Marcos Alécio Spalenza

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

Vitória, ES

2021

Marcos Alécio Spalenza

*p*Nota: Uma Análise das Estruturas 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. A aplicação frequente permite 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 estabelecem meios para reformulação e controle do processo de ensino-aprendizagem (BARREIRA; BOAVIDA; ARAÚJO, 2006). Portanto, através das atividades identificamos a proficiência dos estudantes sobre determinado domínio e sua capacidade de realizar inferências sobre o assunto. O papel da avaliação, portanto, é diagnosticar, apreciar e verificar o aprendizado dos alunos para que o professor atue no processo de formação de modo a consolidar seu método de ensino (OLIVEIRA; SANTOS, 2005).

O modelo de ensino-aprendizem é a ferramenta que torna possível observar os problemas com o aprendizado dos alunos e as ações 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 (BIGGS, 1998). Portanto, é através das atividades que criamos o modelo para mensurar o conhecimento individual dos alunos. Para isso, a mediação tecnológica consolidou-se para aplicação das atividades em quantidade e qualidade. Deste modo, os Ambientes Virtuais de Aprendizagem (AVA) (MAQUINÉ, 2020) se tornaram modelos virtuais para suporte das aulas para turmas presenciais e a distância (RAES et al., 2020). Com a mediação tecnológica, apoiamos o professor na criação, avaliação, recomendação e visualização de dados educacionais impactando diretamente no acompanhamento do currículo do aluno (PAIVA et al., 2012). Deste modo, é com as ferramentas de apoio que o tutor verifica a aptidão dos estudantes, de forma individual ou coletiva, para melhorar a adaptação e a experiência da 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., 2016b).

No primeiro modelo, nas questões fechadas, a resposta esperada é dada de forma direta e não contextualizada, sendo que este é uma premissa já apresentada pelo enunciado da atividade. Por outro lado, nas questões fechadas, a resposta é descritiva onde o estudante pode contextualizar o conteúdo de sua resposta em texto. Tais modelos têm diferentes intúitos na aquisição de informação sobre o conhecimento do estudante sobre o tema e, por consequência, seu conteúdo deve ser analizado de formas específicas. Portanto, por definição, a chave para discriminar tais atividades é a necessidade de maior ou menor conhecimento factual, impactando diretamente na liberdade de criação e escrita do estudante em seu conjunto de resposta.

1.1 Problema

Dentro da literatura da avaliação de respostas discursivas curtas, em inglês *Short-Answer Graders* (SAG), encontramos determinados problemas listados por autores durante os anos de evolução da pesquisa. Os problemas são amplos, onde se busca aprimorar gradativamente os modelos avaliativos para conseguir resultados cada vez mais adequados aos do professor (PADÓ; PADÓ, 2021). Apesar de ser um estudo já realizado há décadas, ainda encontramos nos modelos SAG uma série de desafios com demandas importantes e pouco estudadas até o momento.

Nos primeiros sistemas, a modelagem de questões discursivas era um trabalho realizado com o texto bruto (PÉREZ-MARÍN; PASCUAL-NIETO; RODRÍGUEZ, 2009). 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 (FILIGHERA; STEUER; RENSING, 2020).

Na principal revisão da literatura sobre os sistemas SAG (BURROWS; GUREVYCH; STEIN, 2015), os autores reúnem 37 trabalhos realizados na área. Durante esse estudo, o autor destaca o problema da profundidade do aprendizado, em tradução literal para "*depth of learning*", separando as atividades em dois grupos: de reconhecimento e de recuperação. No Brasil, conhecemos discriminamos os dois tipos de questão como abertas e fechadas (GUNTHER; LOPES-JÚNIOR, 2012). Então, é característica dos modelos SAG a produção de modelos complexos de correção de questões de recuperação, interpretando computacionalmente o conteúdo de respostas curtas em textos de escrita livre.

Um problema diretamente associado ao modelo de interpretação textual, entretanto, é a busca de convergência entre as respostas. É uma dificuldade a extração do viés de

resposta em questões discursivas factuais com múltiplos contextos. Além disso, existe ainda maior complexidade para lidar com questões que resultam em respostas opinativas, individuais ou subjetivas (BAILEY; MEURERS, 2008). Apesar do conteúdo, é esperado do sistema que, independente do conteúdo recebido do professor, lide com a liberdade de escrita do estudante e analize a convergência entre respostas na tentativa de recuperar padrões compatíveis (SAHA et al., 2018).

