Avaliação de Juízes: Um Modelo Estatístico para Perfilação de Avaliadores

James Alves [1], **Elias de Oliveira**[1]

[1] Programa de Pós-Graduação em Informática Universidade Federal do Espírito Santo (UFES) - 29.075-910 - Vitória - ES - Brasil

{james,elias}@lcad.inf.ufes.br

Abstract. *In this article we present a model to audict judges behaviour in the evaluation of essays according to ENEM's principles. Our model quantifies the reliability and the concordance among judges, based on the Pearson correlation coefficient applied both the general notes and, also, on the skills of the correction grid. In the obtained results we highlight the tendency of the judges to be more judicious in Competence 2, with a high degree of concordance, and with a low agreement degree, -0.58, among the judges for Competence 3. Competencies that often denote a sharp discrepancy are a sign of the need for training to align the evaluators.*

Resumo. *Neste artigo apresentamos um modelo para monitoramento de professores na avaliação de redações de acordo com as competências da redação do ENEM/INEP. Nosso modelo quantifica a confiabilidade e a concordância entre juízes baseado no coeficiente de correlação de Pearson aplicado às notas gerais, e também sobre as competências da grade de correção. Como resultado da aplicação do modelo observou-se a divergência entre dois professores na Competência 3, com um fator de correlação de* $\rho = -0.58$*. Ainda sim alguns avaliadores obtiveram uma alta concordância com* $\rho = 1$*, demostrando um alinhamento na forma de avaliar a Competência 2. As Competências que denotam frequentemente uma discrepância acintosa é um sinal da necessidade de treinamento para alinhamento dos avaliadores.*

1. Introdução

A quantificação da capacidade avaliativa entre professores é um desafio no meio da gestão educacional. Atualmente tem-se atenção especial a avaliação de redações do Exame Nacional do Ensino Médio (ENEM), exame promovido pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP). Para avaliar o conhecimento obtido pelo aluno ao longo de seu progresso na academia o ENEM/INEP tem como uma de suas etapas a escrita de uma redação com temas que exigem do aluno conhecimento de acontecimentos e situações diversas.

A prova de redação do ENEM/INEP requer do aluno um bom nível da língua portuguesa e tem peso no resultado geral do exame. O ENEM é um fator preponderante para ingresso em instituições de ensino superior, logo a forma como a redação é avaliada é de suma importância para o indivíduo avaliado. Portanto fica evidente a relevância de que o padrão da avaliação seja criterioso e que obtenha-se resultados de altíssima confiança e com mínimo de viés por parte dos avaliadores.

DOI: 10.5753/cbie.sbie.2019.199

A correção da redação do ENEM/INEP é um processo que atribui uma nota de 0 à 200 em cada uma das 5 competências, perfazendo um total máximo de 1000 pontos. Cada redação é avaliada por dois professores, a diferença entre a nota total dada por eles não pode ser superior a 100 pontos e nas competências 80 pontos. Existindo alguma discrepância um terceiro avaliador também fará a correção, prevalecendo assim as notas dos professores mais aproximadas, caso persista a diferença entre as avaliações um novo grupo de avaliadores é convocado [DAEB 2017].

Os custos de todo processo de correção do ENEM/INEP no Contrato nº 12/2016 foi de R$117.419.455,93 [Romão 2017]. A julgar pelo alto custo desse processo a aplicação de técnicas de monitoramento dos professores torna-se uma importante ferramenta para qualificar os avaliadores. A identificação prévia de discrepâncias entre o entendimento dos avaliadores torna esse processo mais confiável e com menores riscos de reavaliações, visto que a cada novo avaliador aumenta-se os custos com a correção.

