Avaliação por pares como ferramenta colaborativa na correção de redações: Um experimento com o ambiente educacional MeuTutor

Thyago Tenório[1], **Ig Ibert Bittencourt**[2]

[1]Universidade Federal de Alagoas (UFAL) – Campus Arapiraca/Pólo Penedo
Av. Beira Rio – 57.200-000 – Penedo – AL – Brasil

[2]Instituto de Computação (IC) – Universidade Federal de Alagoas (UFAL)
Av. Lourival Melo Mota, s/n – 57.072-900 – Maceió – AL – Brasil

thyago.oliveira@penedo.ufal.br, ig.ibert@ic.ufal.br

Abstract. *While on-line learning environments provide scalable ways to present content, forums and evaluate student progress automatically, they are still limited in their ability to evaluate complex activities such as essays correction. In this kind of activity, the manually correction by teachers / tutors would quickly become infeasible due to the increased number of students. This paper presents a peer assessment mechanism and applies it in essays correction in MeuTutor educational environment. In the experiment conducted, it was concluded that there were no statistical variations between the grades obtained compared to the experts, the correction time was reduced, and the cost involved decreased 72.4 %.*

Resumo. *Embora os ambientes de aprendizagem online ofereçam maneiras escaláveis para apresentar um conteúdo, fóruns e avaliar o progresso dos alunos automaticamente, eles continuam limitados na capacidade de avaliar atividades complexas, como correção de redações. Nesse tipo de atividade, a correção de forma manual pelos professores/tutores se tornaria rapidamente inviável a medida que o número de atividades cresce. Este trabalho apresenta um mecanismo de avaliação por pares e aplica-o no contexto de redações no ambiente educacional MeuTutor. Com o experimento realizado, concluiu-se que não houve variações estatísticas entre as notas obtidas se comparado à especialistas, o tempo de correção foi reduzido, assim como o custo envolvido diminuiu 72.4%.*

1. Introdução

A aplicação de tecnologias da informação e comunicação (TIC) na educação tem sido cada vez mais destacada no processo de ensino-aprendizagem. Nos últimos anos, diversos países têm adaptado suas abordagens educacionais para promover e suportar a utilização de tecnologias, tanto em cursos presenciais quanto em cursos a distância. Contudo, enquanto novas tecnologias da Web permitem formas escaláveis para fornecer conteúdo de vídeo conferência, implementar fóruns sociais e acompanhar o progresso do aluno nestes ambientes, continuamos limitados em nossa capacidade de avaliar e dar feedback para trabalhos complexos dos estudantes e muitas vezes abertos (avaliações escritas), tais como provas matemáticas, projeto de problemas e redações [Piech et al. 2013].

Diante de tal dificuldade, estes ambientes educacionais online frequentemente oferecem soluções mais simples aos seus usuários, como, por exemplo, questões de múltipla escolha, uma vez que este tipo de problema pode ser processado de forma automática pelo computador. Com o crescimento do número de estudantes nestes ambientes e consequentemente, um maior número de atividades sendo feitas a cada momento, a correção delas pelos professores/tutores se tornaria rapidamente inviável. Para suprir este problema pode-se utilizar um conceito chamado de avaliação por pares. Avaliação por pares é um processo pelo qual os alunos ou seus pares atribuem notas com base em parâmetros de referência pré-definidos pelo professor [Sadler and Good 2006]. A prática é utilizada para salvar o tempo dos professores e melhorar a compreensão dos estudantes a respeito dos materiais do curso, bem como melhorar as suas habilidades meta-cognitivas [Malehorn 1994].

No contexto de ambientes de EAD, o sistema gerencia as correções feitas entre os próprios estudantes, isto é, os próprios alunos corrigem as atividades escritas de seus pares. Existem diversas abordagens de avaliação por pares publicados na literatura, como as abordagens formativas [Orsmond* et al. 2004], modelos probabilísticos [Piech et al. 2013] e até mesmo modelos usando redes bayesianas [Wang and Vassileva 2003]. Como consequência, o uso desta técnica resulta em um aumento da aprendizagem do aluno e, principalmente, a redução de sobrecarga do professor [Sadler and Good 2006]. No entanto, o uso de modelos de avaliação por pares aplicados em ambientes educacionais online com o propósito de correções de atividades escritas (redações) ainda é limitado. Os resultados de sua aplicação são fortemente ligados ao contexto envolvido e dependem de vários fatores externos, incluindo, por exemplo, o nível educacional dos avaliadores e as atividades que serão corrigidas.

Neste sentido, a solução proposta visa criar um mecanismo de avaliação por pares e aplicá-lo no contexto de provas discursivas nestes ambientes. Usando este conceito em um ambiente educacional EAD, o sistema ficará responsável pelo gerenciamento das correções que serão feitas pelos próprios estudantes. Usando avaliação por pares é possível desatrelar o custo das correções do número de usuários, e sem necessariamente aumentar o trabalho do professor. Ao se utilizar essa técnica, surgem vários problemas para o bom funcionamento da colaboração entre esses alunos, como por exemplo ”como selecionar os pares mais adequados para a correção?”ou ”como garantir um nível alto de confiança e reputação geral dos avaliadores?”. Na solução aqui apresentada é incluído mecanismos para solucionar tais problemas dentro do modelo proposto.

Um experimento foi conduzido com o objetivo de avaliar o mecanismo de avaliação por pares criado e apresentado neste trabalho. Dessa forma, este foi implementado dentro de um ambiente educacional chamado MeuTutor. Nesse experimento, cerca de 30 alunos tiveram acesso ao sistema e tinham como objetivo realizar algumas redações que foram liberadas no ambiente para serem feitas e, em seguida, corrigir tais redações utilizando o mecanismo de avaliação por pares proposto, assim como o método tradicional (em que as correções foram feitas por professores especialistas). Como resultado, não houve variações estatísticas significativas entre as notas obtidas em ambos os modelos, o tempo de correção usando o modelo proposto se mostrou mais rápido que o tempo de correção por especialista e houveram evidências estatísticas suficientes que indicaram que o custo envolvido do modelo proposto se mostrou menor que o custo do modelo tra-

dicional. Através da criação de um modelo de regressão apresentado, mostrou-se que a redução no custo é de cerca de 72.4%.