Em geral, para além do reconhecimento de padrões de resposta, ainda existe o alinhamento entre o conteúdo das respostas e o critério avaliativo. Esse fator se destaca pela referência utilizada na criação do modelo avaliativo (KRITHIKA; NARAYANAN, 2015). O professor, no papel de especialista, deve ser seguido segundo seu padrão avaliativo na tentando imitá-lo (JORDAN, 2012; FUNAYAMA et al., 2020). Para ilustrar esse aspecto podemos usar como exemplo a avaliação de plágios de forma negativa. Como é de interesse dos sistemas SAG seguir o modelo avaliativo do professor, mesmo que a essência do conteúdo seja coincidente com respostas próximas, o padrão específico reconhecido negativamente deve receber avaliação equivalente por conta do plágio segundo o método avaliativo do professor. Assim, apesar da escolha de palavras alinhadas com um determinado modelo de resposta, é essencial que o sistema forme vínculos entre padrões de avaliação e de respostas para criação de modelos avaliativos robustos (HIGGINS et al., 2014). Nesse aspecto destacamos que é fundamental para além de um sistema tradicional de reconhecimento de padrões a extração do critério avaliativo do professor.

Sabendo disso, o critério avaliativo deve atrelar componentes textuais ao método avaliativo. A aquisição desse modelo deve ser feita através da identificação da forma que o professor avalia uma série de respostas. Porém, uma série de trabalhos utiliza descritores do padrão avaliativo para representar a forma que o professor interpreta modelos de resposta dentre expressões regulares e regras até o quadro de *rubrics* (BUTCHER; JORDAN, 2010; MOHLER; BUNESCU; MIHALCEA, 2011; RAMACHANDRAN; FOLTZ, 2015). Porém, isso contrapõe a proposta de reduzir o esforço avaliativo do professor, se considerarmos a necessidade de produção de qualquer conteúdo extra sobre a avaliação (ZESCH; HEILMAN; CAHILL, 2015; HORBACH; PINKAL, 2018). A partir daí para remontar o critério avaliativo do professor, deve priorizar o uso de padrões de avaliação sem requisitar descritores ou chaves de resposta.

Dentro desse aspecto, durante a análise da compatibilidade entre os modelos do sistema *FreeText Author* e do professor (BUTCHER; JORDAN, 2010), os autores elencaram seis problemas. O primeiro é a omissão dos padrões de avaliação. O segundo é a identificação da associação entre palavras e a sua conexão com o modelo avaliativo. O terceiro é a necessidade de identificação estrutural da sentença. O quarto é o tratamento de classificações incorretas, em especial por parte do especialista. O quinto é o conflito entre padrões corretos e incorretos. E, por fim, o sexto problema listado pelos autores é

diretamente relacionado aos demais, indicando o problema de confiabilidade do sistema como avaliador e a interpretação inconsistente do conteúdo textual.

Na perspectiva da omissão dos padrões avaliativos, detalhamos a diferença entre o sistema ter o conhecimento dos padrões textuais e a avaliação de padrões desconhecidos. Padrões desconhecidos podem ter *outliers* que estritamente recebem um modelo próprio de avaliação (FILIGHERA; STEUER; RENSING, 2020). Entretanto, para além dos métodos aleatórios de amostragem, a avaliação de questões discursivas curtas *a priori* indica um problema de classificação desbalanceada (DZIKOVSKA; NIELSEN; BREW, 2012). Deste modo, destacamos a importância dos métodos de anotação direcionada a diversidade textual e uso de métodos de verificação da distribuição das amostras (MARVANIYA et al., 2018).

O segundo listado, em uma outra esfera avaliativa, estabelece a criação de modelos robustos entre termos e classes (RAMACHANDRAN; FOLTZ, 2015). Portanto, torna-se característico segundo os autores que os modelos SAG devem incorporar detalhes sutis da avaliação (HORBACH; PINKAL, 2018). Portanto, a relação termo-classe deve ser dinâmica e, apesar da avaliação ser passível de revisões e ajustes a qualquer momento, sempre extrair o modelo que melhor atenda às expectativas do professor (SPALENZA et al., 2016b).

Na sequência, o terceiro problema listado é de aspecto estrutural, observando cada resposta segundo os detalhes de sua construção. Por consequência, além da análise detalhada do modelo textual, é fundamental uma extensa capacidade analítica do conteúdo (SAHA et al., 2018). Portanto, além do nível textual desejamos que a análise seja feita em vários níveis, incluindo verificação léxica, morfológica, semântica e sintática de cada resposta (SAKAGUCHI; HEILMAN; MADNANI, 2015; RIORDAN; FLOR; PUGH, 2019; SAHU; BHOWMICK, 2020). Deste modo, incluimos neste aspecto, além de formas de maximizar a aquisição de informações em texto, a análise estrutural para compreensão da escrita das respostas. Somado a isso, algumas abordagens vão além e ainda exploram a conexão semântica entre respostas, questões e domínios (DZIKOVSKA et al., 2013; SAHA et al., 2019).

