Marcos Alécio Spalenza

*p*Nota: Uma Análise das Componentes Textuais para Avaliação de Respostas Discursivas Curtas

Vitória, ES

2021

Marcos Alécio Spalenza

*p*Nota: Uma Análise das Componentes Textuais para Avaliação de Respostas Discursivas Curtas

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. Ph.D. Elias de Oliveira

Vitória, ES

2021

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Duis malesuada laoreet leo at interdum. Nullam neque eros, dignissim sed ipsum sed, sagittis laoreet nisi.

Agradecimentos

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Duis malesuada laoreet leo at interdum. Nullam neque eros, dignissim sed ipsum sed, sagittis laoreet nisi. Duis a pulvinar nisl. Aenean varius nisl eu magna facilisis porttitor. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Ut mattis tortor nisi, facilisis molestie arcu hendrerit sed. Donec placerat velit at odio dignissim luctus. Suspendisse potenti. Integer tristique mattis arcu, ut venenatis nulla tempor non. Donec at tincidunt nulla. Cras ac dignissim neque. Morbi in odio nulla. Donec posuere sem finibus, auctor nisl eu, posuere nisl. Duis sit amet neque id massa vehicula commodo dapibus eu elit. Sed nec leo eu sem viverra aliquet. Nam at nunc nec massa rutrum aliquam sed ac ante.

Vivamus nec quam iaculis, tempus ipsum eu, cursus ante. Phasellus cursus euismod auctor. Fusce luctus mauris id tortor cursus, volutpat cursus lacus ornare. Proin tristique metus sed est semper, id finibus neque efficitur. Cras venenatis augue ac venenatis mollis. Maecenas nec tellus quis libero consequat suscipit. Aliquam enim leo, pretium non elementum sit amet, vestibulum ut diam. Maecenas vitae diam ligula.

Fusce ac pretium leo, in convallis augue. Mauris pulvinar elit rhoncus velit auctor finibus. Praesent et commodo est, eu luctus arcu. Vivamus ut porta tortor, eget facilisis ex. Nunc aliquet tristique mauris id sollicitudin. Donec quis commodo metus, sit amet accumsan nibh. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus.

“Lorem ipsum dolor sit amet, consectetur adipiscing elit.
Duis malesuada laoreet leo at interdum. Nullam neque eros, dignissim
sed ipsum sed, sagittis laoreet nisi.
(Lipsum generator)

Resumo

O processo de avaliação é uma etapa muito importante para a verificação de aprendizagem e manutenção do 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. Conforme é ampliado o acesso à educação, é importante que os métodos avaliativos também sejam adequados para não representarem um fator limitante. Nesse aspecto é importante ressaltar que, apesar da pequena quantidade de texto produzido, é necessário que o professor avalie cautelosamente todos os alunos para identificar possíveis problemas no aprendizado. Além disso, o tempo concorrente entre a análise de desempenho dos alunos, planejamento das aulas e atualização dos materiais impossibilita o acompanhamento detalhado do aluno em classe. Portanto, a adesão de métodos de suporte educacional é fundamental para melhorar a qualidade dos materiais e impactar diretamente no desenvolvimento do aluno. Neste trabalho, apresentamos uma ferramenta de apoio ao tutor na análise, correção e produção de *feedbacks* para o método avaliativo de respostas discursivas curtas. Através de técnicas de aprendizado semi-supervisionado em *Machine Learning*, o sistema auxilia o tutor na identificação principais respostas para reduzir o esforço de correção. Com os modelos avaliativos em meio computacional, o professor audita os resultados produzidos pelo sistema e acompanha seu processo de decisão. Deste modo, apresentamos a robustez do modelo avaliativo produzido pelo sistema através de diferentes *datasets* da literatura, alcançando correlação de X% em relação aos avaliadores humanos no coeficiente Kappa.

Palavras-chaves: Avaliação Automática de Questões Discursivas. Aprendizado Semi-Supervisionado. Sistemas de Apoio ao Tutor. Processamento de Linguagem Natural. Classificação de Texto.

Abstract

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Duis malesuada laoreet leo at interdum. Nullam neque eros, dignissim sed ipsum sed, sagittis laoreet nisi. Duis a pulvinar nisl. Aenean varius nisl eu magna facilisis porttitor. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Ut mattis tortor nisi, facilisis molestie arcu hendrerit sed. Donec placerat velit at odio dignissim luctus. Suspendisse potenti. Integer tristique mattis arcu, ut venenatis nulla tempor non. Donec at tincidunt nulla. Cras ac dignissim neque. Morbi in odio nulla. Donec posuere sem finibus, auctor nisl eu, posuere nisl. Duis sit amet neque id massa vehicula commodo dapibus eu elit. Sed nec leo eu sem viverra aliquet. Nam at nunc nec massa rutrum aliquam sed ac ante.

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

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

ML	*Machine Learning* (Aprendizado de Máquina)
NLP	*Natural Language Processing* (Processamento de Linguagem Natural)
SAG	*Short Answer Grader* (Avaliação de Questões Discursivas Curtas)
EDM	*Educational Data Mining* (Mineração de Dados Educacionais)
PR	*Pattern Recognition* (Reconhecimento de Padrões)
IR	*Information Retrieval* (Recuperação de Informação)
CAA	*Computer-Assisted Assessment* (Avaliação Assistida por Computadores)

Sumário

1 Introdução

As avaliações de aprendizado são fundamentais para todos os níveis de ensino. É por meio do método avaliativo que o professor observa o desempenho da turma e seu progresso nos conteúdos. Com aplicações frequentes, as atividades permitem ao professor interagir com os alunos e com os materiais pedagógicos para reformulação e aperfeiçoamento da sua metodologia. Desse modo, é com o acompanhamento da disciplina e o apoio ao educando que as atividades permitem a reformulação e controle do processo de ensino-aprendizagem (BARREIRA; BOAVIDA; ARAÚJO, 2006). Por meio das atividades podemos identificar o domínio dos estudantes sobre o contexto e sua capacidade de realizar inferências sobre o assunto. O papel da avaliação, portanto, é diagnosticar, apreciar e verificar a proficiência dos alunos para que o professor atue no processo de formação de modo a consolidar o aprendizado (OLIVEIRA; SANTOS, 2005).