O objetivo desse trabalho é propor um modelo estatístico para análise de confiabilidade e concordância entre avaliadores de redação em um ambiente virtual de aprendizado (AVA), vislumbrando identificar e mitigar as diferenças de entendimento avaliativo entre os professores. O trabalho de [Alves et al. 2018] traça uma estratégia para automatização do processo de análise de avaliação por pares e autoavaliação, entretanto neste trabalho usaremos apenas as notas atribuídas, em cada competência, pelos professores, às respectivas redações dos alunos. Para avaliação de redações utilizamos rubricas para formatar os critérios de juízo, dando aos avaliadores e avaliados uma visão clara de suas características que pautam a correção das redações [Arter and Chappuis 2006].

O trabalho está estruturado em 5 Seções. Na Seção 2 apresentamos os trabalhos relacionados com análise entre avaliadores, construção de rubricas e revisão por pares. A Seção 3 apresenta a metodologia utilizada no desenvolvimento da solução proposta neste trabalho. Na Seção 4 são explicados os resultados obtidos com a aplicação do modelo em um caso concreto, seguida pela Seção 5 que será discutido as considerações parciais e trabalhos futuros.

2. Trabalhos Relacionados

Para clarificar aos avaliados e avaliadores o que se espera que seja apresentado na redação propomos a utilização rubricas. Para [Arter and Chappuis 2006] a rubrica ajuda o avaliado a entender o quão bom é o pensamento crítico, e também para uma avaliação geral do objeto avaliado. Os autores [Arter and Chappuis 2006] advogam que para que o aprendizado possa ser reforçado é necessário que os objetivos a serem avaliados devam ser claros e que a rubrica reflita tais objetivos e que seus critérios devam ser descritos de acordo com a necessidade da avaliação.

Juntamente com a utilização de rubricas adotamos a técnica de revisão por pares para comparar as notas dadas pelos avaliadores para as redações dos alunos. Para [Smith 2006] esse tipo de avaliação tem um custo elevado, os avaliadores apresentam uma série de inconsistências e irregularidades ao avaliarem e, também, com frequência a avaliação segue alguma tendência pessoal, um viés, de quem avalia. Para mitigar a idiossincrasia dos avaliadores apresenta-se como solução a aplicação de treinamento prático com avaliadores mais experientes.

A necessidade de medir a concordância entre avaliadores se expressa direta-

mente no trabalho de [Cohen 1960] que quantifica a correspondência entre avaliações psicológicas para o diagnóstico de doenças mentais. Nesse caso o diagnóstico é realizado por psicólogos através de observação comportamental. Essa abordagem permitiu aos profissionais da área alinharem critérios patológicos de doenças de transtorno mental.

[Matos 2014] observa a aplicação das técnicas de avaliação de concordância e confiabilidade de juízes no Brasil. Aplicando a técnica de coeficiente de correlação intraclasse explicitou o descompasso significativo no entendimento entre avaliadores de redações mesmo utilizando critérios de objetivos para a correção. A falta de um processo automático e contínuo de qualificação concorre com a imprescindibilidade de equalizar o juízo dos avaliadores.

Por sua vez [Oliveira and Spalenza 2017] propõe um modelo que busca explicitar através de uma medida estatística a avaliação e autoavaliação de atividades entre os discentes e os docentes, onde professores e alunos avaliam todas as respostas submetidas no AVA. O método proposto explicita a eficácia do aprendizado e os alunos que têm alguma carência de atenção por não diferenciar as repostas certas de respostas erradas.

A abordagem proposta por [Alves et al. 2018] utiliza a correlação para perfilar alunos em uma ação automática a partir de dados obtidos através do AVA. Para analisar os alunos de acordo com seu perfil traça-se um agrupamento de acordo com a semelhança entre a nota dada pelo aluno e o gabarito. Entende-se como gabarito a avaliação da atividade realizada pelo professor, assim perfilando em grupos os alunos que não entenderam e grupo de alunos com alto grau de confiabilidade com o professor.