O quarto problema inclui o tratamento de classificações incorretas. É extremamente relevante aos sistemas SAG a construção de justificativas com fundamentos em referências textuais (FUNAYAMA et al., 2020). Assim, é recorrente a possibilidade de remontar as componentes que levam a correção de cada resposta, sejam regras de associação de respostas, padrões de expressões regulares ou a extração de características textuais (CHAKRABORTY; ROY; CHOUDHURY, 2017; KUMAR et al., 2019). Para além da necessidade de justificativa, para cada nota atribuída, ainda ressaltamos a capacidade de identificar *outliers* (DING et al., 2020) e garantir que não se torne uma influência ao método avaliativo. Nessa linha, é importante que os modelos compreendam o conteúdo sem

avaliações tendenciosas (AZAD et al., 2020), realizando uma análise ampla do conteúdo anotado.

De forma contínua ao quarto, o quinto problema compreende a identificação de incoerências nas avaliações. Entretanto, a incoerência é algo esperado desde que a divergência existe mesmo que entre dois humanos especialistas (ARTSTEIN; POESIO, 2008; PADÓ; PADÓ, 2021). Mas é essencial minorar a diferença cada vez mais entre o modelo do especialista e o modelo de avaliação automática (CONDOR, 2020). Nessa dinâmica, ressaltamos a importância em isolar comportamentos anômalos do método avaliativo para que não influencie no comportamento geral do modelo automático.

Por fim, a confiabilidade do sistema, tangenciando todos os demais citados, é tratada no último item. Superficialmente podemos associar este problema a divergência de notas entre avaliadores. Porém, em um aspecto amplo, a confiabilidade do sistema passa do reconhecimento do critério avaliativo à criação de justificativas de nota através de modelos descritivos de *feedback* (KUMAR et al., 2019). O papel dos modelos de *feedback* vai além de descrever o que o sistema observou na avaliação. Este declara a todos os participantes a relação entre as respostas, o reconhecimento do critério de avaliativo e cada anotação do professor (MARVANIYA et al., 2018). Portanto, a confiabilidade do sistema passa por todos os níveis, desde a aquisição de um critério de avaliação coerente até a representação do conhecimento.

Para além disso 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. Assim, neste trabalho identificamos, testamos e descrevemos uma série de *datasets* na avaliação do método proposto.

1.2 Proposta

Neste trabalho apresentamos um método de avaliação de respostas discursivas curtas através da análise da estrutura textual para produzir modelos avaliativos complexos. Para seu desenvolvimento, identificamos os problemas mais comuns descritos na literatura como deficiências dos sistemas SAG, apresentando uma proposta de solução. Cada um destes problemas é detalhadamente descrito na Seção 1.1. Portanto, a ideia é desenvolver uma estrutura de reconhecimento do critério avaliativo do professor estabelecendo a relação entre as respostas e as notas atribuídas.

Para atender as demandas encontradas nos trabalhos em SAG utilizamos de técnicas clássicas de *Educational Data Mining* (EDM) (ROMERO et al., 2010), *Machine Learning*

(ML) (HAN; PEI; KAMBER, 2011) e *Natural Language Processing* (NLP) (JURAFSKY; MARTIN, 2009). Apesar do método ter fundamento em modelos linguísticos complexos e comportar questões em diversas linguagens, o avaliamos nas principais bases de dados em *inglês* e *português* da literatura. Dentre os *datasets* observamos 3 tipos de avaliações: notas ordinais, notas discretas e notas contínuas (MORETTIN; BUSSAB, 2010). 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 (BAEZA-YATES; RIBEIRO-NETO, 2011). 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 (ROMERO et al., 2010). 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 (EVERITT et al., 2011). 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 tem por premissa atender as expectativas do professor (PADÓ; PADÓ, 2021). 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 (ARTSTEIN; POESIO, 2008). 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 encontrado em base de dados de questões discursivas curtas. Para atingir o objetivo geral descrevemos os seguintes objetivos específicos:

- Organizar os *datasets* públicos da literatura para estabelecer uma comparação com resultados obtidos em estudos correlatos (BURROWS; GUREVYCH; STEIN, 2015);
- Estudar o impacto das técnicas de Processamento de Linguagem Natural e Recuperação da Informação para a identificação da relação termo-classe de forma gramatical, léxica, morfológica, semântica, sintática, estatística ou espacial (GALHARDI; BRANCHER, 2018; KUMAR et al., 2019; SAHU; BHOWMICK, 2020);
- Alinhar modelos de respostas dadas por professores e alunos, observando a frequência de ocorrência e co-ocorrência de termos segundo sua relevância (JORDAN, 2012; SAHA et al., 2018; DING et al., 2020);
- Criar modelos avaliativos robustos através do reconhecimento de padrões de resposta em categorias de dados discretos e contínuos (BUTCHER; JORDAN, 2010; HEILMAN; MADNANI, 2015; 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 (ZESCH; HEILMAN; CAHILL, 2015; CONDOR, 2020; PADÓ; PADÓ, 2021);
- Identificar a relação da avaliação com o comportamento textual da classe para remoção de *outliers* e manter a consistência da classificação (DING et al., 2020; FILIGHERA; STEUER; RENSING, 2020);
- Apresentar avaliações adequadas ao formato de correção do professor (HIGGINS et al., 2014; FUNAYAMA et al., 2020; PADÓ; PADÓ, 2021);
- Gerar *feedbacks* que colaborem com o processo avaliativo, como o quadro de *rubrics*, de forma a contribuir com a discussão de resultados e a representação do critério de correção (MARVANIYA et al., 2018; MIZUMOTO et al., 2019; SÜZEN et al., 2020).

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. As informações são essenciais para o acompanhar do aprendizado dos alunos, verificar a necessidade de reforço do conteúdo e monitorar o cumprimento do curricular. Tradicionalmente essa dinâmica faz parte dos métodos de ensino-aprendizagem empregados pelos professores, porém, superam a capacidade analítica dos mesmos (MADERO, 2019). 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 EDM (SIEMENS; BAKER, 2012; ROMERO et al., 2010).

Em EDM, métodos de extração de informação são aplicados aos dados da classe de alunos para a aquisição de conhecimento, apoio ao tutor e acompanhamento do ensino (FERREIRA-MELLO et al., 2019). Através de técnicas de 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 estudantes. Neste ponto, dentro dos métodos de EDM, um nicho de sistemas que tange diretamente essa demanda são os sistemas SAG (BURROWS; GUREVYCH; STEIN, 2015). 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 (OLIVEIRA; CIARELLI; OLIVEIRA, 2013). Ao aluno, este tipo de questão é fundamental para prática da 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 (JOHNSTONE; ASHBAUGH; WARFIELD, 2002).

Apesar da relevância das questões discursivas curtas, sua aplicação é gradativamente reduzida pela alta carga-horária do professor em sala (BILGIN; ROWE; CLARK, 2017). 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 (MING, 2005). O nicho dos métodos computacionais de apoio aos métodos avaliativos são conhecidos também por CAA (PÉREZ-MARÍN; PASCUAL-NIETO; RODRÍGUEZ, 2009). 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 (BUTCHER; JORDAN, 2010).

Para o uso dos métodos de SAG, é importante que a questão seja elaborada com o objetivo de analisar os conhecimentos dos estudantes segundo um domínio específico (BAILEY; MEURERS, 2008). Comumente separamos as questões em discursivas e objetivas, de acordo com o modelo de resposta esperada. As questões discursivas (BEZERRA, 2008) 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 1 caracteriza as atividades segundo os modelos de resposta (BURROWS; GUREVYCH; STEIN, 2015).