É através do modelo de ensino-aprendizagem, que o professor observa problemas e age para contorná-los. Essa identificação de problemas e sua rápida solução torna a estrutura curricular personalizada, alinhando a turma de acordo com os objetivos da disciplina. É através das atividades, portanto, que é possível mensurar o conhecimento individual dos alunos. Um modo de aperfeiçoar a aplicação das atividades em quantidade e qualidade é dada através da mediação tecnológica. A mediação tecnológica na criação, avaliação, recomendação e visualização em dados educacionais apoia o professor na melhoria e no acompanhamento do currículo do aluno (PAIVA et al., 2012). É com as ferramentas de apoio, então, que o tutor pode verificar a aptidão dos estudantes, de forma individual ou coletiva, para melhorar a adaptação e a experiência com a disciplina.

Na literatura da Avaliação Assistida por Computadores (*Computer Assisted Assessment* - CAA em inglês), existe uma extensa pesquisa por métodos para avaliação de questões discursivas. Sabendo que existe um critério formulado pelo professor para correção das respostas discursivas, propomos uma abordagem de reconhecimento dos padrões de textuais. Assim, neste trabalho descrevemos um modelo semi-supervisionado para reconhecimento do método avaliativo, extração de padrões textuais, classificação da base de dados e produção de *feedbacks*. Considerando a liberdade textual característica das respostas discursivas, verificamos a similaridade entre respostas e os grupos de termos referenciais para atribuir notas de forma equivalente ao avaliador humano. Com os modelos de SAG, esperamos também demonstrar o método avaliativo com a criação de *feedbacks*, como o quadro de *rubrics* (ARTER; CHAPPUIS, 2006) e, consequentemente, melhorar os métodos de avaliação automática (SPALENZA et al., 2016).

1.1 Problema

Dentro da literatura da avaliação de respostas discursivas curtas, em inglês *Short Answer Grader (SAG)*, encontramos determinados problemas listados pelos autores para criação de melhores modelos avaliativos. Apesar de ser um estudo já realizado há décadas, em SAG encontramos desafios observados durante a aplicação da avaliação automática como demandas importantes e pouco estudadas até o momento. Nos primeiros sistemas, a modelagem de questões discursivas era um trabalho realizado com o texto bruto. A partir disso, a busca por equivalência entre a resposta esperada e o texto dos estudantes falhou por inúmeras vezes na padronização dos documentos e na identificação de sinônimos (LEFFA, 2003). O estudo dessa pesquisa fomentou inúmeras discussões em torno da identificação do conhecimento obtido pelas respostas escritas pelo aluno. A robustez dessa análise é parte fundamental de boa parte dos algoritmos atuais em SAG. Portanto, para a recuperação da relação entre o conteúdo e a nota atribuída são aplicadas diversas técnicas entre Aprendizado de Máquina, Estatística, Processamento de Linguagem Natural, Reconhecimento de Padrões, dentre outras.

Em uma revisão da literatura sobre os sistemas SAG (BURROWS; GUREVYCH; STEIN, 2015), os autores reúnem 37 trabalhos realizados na área. Durante essa revisão, o autor destaca o problema da profundidade do aprendizado, tradução literal de "*depth of learning*", separando as atividades em dois grupos: de reconhecimento e de recuperação. Tais modelos têm diferentes intúitos na aquisição de informação do aluno o que gera diferentes modos de processamento. No Brasil, conhecemos essas por questões abertas e fechadas, nomenclaturas também citadas pelos autores. Essa divisão estabelece a diferença entre as atividades que exploram apenas a necessidade de identificação e organização de conteúdo e as que dependem de construção de ideias visando respostas próprias e originais. Portanto, por definição, a chave para separar atividades que necessitem de maior ou menor conhecimento factual, ou questões abertas e fechadas, é a liberdade que o aluno possui para criação do seu conjunto de resposta.

Um problema com esse viés é reagir às questões discursivas factuais e opinativas da forma adequada (BAILEY; MEURERS, 2008). É esperado que o sistema lide com a liberdade de escrita do aluno recuperando o conteúdo. A forma aqui proposta para contornar esse problema de identificação do critério avaliativo parte de interações com o especialista. Essas interações são requisições de correção para buscar avaliações específicas de padrões textuais das respostas. Esse processo seleciona documentos que indiquem certo grau de distinção por grupo, resultante de uma clusterização inicial (OLIVEIRA et al., 2014). Coletamos assim a classificação dada para os itens elencados como relevantes pelo sistema em cada *cluster* para continuidade do processo avaliativo.

Desta forma, a modelagem de classificadores de forma semi-supervisionada permite que encontremos grupos de respostas através da clusterização. A partir daí devemos

encontrar qual é o critério avaliativo do professor, sem requisitar exemplos e chaves de resposta, dado até o momento como necessário (BUTCHER; JORDAN, 2010; MOHLER; BUNESCU; MIHALCEA, 2011; RAMACHANDRAN; CHENG; FOLTZ, 2015). Basicamente, a análise de distribuição de características e a seleção de respostas visa remontar o critério do professor, otimizando o processo de correção automática.

O problema passa da necessidade de avaliação para o aumento do número de padrões à serem avaliados pelo professor para relacionar conteúdo com classificação. Um dos principais sistemas, *FreeText Author*, apresenta problemas importantes para um efetivo processo avaliativo automático (BUTCHER; JORDAN, 2010). Os autores, em uma análise detalhada deste sistema, listaram seis problemas. O primeiro que podemos destacar é a omissão dos padrões de avaliação. O segundo, a identificação de associações entre palavras e o método avaliativo de forma inconsistente. O terceiro problema, conforme os autores, é a necessidade de identificação estrutural da sentença. A quarta dificuldade listada é o tratamento de classificações incorretas por parte do especialista. O quinto problema é o conflito de um padrão correto de resposta com uma avaliação dada como incorreta. Por fim, seguindo a linha do quarto e quinto problema citado, um sexto problema é dado pela confiabilidade do sistema como avaliador em relação a interpretação textuais inconsistentes.