3. Ferramentas e Métodos

Para se obter os indicadores da avaliação fez-se necessário a adoção de um processo automatizado para receber as redações e corrigi-las dentro do AVA. Primeiramente o aluno submete a redação através do AVA que por sua vez que replica desta e é enviada para todos os avaliadores envolvidos no processo de correção juntamente com a rubrica que foi vinculada previamente a atividade. Os avaliadores corrigem as redações utilizando a rubrica e ao final são consolidadas todas as notas dos professores envolvidos e a análise do processo estatístico é iniciado.

Com as análises do modelo em mãos será possível caracterizar a concordância ou discordância entre os avaliadores. Portanto a combinação do ferramental estatístico com rubricas torna-se possível pontuar estritamente as dimensões em que ocorrem diferenças de entendimento entre os avaliadores.

Para este trabalho foi utilizado como AVA o sistema Moodle[1], que oferece nativamente suporte a utilização de rubricas como método avaliativo de atividades. Para extração dos dados do AVA utilizamos o Plugin[2], uma ferramenta proposta por [Spalenza et al. 2018]. Desenvolvemos um programa para aplicação do modelo de monitoramento utilizando a linguagem de programação R[3] devido seu ferramental estatístico acessível.

[1]https://moodle.org/
[2]https://gitlab.com/rii_lcad/plugin
[3]https://cran.r-project.org/

3.1. Rubricas

É esperado que uma rubrica descreva o que deve ser alcançado em termos de qualidade do desempenho pretendido e também o indicador numérico de cada nível de desempenho. De maneira geral rubricas tem a dimensão de qualidade, que é o que pretende-se avaliar, e a dimensão dos qualificadores, que é a escala de pontuação para cada item de qualidade [Carvalho and Fernandes 2012]. Adotamos como dimensão de qualidade as competências e os qualificadores são os critérios avaliativos de cada competência.

Para formular os critérios avaliativos desse trabalho tomamos como base o manual de correção de redação do ENEM/INEP. Para inclusão do modelo no AVA o adaptamos reduzindo a faixa dos critérios de 0 à 200 para 0 à 20. O valor máximo que uma redação pode chegar será de até 100 pontos. Além disso, nesse trabalho não descreveremos os critérios de cada nota, estas descrições estão disponíveis diretamente no manual de correção disponibilizado pelo INEP/ENEM[4].

A Tabela 1 mostra os aspectos que serão avaliados. Para cada redação as competências podem receber o seguinte espectro de notas para as características esperadas: 0 para muito baixa ou ausente, 5 para baixa, 10 quando boa, 15 muito boa e 20 para redação excelente. Após a avaliação de todos os itens as notas são somadas para a formulação da nota final da redação.

Os avaliadores não conhecem as respostas de seus pares antes da análise pelo modelo. Com esse princípio garantimos que não existem influências de terceiros nas avaliações realizadas por esses avaliadores.

Competências	**Faixa de Notas**
Demonstrar um bom nível de escrita formal da língua portuguesa.	0 - 20
Apresenta domínio precário do texto dissertativo-argumentativo apresentando apenas duas características do texto dissertativo – argumentativo, introdução, desenvolvimento ou conclusão. Também o texto apresenta, na maior parte das linhas, apenas orações além de não conter um repertório técnico – científico e os argumentos serem embasados no senso comum	0 - 20
Definir, argumentar, hierarquizar e organizar informações com o intuito de responder o problema de pesquisa apresentado	0 - 20
Elementos de coesão	0 - 20
Resultado da Proposta de Solução	0 - 20
Nota final da redação	0 - 100

Tabela 1. Rubricas para avaliação de redações.