Este documento está estruturado da seguinte forma: A Seção 2 apresenta alguns trabalhos relacionados. Em seguida, na Seção 3 é apresentado de forma resumida o ambiente educacional MeuTutor. Logo após, a Seção 4 apresenta o mecanismo de avaliação por pares criado. Finalmente, a Seção 5 apresenta o planejamento e execução do experimento que foi conduzido com o objetivo de avaliar o mecanismo criado e a Seção 6 apresenta as nossas conclusões, limitações e alguns trabalhos futuros.

2. Trabalhos relacionados

Os trabalhos discutidos aqui envolvem o uso de avaliação por pares em ambientes educacionais. Podemos citar alguns trabalhos como em [Dominguez et al. 2012] em que é aplicada a técnica para avaliação de trabalhos em um curso de engenharia. [Chang et al. 2012] aplica com o objetivo de avaliar portfólios em estudantes do ensino médio que cursam computação. Finalmente em [Kawai 2006] é utilizado a técnica para avaliação de mensagens de voz e cartas para alunos estudantes de língua estrangeira. Neste trabalho, a técnica é utilizada de maneira abstrata, no qual um modelo de utilização é proposto e que pode ser utilizado para correções de avaliações escritas em geral, independente de qual tipo de avaliação específica será realizada pelo professor.

Em [Sterbini and Temperini 2013] é apresentada uma ferramenta chamada OpenAnswer que tem como objetivo permitir o suporte a avaliações escritas usando técnicas de avaliações por pares, algo semelhante ao proposto aqui nesse trabalho. Contudo, essa ferramenta não é integrada ao ambiente, o que implicaria que os usuários tivessem que utilizar mais uma plataforma só para as atividades de avaliações escritas. Em nossa proposta, pretendemos que o uso de avaliação por pares seja incluído direto no ambiente, sem essa necessidade de outros sistemas. Dessa forma o trabalho aqui proposto não apresenta uma ferramenta que utiliza avaliação por pares e sim um modelo de avaliação por pares que poderá ser utilizado nos ambientes educacionais.

Em [Kahiigi Kigozi et al. 2012], os autores projetaram e modelaram um processo de revisão baseado em avaliação por pares para aprendizagem colaborativa. O modelo é baseado no uso do estudante, apoiando-o pedagogicamente sua aprendizagem. Em [Tosic and Nejkovic 2010], é proposto um novo método para avaliação por pares dos alunos com base no conceito de confiança. Outros trabalhos propõe modelos baseados em redes bayesianas como em [Sterbini and Temperini 2013] e [Wang and Vassileva 2003]. Todos esses modelos são implementados de forma própria e utilizados de acordo com suas necessidades, porém não tem como objetivo se integrar com os diversos ambientes educacionais e, portanto, se diferem desse trabalho nesse aspecto.

3. O ambiente educacional MeuTutor

O modelo foi implementado e integrado ao ambiente educacional MeuTutor. Podemos defini-lo como um sistema tutor inteligente, que tem como objetivo acompanhar a aprendizagem dos alunos de forma personalizada, garantindo uma qualidade no ensino e melhorando o desempenho dos seus usuários. O ambiente visa auxiliar alunos do ensino médio a se prepararem para o Exame Nacional do Ensino Médio (ENEM). Com isso, oferece cursos referentes às disciplinas do ensino médio.

O ambiente MeuTutor-ENEM foi construído tendo como ponto forte três bases: ambiente gamificado, aprendizagem personalizada e experiência social. O primeiro tem como objetivo motivar o aluno a continuar seus estudos no ambiente. O objetivo principal da Gamificação é aumentar o engajamento dos usuários por meio do uso de técnicas semelhantes àquelas presentes em jogos, fazendo com que os usuários se sintam no controle de suas ações e se motivem com as tarefas [Pavlus 2010]. Com isso, é possível observar na Figura 1 alguns elementos de Gamificação (jogos) presentes na plataforma, que foi projetada como se fosse um jogo para motivar a aprendizagem e o estudo do aluno. A plataforma disponibiliza um mecanismo de pontos e níveis personalizados, para o aluno evoluir em cada disciplina e no geral (na figura o aluno está no nível 5).

Figura 1. Elementos de gamificação do MeuTutor

Sob o segundo aspecto, a aprendizagem personalizada, no MeuTutor-ENEM as questões e vídeo-aulas se adequam aos alunos. Por meio de uma base de questões no estilo ENEM, o estudante pode praticar seu conhecimento resolvendo vários problemas de cada assunto nas disciplinas disponíveis no sistema. À medida que ele resolve corretamente um problema, seu progresso no assunto cresce. Se errar a questão, uma vídeo-aula será recomendada para ele suprir sua deficiência.

Por fim, no terceiro aspecto (experiência social), o MeuTutor mantém a preferência dos alunos tradicionais (discutir e estudar um determinado assunto ou conteúdo juntos, construindo conhecimento compartilhado), fornecendo recursos poderosos dentro do ambiente por meio de técnicas da aprendizagem colaborativa em sistemas computacionais. Entre os recursos colaborativos providos aos estudantes está a ideia de grupos de estudo. Dessa forma, um aluno pode convidar outros amigos para praticar um determinado assunto. A Figura 2 mostra a tela de grupo de estudos, em que os alunos respondem questões juntos depois de debater e entrar em um determinado consenso.

