de subespaços complexos para cada nota. Assim, foi encontrada uma diferença de 17,32% entre GBC e KNN, sendo o primeiro o melhor desempenho com ACC de 69,11% e F1-ponderado de 67,14%. Entretanto, para comparar com a literatura foram selecionados os trabalhos que apresentam além do coeficiente *Kappa* o nível de erro encontrado na avaliação, com

Tabela 11 – Resultados de classificação para o *ASAP-SAS*.

ASAP-SAS					(4 Categorias)
	Métricas				
	ACC	PRE	REC	F1(m)	F1(w)
DTR	61,03%	46,54%	47,01%	46,39%	61,02%
GBC	**69,11%**	**56,16%**	**50,54%**	**50,92%**	**67,14%**
KNN	51,79%	38,06%	40,36%	36,11%	52,17%
RDF	63,71%	51,27%	42,19%	39,20%	57,85%
SVM	62,01%	44,69%	40,29%	35,44%	54,09%
WSD	52,72%	44,65%	37,39%	30,56%	48,23%

MAE, MSE e RMSE, mesmo sendo um *dataset* com notas discretas. Assim, o nível de erro apresentado na literatura são descritos na Figura 18.

Como é destacado na Figura 18, o nível de erro apresentado pelo *p*Nota é de 0,3558 pontos de MAE e 0,4527 pontos de MSE. As notas desse *dataset* vão de 0 até 3 pontos. Portanto, o erro resultante leva em conta que não existe gradação entre as quatro classes de nota. O melhor resultado é apresentado por (STEIMEL; RIORDAN, 2020), com MSE de 0,2055 pontos. Os autores aplicaram DL com *BERT*, uma *bidirectional transformer* com 12 *layers*. A principal mudança aplicada foi o teste de *mean* e *max-pooling* no lugar do padrão aplicado pela rede. Esse trabalho apresenta um ganho interessante em relação ao trabalho anterior dos autores (RIORDAN; FLOR; PUGH, 2019). Neste primeiro, foi utilizada uma *bidirectional Gated Recurrent Unit*, pré-treinada em uma *word-embeddings*. O principal diferencial desse estudo é a combinação de representações dos documentos (palavras e caracteres) em *Multilayer Perceptron Attention*.

Outro trabalho realizado foi apresentado por (HEILMAN; MADNANI, 2015). Nesse estudo foi aplicado *Support Vector Regression*, combinando estruturas sintáticas e semânticas. Para entender a influência de algumas respostas no treinamento, foram replicadas as amostras como forma de reforço. No entanto, os resultados obtidos indicam MAE de 0,64 pontos e um RMSE de 0,80 pontos.

Na mesma linha das avaliações contínuas também há o *dataset* da *University of North Texas*. Esse *dataset* é bem diferente, com apenas 30 respostas por questão. Cada questão foi avaliada por dois avaliadores, *Avaliador1* e *Avaliador2*, em notas de 0 a 5 discretas. Porém, o objetivo é minimizar o erro para a *Média* extraída entre eles. Na Tabela 12 são detalhados os resultados obtidos via técnicas de regressão para cada uma das três avaliações.

Na Tabela 12 é apresentado o nível de erro para cada avaliação, com destaque para a *Média*. O *p*Nota apresenta MAE de 0,5058, MSE de 0,5476 e RMSE de 0,6199 pontos. Por conta dessa baixa quantidade de amostras, há um problema na comparação com os demais estudos da literatura. Alguns métodos utilizam até 12-*Fold* durante a avaliação

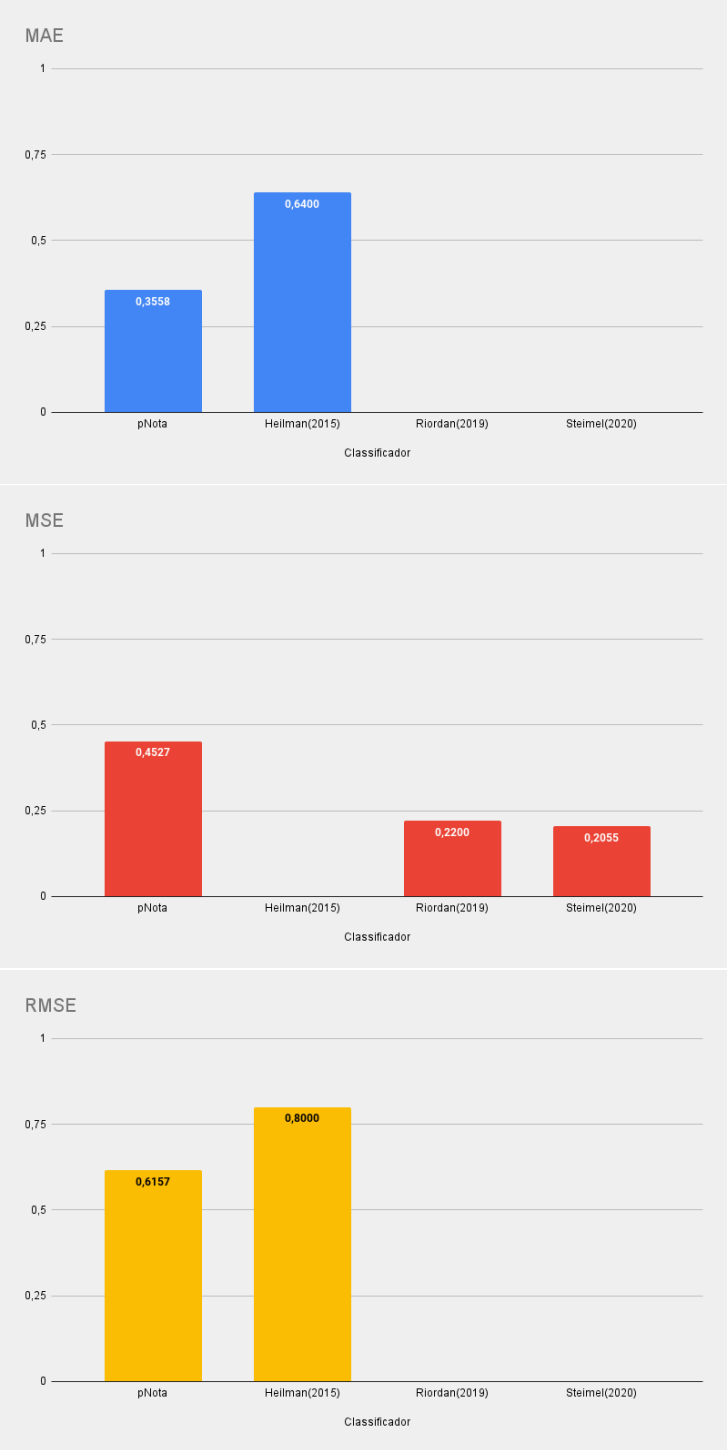


Figura 18 – Resultados dos classificadores com dados do *dataset ASAP-SAS*.

(KUMAR; CHAKRABARTI; ROY, 2017; SAHA et al., 2018). No caso do *p*Nota, com treino fixado pelo método de *clusterização*. Porém, foi designado 75% das amostras para treinamento tal qual encontrado em outros *datasets* da literatura.

Na Figura 19 é mostrado um detalhe interessante na evolução dos sistemas SAG. Os sistemas mais recentes, em ordem da esquerda para a direita, aumentaram consideravelmente o erro nesse *dataset*. Nesse cenário, o menor erro encontrado foi obtido pelo *p*Nota,

Tabela 12 – Índices de erro obtidos em cada um dos cenários de avaliação do *dataset* da *University of North Texas*.