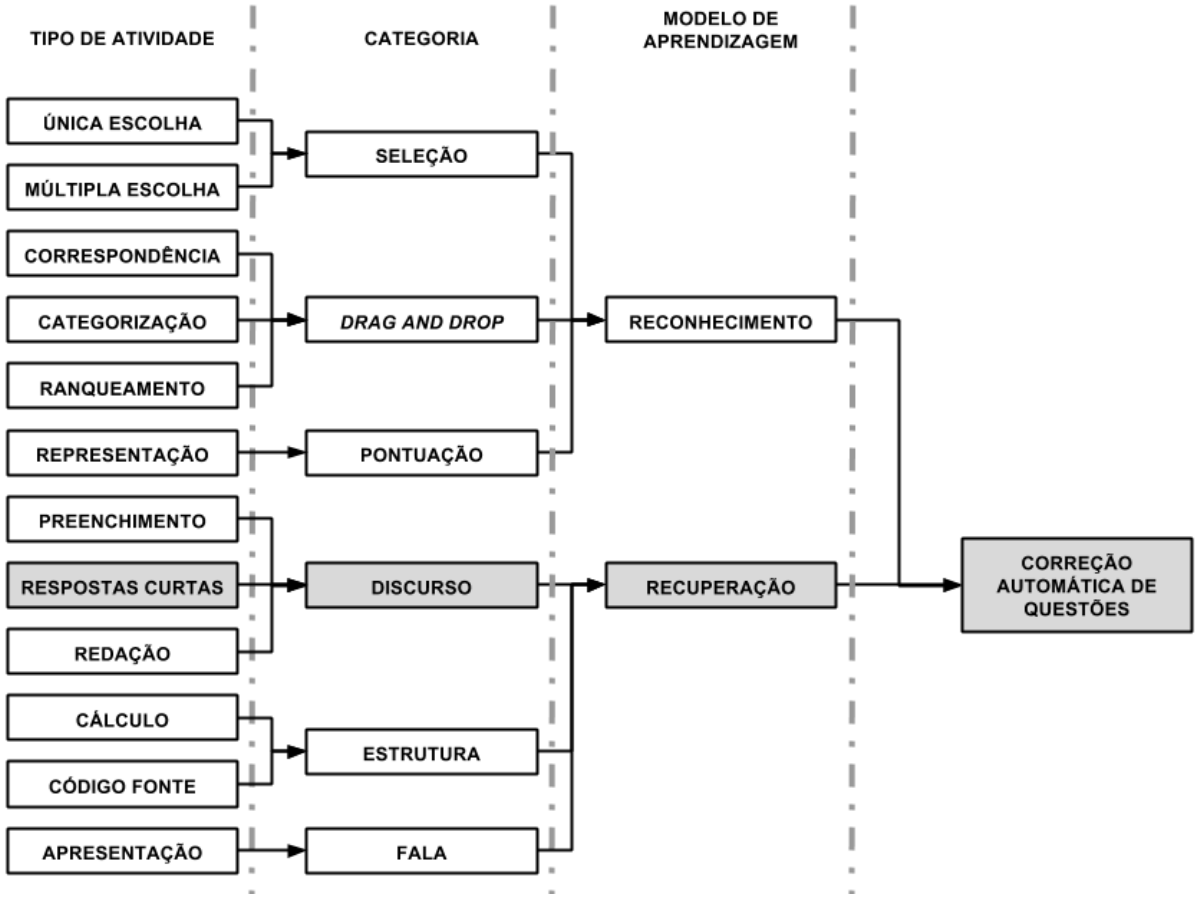

Figura 1 – A extração da informação e os tipos tradicionais de atividade aplicadas no cotidiano de sala de aula.

Como apresentado na Figura 1 o professor dispõe de alguns modelos de atividades que, refletem diferentes aspectos do aprendizado. 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 (BAILEY; MEURERS, 2008). 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 (SIDDIQI; HARRISON, 2008). Por outro lado, a resposta deve representar a informação completa da questão, dando ao sistema embasamento para correção, evitando informações restritas ou codificadas (DING et al., 2020). A Figura 2 demonstra como o espectro de questões trabalhados através das respostas discursivas curtas.

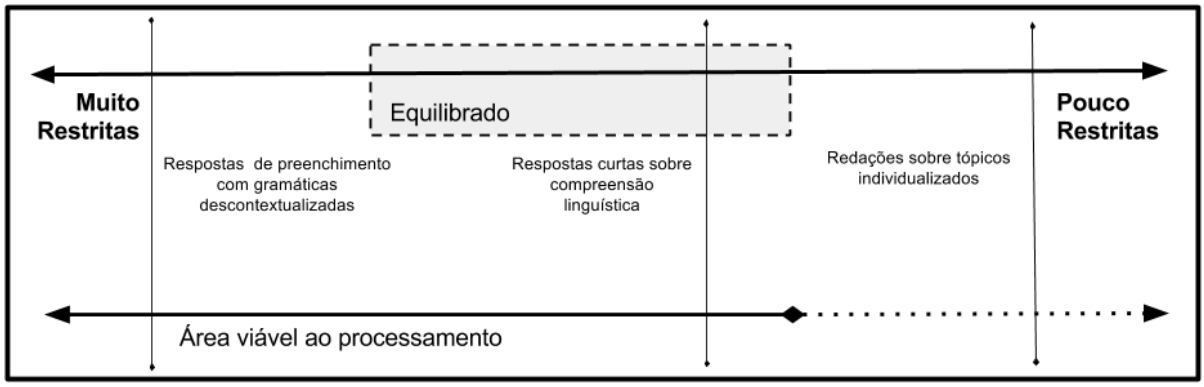

Figura 2 – Extração de informação em questões discursivas: entre respostas pequenas não-convergentes e a subjetividade das competências na avaliação de redações.

A Figura 2 caracteriza exatamente, dentro do nicho de questões discursivas, a dinâmica de uso do processamento computacional das respostas por parte do professor. O ideal é que a questão direcione o aluno a uma ou poucas respostas, evitando várias respostas corretas divergentes (SÜZEN et al., 2020) ou questões longas e subjetivas como redações (ALMEIDA-JÚNIOR; SPALENZA; OLIVEIRA, 2017). Quando as respostas não são únicas e apresentam um conhecimento comum não abstrato, é o ideal para uso 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 (FILIGHERA; STEUER; RENSING, 2020). Para isso, é fundamental que o sistema realize três passos. O primeiro é o aprendizado do modelo de respostas do aluno (RAMACHANDRAN; CHENG; FOLTZ, 2015). O segundo é através do modelo de respostas reconhecer o padrão avaliativo do professor (FUNAYAMA et al., 2020). Por fim, no terceiro passo, o sistema deve replicar o modelo avaliativo e elaborar *feedbacks* coerentes (FOWLER et al., 2021).