É importante descrever ainda a diferença entre o conjunto de informações selecionadas para a classificação correta das respostas e a interpretação avaliativa (RAMACHANDRAN; CHENG; FOLTZ, 2015). A separação do cunho interpretativo do interlocutor deve ser estritamente analisada conforme o modelo avaliativo. Se os padrões de nota e de resposta estiverem em conformidade com a avaliação do especialista, a redução de dimensionalidade dada pela otimização deve indicar uma boa interpretação da relação termo-classe. Assim, encaramos a avaliação como um processo constante e passível de revisões, para que o ajuste do sistema torne o modelo extraído cada vez mais próximos das expectativas do professor.

Por fim, podemos ainda citar dificuldade em encontrar os *datasets* utilizados por trabalhos da literatura (BURROWS; GUREVYCH; STEIN, 2015). É muito comum encontrar trabalhos no qual os autores coletaram dados na própria universidade e não as tornam públicas. Além disso, em SAG uma base de dados adequada deve caracterizar o processo avaliativo do professor e constar com relevantes resultados na literatura. Entretanto, o intúito deste trabalho é propor algumas soluções para tais problemas. Assim, buscamos avanços desde a descrição dos conjuntos de dados disponíveis até a apresentação de uma possível solução para a associação entre a atribuição de notas e o aspecto textual da resposta.

1.2 Proposta

Neste trabalho apresentamos um método de avaliação de respostas discursivas curtas através de modelos avaliativos complexos. Para seu desenvolvimento, buscamos identificar problemas mais comuns descritos na literatura como deficiências dos sistemas *Short-Answer Graders* (SAG) para apresentar uma proposta de solução. A ideia é compor um sistema para reconhecer a relação entre as respostas e as notas atribuídas de acordo com o método avaliativo do professor. Com isso, esperamos atender a demandas atuais dos trabalhos em SAG por meio de técnicas de *Educational Data Mining* (EDM), *Machine Learning* (ML) e *Natural Language Processing* (NLP). Para a criação dos modelos selecinamos diferentes bases de dados de respostas discursivas curtas disponíveis em *inglês* e *português*. Dentre os *datasets* observamos 3 tipos de avaliações: notas discretas, notas graduais e notas contínuas. Portanto, neste trabalho, estudamos estruturas para identificação das principais respostas do conjunto, reconhecimento do método avaliativo do professor (especialista) e elaboração *feedbacks*.

Para identificação das principais respostas apresentamos um modelo de aprendizado semi-supervisionado. No aprendizado semi-supervisionado o especialista ativamente passa o conhecimento para o algoritmo de classificação. O algoritmo, por sua vez, utiliza o as informações passadas para criar um modelo que imite o especialista na tarefa. Neste caso, o professor ensina ao sistema seu método avaliativo e, através da atribuição de notas, é formado um modelo que tenta replicar o método para as demais respostas da atividades. Cada uma das respostas enviadas para atividade é considerada uma amostra para o sistema. Dentre todas as amostras, é fundamental que o sistema aprenda cada uma das características das respostas, selecionando as principais por representatividade. Para essa seleção o sistema utiliza de técnicas de otimização e clusterização. As respostas selecionadas são denominadas de treinamento, pois serão utilizadas para produção dos modelos, enquanto as demais são o conjunto de teste.

No reconhecimento do método avaliativo do professor, modelos são criados para classificação das respostas discursivas. A categorização deve se aproximar ao máximo da tarefa realizada pelo professor, analisando detalhes parecidos na resposta. Portanto, o modelo avaliativo do sistema objetiva atender as expectativas do professor. Quanto menor a diferença entre a nota dada pelo sistema e a nota atribuída pelo professor, melhor o modelo criado. Consequentemente, os melhores modelos representam melhor a diversidade de notas e respostas com tendência menor de erros. Na gradação das notas, quanto maior a discrepância entre as notas mais críticos são os erros. Sabendo que, entre avaliadores humanos também existe esse erro. Os dados selecionados para treino do classificador ditam o conhecimento da gradação de notas distribuídas por ele. Portanto, o classificador recebe as características de cada resposta e a sua respectiva avaliação e as compara com as amostras de teste, com notas não conhecidas. 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* e relatórios é fundamental para o suporte ao professor. Em sala de aula, os *feedbacks* são um material que detalha a avaliação para professores e alunos e descrevem o método avaliativo de forma a sanar qualquer dúvida e evidenciar qualquer problema no aprendizado. Por outro lado, na perspectiva da interação do professor com o sistema, os *feedbacks* caracterizam a decisão, descrevem o modelo textual e a equivalência entre respostas. Portanto, em todos os ciclos do sistema esperamos reduzir o esforço de correção do tutor, apresentar resultados de alto nível com o modelo avaliativo e gerar materiais explicativos e completementares de qualidade.

1.3 Objetivos

O objetivo deste trabalho, portanto, é ajustar o modelo de correção criado pela máquina aos padrões estabelecidos pelo professor através da sua avaliação. Para isso, os modelos avaliativos devem compreender o método aplicado pelo professor, categorizando as respostas em classes, níveis ou intervalos contínuos de nota. Segundo a consistência de cada grupo, buscamos reduzir o esforço de correção do professor com a avaliação das respostas que apresentem apenas as principais caracteristicas textuais. Através de padrões bem definidos, esperamos reproduzir o critério avaliativo da questão justificando a classe atribuída através do seu respectivo sumário. Tal sumário, então, são os padrões de cada classe de nota partindo do agrupamento *a priori* das questões. É através desse sumário por nota que recuperamos um possível critério de correção. Desta forma, através do *p*Nota, esperamos que o professor esteja apto para gerenciar o seu método avaliativo em um tempo menor para concentrar-se na verificação de aprendizagem do aluno.

Portanto, temos como âmbito principal a criação de modelos para aproximar o critério avaliativo aplicado ao aluno da definição de padrões de correção e a criação de *feedbacks*. Para isso, estudamos os padrões avaliativos do professor e os métodos de representação do conhecimento em base de dados de questões discursivas curtas. Para atingir o objetivo geral descrevemos os seguintes objetivos específicos:

- Organizar *datasets* públicos e locais para comparação direta 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 léxica, morfológica, semântica, sintática, estatística ou espacial (BURROWS; GUREVYCH; STEIN, 2015; BUTCHER; JORDAN, 2010).
- Unificar padrões de respostas dadas por professores e alunos, observando a frequência de ocorrência e co-ocorrência de termos segundo sua relevância (BUTCHER;

JORDAN, 2010).