University of North Texas			(Notas 0 - 5)
		Métricas	
	Avaliador1		
	MAE	MSE	RMSE
LINR	1,0066	**2,5069**	**1,1955**
LSSR	1,3273	3,1713	1,4712
KNRG	0,9366	2,9032	1,2557
DTRG	**0,9233**	3,7338	1,4482
WSRG	1,2832	3,0113	1,4240
	Avaliador2		
	MAE	MSE	RMSE
LINR	**0,4752**	**0,6099**	**0,6119**
LSSR	0,6502	0,8605	0,7640
KNRG	0,4917	0,7550	0,6658
DTRG	0,5121	1,2002	0,7856
WSRG	0,6523	0,8839	0,7680
	Média		
	MAE	MSE	RMSE
LINR	**0,5058**	**0,5476**	**0,6199**
LSSR	0,7299	0,8464	0,8170
KNRG	0,5055	0,6804	0,6765
DTRG	0,5811	1,1244	0,8372
WSRG	0,7024	0,8088	0,7920

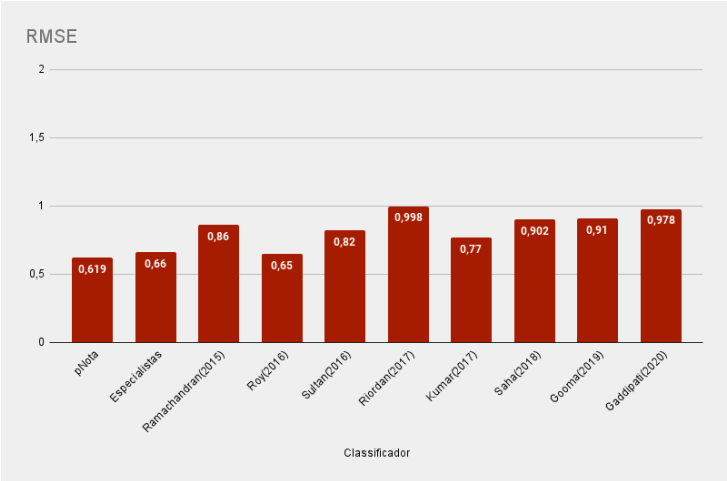


Figura 19 – Resultados dos avaliadores com dados do *dataset University of North Texas*.

com apenas 0,619 pontos de RMSE para a média dos avaliadores. Isso acontece com os métodos já descritos (RIORDAN et al., 2017; SAHA et al., 2018; GOMAA; FAHMY, 2019; RAMACHANDRAN; CHENG; FOLTZ, 2015).

A divergência observada entre os especilistas nesse *dataset* foi de 0,66 pontos. Apenas o trabalho descrito por (ROY et al., 2016) chegou a esse nível, com RMSE de 0,65 pontos. Seu diferencial, como comentado anteriormente, foi uma métrica de sobreposição entre respostas, calculando equivalência e divergências entre as respostas. Em um trabalho mais elaborado, (KUMAR; CHAKRABARTI; ROY, 2017) atingiu 0,77 pontos de RMSE com uma técnica específica de *pooling* aplicada em *Siamese LSTMs*. Para reforço do aprendizado com poucas amostras, os autores utilizaram *data aumentation*. Enquanto uma LSTM trabalha no *encodding* da resposta candidata, a segunda realiza o *encodding* das respostas dos alunos. O *layer* que avalia a compatibilidade utiliza a *Earth mover distance pooling*. Essa distância calcula a transferência mínima para aproximar os dois vetores na comparação de sequencias em *word embeddings*.

O trabalho mais recente de (GADDIPATI; NAIR; PLÖGER, 2020) mostra uma comparação do desempenho de modelos de *Transfer Learning*. Entre eles há o ELMo, GPT, GPT-2 e BERT. Os autores utilizaram 70% dos dados para treino e 30% para teste. Os autores compararam os resultados com outros métodos de análise vetorial tradicionais como TF-IDF, *Word2Vec*, *GloVe* e *FastText*. O melhor resultado obtido foi com o ELMO, com 0,978 pontos de RMSE. Ainda nesse estudo os autores destacam o problema do desbalanceamento das notas, com média de 4,17 pontos por resposta, tornando bem complexa a tarefa de aprendizado para níveis menores de nota.

Outros dois *datasets* aplicados foram com dados locais. O *dataset VestUFES* foi utilizado durante os anos para o desenvolvimento do *p*Nota (**??**). O primeiro é aplicado com dados do vestibular da universidade. Ele contém notas normalizadas para a escala de 0 a 10, mas foi avaliador inicialmente em notas contínuas entre 0 e 2 pontos. Os resultados de avaliação do *p*Nota são apresentados na Tabela 13.

Na Tabela 13 são mostrados resultados distintos entre as notas individuais dos avaliadores e o resultado médio. Tanto para o *Avaliador 1* quanto para o *Avaliador 2*, o menor nível de erro observado em MAE foi dado pelo DTRG. No entanto, para a nota média o KNRG apresentou melhor desempenho com MAE de 1,7667 pontos.

O segundo *dataset* local foi desenvolvido em 2019 para retratar um pouco melhor os resultados em dados nacionais. O *dataset* do *Projeto Feira Literária* foi coletado em conjunto com os autores para descrição da aplicação do *p*Nota e seu uso por professores (NASCIMENTO; KAUARK; MOURA, 2020). Este caracteriza-se pela presença de erros de escrita e por conteúdos fora de tópico, sendo fatores avaliados negativamente pelo tutor. Na Tabela 14 é apresentado o desempenho de cada um dos seis classificadores e seus resultados alcançados na avaliação das questões do projeto.

Conforme é caracterizado na Tabela 14, há alta qualidade de três classificadores RDF, DTR e GBC. O último, em especial, teve o maior desempenho com 78,58% de ACC. Com 70 respostas e a pluralidade de estruturas textuais encontradas, destaca-se

Tabela 13 – Índices de erro obtidos em cada um dos cenários de avaliação do *dataset* do *VestUFES*.

VestUFES			(Notas 0 - 10)
		Métricas	
	Avaliador1		
	MAE	MSE	RMSE
LINR	1,8614	**6,0562**	**2,3261**
LSSR	2,4570	9,5859	2,9596
KNRG	1,9058	6,8529	2,5057
DTRG	**1,6348**	7,0130	2,5333
WSRG	2,4665	9,6489	2,9515
	Avaliador2		
	MAE	MSE	RMSE
LINR	1,7943	**5,2526**	**2,1898**
LSSR	2,4358	8,5399	2,8264
KNRG	1,7290	6,3471	2,3781
DTRG	**1,6783**	7,7696	2,6739
WSRG	2,4282	8,5112	2,8111
	Final		
	MAE	MSE	RMSE
LINR	1,8556	**5,4816**	**2,2326**
LSSR	2,5365	9,5175	2,9435
KNRG	**1,7667**	6,7042	2,3789
DTRG	1,8457	8,3614	2,6566
WSRG	2,5256	9,4779	2,9198

Tabela 14 – Resultados de classificação para o *Projeto Feira Literária*.

Projeto Feira Literária					(4 Categorias)
			Métricas		
	ACC	PRE	REC	F1(m)	F1(w)
DTR	77,86%	59,99%	59,29%	58,42%	76,84%
GBC	**78,58%**	59,58%	59,57%	57,07%	**77,74%**
KNN	65,00%	58,72%	58,02%	54,75%	66,91%
RDF	78,57%	**61,82%**	**64,98%**	**61,36%**	76,53%
SVM	72,86%	51,49%	57,85%	49,34%	67,66%
WSD	75,71%	54,77%	61,24%	56,14%	74,57%

a semelhança deste *dataset* com os desafios apresentados por *Beetle* e *SciEntsBank* no ensino de ciências.