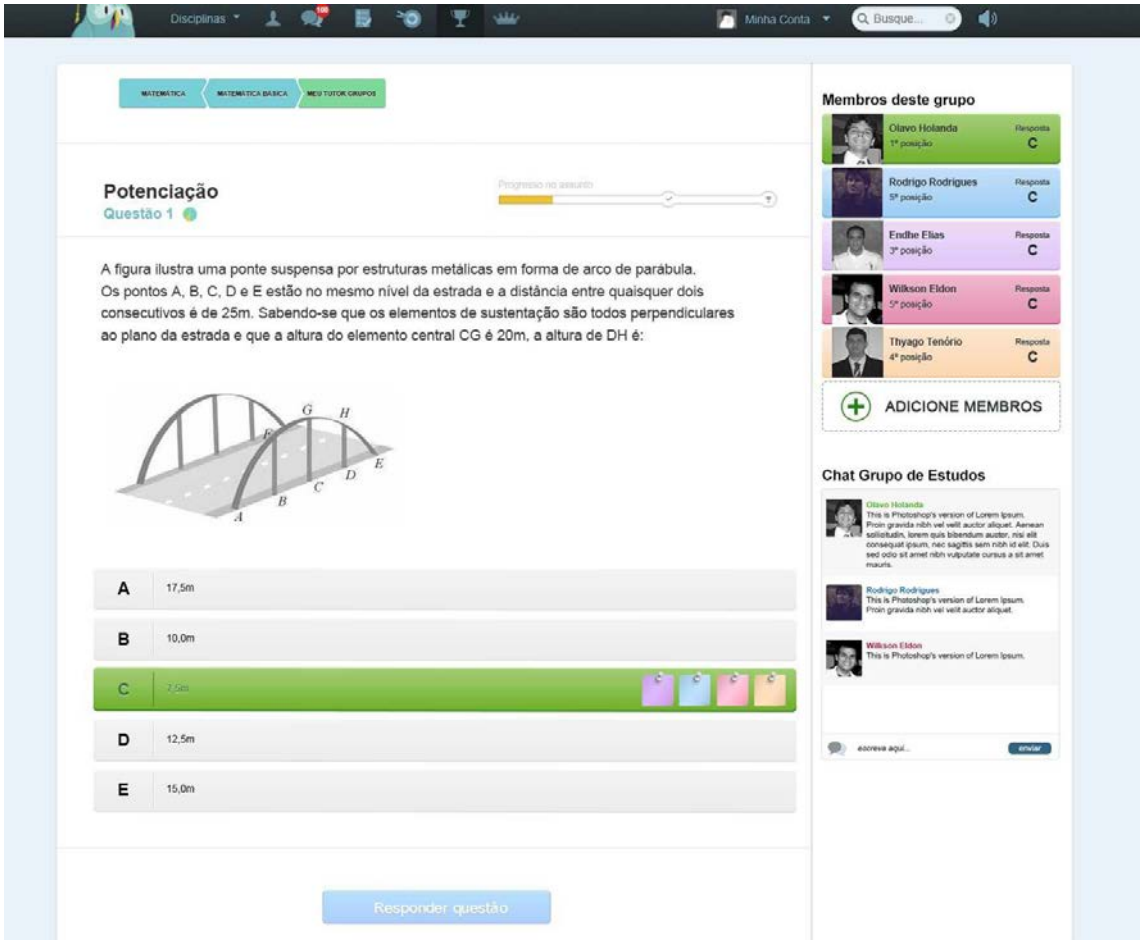

Figura 2. Estudos em grupo - Experiência social do MeuTutor

4. Mecanismo Avaliação por pares

Para entender melhor como o mecanismo se comporta dentro do sistema, a Figura 3 apresenta o fluxo de execução que deve ser incrementado no ambiente educacional.

O primeiro passo é a criação das atividades discursivas feitas pelo professor juntamente com a criação e definição do formulário de avaliação, contendo os critérios no qual as atividades serão avaliadas pelos alunos. As atividades e os formulários deverão, em seguida, serem enviados (cadastrados) ao sistema pelo professor em uma interface específica para isso. O sistema por sua vez irá disponibilizar as atividades aos alunos de acordo com o plano de avaliação pré-definido. A partir da disponibilização da atividade ao aluno para ser respondida, inicia-se o período de submissões - segundo passo. Esse período também está definido no plano de avaliação. Os estudantes responderão as atividades no próprio ambiente online.

Em seguida, o aluno submeterá sua resposta para o ambiente e este será responsável pelo gerenciamento da correção desta. No terceiro passo, em primeiro lugar, a gerência do sistema identifica os alunos que irão corrigir a atividade submetida e, a partir desse momento, o sistema enviará a atividade para estes alunos. Esse processo de escolha dos corretores deve ser feito por meio de uma lista de usuários que fizeram a mesma atividade e também possui atividades pendentes de correção, em ordem crescente

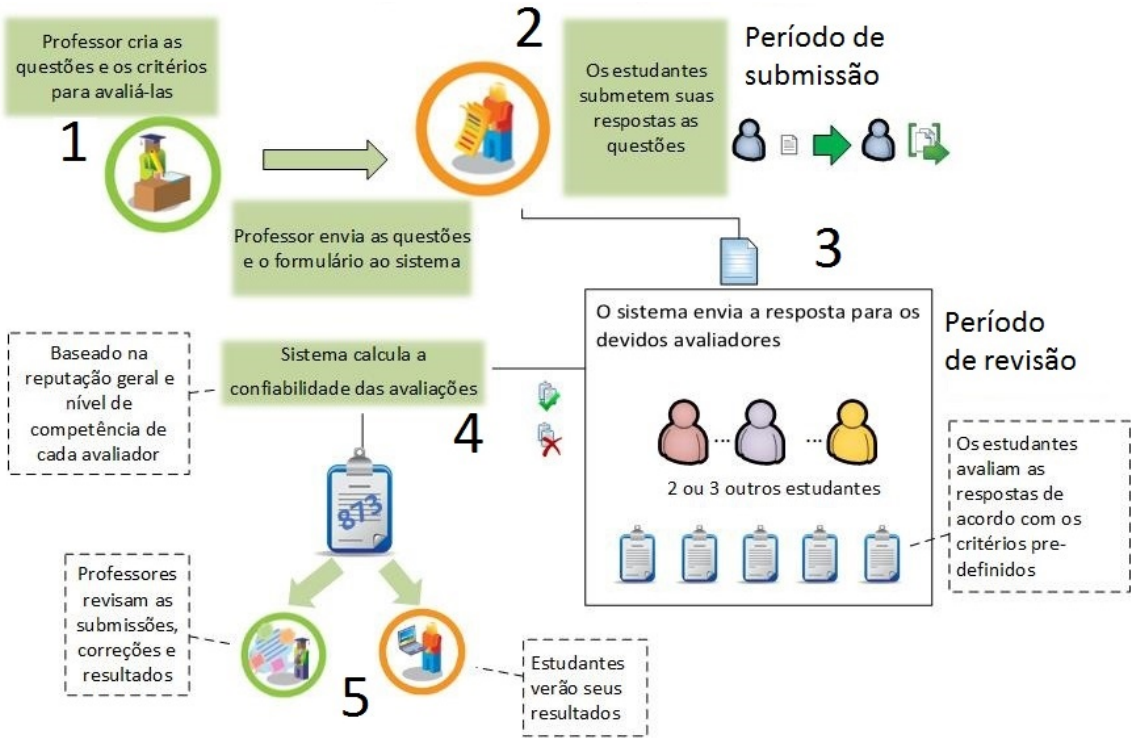

Figura 3. Fluxo de execução do modelo em detalhes

de atividades atribuídas, de acordo com o algoritmo a seguir.