- Criar modelos avaliativos através do reconhecimento de padrões variáveis em categorias em dados discretos e contínuos (BURROWS; GUREVYCH; STEIN, 2015).
- Elaborar e ajustar modelos de acordo com a eficiência do sistema na recuperação da resposta atribuída pelo professor e seu modelo avaliativo (BURROWS; GUREVYCH; STEIN, 2015).
- Identificar a relação da avaliação com o comportamento textual da classe para remoção de *outliers* e manter a consistência da classificação (BUTCHER; JORDAN, 2010).
- Apresentar avaliações adequadas ao formato de correção do professor (BUTCHER; JORDAN, 2010).
- 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 (OLIVEIRA et al., 2010).

1.4 Estrutura do Trabalho

A seguir são apresentados os conteúdos dessa tese. A proposta é discutida em detalhes através de 5 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 oito *datasets* as diferentes formas de apoio avaliativo, 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 sala de aula é um ambiente que produz diariamente grande quantidade de informações. Dentre essas informações podemos citar como essenciais o acompanhamento do aprendizado dos alunos, necessidade de reforço do conteúdo e o cumprimento do curricular. Tradicionalmente essa dinâmica faz parte dos métodos de ensino-aprendizagem empregados pelos professores, porém, superam a capacidade analítica dos mesmos. Por conta disso, para ampliar a verificação do professor em analisar os materiais produzidos em sala, ganharam maior notoriedade e espaço prático os sistemas de *Educational Data Mining* (EDM).

Em EDM, métodos são aplicados aos dados da classe de alunos para a extração de conhecimento, apoio ao tutor e acompanhamento do ensino. Através de técnicas de *Machine Learning* (ML), ocorre a redução da carga do professor para tratamento e acompanhamento do conteúdo ministrado em sala. Deste modo, o professor torna-se responsável pela auditoria, monitoramento e aplicação dos resultados obtidos. Assim, os sistemas apoiam a descoberta de problemas de aprendizado, a personalização do ensino e a acompanhamento coletivo dos alunos em sala.

Portanto, através da mineração de dados, é possível ao professor a análise de todo material produzido pelos alunos, a criação de feedbacks individuais e a aplicação de reforço para determinados grupos de alunos. Neste ponto, dentro dos métodos de EDM, um nicho de sistemas que tange diretamente essa demanda são os sistemas *Short-Answer Graders* (SAG). Os SAG são responsáveis pela verificação em massa das respostas textuais curtas, auxiliando o professor no processo de correção. É característico deste tipo de questão a verificação do aprendizado do aluno segundo o material ministrado em sala. Ao aluno, este tipo de questão é fundamental para prática da dinâmica de escrita, busca de informações e sumarização do conteúdo em poucas palavras. Portanto, este tipo de atividade envolve métodos relevantes para todos os níveis de ensino, principalmente durante o aprendizado e desenvolvimento da escrita.

Apesar da relevância das questões discursivas curtas, sua aplicação é gradativamente reduzida pela alta carga-horária do professor em sala. Assim, torna-se uma demanda secundária o planejamento, a revisão e a análise do material dos alunos. O apoio computacional, reduz o tempo necessário fora da sala para avaliação do conteúdo, com o professor participando parcialmente do processo de avaliação. O nicho dos métodos computacionais de apoio aos métodos avaliativos são conhecidos também por *Computer-Assisted Assessment* (CAA). Neste processo, os resultados obtidos são auditados pelo professor para garantir que o modelo avaliativo foi seguido fielmente para que a representação do conhecimento

das respostas atenda coerentemente as demandas da atividade. Enquanto isso, a aplicação de técnicas de ML reflete que a descrição do modelo de correção utilizado é um potencial *feedback* com aplicação direta em sala.

Para o uso dos métodos de SAG, é importante que a questão seja elaborada para a requisição de conhecimentos gerais dos alunos. Comumente separamos as questões em discursivas e objetivas, de acordo com o modelo de resposta esperada. As questões abertas envolvem a liberdade de escrita do aluno, avaliando sua capacidade de descrição e desenvolvimento textual. Por outro lado as questões objetivas desenvolvem o raciocínio, a leitura e interpretação do material didático e a busca de informações. Sabendo disso, as questões discursivas permeiam ambos os tipos de habilidades do aluno. A Figura X caracteriza as atividades segundo os modelos de resposta.

Como apresentado na Figura X o professor dispõe de alguns modelos de atividades que, refletem diferentes aspectos do aprendizado. Dentre as redações de cunho aberto e irrestrito e as respostas diretas com opções elencadas no enunciado, as respostas discursivas encontram-se em âmbito intermediário. As respostas curtas buscam que o aluno estabeleça relação entre o aprendizado com material didático e a sua descrição textual. Assim, dentre os conhecimentos gerais, a questão deve evitar abordar temas de cunho interpretativo e que tangenciam experiências específicas de cada aluno. Por outro lado, a resposta deve representar a informação completa da questão, dando ao sistema embasamento para correção, evitando informações restritas ou codificadas. A Figura X demonstra como o espectro de questões trabalhados através das respostas discursivas curtas.

A Figura X caracteriza exatamente, dentro do nicho de questões discursivas, a dinâmica de uso do processamento computacional das respostas por parte do professor. Quando as respostas não são únicas nem abstratas e apresentam um conhecimento comum, é o uso ideal dos métodos de correção automática. Portanto, é fundamental a convergência das respostas, para que as respostas apresentem uma ou poucas direções a serem abordadas pelos estudantes. Para isso, é fundamental que o sistema realize três passos. O primeiro é o aprendizado do modelo de respostas do aluno. O segundo é através do modelo de respostas reconhecer o padrão avaliativo do professor. Por fim, no terceiro passo, o sistema deve replicar o modelo avaliativo e elaborar *feedbacks* coerentes.