4.3 Discussão de Resultados

A avaliação de questões discursivas é uma área que evoluiu muito em relação as propostas iniciais dos anos 60. O volume de dados e a capacidade de processamento e

suporte computacional foram fundamentais para incorporar isso na rotina avaliativa do professor. Dos sistemas mais rígidos, que operam com regras e padrões pré-fixados, até hoje com maior profundidade dos modelos com análises de NLP. Isso inclui técnicas de ML e DL aparecendo nos trabalhos mais recentes.

Um paradigma que entra em conflito com os requisitos e autonomia de técnicas mais robustas de classificação é o volume limitado de dados segundo o tema. Para isso, muitos métodos consideram o *dataset* como um todo para treinamento, agrupando toda série de atividades para compreender, em nível macro, cada nível de nota. Isso inibe parcialmente a falta de amostras mas presume *a priori* uma equivalência contextual. Essa equivalência nem sempre existe. Certas situações podem afetar diretamente a produção textual que, se não analisadas separadamente, causam problemas na interpretação entre a equivalência e divergência contextual. Isolando essas situações, podemos identificar alterações em notas, experiências, plágios e cenários avaliativos.

Uma situação que pode causar isso é a aplicação de uma questão com turmas que tiveram experiências distintas. Se uma turma teve contato com um objeto da questão enquanto outra não teve, possivelmente temos impacto direto na forma de resposta. Quando tal situação é explorada nas atividades disponibilizadas pelo professor, esse tipo de situação ganha ainda mais notoriedade.

Por conta disso, os estudos do *p*Nota envolvem comprender os diferentes relatos dos estudantes para assimilar detalhes sobre o contexto. Então, mesmo nesse caso, presume-se a coerência e a consistência textual, mas apenas em no nível das atividades. É interessante a partir daí, que ocorra o enriquecimento textual. Por isso, os trabalhos mais recentes combinam várias formas de interpretação do texto, atingindo ao menos níveis semânticos e sintáticos.

Assim, na criação do modelo linguístico, de uma forma geral os sistemas trabalham o alinhamento entre o conjunto de respostas e as respostas candidatas. A proposta apresentada neste trabalho busca iterações com o professor para evolução do modelo criado. Para além da análise textual o sistema prioriza a conexão entre conteúdo e critério avaliativo. Dessa maneira, as nuances do contexto são interpretadas pelo tutor durante a avaliação, destacada com o critério de atribuição de notas. Superdimensionados para a avaliação de uma determinada disciplina ou tema, os modelos rígidos divergem bastante da aplicação dos sistemas SAG no cotidiano do professor. Assim, o professor espera que o sistema seja capaz de reduzir o esforço de correção, dando suporte ao seu método assim que requisitado. Independentemente do cenário ao qual é aplicado, o sistema SAG deve lidar com as respostas como parte do ensino-aprendizagem, buscando minorar a demanda de verificação do conteúdo.

De qualquer forma, pela dinâmica dos trabalhos na área, vê-se que é muito complexa a tarefa de se adequar a todos os cenários. Nessa perspectiva, é identificado com o *p*Nota um

ótimo desempenho nos diferentes *datasets*, sendo equiparável aos resultados da literatura em todos os testes. Além disso, o vínculo do método avaliativo com as características textuais encontradas no *p*Nota tem maior capacidade de resolver questões do que boa parte dos sistemas mencionados. Isso acontece porque a proposta aqui descrita busca compreender diferenças encontradas nos dados antes do processo avaliativo. De forma geral, a preocupação com a modelagem do critério avaliativo torna mais efetiva a contextualização sobre o tema. Consequentemente, um sistema contextualizado tende a apresentar bom desempenho como avaliador.

O *p*Nota também precisa de apenas algo entre seis minutos e uma hora para processamento em um computador comum. Essa variação é dada especificamente pelo número de testes executados, vinculado principalmente às características de cada atividade. Essa *performance* foi observada em um computador com processador Intel Core i7-8700 (3,2GHz x 12) com 16 GB de memória, portando uma placa de vídeo NVIDIA GeForce GTX 1070. O tempo de execução é definido segundo o número de etapas e o número de características encontradas nas amostras do conjunto de dados. Esse tempo é resultado de um refinamento dos processos e redução de testes que não adicionavam valor aos processos. Porém, essa avaliação considera apenas o tempo do sistema, desconsiderando as interações e o tempo de anotação com o professor durante a execução. Apesar de enquadrar-se bem no tempo que o professor demanda para avaliar cada uma das respostas, ainda são válidos os esforços para otimizar as etapas parciais de cada um dos processos para a melhoria da *performance* geral. Logo, esse tipo de sistema deve atender ao problema e, simultaneamente, ser associado à rotina de avaliação do professor.

Ainda é importante salientar que, entre todas as comparações criadas com os *datasets* disponíveis, o modelo proposto neste trabalho foi o mesmo aplicado para todas as questões. Os processos de otimização que compõe o modelo que realizam a calibração do que deve ser entregue ao professor. Portanto, o *p*Nota não aprensenta rigidez e contextualização que são necessários a parte dos modelos que atuam em domínios específicos. Assim, inclui-se na qualidade dos resultados obtidos a capacidade do próprio sistema de adaptação ao tema e ao modelo de avaliação, sendo parte das etapas do *p*Nota a calibração dos processos.

A escolha dos modelos é realizada pela paridade com os resultados obtidos nas amostras com avaliação do especialista. Por conta disso, o *p*Nota conta com flexibilidade para estudar o contexto aplicado a cada questão e otimizar o aprendizado dos algoritmos. A tendência, então, é que os métodos selecionados apresentem o melhor reconhecimento dos grupos de nota, seja via técnicas de classificação ou de regressão.

Por fim, para avaliar os resultados obtidos com tal processo de classificação, foi testada a hipótese de que o classificador GBC apresenta resutados superiores, dado que o mesmo mostrou-se melhor na maioria dos cenários. Para isso, o primeiro teste realizado foi o Teste de Shapiro-Wilk, onde é avaliado se os resultados seguem ou não uma distribuição

normal. A normalidade da distribuição foi rejeitada com *p-value* de 0.0038. Um segundo teste, Teste de Levene, foi aplicado para verificar se os resultados apresentam variâncias iguais. Com um *p-value* de 0.9419 os resultados mostram que as amostras são homogêneas, sem diferenças significativas entre as variâncias. Por fim, o terceito teste aplicado foi o Teste de Kruskal-Wallis, onde foi avaliado se, dado a homogeniedade da variância, as amostras têm origem em uma mesma distribuição. Entretanto, o fato da hipótese não ser rejeitada corrobora as principais contribuições deste trabalho, que indicam não haver diferença significativa entre os classificadores após o ganho obtido via *Active Learning*.

5 Considerações Finais

O processo de avaliação de questões discursivas compreende um longo ciclo, da formulação das atividades até a análise de desempenho. Por meio das avaliações que há o redimensionamento das práticas de ensino e aprendizagem. Assim, a dinâmica avaliativa do professor permite identificação e tratar os *gaps* de aprendizagem. A avaliação também é um indicativo para possíveis refinamentos da disciplina. Conforme o desempenho de cada estudante e cada turma são identificadas as melhores formas para aprimorar as técnicas e apresentar os resultados obtidos em sala. Entretanto, uma premissa para alcançar tais resultados de produção técnica é que sejam reduzidos os esforços da aplicação das atividades discursivas nas rotinas de ensino.