É de salientar que o número mínimo de correções necessárias (minCorrection do Passo 1) para uma atividade é muito dependente do tipo de atividade que está sendo executada. Teoricamente, um maior número torna mais confiável os resultados da avaliação. Por outro lado, um maior número gera mais trabalho para os alunos, o que pode ser prejudicial para o resultado final do modelo. Portanto, há um trade-off na escolha deste número. O número de margem de segurança (safetyMargin a partir do Passo 2) serve como um armazenamento de usuários que serão recuperados para análise. Quanto maior o número de alunos, maior a precisão e mais lento o algoritmo. Por sua vez, o número máximo de atividades de correção pendentes a um único utilizador (maxPendentCorrection a partir do Passo 3) também deve ser escolhido cuidadosamente, uma vez que pode sobrecarregar os estudantes envolvidos.

O número mínimo de estudantes selecionado foi 2. A razão foi que, com duas correções é possível avaliar os resultados que sejam semelhantes. Se houver discordância, um terceiro estudante será associado. A nota final é a média das duas notas mais próximas dos alunos. Por outro lado, para que o modelo funciona corretamente o número máximo de atividades pendentes foi escolhido como 3, porque esperamos que casa usuário corrija, pelo menos, três redações. Em seguida, no quarto passo, os estudantes, de acordo com a distribuição na etapa anterior, por sua vez, irão avaliar as respostas para as atividades apresentadas para serem corrigidas, baseado no formulário de avaliação pré-definido e disponibilizado pelo professor. O prazo definido para a revisão (correção) das atividades no sistema é chamado de período de avaliação. É neste período que os alunos avaliarão seus pares e enviarão as notas para o sistema.

Algorithm 1 Recuperar usuários para correção de uma atividade

function RETRIEVEUSERS(minCorrection, safetyMargin, maxPendentCorrection)

/ minCorrection - Número mínimo de correção por estudantes; safetyMargin - Margem de segurança para buscar usuários; maxPendentCorrection - Número máximo de atividades pendentes de correção para um único usuário/

List<User> users = retrieveUsersMadeActivity(idActivity, minCorrection + safetyMargin, "Ascending order");
for i **do** 1 until size(users)
$qtdPendentCorrection$ = getAmountPendentCorrectionByUser(users.get(i));
if $qtdPendentCorrection > maxPendentCorrection$ **then**
users.remove(user.get(i));
end if
end for
if $size(users) >= minCorrection$ **then**
return users.sublist(0,minCorrection-1); ▷ Retorna os primeiros minCorrection elementos da lista
else
List<User> moreUsers = retrieveUsersHighProbabilityCorrect(minCorrection - size(users));
users.addAll(moreUsers);
return users;
end if
end function

Para atribuir sua nota, o estudante receberá as respostas de um determinado aluno e o formulário de avaliação com os critérios definidos. Em seguida, irá analisar a atividade sob o ponto de vista dos critérios (assim como um professor faria) e em seguida basta preencher o formulário com suas respostas. Esse processo será repetido enquanto houver atividades existentes para a correção para aquele aluno. De acordo com o algoritmo apresentado, esse número tende a ser um média entre os alunos, para não sobrecarregar nenhum. Porém, é possível que haja pequenas variações.

No quinto passo, para se ter uma maior confiança na nota final, deverão ser avaliados a reputação geral de cada avaliador e o nível de competência de cada avaliador em cada critério. O cálculo da reputação geral de um avaliador deverá levar em consideração as seguintes atividades realizadas por um aluno (em que + significa um aspecto positivo e - significa um aspecto negativo): + Acesso frequente ao sistema (A); + Alto número de atividades discursivas realizadas (QP); + Alto número de correções de atividades feitas (QC); − Baixa frequência de acesso ao sistema;− Alto número de correções pendentes (PC); − Alto número de atividades feitas com baixo número de correções.

Para se calcular o nível de competência do avaliador em um determinado critério deve-se olhar seu histórico de atividades feitas que utilizaram aquele critério. Caso não possua dados, seu nível será considerado neutro. Caso possua dados no histórico, este de-

verá ser utilizado para determinar se o avaliador possui um aspecto positivo e/ou negativo a depender das notas dos critérios. A partir dos cálculos definidos, o sistema calculará a confiabilidade de cada avaliação, o nível de competência do avaliador nos critérios e poderá então calcular o resultado final usando a reputação geral (OR) de cada avaliador. Este processo de cálculo da reputação geral do aluno X acontece com as seguintes equações:

$$FA = \frac{A(X) - AVG(A)}{MAX(A)} \qquad FQP = \frac{QP(X) - AVG(QP)}{MAX(QP)} \qquad FQC = \frac{QC(X) - AVG(QC)}{MAX(QC)}$$

$$OR = \frac{FA*1 + FQP*2 + FQC*2}{5}$$

O intervalo para as funções FA, FQP e FQC é $-1 <= F <= 1$. Nessas funções, quanto mais alto os valores para a quantidade de acesso, quantidade de redações realizadas e quantidade de redações corrigidas, respectivamente, mais próximos os valores de sua função será 1. Da mesma forma, valores muito baixos para tais quantidades, implicará em valores próximos a -1. Se um estudante estiver na média, o resultado será 0 (neutro). Note que a função de reputação geral (OR) tem um comportamento similar, no intervalo [-1,1], sendo definida como a média ponderada das funções anteriores.