2.1 Avaliação Semi-Supervisionada

O método de aprendizado dita a forma de aquisição de informações do sistema e a criação de modelos com desempenho similar ao humano. Neste trabalho apresentamos um método de amostragem semi-supervisionado de aprendizado através da anotação do professor em itens selecionados através da clusterização. Porém, a requisição de anotação do professor para amostras de respostas não é o método tradicional para modelo avaliativo.

A grande maioria dos trabalhos utiliza amostragem através do particionamento entre treino e teste dos dados, previamente selecionado nos *datasets*. Considerando cada resposta de um aluno 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. Esse modelo clássico permite ao sistema observar apenas uma parcela dos dados, onde o sistema deve realizar a inferência em dados ainda desconhecidos. Assim, o sistema deve absorver o modelo avaliativo do conjunto de treino e replicar o método avaliativo no conjunto de teste, pressupondo a equivalência dos mesmos. Porém, o modelo não necessariamente é similar ao de teste, não refletindo diretamente a aplicação de um sistema SAG em conjunto com o professor.

Outros métodos, mais próximos da demanda do professor, utilizam de exemplos anotados de respostas para criação de modelo. Tais exemplos são denominadas respostas candidatas. 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. Porém, este tipo de treinamento gera uma tendência na avaliação, com reduzida interpretação das respostas dos alunos. O modelo criado não é capaz de identificar múltiplos contextos e as referências apresentadas pelo aluno. Portanto, as limitações da informação passada são um contraponto à liberdade textual esperada das atividades de escrita livre. Além de tornar-se engessada, não necessariamente são bons representantes dos demais documentos do *dataset*.

Para contornar as limitações, ainda existem alguns métodos utilizados para ampliar a capacidade de interpretação do sistema. O primeiro método visa maximizar o uso de informações das atividades. Nesta proposta, os sistemas são treinados com conteúdos adjacentes à questão, como o enunciado, o material de apoio e o quadro de *rubrics* utilizado pelo professor na correção. O enunciado e o material de apoio adicionam ao sistema conhecimento externo sobre o tema. Enquanto as respostas candidatas e o quadro de *rubrics* são materiais descritivos do modelo avaliativo do professor para todos, inclusive o sistema. Por outro lado, existem sistemas que demandam modelos mais complexos do método avaliativo, como regras de avaliação e filtros de conteúdo feitos manualmente.

Outra estratégia é o uso de aumento de dados. Com aumento de dados 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. Assim, a quantidade de amostras para treinamento e variações para cada modelo de resposta torna-se muito superior à quantidade dada inicialmente. Outras formas incomuns ainda compreendem métodos de associação entre respostas com descoberta de padrões através de aprendizado não-supervisionado. Neste conjunto de

técnicas, destacam-se os métodos de clusterização. Com a clusterização os documentos de resposta são agrupados pelo coeficiente de similaridade e associados diretamente à uma determinada nota para o conjunto. Portanto, torna-se função do professor avaliar grupos de resposta segundo os componentes identificados como equivalentes.

De forma diferente das estratégias citadas, o aprendizado semi-supervisionado proposto combina os métodos de clusterização e classificação. A clusterização é um conjunto de técnicas responsáveis por identificar de forma não-supervisionada um determinado número de agrupamentos de respostas pela similaridade. Os grupos, denominados clusters, indicam que os itens compartilham características equivalentes. Porém, nessa dinâmica, através do reconhecimento da distribuição dos documentos, grupos são formados para amostragem partindo dos clusters. Essa amostragem visa identificar itens representativos dos agrupamentos, associando as principais características textuais diretamente com o método avaliativo do professor.

2.2 Classificação de Documentos

Uma tradicional área em ML, a classificação de documentos, possui inúmeras subdivisões segundo a especialização, motivação e conteúdo do conjunto de documentos. Cada conjunto de documentos é conhecido como *dataset*, base de dados ou *corpus*. A coleção destes, porém, é denominada *corpora*. A classificação de documentos envolve treinar algoritmos de classificação com exemplos rotulados para replicar métodos de identificação de conteúdo e rotulação feitos por um especialista. Portanto, para além da origem e conteúdo dos documentos, o algoritmo deve se adaptar para especialização na triagem dos documentos de acordo com suas características.

O especilista realiza uma leitura dos documentos e identifica informações específicas que justificam a categoria atribuída. Para replicar tal tarefa, através da análise do conteúdo, o sistema deve identificar características que estão diretamente relacionadas a cada classe de documentos. Dependendo da característica dos documentos, o conteúdo relevante de um documento para categorização pode incluir a identificação de poucas palavras-chave até a formação de modelos linguísticos complexos. Por exemplo, na triagem de documentos pré-formatados as informações básicas como título, autor e organizações ou setores responsáveis podem ser descritores diretos da classe a ser atribuída. Por outro lado, em modelos como SAG, é necessário que relações textuais complexas sejam avaliadas para atribuição de notas.

Deste modo, a atribuição de notas torna dos sistemas SAG uma complexa tarefa de classificação de documentos. É essencial a adaptação do algoritmo de acordo com o método de classificação utilizado pelo especialista. Portanto, apesar do conteúdo textual, a subjetividade do critério de avaliação deve ser levada em consideração pelo sistema. Assim,

a combinação entre o reconhecimento do modelo avaliativo e o reconhecimento do modelo textual deve atender às expectativas do professor. Enquanto em parte das situações as notas fortemente correlacionadas com a ocorrência dos termos, em outras o critério do professor pode ter baixa correlação com os termos e apresentar diferentes nuances na atribuição de notas. Deste modo, é determinante que o sistema compreenda a essência do conteúdo do documento enviado por cada aluno para reconhecimento da relação com as respectivas notas atribuídas.

2.3 Processamento de Linguagem Natural

Para criação de um modelo linguístico, os sistemas utilizam estratégias de aquisição de informação com técnicas de NLP. As primeiras técnicas de SAG da literatura e os primeiros sistemas propostos utilizavam descritores. Os descritores são caracteríticas extraídas do texto, elencando estruturas de cada resposta. Em geral, são formados por características pré-definidas, identificadas no texto da resposta do aluno. Dentre os descritores, os mais comuns eram a contagem de erros da linguagem, a quantidade de palavras e a frequência de certas classes gramaticais. Porém, as características pré-definidas, consequentemente, não atendem a uma grande quantidade de respostas, criando modelos linguísticos com pouca aderência ao conteúdo.