Em especial, a aplicação de atividades que contribuem para melhoria da leitura e da escrita é essencial para todos os níveis de instrução dos estudantes. Assim, a continuidade do processo avaliativo tem benefícios para todos os participantes do ciclo. Cientes disso, foi apresentado neste trabalho um estudo das estruturas textuais combinados com técnicas de *Active Learning* para suporte a avaliação de respostas discursivas. Como destaques dos resultados obtidos com este trabalho foram listadas uma série de contribuições que compõem este estudo:

- Análise contextual do modelo linguístico através de *Active Learning* com refinamento por ciclo avaliativo;
- Análise da distribuição de amostras combinando processos de clusterização e classificação;
- Otimização da seleção de hiperparâmetros, identificando a composição espacial dos *clusters*;
- Amostragem por representatividade intra-*cluster* com anotação guiada do usuário para extração de relações de equivalência e divergência contextual;
- Construção do modelo com padrões de diferentes camadas da linguagem, para caracterizar formas distintas de resposta em nível de termos e sequências.
- Reconhecimento de padrões que combinam a interação entre estudantes, o professor e o sistema na avaliação, detalhando o critério segundo o alinhamento ao tema;
- Criação de formatos próprios para relatórios e *feedbacks*, com nível de explicabilidade entre termos e notas, acompanhando didático e auditoria dos resultados.

- Suporte para uma série de pesquisas pedagógicas, colaborando com duas Dissertações de Mestrado do Mestrado Profissional em Química em Rede Nacional (ProfQui) do Instituto Federal do Espírito Santo.

5.1 Conclusões

O suporte computacional do processo avaliativo compõe importante parte da integração do ensino em meio digital. O meio digital possibilita a aplicação em massa, tornando prático e rápido o ciclo de cada avaliação. Desse modo, o professor em um mesmo ambiente consegue interagir com todos os seus alunos e acompanhar seu desempenho na disciplina.

Por meio dessas plataformas de ensino, possibilita-se o emprego de múltiplas técnicas de EDM para análise do conteúdo e, consequentemente, o aumento da capacidade avaliativa. Ao longo deste trabalho foi descrito o *p*Nota. O *p*Nota é um sistema para construção de modelos avaliativos que interage diretamente com o tutor para compreender seu critério avaliativo sobre um conjunto de resposta. Portanto, neste trabalho, está descrito um estudo profundo do ciclo avaliativo para integração humano-máquina na correção de respostas discursivas via *Active Learning*. Como descrito neste trabalho, integram o *p*Nota uma série módulos que foram adaptados para os ciclos avaliativos observados entre os professores, realizando o reconhecimento de padrões e a identificação contextual.

Como destaque aos resultados observados, temos níveis de desempenho similares aos observados entre dois humanos na atribuição de notas, com F1-ponderado médio de 78% e ACC de 79%. Destaca-se que tais valores refletem que, das 255 atividades de classificação, 137 foram avaliadas com mais de 75% de ACC. O mesmo desempenho também se reflete nas atividades de regressão. Os resultados obtidos com o sistema, indicam que é possível que o mesmo faça parte da rotina pedagógica do professor. Assim, com o *p*Nota o professor pode corrigir atividades e utilizar os resultados em função do desenvolvimento de seus métodos de ensino. Adicionalmente, com os *feedbacks*, esperamos compor materiais que auxiliem várias etapas do método avaliativo, incluindo a discussão de resultados em sala.

5.2 Trabalhos Futuros

Em uma perspectiva de próximos estudos em torno do modelo proposto, destacam-se alguns trabalhos ainda de refinamento do modelo. O ganho analítico dos modelos de linguagem e domínio, em especial com *word embeddings* e *Large Language Models*, podem proporcionar ao sistema novos níveis de conhecimento. Unindo esses modelos pode-se ter um ganho importante com contextualização. Em uma outra vertente, também são válidos os esforços para entender as variações textuais encontradas em novas iterações de

aplicação das atividades. Inclusive, são necessários estudos que se aprofundem no ganho de informação com a evolução do aprendizado. Assim, para além do nível linguístico, é essencial mensurar a capacidade de vincular termos e notas, aprendendo as componentes da avaliação.

Outra característica importante para evolução do modelo é a integração de critérios mais sofisticados na otimização dos resultados, para clusterização, classificação e regressão. A otimização pode ser relevante para alcançar ainda mais qualidade na atribuição de notas, extraindo detalhes da formação de cada conjunto de respostas. Nessa mesma linha, são fundamentais o acompanhamento do sistema em características de usabilidade. Portanto, além de conhecer a prática e os ciclos avaliativos também devemos considerar ter mais proximidade nas iterações sob a ótica do professor. Em especial, é necessário contemplar em detalhes o acompanhamento de escolas, turmas ou grupos de alunos sob a concepção das técnicas de ensino-aprendizagem. Nesse aspecto, é desejável a melhoria da contextualização dos documentos, com o detalhamento da produção escrita e dos ciclos de aplicação das atividades.

Referências

AGARWAL, D.; GUPTA, S.; BAGHEL, N. ScAA: A Dataset for Automated Short Answer Grading of Children's free-text Answers in Hindi and Marathi. In: *Proceedings of the 17th International Conference on Natural Language Processing (ICON)*. Patna, India: NLP Association of India (NLPAI), 2020. v. 17, p. 430–436. Citado na página 69.

ALEKSANDER, I.; THOMAS, W. V.; BOWDEN, P. A. Wisard - a radical step forward in image recognition. *Sensor Review*, MCB UP Ltd, v. 4, n. 3, p. 120–124, 1984. Citado na página 60.

ALMEIDA-JÚNIOR, C. R. C.; SPALENZA, M. A.; OLIVEIRA, E. de. Proposta de um Sistema de Avaliação Automática de Redações do ENEM Utilizando Técnicas de Aprendizagem de Máquina e Processamento de Linguagem Natural. In: *Computer on the Beach*. Florianópolis (SC), Brasil: Universidade do Vale do Itajaí - UNIVALI, 2017. v. 8, p. 474–483. Citado 2 vezes nas páginas 37 e 65.

ARTER, J. A.; CHAPPUIS, J. *Creating & Recognizing Quality Rubrics*. 1st. ed. New York (NY), USA: Pearson Education, 2006. (Assessment Training Institute, Inc Series). Citado na página 27.

ARTSTEIN, R.; POESIO, M. Inter-Coder Agreement for Computational Linguistics. *Computational Linguistics*, MIT Press, v. 34, n. 4, p. 555–596, 2008. Citado 3 vezes nas páginas 29, 30 e 60.

AZAD, S. et al. Strategies for Deploying Unreliable AI Graders in High-Transparency High-Stakes Exams. In: *Proceedings of the 21st International Conference on Artificial Intelligence in Education*. Ifrane, Morocco: Springer International Publishing, 2020. (AIED' 2020, v. 21), p. 16–28. Citado 3 vezes nas páginas 28, 35 e 42.

BAEZA-YATES, R.; RIBEIRO-NETO, B. *Modern Information Retrieval: The Concepts and Technology behind Search*. 2nd. ed. Boston (MA), USA: Addison-Wesley Publishing Company, 2011. Citado 6 vezes nas páginas 35, 40, 46, 48, 52 e 53.

BAILEY, S.; MEURERS, D. Diagnosing Meaning Errors in Short Answers to Reading Comprehension Questions. In: *Proceedings of the Third Workshop on Innovative Use of NLP for Building Educational Applications*. Columbus (OH), USA: Association for Computational Linguistics, 2008. (EANL '08, v. 3), p. 107–115. Citado 3 vezes nas páginas 28, 36 e 37.

BANJADE, R. et al. Evaluation Dataset (DT-Grade) and Word Weighting Approach Towards Constructed Short Answers Assessment in Tutorial Dialogue Context. In: *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*. San Diego (CA), USA: Association for Computational Linguistics, 2016. v. 11, p. 182–187. Citado na página 40.