Finalmente, no sexto passo, o sistema deverá gerar um relatório contendo o resultado final da avaliação que será apresentado ao aluno que submeteu a atividade. Este relatório deverá ser detalhado, apresentando as notas de cada critério, bem como a nota final da atividade. É possível a inclusão de comentários e possíveis alterações nas notas por parte do professor (opcional) antes de sua apresentação ao aluno. Este por sua vez irá identificar os pontos fracos e usar o sistema para aprender mais, através dos materiais do curso, realizando novas atividades e/ou corrigindo as atividades dos outros estudantes.

5. Experimento

Em ordem para avaliar a efetividade, nós projetamos e realizamos um experimento controlado no ambiente MeuTutor. Inicialmente, o MeuTutor não suportava redações. Por esta razão, os criadores do ambiente incluíram o mecanismo aqui apresentado, permitindo que os estudantes pudessem realizar redações dentro do ambiente. Essa inclusão do modelo nos levou a questionar a qualidade das correções com a seguinte pergunta: **"Como nós podemos avaliar a qualidade das correções feitas pelos alunos?"** e, principalmente, **"essas avaliações podem ser comparadas com uma avaliação de um especialista (professor)?"**. Além disso, também nos questionamos sobre **"Será que o modelo aplicado nesse contexto realmente soluciona os problemas citados?"** (Redução do custo e sobrecarga do professor, entre outros). Assim, surgiram as seguintes questões de pesquisa:

P1 - O uso do modelo de avalição por pares proposto apresenta a mesma eficiência, isto é, apresenta resultados semelhantes se comparado ao método tradicional de correções? O que nos leva às seguintes hipóteses: **H1-0**: O uso do modelo de avaliação por pares proposto é equivalente ao método tradicional; **H1-1**: O uso do modelo de avaliação por pares proposto não é equivalente ao método tradicional

P2 - O uso do modelo de avaliação por pares proposto apresenta diferenças nas métricas de tempo em relação a não usar esse modelo? O que nos leva às seguintes hipóteses: **H2-0**: O uso do modelo de avaliação por pares proposto não traz diferenças de tempo em relação ao método tradicional; **H2-1**: O uso do modelo de avaliação por pares proposto traz diferenças significativas de tempo em relação ao método tradicional.

P3 - O uso do modelo de avaliação por pares proposto apresenta diferenças nas métricas de custo em relação a não usar esse modelo? O que nos leva às seguintes hipóteses: **H3-0**: O uso do modelo de avaliação por pares proposto não traz diferenças de custo em relação ao método tradicional; **H3-1**: O uso do modelo de avaliação por pares proposto traz diferenças significativas de custo em relação ao método tradicional.

Formalmente, as hipóteses descritas anteriormente podem ser definidas conforme a Tabela 1. As funções N, T e C, apresentadas na tabela, retornam respectivamente, o valor da nota final, o tempo gasto e o custo envolvido, com relação a utilização do modelo tradicional M1 ou a utilização do modelo proposto M2, quando aplicado no ambiente educacional MeuTutor E1 ou quando aplicado sem ambiente educacional E2. (N, T e C são as métricas e M (M1 = Modelo tradicional / M2 = Modelo proposto) e E (E1 = MeuTutor / E2 = Sem ambiente) são os fatores do nosso experimento.

Tabela 1. Definição formal das hipóteses de pesquisa.

Hipótese	Hipótese Nula	Hipótese Alternativa
H1	H1-0: N(M1,E1) = N(M2,E1)	H1-1: N(M1,E1) $\neq$ N(M2,E1)
H2	H2-0: T(M1,E1) = T(M2,E1)	H2-1: T(M1,E1) $\neq$ T(M2,E1)
H3	H3-0: C(M1,E1) = C(M2,E1)	H3-1: C(M1,E1) $\neq$ C(M2,E1)

No nosso caso, como temos fatores com apenas 2 níveis cada, podemos utilizar um experimento fatorial 2^k sem repetição, retirando uma combinação que é impossível de ser realizada (que é a utilização do modelo proposto M2 em nenhum ambiente educacional E2). Cada execução do modelo tem um custo relativamente alto além de exigir bastante tempo para preparação do ambiente, alunos e demais esforços necessários para sua conclusão. Por este motivo, foi escolhido um projeto fatorial sem repetição. Nesse sentido temos apenas 3 ensaios possíveis, sendo executado cada um apenas uma vez, totalizando três tratamentos possíveis e consequentemente três execuções. Os tratamentos são: T1 (M1 e E1), T2 (M2 e E1) e T3 (M1 e E2).

Após a configuração do ambiente MeuTutor, o próximo passo foi a escolha dos alunos que iriam participar do experimento. Para isso, foram selecionados 30 usuários do MeuTutor que gostariam de participar desse experimento. Vale ressaltar que todos os usuários possuem o mesmo conhecimento na disciplina abordada neste experimento. Em seguida, foram convidados dois especialistas em correção de redações, indicados aqui por especialista 1 e especialista 2. Os usuários executaram seus devidos tratamentos e os especialistas corrigiram as redações em T1 e T3.

5.1. Análise de dados

A seguir, serão analisados os dados de cada uma das variáveis estudadas. A primeira variável analisada é a nota final (N). A Figura 4 apresenta o diagrama de caixa para a métrica da nota final N com relação aos tratamentos T1, T2 e T3.