Posteriormente, observando os diferentes propósitos das questões discursivas curtas e sua aplicação multidiciplinar, surgiram estruturas para maior aquisição de informação e modelagem linguística. Os modelos linguísticos ampliaram a aderência do sistema ao tema das atividades. Assim, através do conjunto de respostas, cada sistema elabora modelos linguísticos com contexto suficiente para encontrar associações entre palavras. Através dessas associações, os sistemas estabeleceram relações diretas entre os termos de cada resposta e o método de atribuição de nota do professor.

As estratégias voltadas na análise do texto por completo, adicionaram muita informação aos sistemas. Porém, tais informações não necessáriamente são relevantes para o método avaliativo. Como consequência, ocorreu a evolução e desenvolvimento de técnicas próprias de ponderação, seleção de características e identificação de padrões textuais. Para ponderação é mais utilizado o modelo Term Frequency - Inverse Document Frequency (TF-IDF). 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*. Por outro lado, dentre as técnicas de seleção de características que se destacam, o *Latent Semantic Analysis* (LSA) é uma das mais utilizadas na literatura. O uso desta técnica compreende identificar relações semânticas dentro do conjunto de respostas. Assim, através do LSA, os sistemas reunem o conteúdo que potencialmente contém maior significância no tema.

Entretanto, os modelos linguísticos criados através da 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 propõe maior extração de informação textual, ainda que em textos curtos, para formação de componentes linguísticos mais robustos.

Em especial, trabalhos mais recentes reportam o ganho de informação com o uso de *embeddings*.

2.4 Avaliação de Questões Discursivas Curtas

3 Método

O mapa de características foi uma técnica desenvolvida com base no método de avaliação de respostas discursivas (SPALENZA et al., 2016). Com esta ferramenta, o professor corrige cada questão com notas que serão interpretadas como classes no sistema para, automaticamente, identificar trechos relevantes para cada uma delas. Em geral o processo segue as etapas de coleta de dados, padronização, agrupamento, seleção de características e elaboração de relatórios.

Para a coleta de dados utilizamos o Ambiente Virtual de Aprendizagem - AVA Moodle[1]. O uso dessa plataforma é muito conhecido nos cursos EaD e conta também com várias aplicações de sucesso em turmas presenciais. A submissão de tarefas em texto *online* no AVA permite que a atividade, quando finalizada, seja coletada pelo *software* de extração de informações descrito por (PISSINATI, 2014) para ser processado pelo sistema.

Após a coleta o sistema realiza processos de padronização textual como a eliminação de sinais gráficos, *stemming* (radicalização), contagem de *n-grams* e remoção de *stopwords*. Essa etapa de padronização é uma tentativa de tratamento do conteúdo do documento para garantir equivalência entre as informações apresentadas. Após esse pré-processamento ocorre a vetorização de documentos para interpretação computacional do seu conteúdo. A vetorização adotada usa a frequência dos termos indexados a partir dos textos, ou *Term Frequency - TF* em inglês. Esse modelo se baseia na técnica de processamento de textos "*bag of words*" onde cada índice é associado a um termo distinto encontrado no conjunto de documentos.

No modelo vetorial, cada documento d de um conjunto $D = \{d_1, d_2, d_3, \ldots, d_{|D|}\}$ de documentos é representado como um vetor de termos t. Durante a vetorização. é verificada a frequência de ocorrência (TF) de cada um dos t termos em cada documento d em $|D|$, sendo $|D|$ o total de documentos desse conjunto. Para cada d, portanto, é computada a frequência individual dos $t = \{t_1, t_2, t_3, \ldots, t_k\}$ termos que existem na base de dados, sendo k o número de termos distintos encontrados na coleção. Assim, cada documento d é representado como um vetor com as frequências individuais n de cada um dos k termos

[1] *Modular Object Oriented Distance LEarning* - Moodle - <moodle.org>

distintos, como o exemplo visto na Equação 3.1.

$$\begin{aligned} d_0 &= \{n_{t_1}, n_{t_2}, n_{t_3}, \ldots, n_{t_k}\} \\ d_1 &= \{n_{t_1}, n_{t_2}, n_{t_3}, \ldots, n_{t_k}\} \\ &\vdots \\ d_{|D|} &= \{n_{t_1}, n_{t_2}, n_{t_3}, \ldots, n_{t_k}\} \end{aligned} \tag{3.1}$$

Depois do processo de vetorização ocorre a análise de similaridade dos documentos. Durante esse processo os documentos são clusterizados, particionados em treino e teste conforme os principais padrões selecionados nos grupos resultantes do algoritmo de *clustering*. Definimos o número de grupos *a priori* pelo agrupamento de melhor *silhouette score* (ROUSSEEUW, 1987). Após a amostragem, os itens são avaliados pelo especialista e cada nota é considerada uma classe.

Para seleção das informações que representem a essência das tarefas, o mapa de características busca a resposta mínima de cada classe. Sabendo que o professor já efetuou a correção, o sistema agrupa as respostas com notas equivalentes. Em cada grupo o conjunto mínimo de resposta é recuperado com a análise iterativa da densidade do agrupamento com um Algoritmo Genético. Nesse método de otimização tentamos reduzir o número de termos selecionados avaliando o conjunto de informações da classe. Assim, a cada ciclo do algoritmo, o *fitness* é calculado pela razão entre o número de características selecionadas para esses documentos e a densidade de similaridade.

As características selecionadas são utilizadas para representar as partes significativas de cada grupo de documentos, sumarizando-os. Com um peso atribuido pela frequência de ocorrência do termo na classe, uma visualização em HTML é gerada para discussão colaborativa dos resultados.