3.2. Modelo de Monitoramento de Concordância

Inicialmente o modelo consiste em representar as avaliações realizadas pelos professores de forma matricial. Em (1) a matriz *N* é criada para cada competência c (descrito como: N_c), onde cada professor *i* atribui uma nota n ao aluno a. As notas finais também são

[4]http://download.inep.gov.br/educacao_basica/enem/guia_participante/2016/manual_de_redacao_do_enem_2016.pdf

transformadas nesse modelo. Logo faz-se uma comparação entre conjunto de avaliadores experientes contra os novatos para uma mesma competência.

$$N_c = \begin{bmatrix} n_{11} & n_{21} & n_{31} & \dots & n_{i1} \\ n_{12} & n_{22} & n_{32} & \dots & n_{i2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ n_{1a} & n_{2a} & n_{3a} & \dots & n_{ia} \end{bmatrix} \tag{1}$$

Sobre cada N_c aplica-se os métodos de correlação. Formalmente podemos descrever correlação como uma medida da proporção de mudanças entre duas classes, a classe para esta abordagem são todas as notas imputadas por um professor $n_i = \{n_{i1}, n_{i2}, n_{i3}, ..., n_{ia}\}$ para a competência c. Existem diversos modelos de correlação, este trabalho aplica o modelo de correlação de Pearson [Bussab and Morettin 2013]. Logo, em 2, obtêm-se a quantificação da associação ρ entre os docentes n_i e n_j onde $i \neq j$ para os a alunos avaliados. O valor da correlação será $-1 \leq \rho \leq 1$.

$$\rho = \frac{\mathrm{cov}(n_i, n_j)}{\sqrt{\mathrm{var}(n_i) \cdot \mathrm{var}(n_j)}} \tag{2}$$

A Tabela 2 é um exemplo de como as notas de uma competência ficam organizadas na representação matricial. A correlação positiva próximo do valor 1 significará que as notas atribuídas por um par de professores estão bem correlacionadas, ao passo que valores negativos de correlação apontarão que um par de professores divergem em suas respectivas formas de dar notas.

Os avaliadores n1 e n2 apresentam diferenças nas notas atribuídas aos alunos, porém o comportamento da nota dos dois é o mesmo. O avaliador 1, $n_1 = \{0, 40, 120\}$, é comparado com o avaliador 2, $n_2 = \{120, 160, 200\}$, o que resulta em $\rho = 0.98$. O ρ positivo próximo a 1 mostra uma correlação positiva alta pelo comportamento de notas crescentes entre os pares, mesmo verificando-se uma diferença considerável entre as notas dadas por eles. A mesma comparação entre os avaliadores n1 e n3 apresenta uma relação invertida de entendimento, tendo $\rho = -0.98$ mostra uma correlação negativa, enquanto um professor aumenta as notas o outro diminui.

n1	**n2**	**n3**
0	120	120
40	160	80
120	200	40

Tabela 2. Exemplo de organização de notas para cada competência e total.

Para uma visualização adequada dos resultados utilizamos o gráfico de *heapmap* exemplificado na Figura 1, mostrando de maneira facilitada o grau de confiabilidade das avaliações da Tabela 1. Nessa visualização, quanto mais o quadrante for azul, melhor correlacionado estarão os professores, por exemplo: os avaliadores n1 e n2, na secunda coluna na primeira linha da Figura. Observe que na Tabela 2 ambos deram notas crescente. Nos quadrantes em vermelho estão os avaliadores que estão destoando entre si,

caso dos avaliadores n1 e n3 na célula da 3ª coluna na 2ª linha da Figura. Nesse caso observando novamente a Tabela 2 os avaliadores tem uma perspectiva completamente inversa ao dar notas.

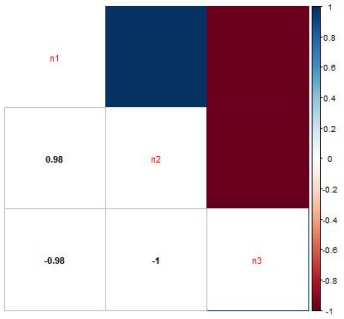


Figura 1. Exemplo de resultado a partir da análise da Tabela 2

3.3. Modelo de Inferência

Ao analisarmos a correlação das notas atribuídas às redações pelos professores mostramos que é possível detectar através da correlação aspectos do comportamento dos avaliadores ao imputar uma nota a redação. Entretanto a correlação é insuficiente para detectar as diferenças entre as notas dos avaliadores segundo as restrições do INEP/ENEM, que conforme visto da Seção anterior mesmo obtendo um valor alto de correlação ($\rho = 0.98$) percebemos que ainda assim a diferença das notas dos avaliadores superaram a limitação de 80 pontos de diferença nas competências.