Podemos ver na figura que a média (média (T1)=615 contra média(T2)=578) e mediana (mediana(T1) = 660 contra mediana(T2)=573), nos dois casos, estão bem próximas. Esta figura sugere que as notas obtidas com a aplicação desses tratamentos possuem variações estatísticas semelhantes, o que nos indica uma certa semelhança nos tratamentos. No entanto, ainda não foram geradas evidências estatísticas para afirmar isso. Tais afirmações só poderão ser feitas quando os testes estatísticos forem realizados. Como

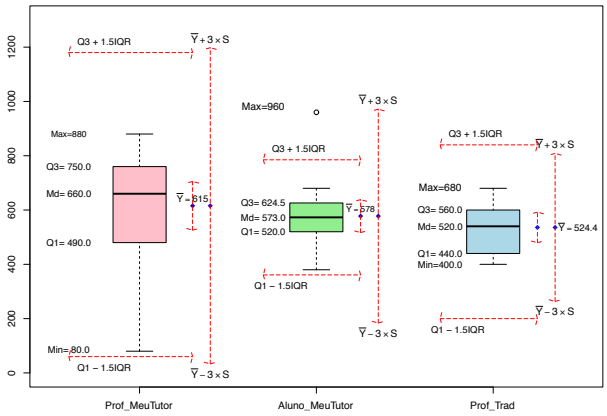

Figura 4. Diagrama de caixa com comparativo da métrica nota para os tratamentos T1, T2 e T3

uma de nossas questões de pesquisa é avaliar as notas obtidas pelo modelo proposto comparadas com as notas obtidas pelo modelo tradicional, então nós subtraímos T1 de T2. Podemos observar na figura que a diferença das notas do tratamento T1 e T2, na média (média(T1 - T2)= 37,8) e mediana (mediana(T1 - T2)= 40), estão bem próximas de 0 (zero), indicando que são bem equivalentes. No Geral, a diferença entre as notas T1 e T2 variou entre -79,75 (o resultado negativo indica que a nota do especialista no tratamento T1 foi menor que a nota do modelo no tratamento T2) a 110 (resultado positivo indica que a nota do modelo foi menor que a nota do especialista). Se considerarmos a escala de notas entre 0 e 1000, temos que as diferenças entre as notas variam entre 7.9% a 11%.

Contudo, ainda precisamos verificar a validade das hipóteses de pesquisa. Enfatizamos que para cada hipótese, deve-se indicar qual tratamento foi melhor. A primeira tarefa para se realizar a verificação de uma determinada hipótese de pesquisa é analisar a normalidade da distribuição dos dados que estão envolvidos em sua resposta. A normalidade dos dados é importante pois determina qual teste estatístico deve ser utilizado na análise. Existem alguns testes estatísticos para verificar a normalidade dos dados, porém o mais recomendado pelos estatísticos é o teste de Shapiro-Wilk [Shapiro and Francia 1972].

A primeira verificação de hipótese será feita com relação a métrica de nota. Os resultados da aplicação do teste de Shapiro-Wilk quando executado sobre os dados da métrica nota nos tratamentos estão apresentados na tabela 2. Neste caso, nós vemos que os dados vem de uma distribuição normal, uma vez que $W_{calculado} > W_{\alpha}$ e $P_{valor} > \alpha$ em todos os casos. Assim, nós aplicamos o T-Teste [Welch 1938], comparando os tratamentos. Os dados da execução do T-test estão apresentados na Tabela 3.

Podemos observar que os valores obtidos com a execução do T-test, são todos maiores que $\alpha = 0.05$. Portanto, com 95% de confiança, estatisticamente, não pode se afirmar que os valores das notas obtidos entre os tratamentos possuem diferenças significativas entre si, isto é, não há evidências estatísticas que mostrem a não equivalência das notas, não sendo possível gerar evidência estatística suficiente para refutar a hipótese

Tabela 2. Resultado da aplicação do teste de Shapiro-Wilk com os dados da métrica Nota

Tratamento	$W_{calculado}$	W_{α}	P_{valor}	α
T1	0.92223	0.897	0.1416	0.05
T2	0.90344	0.897	0.06599	0.05
T3	0.96201	0.842	0.8086	0.05

Tabela 3. Resultado da aplicação do T-test com os dados da métrica Nota

Tratamento	t	df	p-value
T1 x T2	0.85383	17	0.4051
T1 x T3	1.4728	25.546	0.153
T2 x T3	0.98704	24.523	0.3333

nula H1-0, implicando que **O uso do modelo de avaliação por pares proposto é equivalente ao método tradicional**. A próxima verificação de hipótese será feita com relação a métrica Tempo (T). A Figura 5 apresenta um conjunto de diagramas de caixa para a métrica do tempo (T) com relação aos tratamentos T1, T2 e T3.

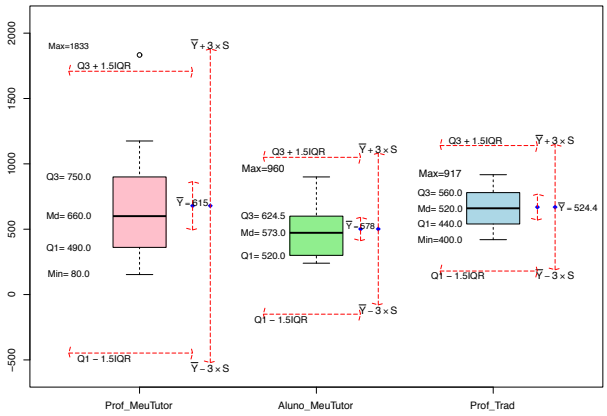

Figura 5. Diagrama de caixa com comparativo da métrica tempo para os tratamentos T1 , T2 e T3

Ao se analisar os dados de média e mediana dos tratamentos em que há um professor especialista envolvido, podemos verificar que os dados são bem semelhantes. Se compararmos esses dados com os dados do tratamento T2, notamos que o tempo no tratamento T2 é significativamente menor que o tempo nos outros dois tratamentos. Porém, como os diagramas se interceptam, é necessário a execução dos testes de hipóteses para ter uma conclusão estatisticamente válida. Nesse caso, aplicou-se o teste de Shapiro-Wilk com os dados desses tratamentos para analisar a normalidade dos dados e o resultado pode ser visto na Tabela 4.

Ao se analisar os dados da tabela, temos um caso especial no tratamento T1, uma vez que $W_{calculado}$(T1) = 0.89637 $<$ W_{α}(T1) = 0.897 e P_{valor}(T1) $=$ 0.04969 $<$ α = 0.05. Com isso, podemos refutar a hipótese nula do teste de Shapiro-Wilk e consequentemente, podemos afirmar com nível de significância de 5% que a amostra não provém de uma

Tabela 4. Resultado da aplicação do teste de Shapiro-Wilk com os dados da métrica Tempo

Tratamento	$W_{calculado}$	W_{α}	P_{valor}	α
T1	0.89637	0.897	0.04969	0.05
T2	0.91724	0.897	0.1155	0.05
T3	0.97249	0.842	0.9129	0.05

população normal. Se analisarmos os dados do tratamento T2 e T3 vemos que esses provêm de uma distribuição normal. Nesse caso, como os dados do tratamento T1 possui uma distribuição não-normal, então quando os dados de comparação envolverem T1 deveremos usar um teste não-paramétrico, como o Teste de Wilcoxon. Nesse caso, apenas a combinação T2 x T3 envolverá o T-test, uma vez que ambas são distribuições normais. Os dados da execução desses testes estão apresentados na Tabela 5.