3.1 Clusterização

3.2 Análise Linguística

3.3 Identificação de Respostas Candidatas

3.4 Classificação

4 Experimentos e Resultados

Esse capítulo apresenta três séries de experimentos. A primeira apresenta a parte fundamental do aprendizado semi-supervisionado do sistema *p*Nota, utilizando clustering para a identificação dos principais itens de resposta em cada base de dados. A segunda apresenta os métodos de classificação, a qualidade do aprendizado do sistema na predição de notas e sua adequação ao modelo esperado pelo tutor. Por fim, o terceiro módulo reflete como os modelos de resposta são formados pelo sistema e apresentados como feedback aos alunos e professores. Os experimentos foram realizados utilizando conjuntos de dados da literatura que apresentam diferentes características.

4.1 Base de Dados

Oito bases de dados foram selecionadas de acordo com a literatura, em português e inglês. Cada base de dados foi utilizada conforme as suas características. As bases de dados foram organizadas segundo o formato da nota, entre discreta, gradual e contínua.

Em bases de dados com notas *discretas* 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 avaliados 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 da variável não definir uma categoria que represente a divergência entre respostas dificulta o aprendizado do modelo avaliativo. Ao sistema, isso torna subjetiva a espectativa de resposta subjetivo. Assim, esse tipo de atividade é avaliada 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 *gradativa* numérica é a mais comum. Esse modelo favorece também os sistemas computacionais na criação da representação de resposta por categoria de nota. Ao tempo que a categoria induz a equivalência de todas as respostas ao qual foi associada. 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 este 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. A Tabela 1 apresenta os detalhes de cada *dataset*, incluindo o número de questões, o total de respostas, o modelo avaliativo aplicado e a linguagem.

Base de Dados	Questões	Respostas	Modelo Avaliativo	Linguagem
Kaggle ASAP-SAS	10	17043	gradual	Inglês
Kaggle PTASAG	15	7473	gradual	Português
Projeto Feira Literária	10	700	gradual	Português
SEMEVAL'2013 Beetle II	47	3941	discreto	Inglês
SEMEVAL'2013 SciEntBank	143	5251	discreto	Inglês
UK Open University	20	23790	gradual	Inglês
University of North Texas	87	2610	contínua	Inglês
VestUFES	5	460	contínuo	Português

Tabela 1 – Bases de dados e suas principais características.

A Tabela 1 descreve os oito *datasets* utilizados nos experimentos deste capítulo. Através das características apresentadas, sabendo que cada *dataset* contém uma quantidade regular de respotas, observamos a grande diversidade de quantidade de respostas por questão. Com questões de 30 até mais de 1800 respostas. No total, esse *corpora* apresenta um total de 337 questões e 61.268 respostas. Cada base de dados e sua descrição completa é apresentada a seguir:

4.1.1 Base de Dados do Concurso ASAP-SAS no *Kaggle (Inglês)*

A base de dados *ASAP - SAS*, *Automated Student Assessment Prize - Short Answer Scoring* é uma competição proposta pela *Hewllet Foundation* na plataforma *Kaggle*. A *ASAP* 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).

O objetivo da competição foi 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.

Neste contexto, a competição apresentou 10 questões multidiciplinares, de ciências à artes. Estão distribuidas 17043 respostas de alunos dentre essas atividades. Para chegar nessa quantidade, foram selecionadas por volta de 1700 respostas dentre 3000 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 apresentou concordância de 90% no coeficiente *Kappa*.

4.1.2 Base de dados *PTASAG* no *Kaggle (Português)*

A *PTASAG - Portuguese Automatic Short Answer Grading Data* é uma base de dados brasileira apresentada por (GALHARDI; SOUZA; BRANCHER, 2020) e disponibilizada na plataforma *Kaggle*. 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 ao Ensino Fundamental para 326 estudantes de 12 a 14 anos do 8º e 9º ano. Somados a estes, também foram aplicados a 333 alunos do Ensino Médio de 14 a 17 anos. As respostas foram avalidas 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;
- Nota 3: Majoritariamente correta apresentando os principais pontos.

No total, participaram 659 estudantes com um total de 7473 respostas. Cada uma das 15 questões apresenta entre 348 e 615 respostas. Apenas 4 questões foram avalidas por mais de um avaliador para verificar a concordância entre avaliadores. O coefficiente *Kappa* observado foi de, em média, 53.25%.

4.1.3 Base de Dados *Beetle II* 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*. 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 3 a 5 horas para preparar materiais, construir e observar circuitos no simulador e interagir com o STI. Esse sistema faz questões aos alunos, avalia as respostas e envia *feedbacks* via *chat*. Na construção deste *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. Equanto 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 3941 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.

4.1.4 Base de Dados *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*, com foco na avaliação de sistemas conforme a sua capacidade de análise e exploração semântica da linguagem. É uma base de dados formadas pela avaliação de 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, dentre 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: *correct*, *partially-correct-incomplete*, *contradictory*, *irrelevant* e *non-domain*. Portanto, o *SciEntsBank* consiste em um conjunto com 143 questões selecionadas e 5251 respostas. No processo de avaliação foi observado o coeficiente *Kappa* com 72.8% de concordância.

4.1.5 Base de Dados 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 Formula Secreta* de David Shephard.

Conforme o livro, os professores formularam 10 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 10 questões, 700 respostas e suas respectivas avaliações.

4.1.6 Base de Dados da *UK Open University (Inglês)*

A base de dados da *UK Open University* é um conjunto de questões coletadas na disciplina de introdução à ciências, denominada *S103 - Discovering Science* (**??**). O foco do conjunto de atividades são abordagens em questões factuais bem concisas, não excedendo 20 palavras. Os alunos receberam as atividades através do ambiente 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 analizando 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 resultado varia muito entre alguns poucos minutos até mais do que um dia.

Dentre as 20 questões, esse *dataset* apresenta diferentes quantidades de respostas entre 511 e 1897. A avaliação é gradual e binária, definindo cada resposta como correta ou incorreta. Não existe notas intermediárias, representando diretamente se o aluno atendeu ou não os requisitos da resposta.

4.1.7 Base de dados da *University of North Texas (Inglês)*

O *dataset* da *University of North Texas - UNT* (MOHLER; BUNESCU; MIHALCEA, 2011), conhecido como *Texas dataset*, é uma coleção de questões discursivas extraída

no curso de Ciência da Computação. Composta por 80 atividades únicas, esse conjunto é composto por 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 2273 respostas de alunos dentre as 80 atividades.