2.1 Avaliação Semi-Supervisionada

O método de aprendizado é o procedimento que dita a forma de aquisição de informações do sistema para 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* (HORBACH; PINKAL, 2018). 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 dos estudantes uma amostra, o particionamento em treino e teste reflete a divisão *a priori* do conjunto de dados em um grupo para criação do modelo e outro para avaliação (HEILMAN; MADNANI, 2015). Esse modelo clássico permite ao sistema observar apenas uma parcela dos dados, onde o sistema realiza a inferência nos demais dados 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 (SUNG et al., 2019).

Outros métodos, mais próximos da demanda do professor, utilizam de exemplos anotados de respostas para criação de modelo (BANJADE; RUS; NIRAULA, 2015; ROY et al., 2016). 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 (KAR; CHATTERJEE; MANDAL, 2017; JIMENEZ; BECERRA; GELBUKH, 2013). Porém, este tipo de treinamento gera uma tendência na avaliação, com limitada interpretação das respostas dos alunos (RAMACHANDRAN; FOLTZ, 2015). 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 (BURROWS; GUREVYCH; STEIN, 2015). Além de tornar-se engessada, não necessariamente são similares aos 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 (RAMACHANDRAN; CHENG; FOLTZ, 2015; WANG et al., 2019). 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 (MIZUMOTO et al.,

2019; MARVANIYA et al., 2018). Por outro lado, existem sistemas que demandam modelos mais complexos do método avaliativo, como regras de avaliação e filtros de conteúdo feitos manualmente (PRIBADI et al., 2017; BUTCHER; JORDAN, 2010).

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 (KUMAR et al., 2019; LUN et al., 2020). 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 (ZHANG; SHAH; CHI, 2016). 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 (BASU; JACOBS; VANDERWENDE, 2013; ZESCH; HEILMAN; CAHILL, 2015).

De forma diferente das estratégias citadas, o aprendizado semi-supervisionado proposto combina os métodos de *clusterização* e classificação (OLIVEIRA et al., 2014). 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 (EVERITT et al., 2011). Entretanto, na associação entre *clusterização* e classificação, os grupos são formados para amostragem, partindo dos *clusters* para reconhecimento da distribuição dos documentos. Essa amostragem visa identificar os itens que melhor descrevem cada agrupamento, 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. As referências a cada conjunto de documentos podem ser dadas também 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 (BAEZA-YATES; RIBEIRO-NETO, 2011). 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 (JURAFSKY; MARTIN, 2009). 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 (PAIVA et al., 2012; YANG et al., 2021).

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 (PADÓ; PADÓ, 2021). Assim, a combinação entre o reconhecimento do modelo avaliativo e o reconhecimento do modelo textual deve atender às expectativas do professor (CONDOR, 2020). 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 (AZAD et al., 2020). 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 (MOHLER; BUNESCU; MIHALCEA, 2011).

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 (GALHARDI; BRANCHER, 2018). Os descritores são caracteríticas simples extraídas segundo o formato da escrita de cada documento. Em geral, são formados por características pré-definidas, de acordo com a estrutura da resposta do aluno, sem levar em consideração a profundidade do conteúdo (MOHLER; MIHALCEA, 2009). Dentre os descritores, os mais comuns eram a contagem de erros da linguagem, a quantidade de palavras e a frequência de certas classes gramaticais (RIORDAN; FLOR; PUGH, 2019; GALHARDI et al., 2018). 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 (KUMAR et al., 2019; SAHA et al., 2018). 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 (TAN et al., 2020). Através dessas associações, os sistemas estabeleceram relações complexas entre os termos de cada resposta e o método de atribuição de nota do professor (SAHU; BHOWMICK, 2020).

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, desenvolvimento e uso de técnicas de ponderação, seleção de características e identificação de padrões textuais (BANJADE et al., 2016). Para ponderação textual o modelo mais comum é o Term Frequency - Inverse Document Frequency (TF-IDF) (BAEZA-YATES; RIBEIRO-NETO, 2011). O TF-IDF é um método clássico que realiza a ponderação de acordo com a frequência dos termos, equilibrando a relevância de cada termo segundo sua ocorrência nos documentos e no *dataset* (SULTAN; SALAZAR; SUMNER, 2016). Por outro lado, dentre as técnicas de seleção de características que se destacam, o *Latent Semantic Analysis* (LSA) (LANDAUER; FOLTZ; LAHAM, 1998) é uma das mais utilizadas na literatura (BASU; JACOBS; VANDERWENDE, 2013; SAHU; BHOWMICK, 2020). O uso desta técnica compreende identificar relações semânticas dentro do conjunto de respostas (MOHLER; MIHALCEA, 2009). 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 (SAHA et al., 2018; ZESCH; HORBACH, 2018). Uma estratégia é a análise estrutural dos termos, observando a construção frasal de cada resposta de aluno. Deste modo, na literatura alguns trabalhos citam a análise da construção gramatical das sentenças (RAMACHANDRAN; CHENG; FOLTZ, 2015; ROY et al., 2016).