Tabela 5. Resultado da aplicação dos testes estatísticos com os dados da métrica Tempo

Tratamento	Teste Utilizado	v/w	t	df	p-value
T1 x T2	Wilcoxon signed rank test with continuity correction	122	-	-	0.1169
T1 x T3	Wilcoxon rank sum test with continuity correction	81.5	-	-	0.7007
T2 x T3	Welch Two Sample t-test	-	-2.4605	21.938	0.02222

Podemos observar, de acordo com a Tabela 5, que como p-value(T1 x T2) = 0.1169 $> \alpha$ = 0.05, então não é possível concluir que há diferenças estatísticas entre esses tratamentos. Contudo, ao analisarmos o p-value dos tratamentos T2 e T3, p-value(T2 x T3) = 0.02222 $< \alpha$ = 0.05, podemos concluir que há diferenças estatísticas significativas e podemos concluir, com 95% de confiança, que o tempo em T2 é menor que o tempo em T3. Como a nossa avaliação se baseia no comportamento dos dados do tratamento T2 (em que o modelo é aplicado), temos estatisticamente que o tempo envolvido nele é menor que o tempo do tratamento T3 (especialista sem ambiente) com 95% de confiança. Há evidências estatísticas suficientes de que o uso do modelo proposto traz diferenças de tempo em relação ao método tradicional e com isso, pode-se refutar a hipótese nula H2-0 (que não haveria diferença) e aceitar a hipótese alternativa H2-1, com 95% de confiança, implicando que **o uso do modelo de avaliação por pares proposto traz diferenças significativas de tempo em relação ao método tradicional**.

Finalmente, a última métrica a ser analisada é o custo da correção (C). Esse custo tem diferentes significados a depender do tratamento envolvido. Quando envolve professores, o custo refere-se ao preço cobrado (em média) pelos especialistas para corrigir cada redação. Isso inclui também dados que o próprio INEP disponibiliza como valor base para o pagamento de correção de redações. Já com relação ao custo do modelo, se aplica aos custos de manter o servidor com a aplicação funcionando mais o processo de configuração do ambiente pelo professor (pago por hora de trabalho). A figura 6 apresenta o diagrama de caixa para a métrica de custo envolvido (C) com relação aos tratamentos executados.

Podemos observar que as médias e medianas nos tratamentos T1 e T3 (média(T1)= R$6.03 / média(T3) = R$4,83) são parecidas, uma vez que ambos os tratamentos utilizam especialistas nas correções. A variação existente entre elas se dá pelo fato de que os especialistas cobraram preços diferentes. Há um espalhamento (variação) maior

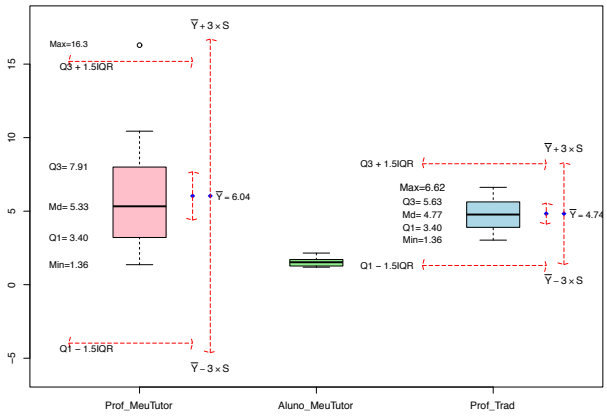

Figura 6. Diagrama de caixa com comparativo da métrica custo entre os tratamentos T1, T2 e T3

nos dados do tratamento T1, enquanto que esse comportamento não ocorre no tratamento T3, indicando que o custo por usuário no tratamento T3 é mais uniforme que o custo por usuário no tratamento T1. Finalizando a verificação de hipóteses, serão analisadas as hipóteses que envolvem a métrica de custo (C). Os resultados da aplicação do teste de Shapiro-Wilk com os dados dos tratamentos T1, T2 e T3 estão ilustrados na Tabela 6.

Tabela 6. Resultado da aplicação do teste de Shapiro-Wilk com os dados da métrica Custo

Tratamento	$W_{calculado}$	W_{α}	P_{valor}	α
T1	0.89635	0.897	0.04965	0.05
T2	0.25269	0.897	1.057e-08	0.05
T3	0.97276	0.842	0.9152	0.05

Analisando os dados da tabela, observamos agora dois casos de anormalidade dos dados. No caso do tratamento T1 e T2, uma vez que $W_{calculado} < W_{\alpha}$ e $P_{valor} < \alpha$. Por outro lado, T3 é normal. Os resultados da aplicação do Teste de Wilcoxon com os dados desses tratamentos estão apresentados na Tabela 7.

Tabela 7. Resultado da aplicação do teste de Wilcoxon com os dados da métrica Custo

Tratamento	Teste Utilizado	v/w	p-value
T1 x T2	Wilcoxon signed rank test with continuity correction	170	0.0002522
T1 x T3	Wilcoxon rank sum test with continuity correction	103	0.5486
T2 x T3	Wilcoxon rank sum test with continuity correction	0	1.111e-06

Com os dados da tabela, podemos observar que o p-value(T1 x T2) = 0.0002522 $< \alpha = 0.05$, implicando que há diferenças estatísticas significativas entre o custo dos tratamentos T1 e T2. Com isso, podemos concluir que o custo do tratamento T2 é menor que o custo do tratamento T1 com uma confiança de 95%. Da mesma forma, se analisarmos o valor de p-value(T2 x T3) = 1.111e-06 $< \alpha = 0.05$, concluímos que o custo também são diferentes, isto é, com 95% de confiança temos que o custo de T2 é menor que o custo de

T3. Por outro lado, o valor de p-value(T1 x T3) = 0.5486 é maior do que α = 0.05, o que implica que não há diferença significativa entre os tempos de T1 e T3.