BANJADE, R.; RUS, V.; NIRAULA, N. B. Using an Implicit Method for Coreference Resolution and Ellipsis Handling in Automatic Student Answer Assessment. In: *The Twenty-Eighth International Flairs Conference*. Hollywood (FL), USA: AAAI Press, 2015. v. 28, p. 150–155. Citado na página 39.

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. Citado na página 21.

BASU, S.; JACOBS, C.; VANDERWENDE, L. Powergrading: A Clustering Approach to Amplify Human Effort for Short Answer Grading. *Transactions of the Association for Computational Linguistics*, MIT Press, v. 1, n. 1, p. 391–402, 2013. Citado 4 vezes nas páginas 38, 41, 73 e 86.

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. Citado 4 vezes nas páginas 30, 32, 33 e 36.

BEZERRA, M. A. Questões Discursivas para Avaliação Escolar. *Acta Scientiarum. Language and Culture*, Universidade Estadual de Maringá, v. 30, n. 2, p. 149–157, 2008. Citado na página 36.

BIGGS, J. Assessment and Classroom Learning: A Role for Summative Assessment? *Assessment in Education: Principles, Policy & Practice*, Routledge, v. 5, n. 1, p. 103–110, 1998. Citado na página 21.

BILGIN, A. A.; ROWE, A. D.; CLARK, L. Academic Workload Implications of Assessing Student Learning in Work-Integrated Learning. *Asia-Pacific Journal of Cooperative Education*, ERIC, v. 18, n. 2, p. 167–183, 2017. Citado na página 36.

BOGARÍN, A.; CEREZO, R.; ROMERO, C. A Survey on Educational Process Mining. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, Wiley Online Library, v. 8, n. 1, p. e1230, 2018. Citado na página 21.

BONTHU, S.; SREE, S. R.; KRISHNA-PRASAD, M. H. M. Automated Short Answer Grading Using Deep Learning: A Survey. In: *Proceedings of the 5th International Cross-Domain Conference for Machine Learning and Knowledge Extraction (CD-MAKE'2021)*. Virtual Event: Springer, 2021. v. 5, p. 61–78. Citado 3 vezes nas páginas 27, 42 e 45.

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. Citado 10 vezes nas páginas 27, 28, 30, 32, 33, 36, 42, 43, 45 e 82.

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. Citado 6 vezes nas páginas 29, 33, 36, 39, 43 e 86.

CALIŃSKI, T.; J., H. A Dendrite Method for Cluster Analysis. *Communications in Statistics*, Taylor & Francis, v. 3, n. 1, p. 1–27, 1974. Citado na página 55.

CAMUS, L.; FILIGHERA, A. Investigating Transformers for Automatic Short Answer Grading. In: *Proceedings of the 21st International Conference on Artificial Intelligence in Education*. Ifrane, Morocco: Springer International Publishing, 2020. (AIED' 2020, v. 21), p. 43–48. Citado 2 vezes nas páginas 42 e 43.

CASIRAGHI, B.; ALMEIDA, L. S. Elaboração de um Instrumento de Avaliação do

Pensamento Crítico em Estudantes Universitários. In: *Atas do V Seminário Internacional Cognição, Aprendizagem e Desempenho*. Braga, Portugal: CIEd-Universidade do Minho Portugal, 2017. v. 5, p. 30–41. Citado na página 21.

CHAKRABORTY, U. K.; ROY, S.; CHOUDHURY, S. A Fuzzy Indiscernibility Based Measure of Distance between Semantic Spaces Towards Automatic Evaluation of Free Text Answers. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, World Scientific, v. 25, n. 6, p. 987–1004, 2017. Citado 2 vezes nas páginas 28 e 42.

COHEN, J. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, SAGE Publishing, v. 20, n. 1, p. 37–46, 1960. Citado na página 60.

CONDOR, A. Exploring Automatic Short Answer Grading as a Tool to Assist in Human Rating. In: *Proceedings of the 21st International Conference on Artificial Intelligence in Education*. Ifrane, Morocco: Springer International Publishing, 2020. (AIED' 2020, v. 21), p. 74–79. Citado 3 vezes nas páginas 29, 33 e 35.

CONDOR, A.; LITSTER, M.; PARDOS, Z. Automatic Short Answer Grading with SBERT on Out-of-Sample Questions. In: *Proceedings of the 14th International Conference on Educational Data Mining (EDM)*. Virtual Event: ERIC, 2021. v. 14, p. 345–352. Citado 3 vezes nas páginas 29, 30 e 69.

DAVIES, D. L.; BOULDIN, D. W. A Cluster Separation Measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, IEEE, v. 1, n. 2, p. 224–227, 1979. Citado na página 55.

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. Citado 6 vezes nas páginas 29, 33, 37, 43, 52 e 69.

DUNLAP, J. C. Workload Reduction in Online Courses: Getting Some Shuteye. *Performance Improvement*, Wiley Online Library, v. 44, n. 5, p. 18–25, 2005. Citado na página 22.

DZIKOVSKA, M. et al. SemEval-2013 Task 7: The Joint Student Response Analysis and 8th Recognizing Textual Entailment Challenge. In: *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*. Atlanta (GE), USA: Association for Computational Linguistics, 2013. v. 7, p. 263–274. Citado 3 vezes nas páginas 30, 71 e 80.

DZIKOVSKA, M. O.; NIELSEN, R. D.; BREW, C. Towards Effective Tutorial Feedback for Explanation Questions: A Dataset and Baselines. In: *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Montréal, Canada: Association for Computational Linguistics, 2012. v. 11, p. 200–210. Citado 2 vezes nas páginas 29 e 71.

EVERITT, B. S. et al. *Cluster Analysis*. 5th. ed. Chichester, United Kingdom: John Wiley, 2011. Citado 3 vezes nas páginas 31, 40 e 55.

FERREIRA-MELLO, R. et al. Text Mining in Education. *WIREs Data Mining and Knowledge Discovery*, Wiley Online Library, v. 9, n. 6, p. e1332.1–e1332.49, 2019. Nenhuma citação no texto.

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. Citado 8 vezes nas páginas 27, 29, 33, 37, 42, 43, 52 e 84.

FOWLER, M. et al. Autograding "Explain in Plain English" Questions Using NLP. In: *Proceedings of the 52nd ACM Technical Symposium on Computer Science Education*. New York (NY), USA (Virtual Event): Association for Computing Machinery, 2021. (SIGCSE'21, v. 52), p. 1163–1169. Citado 2 vezes nas páginas 38 e 43.

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. Citado 4 vezes nas páginas 28, 33, 38 e 42.

GADDIPATI, S. K.; NAIR, D.; PLÖGER, P. G. *Comparative Evaluation of Pretrained Transfer Learning Models on Automatic Short Answer Grading*. Sankt Augustin, Germany, 2020. Citado na página 93.

GALHARDI, L.; SOUZA, R. de; BRANCHER, J. Automatic Grading of Portuguese Short Answers Using a Machine Learning Approach. In: *Anais do XVI Simpósio Brasileiro de Sistemas de Informação (SBSI'2020)*. São Bernardo do Campo (SP), Brazil (Online Event): Sociedade Brasileira de Computação, 2020. v. 16, p. 109–124. Citado 2 vezes nas páginas 41 e 88.

GALHARDI, L. B. et al. Portuguese Automatic Short Answer Grading. In: *Proceedings of the 29th Brazilian Symposium on Computers in Education (SBIE'2018)*. Fortaleza (CE), Brazil: Sociedade Brasileira de Computação, 2018. v. 29, p. 1373–1382. Citado 4 vezes nas páginas 40, 43, 74 e 88.

GALHARDI, L. B.; BRANCHER, J. D. Machine Learning Approach for Automatic Short Answer Grading: A Systematic Review. In: *Proceedings of the 16th Ibero-American Conference on Artificial Intelligence - IBERAMIA 2018*. Trujillo, Peru: Springer International Publishing, 2018. (IBERAMIA 2018, v. 16), p. 380–391. Citado 2 vezes nas páginas 33 e 40.