A avaliação foi feita com cinco notas gradativas, 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 é a resultante da média ente os avaliadores, em intervalo contínuo. Dentre as notas atribuídas, 57,7% das respostas receberam a mesma nota. 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 atribuia maiores notas 76% das vezes.

4.1.8 Base de Dados 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 5 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 a 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 pontos com desvio padrão de 1,75.

4.2 Experimentos de Clusterização

4.3 Experimentos de Classificação

4.4 Discussão de Resultados

5 Considerações Finais

5.1 Trabalhos Futuros

5.2 Conclusões

Referências

ARTER, J. A.; CHAPPUIS, J. *Creating & Recognizing Quality Rubrics*. Nee York, USA: Pearson Education, 2006. (Assessment Training Institute, Inc Series). Citado na página 19.

BAILEY, S.; MEURERS, D. Diagnosing Meaning Errors in Short Answers to Reading Comprehension Questions. In: *Proceedings of the 3rd Workshop on Innovative Use of NLP for Building Educational Applications, held at ACL 2008*. Columbus, Ohio, USA: Association for Computational Linguistics, 2008. p. 107–115. Citado na página 20.

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 19.

BURROWS, S.; GUREVYCH, I.; STEIN, B. The Eras and Trends of Automatic Short Answer Grading. *International Journal of Artificial Intelligence in Education*, v. 25, n. 1, p. 60–117, 2015. Citado 4 vezes nas páginas 20, 21, 23 e 24.

BUTCHER, P. G.; JORDAN, S. E. A Comparison of Human and Computer Marking of Short Free-Text Student Responses. *Computers & Education*, v. 55, n. 2, p. 489 – 499, 2010. Citado 3 vezes nas páginas 21, 23 e 24.

DZIKOVSKA, M. O.; NIELSEN, R. D.; BREW, C. Towards Effective Tutorial Feedback for Explanation Questions: A Dataset and Baselines. In: . USA: Association for Computational Linguistics, 2012. (NAACL HLT '12), p. 200–210. Citado na página 36.

GALHARDI, L.; SOUZA, R.; BRANCHER, J. Automatic Grading of Portuguese Short Answers Using a Machine Learning Approach. In: *Anais Estendidos do XVI Simpósio Brasileiro de Sistemas de Informação*. Porto Alegre-RS, Brasil: SBC, 2020. p. 109–124. Citado na página 35.

GOMAA, W. H.; FAHMY, A. A. Short Answer Grading Using String Similarity And Corpus-Based Similarity. *International Journal of Advanced Computer Science and Applications(IJACSA)*, v. 3, n. 11, 2012. Nenhuma citação no texto.

LEFFA, V. J. Análise Automática da Resposta do Aluno em Ambiente Virtual. *Revista Brasileira de Linguística Aplicada*, SciELO, v. 3, p. 25 – 40, 00 2003. Citado na página 20.

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 - Volume 1*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2011. (HLT '11), p. 752–762. Citado 2 vezes nas páginas 21 e 37.

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*. Vila Velha-ES, Brasil:

Instituto Federal do Espírito Santo, 2020. v. 9. (Série Guia Didático de Ciências/Química, v. 9). Citado na página 37.

NOORBEHBAHANI, F.; KARDAN, A. A. The automatic assessment of free text answers using a modified bleu algorithm. *Computers & Education*, Elsevier Science Ltd., Oxford, UK, UK, v. 56, n. 2, p. 337–345, feb 2011. Nenhuma citação no texto.

OLIVEIRA, E. et al. Combining Clustering and Classification Approaches for Reducing the Effort of Automatic Tweets Classification. In: *KDIR 2014 - Proceedings of the International Conference on Knowledge Discovery and Information Retrieval*. Rome, Italy: ScitePress, 2014. v. 6, p. 465–472. Citado na página 20.

OLIVEIRA, E. et al. Uma Tecnologia de Agrupamento de Respostas para Redução de Esforço de Correção de Atividades em Sistema Online de Apoio à Avaliação Formativa em Indexação. In: *XI Encontro Nacional de Pesquisa em Ciência da Informação*. Rio de Janeiro, RJ, Brasil: Associação Nacional de Pesquisa e Pós-Graduação em Ciência da Informação (Ancib), 2010. Citado na página 24.

OLIVEIRA, K. L. d.; SANTOS, A. A. A. Compreensão em Leitura e Avaliação da Aprendizagem em Universitários. *Psicologia: Reflexão e Crítica*, SciELO, v. 18, p. 118 – 124, 04 2005. Citado na página 19.

OLIVEIRA, M. G.; CIARELLI, P. M.; OLIVEIRA, E. Recommendation of Programming Activities by Multi-label Classification for a Formative Assessment of Students. *Expert Syst. Appl.*, Pergamon Press, Inc., Tarrytown, NY, USA, v. 40, n. 16, p. 6641–6651, nov 2013. Nenhuma citação no texto.

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 na página 19.

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. Nenhuma citação no texto.

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, Setembro 2014. Citado 2 vezes nas páginas 31 e 38.

RAMACHANDRAN, L.; CHENG, J.; FOLTZ, P. W. Identifying patterns for short answer scoring using graph-based lexico-semantic text matching. In: *BEA@NAACL-HLT*. Denver, Colorado: Association for Computational Linguistics, 2015. Citado na página 21.

RAMACHANDRAN, L.; FOLTZ, P. W. Generating reference texts for short answer scoring using graph-based summarization. In: *BEA@NAACL-HLT*. Denver, Colorado: Association for Computational Linguistics, 2015. Nenhuma citação no texto.

ROUSSEEUW, P. J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, v. 20, p. 53 – 65, 1987. Citado na página 32.

SPALENZA, M. A. et al. Uso de Mapa de Características na Avaliação de Textos Curtos nos Ambientes Virtuais de Aprendizagem Classes de Respostas Discursivas. In: *Simpósio Brasileiro de Informática na Educação (SBIE)*. Uberlândia, MG, Brasil: Sociedade Brasileira de Computação, 2016. v. 27. Citado 2 vezes nas páginas 19 e 31.

Apêndices