Supomos ter avaliadores que são especialistas na correção da redação segundo o modelo INEP/ENEM e que satisfaçam as exigências de correção. Tendo eles corrigido as mesmas redações em uma competência c teremos um conjunto de notas $m_{\cdot a}$ que tem média μ_a e desvio padrão σ_a. Caso outro avaliador dê uma nota $n_{\cdot a}$ para a mesma redação gostaríamos saber a probabilidade desse novo avaliador dar nota conforme a distribuição de $m_{\cdot a}$.

Para obter a probabilidade formularemos uma regra de decisão que facilitará a tomada de decisão. Sabe-se que existem duas possibilidades de erros para o modelo. Enumeramos esses erros da seguinte maneira:

Erro tipo I : dizer que a nota é padrão quando não é.
Erro tipo II : dizer que a nota não é padrão quando é.

Também podemos enumerar as hipóteses que temos sobre a nota dada pelo professor:

H_0 :A nota é padrão.
H_1 :A nota não é padrão.

Assim podemos definir pontos x_1 e x_2 para delimitar uma região de aceitação ra onde H_1 será rejeitado, dado por $ra = \{n_{\cdot a} \in \mathbb{R} | x_1 \leq n_{\cdot a} \leq x_2\}$. Os valores que podem

ser assumidos por $x.$ estão entre 0 e 200. Agora podemos formular a probabilidade de cada erro em função da ra postulada:

$$P(\textit{Erro I}) = P(n_{.a} \notin ra | H_0 \text{ é verdade}) = \alpha$$
$$P(\textit{Erro II}) = P(n_{.a} \in ra | H_1 \text{ é verdade}) = \beta$$

Para confirmar que um professor dá nota conforme um especialista precisamos determinar a probabilidade β. Aplicamos a proposta de [Bussab and Morettin 2013] que define um $\alpha = 0.05$ afim de determinar os pontos que delimitam ra que serão usados como parâmetros para obter-se $\beta(n_{.a})$. A conhecer que $\beta = 1-\alpha$, então quando $\beta(n_{.a}) < 0.5$, $n_{.a}$ não deu nota conforme os especialistas para a redação observada.

Usaremos a transformação normal padrão (3) para obtermos x, onde Z é uma variável aleatória com $\mu = 0$ e $\sigma = 1$. A transformação (3) é fundamental para calcular probabilidades relativas de qualquer distribuição normal.

$$Z = \frac{m_{ic} - \mu}{\sigma} \tag{3}$$

Exemplificando todo o processo observamos a Tabela 2. Consideramos que todas as notas de n1 são médias de todos os avaliadores especialistas e tenhamos calculado previamente o $\sigma_1 = \{40.82, 40.82, 40.82\}$. Selecionamos $a = 2$, logo $\mu_c = 40$ e $\sigma_c = 40.82$ e então encontramos os ponto x_1 e x_2 para $P(Z) = \alpha \div 2$ utilizando 3, observando a simetria da curva gaussiana para x_1 teremos $P(Z) = -\alpha \div 2$ logo:

$$\begin{aligned} P(Z) =& (x_1 - 40) \div 40.82 \\ x_1 =& (-1.96 \times 40.82) + 40 \\ x_1 =& -40.0 \end{aligned}$$