GALHARDI, L. B. et al. Exploring Distinct Features for Automatic Short Answer Grading. In: *Anais do XV Encontro Nacional de Inteligência Artificial e Computacional*. São Paulo (SP), Brazil: Sociedade Brasileira de Computação, 2018. (XV ENIAC, v. 15), p. 1–12. Citado 2 vezes nas páginas 81 e 83.

GHAVIDEL, H.; ZOUAQ, A.; DESMARAIS, M. Using BERT and XLNET for the Automatic Short Answer Grading Task. In: *Proceedings of the 12th International Conference on Computer Supported Education (CSEDU)*. Prague, Czechia (Virtual Event): SciTePress, 2020. (CSEDU 2020, v. 12), p. 58–67. Citado na página 41.

GOLDBERG, Y.; HIRST, G. *Neural Network Methods in Natural Language Processing*.

1st. ed. San Rafael (CA), USA: Morgan & Claypool Publishers, 2017. Citado na página 41.

GOMAA, W. H.; FAHMY, A. A. Ans2vec: A Scoring System for Short Answers. In: *The International Conference on Advanced Machine Learning Technologies and Applications (AMLTA'2019)*. Cairo, Egypt: Springer, 2019. v. 4, p. 586–595. Citado 3 vezes nas páginas 69, 83 e 92.

GOMES-ROCHA, F. et al. Supervised Machine Learning for Automatic Assessment of Free-Text Answers. In: *Proceeding of the 20th Mexican International Conference on Artificial Intelligence (MICAI'2021)*. Mexico City, Mexico (Online Event): Springer International Publishing, 2021. v. 20, p. 3–12. Citado na página 89.

HALLER, S. et al. *Survey on Automated Short Answer Grading with Deep Learning: from Word Embeddings to Transformers*. Twente, Netherlands, 2022. Citado 3 vezes nas páginas 41, 42 e 45.

HAN, J.; PEI, J.; KAMBER, M. *Data Mining: Concepts and Techniques*. 3rd. ed. Waltham (MA), USA: Elsevier, 2011. Citado 2 vezes nas páginas 49 e 56.

HEILMAN, M.; MADNANI, N. The Impact of Training Data on Automated Short Answer Scoring Performance. 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. 81–85. Citado 3 vezes nas páginas 33, 38 e 90.

HIGGINS, D. et al. *Is Getting the Right Answer Just About Choosing the Right Words? The Role of Syntactically-Informed Features in Short Answer Scoring*. Princeton (NJ), USA, 2014. Citado 2 vezes nas páginas 28 e 33.

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. 4065–4071. Citado 3 vezes nas páginas 29, 30 e 38.

ISHIOKA, T.; KAMEDA, M. Overwritable Automated Japanese Short-Answer Scoring and Support System. In: *Proceedings of the International Conference on Web Intelligence (WI'2017)*. Leipzig, Germany: Association for Computing Machinery, 2017. v. 1, p. 50–56. Citado na página 69.

JIMENEZ, S.; BECERRA, C.; GELBUKH, A. SOFTCARDINALITY: Hierarchical Text Overlap for Student Response Analysis. In: *Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*. Atlanta (GA), USA: Association for Computational Linguistics, 2013. v. 7, p. 280–284. Citado 3 vezes nas páginas 39, 41 e 42.

JOHNSTONE, K. M.; ASHBAUGH, H.; WARFIELD, T. D. Effects of Repeated Practice and Contextual-Writing Experiences on College Students' Writing Skills. *Journal of Educational Psychology*, American Psychological Association, v. 94, n. 2, p. 305–315, 2002. Citado na página 36.

JORDAN, S. Student Engagement with Assessment and Feedback: Some Lessons from Short-Answer Free-Text e-Assessment Questions. *Computers & Education*, Elsevier, v. 58, n. 2, p. 818–834, 2012. Citado 5 vezes nas páginas 28, 33, 42, 43 e 73.

JURAFSKY, D.; MARTIN, J. H. *Speech and Language Processing*. 2nd. ed. Upper Saddle River (NJ), USA: Prentice-Hall, Inc., 2009. Citado 2 vezes nas páginas 35 e 41.

KAR, S. P.; CHATTERJEE, R.; MANDAL, J. K. A novel automated assessment technique in e-learning using short answer type questions. In: *Proceedings of the 1st International Conference on Computational Intelligence, Communications, and Business Analytics*. Kolkata, India: Springer Singapore, 2017. (CICBA 2017, v. 1), p. 141–149. Citado 2 vezes nas páginas 39 e 42.

KRITHIKA, R.; NARAYANAN, J. Learning to grade short answers using machine learning techniques. In: *Proceedings of the Third International Symposium on Women in Computing and Informatics*. Kochi, India: Association for Computing Machinery, 2015. (WCI '15, v. 3), p. 262–271. Citado na página 28.

KUMAR, S.; CHAKRABARTI, S.; ROY, S. Earth Mover's Distance Pooling over Siamese LSTMs for Automatic Short Answer Grading. In: *Proceedings of the 26th International Joint Conference on Artificial Intelligence*. Melbourne, Australia: AAAI Press, 2017. (IJCAI'17, v. 26), p. 2046–2052. Citado 3 vezes nas páginas 41, 91 e 93.

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. Citado 8 vezes nas páginas 26, 28, 30, 33, 39, 40, 42 e 43.

LANDAUER, T. K.; FOLTZ, P. W.; LAHAM, D. An Introduction to Latent Semantic Analysis. *Discourse Processes*, Routledge, v. 25, n. 2, p. 259–284, 1998. Citado na página 41.

LEFFA, V. J. Análise Automática da Resposta do Aluno em Ambiente Virtual. *Revista Brasileira de Linguística Aplicada*, SciELO, v. 3, n. 2, p. 25–40, 2003. Citado na página 27.

LIMA-FILHO, A. S. et al. *wisardpkg – A Library for WiSARD-Based Models*. Rio de Janeiro (RJ), Brazil, 2020. Citado na página 60.

LUI, A. K.-F.; NG, S.-C.; CHEUNG, S. W.-N. A Framework for Effectively Utilising Human Grading Input in Automated Short Answer Grading. *International Journal of Mobile Learning and Organisation*, Inderscience Publishers, v. 16, n. 3, p. 266–286, 2022. Citado 4 vezes nas páginas 28, 29, 33 e 87.

LUN, J. et al. Multiple Data Augmentation Strategies for Improving Performance on Automatic Short Answer Scoring. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. New York (NY), USA: AAAI Press, 2020. v. 34, n. 09, p. 13389–13396. Citado na página 39.

MADERO, C. Secondary Teacher's Dissatisfaction with the Teaching Profession in Latin America: The Case of Brazil, Chile, and Mexico. *Teachers and Teaching*, Routledge, v. 25, n. 3, p. 358–378, 2019. Citado na página 36.

MAIMON, O.; ROKACH, L. *Data Mining and Knowledge Discovery Handbook*. 1st. ed. New York (NY), USA: Springer, 2005. Citado na página 55.

MALIK, A. et al. Generative Grading: Near Human-level Accuracy for Automated

Feedback on Richly Structured Problems. In: *Proceedings of The 14th International Conference on Educational Data Mining (EDM)*. Online Event: ERIC, 2021. v. 14, p. 275–286. Citado na página 87.

MANNING, C. D.; RAGHAVAN, P.; SCHUTZE, H. *Introduction to Information Retrieval*. 39th. ed. Cambridge (MA), USA: Cambridge University Press, 2008. Citado na página 61.