Outras propostas porém, remontam o contéudo das respostas sob a perspectiva sequencial da construção textual (KUMAR; CHAKRABARTI; ROY, 2017). A análise, com a seleção de n termos de cada sentença da resposta é denominada *n-grams* (MANNING; SCHUTZE, 1999). Os sistemas avaliam as respostas através da vetorização das respostas com análise de compatibilidade entre essas sequências (SAKAGUCHI; HEILMAN; MADNANI, 2015; SULTAN; SALAZAR; SUMNER, 2016). Essas sequências subdividem cada resposta em pequenos trechos que contém de 1 a n termos para aplicar na análise de

equivalência e sobreposição entre respostas (JIMENEZ; BECERRA; GELBUKH, 2013).

Ainda nestes modelos, destacam-se propostas de seleção de características e filtragem de conteúdo (HIGGINS et al., 2014; SPALENZA et al., 2016b). Para filtragem de conteúdo, a identificação de termos comuns ou de baixa frequência representam um refinamento no modelo para análises mais consistentes do conteúdo (ZHANG; LIN; CHI, 2020; MARVANIYA et al., 2018). Termos comuns da linguagem em geral podem ser encontrados como *stopwords*, organizados em listas, são conectivos linguísticos muito utilizados que não têm aderência ao tema (JURAFSKY; MARTIN, 2009). Entretanto, em situação oposta, palavras com baixa frequência, com uso específico e, em geral, não são fundamentais para a resposta do aluno. Em ambos os casos, a filtragem propõe que termos de baixa correlação com o tema sejam removidos. Com uma proposta diferente, a seleção de características interpreta o conjunto de documentos em busca de termos correlatos. De acordo com a frequência de ocorrência e associação dentro do conjunto de respostas, termos são selcionados visando ampliar a capacidade do modelo avaliativo (KRITHIKA; NARAYANAN, 2015; SPALENZA et al., 2016a; HORBACH; PINKAL, 2018). Portanto, o intúito da seleção de características é diretamente relacionado ao modelo linguístico e avaliativo da base de conhecimento. Nessa perspectiva, apenas os termos selecionados são utilizados para representar o conjunto de respostas.

Recentemente, algo um pouco mais robusto do que a análise de vizinhança de termos vêm sendo empregada para avaliar a linguagem segmentos de resposta. Para isso, cada termo é avaliado por similariade no contexto ao qual é empregado. Um método em especial aplicado nesta proposta é denominado *word embeddings* (SUNG; DHAMECHA; MUKHI, 2019; GHAVIDEL; ZOUAQ; DESMARAIS, 2020). As *embeddings* são modelos linguísticos de grande dimensionalidade adquiridos de uma coleção de documentos (GOLDBERG; HIRST, 2017). Esses modelos relacionam o emprego de cada par de termos encontrados em coleções de larga escala. Assim, os sistemas avaliam a correspondência do emprego dos termos em cada sequência de forma pareada. Deste modo, os sistemas avaliam proximidade entre diferentes termos, frases e contextos de uso para cada resposta dos estudantes (RIORDAN et al., 2017).

2.4 Avaliação de Questões Discursivas Curtas

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

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

O método de mapeamento de conceitos consiste em um processo de detecção de determinado conteúdo nas respostas produzidas pelos estudantes. O reconhecimento de conteúdo, portanto, é realizado com análise de alinhamento entre termos de respostas (JIMENEZ; BECERRA; GELBUKH, 2013). Deste modo, é fundamental neste método avaliativo, identificar a existência dos principais conceitos nas respostas para a atribuição de notas (KAR; CHATTERJEE; MANDAL, 2017; CHAKRABORTY; ROY; CHOUDHURY, 2017). Porém, mesmo com a construção automática de padrões através da amostragem, não é garantida a consistência dos modelos produzidos (AZAD et al., 2020). Deste modo, o principal fator destes sistemas é a busca por compatibilidade entre respostas, tornando o sistema muito dependente do objetivo da questão e o conteúdo enviado nas respostas (FILIGHERA; STEUER; RENSING, 2020).