Observando as restrições do intervalo de notas possíveis então $x_1 = 0$ e pela simetria da curva gaussiana teremos $x_2 = 120$. Dado os pontos do intervalo então temos $ra = \{n_{.a} \in \mathbb{R} | 0 \leq n_{.a} \leq 120\}$. Agora precisamos calcular a probabilidade $\beta(n_{.a}) = P(n_{.a} \in ra | H_0 \text{ é verdade})$ dado um valor de $n_{.a}$, que será usado como μ da função característica de β dado por: $\beta(\mu) = P(\bar{X} \in ra | \mu)$, e para todos os teste dessa distribuição teremos $\sigma = 40.82$. Para esta redação podemos testar se outras notas pertencem a distribuição. Para a nota n_{22} teremos:

$$\begin{aligned} \beta(160) &= P(X \notin RC | \mu = 160) \\ &= P(-1.96 \leq Z \leq 1.96) = 16.35\% \end{aligned}$$

Com o resultado obtido podemos afirmar que com apenas $\beta(160) = 16.35\%$ o avaliador n_2 não dá notas conforme os especialistas para $a = 2$. Aplicando o mesmo modelo a n_{32} temos $\beta(80) = 83.48\%$, logo n_2 tem maior probabilidade de dar notas como os especialistas para $a = 2$.

Para visualizar o resultado da execução do procedimento para todas as notas da Tabela 2 utilizaremos um gráfico de radar, disposto na Figura 2. Na Figura quanto mais próximo a nota da redação esta do circulo do centro maior a probabilidade do avaliador dar notas conforme o especialista em uma determinada redação.

Figura 2. Exemplo Probabilidades de dar nota como um especialista.

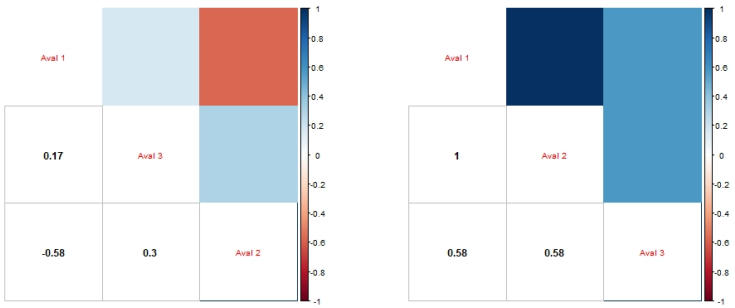
4. Experimento e Resultados

Em uma turma escolar durante o segundo semestre de 2018, foram indicados 6 temas de redações. Ao todo foram coletados 44 redações para correção e participaram do experimento 3 avaliadores, cada um analisou de maneira independente as redações submetidas.

4.1. Correlação entre avaliadores

Sobre a amostra selecionada analisamos os dados com o modelo proposto e obtivemos como resultado a Figura 3 para comparamos os resultados. Na Figura 3(a) vemos que na Competência 3 existe uma divergência entre o entendimento dos avaliadores, evidenciado na célula da coluna 2 na linha 1. No caso os avaliadores aval1 e aval3 discordam nas notas dadas na Competência 3. Entre dois avaliadores há alguma concordância, mesmo que baixa. No caso o Aval 2 comporta-se como o avaliador mais bem correlacionado com os outros avaliadores, visto que os quadrantes que o compara com outros avaliadores se encontram em azul.

Seguindo para a competência 2 na Figura 3(b) mostra uma melhor sintonia entre os avaliadores aja visto que todas as células estão em azul. Observando estes comportamentos é possível dizer que é necessário uma atenção especial na explicação de como avaliar a competência 3. Mesmo os avaliadores mais bem correlacionados denotam pouca convergência ao avaliarem as redações.



(a) (b)

Figura 3. Heapmap da correlação por competência

4.2. Inferência de Avaliadores

Adaptamos o modelo para este trabalho utilizando o Moodle como ferramentar para receber as redações. Como exposto da Seção 3.1 transformamos a faixa possíveis para valores de 0 à 20, também adaptamos a restrição de diferença aceitável entre os avaliadores. A diferença entre as notas das competências deve ser inferior a 5 e a diferença entre as notas finais deve ser inferior a 10.