MANNING, C. D.; SCHUTZE, H. *Foundations of Statistical Natural Language Processing*. 1st. ed. Cambridge (MA), USA: MIT Press, 1999. Citado na página 60.

MAO, L. et al. Validation of Automated Scoring for a Formative Assessment that Employs Scientific Argumentation. *Educational Assessment*, Routledge, v. 23, n. 2, p. 121–138, 2018. Citado na página 42.

MAQUINÉ, G. Recursos para Avaliação da Aprendizagem: Estudo Comparativo entre Ambientes Virtuais de Aprendizagem. In: *Anais do XXVI Workshop de Informática na Escola*. Natal (RN) (Online), Brasil: Sociedade Brasileira de Computação, 2020. v. 26, p. 299–308. Citado na página 22.

MARNEFFE, M.-C. et al. Universal Dependencies. *Computational Linguistics*, MIT Press, v. 47, n. 2, p. 255–308, 2021. Citado na página 52.

MARVANIYA, S. et al. Creating Scoring Rubric from Representative Student Answers for Improved Short Answer Grading. In: *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*. Torino, Italy: Association for Computing Machinery, 2018. (CIKM '18, v. 27), p. 993–1002. Citado 4 vezes nas páginas 29, 30, 38 e 39.

MENINI, S. et al. Automated Short Answer Grading: A Simple Solution for a Difficult Task. In: *Proceedings of the Sixth Italian Conference on Computational Linguistics*. Bari, Italy: CEUR-WS, 2019. (CLiC-it, v. 6), p. 48.1–48.7. Citado na página 43.

MILLER, B.; LINDER, F.; MEBANE, W. R. Active Learning Approaches for Labeling Text: Review and Assessment of the Performance of Active Learning Approaches. *Political Analysis*, Cambridge University Press, v. 28, n. 4, p. 532–551, 2020. Citado 2 vezes nas páginas 31 e 38.

MING, L. S. Reduction of Teacher Workload in a Formative Assessment Environment through use of Online Technology. In: *6th International Conference on Information Technology Based Higher Education and Training*. Santo Domingo, Dominican Republic: IEEE, 2005. v. 6, p. 18–21. Citado na página 36.

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. Citado 2 vezes nas páginas 33 e 39.

MOHAPATRA, S.; MOHANTY, R. Adopting MOOCs for Afforable Quality Education. *Education and Information Technologies*, Springer, v. 22, n. 5, p. 2027–2053, 2017. Citado na página 46.

MOHLER, M.; BUNESCU, R.; MIHALCEA, R. Learning to Grade Short Answer

Questions using Semantic Similarity Measures and Dependency Graph Alignments. In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Portland (OR), USA: Association for Computational Linguistics, 2011. v. 10, p. 752–762. Citado 5 vezes nas páginas 29, 35, 42, 63 e 74.

MOHLER, M.; MIHALCEA, R. Text-to-text semantic similarity for automatic short answer grading. In: *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*. Athens, Greece: Association for Computational Linguistics, 2009. v. 12, p. 567–575. Citado 2 vezes nas páginas 40 e 41.

MORETTIN, P. A.; BUSSAB, W. O. *Estatística Básica*. 6. ed. Pinheiros (SP), Brasil: Editora Saraiva, 2010. Citado 3 vezes nas páginas 31, 63 e 69.

NASCIMENTO, P. V.; KAUARK, F. S.; MOURA, P. R. G. *Construindo uma Atividade Experimental Problematizada (AEP) e Avaliando Seu Nível Cognitivo de Aprendizagem Através do Software pNota no Contexto do Ensino Fundamental*. 9. ed. Vila Velha (ES), Brasil: Instituto Federal do Espírito Santo, 2020. (Série Guia Didático de Ciências/Química). Citado 3 vezes nas páginas 65, 75 e 93.

OLIVEIRA, E. et al. Combining Clustering and Classification Approaches for Reducing the Effort of Automatic Tweets Classification. In: *Proceedings of the International Conference on Knowledge Discovery and Information Retrieval - KDIR, (IC3K 2014)*. Rome, Italy: SciTePress, 2014. (KDIR '14, v. 6), p. 465–472. Citado na página 40.

OLIVEIRA, K. L. d.; SANTOS, A. A. A. Compreensão em Leitura e Avaliação da Aprendizagem em Universitários. *Psicologia: Reflexão e Crítica*, SciELO, v. 18, n. 1, p. 118–124, 2005. Citado na página 21.

OLIVEIRA, M. G.; CIARELLI, P. M.; OLIVEIRA, E. Recommendation of Programming Activities by Multi-label Classification for a Formative Assessment of Students. *Expert Systems with Applications*, Elsevier, v. 40, n. 16, p. 6641–6651, 2013. Citado na página 36.

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. Citado 6 vezes nas páginas 25, 27, 29, 33, 35 e 60.

PAIVA, R. et al. Mineração de Dados e a Gestão Inteligente da Aprendizagem: Desafios e Direcionamentos. In: *I Workshop de Desafios da Computação Aplicada á Educação (DesafIE!2012)*. Curitiba (PR), Brasil: Sociedade Brasileira de Computação, 2012. v. 1. Citado 2 vezes nas páginas 22 e 35.

PÉREZ-MARÍN, D.; PASCUAL-NIETO, I.; RODRÍGUEZ, P. Computer-Assisted Assessment of Free-Text Answers. *The Knowledge Engineering Review*, Cambridge University Press, v. 24, n. 4, p. 353–374, 2009. Citado 2 vezes nas páginas 27 e 36.

PIROVANI, J. P. C. et al. Adapting NER (CRF+LG) for Many Textual Genres. In: *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019)*. Bilbao, Spain: CEUR-WS, 2019. (IberLEF - SEPLN 2019, v. 35), p. 421–433. Citado na página 52.

PISSINATI, E. M. *Uma Proposta de Correção Semi-Automática de Questões Discursivas e de Visualização de Atividades para Apoio à Atuação do Docente*. Dissertação (Mestrado) — PPGI - Universidade Federal do Espírito Santo, Vitória (ES), Brasil, Set 2014. Citado na página 75.

PRIBADI, F. S. et al. Automatic Short Answer Scoring Using Words Overlapping Methods. In: *AIP Conference Proceedings*. Bandung, Indonesia: AIP Publishing LLC, 2017. v. 1818, p. 020042:1–020042:6. Citado na página 39.

RAES, A. et al. A Systematic Literature Review on Synchronous Hybrid Learning: Gaps Identified. *Learning Environments Research*, Springer, v. 23, n. 3, p. 269–290, 2020. Citado na página 22.

RAMACHANDRAN, L.; CHENG, J.; FOLTZ, P. Identifying Patterns For Short Answer Scoring Using Graph-based Lexico-Semantic Text Matching. 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. 97–106. Citado 6 vezes nas páginas 38, 39, 41, 43, 83 e 92.

RAMACHANDRAN, L.; FOLTZ, P. Generating Reference Texts for Short Answer Scoring Using Graph-based Summarization. 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. 207–212. Citado 3 vezes nas páginas 29, 39 e 83.

RIBEIRO, M. T.; SINGH, S.; GUESTRIN, C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. San Francisco (CA), USA: Association for Computing Machinery, 2016. (KDD '16, v. 22), p. 1135–1144. Citado na página 66.

RIORDAN, B.; FLOR, M.; PUGH, R. How to Account for Mispellings: Quantifying the Benefit of Character Representations in Neural Content Scoring Models. 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. 116–126. Citado 5 vezes nas páginas 30, 40, 42, 43 e 90.

RIORDAN, B. et al. Investigating Neural Architectures for Short Answer Scoring. In: *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*. Copenhagen, Denmark: Association for Computational Linguistics, 2017. v. 12, p. 159–168. Citado 4 vezes nas páginas 73, 82, 83 e 92.