Como nós queremos avaliar é o comportamento dos dados do tratamento T2, temos comprovado estatisticamente que o custo envolvido em T2 é menor que os custos envolvidos tanto no tratamento T1 quanto no tratamento T3 com uma confiança de 95%. Com isso, temos evidências estatísticas suficientes de que o custo são diferentes e portanto, conseguimos refutar a hipótese nula H3-0 (em que não haveria diferença) e aceitar a hipótese alternativa H3-1, com uma confiança de 95%, implicando que **o uso do modelo de avaliação por pares proposto traz diferenças significativas de custo em relação ao método tradicional**. Por meio de um modelo de regressão linear foi mostrado que o custo é 72.4% menor.

6. Conclusão e trabalhos futuros

O trabalho apresentou um mecanismo de avaliação por pares que tem como objetivo prover uma solução para a inclusão de avaliações escritas em ambientes educacionais online de forma eficiente. A necessidade da criação desse modelo surgiu do grande número de estudantes que estavam presentes nos ambientes e da grande dificuldade em prover provas discursivas neles, uma vez que isso gerava um alto custo e uma grande sobrecarga nos professores envolvidos no processo. Dessa forma, as atividades discursivas eram limitadas, dificultando a aprendizagem do aluno no processo.

O modelo proposto atingiu o objetivo, possibilitando a inclusão de avaliações discursivas nos ambientes online de maneira viável, como podemos observar em sua integração com o ambiente educacional MeuTutor e seu uso com provas de redação aplicadas para cerca de 30 alunos. Os experimentos apresentados também deram comprovações estatísticas suficientes de que os resultados obtidos com o modelo proposto são semelhantes aos resultados do modelo tradicional. Os resultados foram bastante satisfatórios, tendo em vista que foi comprovado, estatisticamente, que as notas dos modelos são equivalentes, permitindo que a substituição entre eles possa ser feita, sem comprometer os resultados finais. Da mesma forma, os resultados da métrica Tempo indicaram uma possibilidade de o tempo de correção ser menor usando o modelo proposto, contudo em alguns casos esse tempo seja semelhante. Por fim, os resultados da métrica Custo foram bem favoráveis a aplicação do modelo, tendo em vista que não é necessária uma correção por especialista no modelo proposto. O modelo de regressão criado mostrou que o custo é cerca de 72.4% menor.

Como uma contribuição secundária do modelo proposto, temos os algoritmos de seleção de usuários, bem como os cálculos de confiança e reputação dos usuários. Como trabalho futuro temos o planejamento de implantar o modelo em outros ambientes educacionais diferentes do MeuTutor, como por exemplo, o Moodle. Com a implementação/integração em outros ambientes, as dificuldades que por ventura poderão aparecer servirão como entrada para possíveis melhorias na implementação do modelo, deixando-o mais completo, flexível e compatível com a maioria dos ambientes educacionais.

Além disso, é preciso avaliar novas métricas de comparação do modelo proposto com o modelo tradicional que podem ser relevantes, como por exemplo, o nível de aprendizagem do aluno, melhorias das suas capacidades de julgamento, entre outros benefícios

que um modelo de avaliação por pares pode trazer. Com isso, pretende-se realizar novos experimentos mais completos a fim de avaliar tais métricas. Além disso, é preciso avaliar também questões subjetivas através da aplicação de questionários com os atores envolvidos no processo.

Referências

Chang, C.-C., Tseng, K.-H., and Lou, S.-J. (2012). A comparative analysis of the consistency and difference among teacher-assessment, student self-assessment and peer-assessment in a web-based portfolio assessment environment for high school students. *Computers & Education*, 58(1):303–320.

Dominguez, C., Cruz, G., Maia, A., Pedrosa, D., and Grams, G. (2012). Online peer assessment: An exploratory case study in a higher education civil engineering course. In *Interactive Collaborative Learning (ICL), 2012 15th International Conference on*, pages 1–8. IEEE.

Kahiigi Kigozi, E., Vesisenaho, M., Hansson, H., Danielson, M., and Tusubira, F. (2012). Modelling a peer assignment review process for collaborative e-learning. *Journal of Interactive Online Learning*, 11(2):67–79.

Kawai, G. (2006). Collaborative peer-based language learning in unsupervised asynchronous online environments. In *Creating, Connecting and Collaborating through Computing, 2006. C5'06. The Fourth International Conference on*, pages 35–41. IEEE.

Malehorn, H. (1994). Ten measures better than grading. *The Clearing House*, 67(6):323–324.

Orsmond*, P., Merry, S., and Callaghan, A. (2004). Implementation of a formative assessment model incorporating peer and self-assessment. *Innovations in Education and Teaching International*, 41(3):273–290.

Pavlus, J. (2010). The game of life. *Scientific American*, 303(6):43–44.

Piech, C., Huang, J., Chen, Z., Do, C., Ng, A., and Koller, D. (2013). Tuned models of peer assessment in moocs. *arXiv preprint arXiv:1307.2579*.

Sadler, P. M. and Good, E. (2006). The impact of self-and peer-grading on student learning. *Educational assessment*, 11(1):1–31.

Shapiro, S. S. and Francia, R. (1972). An approximate analysis of variance test for normality. *Journal of the American Statistical Association*, 67(337):215–216.

Sterbini, A. and Temperini, M. (2013). Openanswer, a framework to support teacher's management of open answers through peer assessment. In *Frontiers in Education Conference, 2013 IEEE*, pages 164–170. IEEE.

Tosic, M. and Nejkovic, V. (2010). *Trust-based peer assessment for virtual learning systems*. Springer.

Wang, Y. and Vassileva, J. (2003). Trust and reputation model in peer-to-peer networks. In *Peer-to-Peer Computing, 2003.(P2P 2003). Proceedings. Third International Conference on*, pages 150–157. IEEE.

Welch, B. L. (1938). The significance of the difference between two means when the population variances are unequal. *Biometrika*, pages 350–362.