Também contamos com apenas um avaliador que foi treinado para correção de redações. Logo assumimos que μ_c é a nota desse avaliador. Com base no μ_c proposto calculamos um σ_c que ainda atende as restrições adaptadas adotando (4):

$$P\left(\frac{\mu_c - 5}{\sigma_c} \leq \mu_c \leq \frac{\mu_c + 5}{\sigma_c}\right) = 0.95 \tag{4}$$

Determinado o σ_c aplicamos nosso modelo aplicamos a matriz de cada competência e a matriz de notas finais. Apresentamos a baixo os resultados para $c = 5$. O resultado exposto na Figura 4 mostra que Avaliador 3 na redação 3 expressa uma menor probabilidade de avaliar redações como um especialista.

Figura 4. Probabilidades de dar nota como um especialista para a Competência 1.

Observe que para a Competência 1 o Avaliador 2 não esta fora de rc denotando que esta fora das restrições de avaliações. Esse resultado indica a necessidade de aprofundar-se na avaliação da competência.

5. Considerações Finais e Trabalhos Futuros

Este trabalho apresentou modelo para identificação de correlação entre múltiplos avaliadores de maneira automática. O modelo busca identificar as semelhanças e diferenças entre as notas de uma mesma competência para avaliadores de redações utilizando um modelo de rubricas.

Os professores avaliaram as mesmas redações para que pudéssemos realizar a aplicação do modelo; fixando portanto um mesmo contexto a todos eles. Dentro dessas condições o monitoramento de cada competência da rubrica mostrou claramente os pontos da diferença entre os professores na inserção de notas em redações de acordo com os critérios do ENEM/INEP. Nossos instrumentos de visualização expôs com clareza os pontos onde é necessário a conciliação de entendimento sobre a forma de correção. A abordagem automatizada é um diferencial desse trabalho pois viabiliza a aplicação periódica de testes entre avaliadores, e uma verificação sistemática dos métodos avaliativos esperados.

Para trabalhos futuros se faz necessária a adoção de uma abordagem que identifique com rigor as exatas notas que estão diferentes e a proporção das diferenças de cada aluno avaliado, para que assim as discussões possam ser pautadas estritamente nas redações com diferenças de compreensão dos avaliadores.

Agradecimentos

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001

Referências

Alves, J., Pereira, W., Brito, O., and Oliveira, E. D. (2018). Avaliação em Pares e Autoavaliação: Um Modelo Estatístico Para Perfilação de Alunos. In *Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)*, volume 29, pages 1653–1662.

Arter, J. A. and Chappuis, J. (2006). *Creating & Recognizing Quality Rubrics*. Assessment Training Institute, Inc Series. Pearson Education, Nee York, USA.

Bussab, W. O. and Morettin, P. A. (2013). *Estatística Básica*. Saraiva, São Paulo, 8 edition.

Carvalho, R. S. and Fernandes, C. T. (2012). Easy Rubric: um Editor de Rubricas no Padrão IMS Rubric. *Anais do Workshop do Congresso Brasileiro de Informática na Educação*, pages 10–11.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

DAEB, D. D. A. D. E. B. (2017). Redação No Enem 2017 Cartilha Do Participante.

Matos, D. A. S. (2014). Confiabilidade e concordância entre juízes : aplicações na área educacional. *Estudos em Avaliação Educacional*, 25(59):298–324.

Oliveira, E. and Spalenza, M. (2017). Self and peer assessment strategies. *Anais do Computer on the Beach*.

Romão, C. (2017). Proposta de um sistema automático de avaliação de redações do enem, foco na competência 1: Demonstrar domíno da modalidade escrita formal da língua portuguesa. Master's thesis, Universidade Federal do Espírito Santo.

Smith, R. (2006). Peer review: A flawed process at the heart of science and journals. *Journal of the Royal Society of Medicine*, 99(4):178–182.

Spalenza, M. A., Nogueira, M. A., de Andrade, L. B., and de Oliveira, E. (2018). Uma Ferramenta para Mineração de Dados Educacionais: Extração de Informação em Ambientes Virtuais de Aprendizagem. *Anais do Computer on the Beach*, pages 741–750.