ROMERO, C. et al. *Handbook of Educational Data Mining*. 1st. ed. Boca Raton (FL), USA: CRC Press, 2010. Citado 2 vezes nas páginas 31 e 36.

ROUSSEEUW, P. J. Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis. *Journal of Computational and Applied Mathematics*, Elsevier, v. 20, n. 1, p. 53–65, 1987. Citado 2 vezes nas páginas 55 e 57.

ROY, S. et al. Wisdom of Students: A Consistent Automatic Short Answer Grading Technique. In: *Proceedings of the 13th International Conference on Natural Language Processing*. Varanasi, India: NLP Association of India, 2016. v. 13, p. 178–187. Citado 4 vezes nas páginas 39, 41, 83 e 93.

ROY, S.; RAJKUMAR, A.; NARAHARI, Y. Selection of Automatic Short Answer Grading Techniques Using Contextual Bandits for Different Evaluation Measures. *International Journal of Advances in Engineering Sciences and Applied Mathematics*, Springer, v. 10, n. 1, p. 105–113, 2018. Citado na página 43.

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. Citado 8 vezes nas páginas 28, 30, 33, 40, 41, 83, 91 e 92.

SAHA, S. et al. *Joint Multi-Domain Learning for Automatic Short Answer Grading*. New Delhi, India, 2019. Citado 2 vezes nas páginas 30 e 43.

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. Citado 8 vezes nas páginas 30, 33, 40, 41, 52, 67, 81 e 83.

SAKAGUCHI, K.; HEILMAN, M.; MADNANI, N. Effective Feature Integration for Automated Short Answer Scoring. In: *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Denver (CO), USA: Association for Computational Linguistics, 2015. v. 14, p. 1049–1054. Citado 2 vezes nas páginas 30 e 41.

SIDDIQI, R.; HARRISON, C. J. On the Automated Assessment of Short Free-Text Responses. In: *Proceedings of the 34th International Association for Educational Assessment Annual Conference*. Cambridge, United Kingdom: IAEA, 2008. (IAEA Conference, v. 34), p. 1–11. Citado na página 37.

SIEMENS, G.; BAKER, R. S. J. d. Learning Analytics and Educational Data Mining: Towards Communication and Collaboration. In: *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge*. Vancouver, Canada: Association for Computing Machinery, 2012. (LAK '12, v. 2), p. 252–254. Citado na página 36.

SILVA, C.; RIBEIRO, B. On Text-based Mining with Active Learning and Background Knowledge Using SVM. *Soft Computing*, Springer, v. 11, n. 6, p. 519–530, 2007. Citado na página 31.

SPALENZA, M. A. et al. Uma Ferramenta para Mineração de Dados Educacionais: Extração de Informação em Ambientes Virtuais de Aprendizagem. In: *Computer on the Beach*. Florianópolis (SC), Brasil: Universidade do Vale do Itajaí - UNIVALI, 2018. v. 9, p. 741–750. Citado na página 46.

SPALENZA, M. A. et al. Construção de mapas de características em classes de respostas discursivas. In: *Conferência Internacional sobre Informática na Educação (TISE 2016)*. Santiago, Chile: Centro de Computación y Comunicación para la Construcción del Conocimiento (C5), 2016. (TISE 2016, v. 12), p. 630–635. Nenhuma citação no texto.

SPALENZA, M. A. et al. Uso de Mapa de Características na Avaliação de Textos Curtos nos Ambientes Virtuais de Aprendizagem. In: *XXVII Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação*. Uberlândia (MG), Brazil: Sociedade Brasileira de Computação, 2016. (SBIE 2016, v. 27), p. 1165–1174. Citado 4 vezes nas páginas 27, 30, 66 e 69.

SPALENZA, M. A. et al. Using NER + ML to Automatically Detect Fake News. In: *Proceedings of the 20th International Conference on Intelligent Systems Design and*

Applications. Online Event: Springer International Publishing, 2020. (ISDA 2020, v. 20), p. 1176–1187. Citado 2 vezes nas páginas 52 e 53.

SPALENZA, M. A.; PIROVANI, J. P. C.; OLIVEIRA, E. de. Structures Discovering for Optimizing External Clustering Validation Metrics. In: *Proceedings of the 19th International Conference on Intelligent Systems Design and Applications*. Auburn (WA), USA: Springer International Publishing, 2019. (ISDA 2019, v. 19), p. 150–161. Citado 5 vezes nas páginas 31, 39, 55, 56 e 76.

STEIMEL, K.; RIORDAN, B. Towards Instance-Based Content Scoring with Pre-Trained Transformer Models. In: *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence*. New York (NY), USA: AAAI Press, 2020. v. 34, p. 1–3. Citado na página 90.

SULTAN, M. A.; SALAZAR, C.; SUMNER, T. Fast and Easy Short Answer Grading with High Accuracy. In: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. San Diego (CA), USA: Association for Computational Linguistics, 2016. v. 15, p. 1070–1075. Citado na página 41.

SUNG, C. et al. Pre-Training BERT on Domain Resources for Short Answer Grading. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, 2019. v. 9, p. 6071–6075. Citado 2 vezes nas páginas 39 e 43.

SUNG, C.; DHAMECHA, T. I.; MUKHI, N. Improving Short Answer Grading Using Transformer-Based Pre-training. In: *Proceedings of the 20th International Conference on Artificial Intelligence in Education*. Chicago (IL), USA: Springer, 2019. (AIED' 2019, v. 20), p. 469–481. Citado na página 41.

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. Citado 3 vezes nas páginas 33, 37 e 43.

TAN, H. et al. Automatic Short Answer Grading by Encoding Student Responses via a Graph Convolutional Network. *Interactive Learning Environments*, Taylor & Francis, v. 28, n. 1, p. 1–15, 2020. Citado 2 vezes nas páginas 40 e 43.

WANG, T. et al. Inject Rubrics into Short Answer Grading System. In: *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*. Hong Kong, China: Association for Computational Linguistics, 2019. (DeepLo 2019, v. 2), p. 175–182. Citado na página 39.

YANG, S. J. H. et al. Human-Centered Artificial Intelligence in Education: Seeing the Invisible through the Visible. *Computers and Education: Artificial Intelligence*, Elsevier, v. 2, n. 1, p. 100008, 2021. Citado 2 vezes nas páginas 35 e 42.

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. Citado 2 vezes nas páginas 29 e 33.

ZESCH, T.; HORBACH, A. ESCRITO - An NLP-Enhanced Educational Scoring Toolkit. In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki, Japan: European Language Resources Association (ELRA), 2018. v. 11, p. 2310–2316. Citado 2 vezes nas páginas 41 e 43.

ZHANG, Y.; LIN, C.; CHI, M. Going Deeper: Automatic Short-Answer Grading by Combining Student and Question Models. *User Modeling and User-Adapted Interaction*, Springer, v. 30, n. 1, p. 51–80, 2020. Citado 3 vezes nas páginas 39, 42 e 69.

ZHANG, Y.; SHAH, R.; CHI, M. Deep Learning+ Student Modeling+ Clustering: A Recipe for Effective Automatic Short Answer Grading. In: *Proceedings of the 9th International Conference on Educational Data Mining*. Raleigh (NC), USA: ERIC, 2016. (EDM 2016, v. 09), p. 562–567. Citado 2 vezes nas páginas 38 e 43.

ZIAI, R.; OTT, N.; MEURERS, D. Short Answer Assessment: Establishing Links Between Research Strands. In: *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*. Montreal, Canada: Association for Computational Linguistics, 2012. (NAACL HLT '12, v. 7), p. 190–200. Citado 2 vezes nas páginas 43 e 69.

Apêndices