Por outro lado, métodos de extração de informação apresentam características de identificação factual nas respostas dos estudantes. Portanto, compreendem métodos mais robustos de análise do conteúdo, sendo compostos por operações de reconhecimento de padrões e séries de expressões regulares (RAMACHANDRAN; CHENG; FOLTZ, 2015; BUTCHER; JORDAN, 2010). Assim, sistemas SAG com base na extração de informação apresentam modelos de resposta para análise da equivalência de cada resposta com a expectativa de resposta do professor. Deste modo, a associação entre respostas estabelece maior profundidade ao conhecimento do sistema sobre o conteúdo (TAN et al., 2020). Então, o modelo de avaliação utilizado pelo sistema torna-se próximo da observação do professor ao conjunto de respostas, porém, atendendo apenas modelos pré-definidos.

De forma distinta, os métodos baseados em *corpus* traçam análises estatísticas das respostas de cada conjunto de dados (KUMAR et al., 2019). Neste método, os sistemas utilizam de análises da linguagem para validação do alinhamento entre respostas, interpretar variações de uso e caracterizar o conteúdo das respostas (ZIAI; OTT; MEURERS, 2012; MENINI et al., 2019). Para além dos termos utilizados, a adição de informação acrescenta diversidade semântica, tornando modelos mais flexíveis para análise do vocabulário do material (FOWLER et al., 2021).

Apesar da consistência dos modelos anteriores, existem limitações em um âmbito geral da aplicação de cada uma das técnicas de acordo com um base de conhecimento (RIORDAN; FLOR; PUGH, 2019; DING et al., 2020). Em geral, as descrições de modelo

avaliativo do especialista não representam bem o conhecimento para a criação do modelo avaliativo do sistema (FILIGHERA; STEUER; RENSING, 2020). Em contraste aos modelos superficiais, as técnicas de ML foram incorporadas na análise textual para criação de modelos mais robustos, com fundamentação estatística (GALHARDI et al., 2018). Assim, modelos de aprendizado alinham o conteúdo dos documentos, através das diferentes componentes textuais, para reconhecimento dos padrões (SÜZEN et al., 2020). Portanto, os métodos criam estruturas mais complexas que regras, sendo capazes de avaliar formatos distintos de resposta (ZHANG; SHAH; CHI, 2016; SAHA et al., 2019; CAMUS; FILIGHERA, 2020). A robustez destes modelos permite a associação de padrões não convergentes, podendo estabelecer critérios distintos para amostras atribuídas a uma mesma nota.

Em geral, um objetivo dos sistemas SAG, descrito pela literatura, é mesclar os métodos e suas dinâmicas de aprendizado para evolução do modelo avaliativo (BURROWS; GUREVYCH; STEIN, 2015; ZESCH; HORBACH, 2018). Deste modo, é essencial a construção de modelos que comportem padrões avaliativos de alta qualidade e similares ao do especialista, reproduzindo com alta qualidade através de ML (JORDAN, 2012). Apesar das dificuldades e dos detalhes subjetivos da avaliação (ROY; RAJKUMAR; NARAHARI, 2018), o intúito é que o desenvolvimento do modelo avaliativo compreenda a relação entre diferentes características de avaliação e a capacidade de atender diferentes domínios (SUNG et al., 2019; SAHA et al., 2019). Portanto, espera-se o desenvolvimento de sistemas SAG mais robustos, lidando com diferentes combinações entre respostas e avaliações, aprendendo pela demanda do professor o domínio empregado.

3 Método

Neste trabalho, apresentamos um modelo de avaliação de respostas discursivas curtas através da análise da relação entre o conteúdo das respostas dos estudantes e o método avaliativo do professor. Acompanhando o desenvolvimento recente da literatura dos sistemas SAG (BURROWS; GUREVYCH; STEIN, 2015), identificamos pontos sensíveis e problemas descritos nestes estudos. Utilizamos como base os fundamentos de análise documental e modelagem do método avaliativo do tutor para a criação de uma proposta de sistema SAG. Através deste direcionamento, verificamos os principais métodos para análise das componentes textuais para elaborar um conjunto robusto de informações sobre cada resposta. Associamos ao conhecimento das respostas uma descrição do método de correção do professor. Com isso, esperamos construir modelos com o intúito de maximizar os resultados de acordo com o padrão de correções coletados junto ao professor.

Deste modo, apresentamos um sistema composto por quatro módulos. O primeiro módulo é o de coleta de dados, verificação textual, extração de informação e organização do conhecimento. Nesta primeira etapa, o sistema verifica cada resposta individualmente e aplica tratamentos textuais para padronização e extração de características. O resultado desta etapa é o conjunto de vetores de documentos padronizado para processamento. O segundo módulo é composto pelo particionamento de forma semi-supervisionada das amostras para reconhecimento dos padrões de resposta. Tal método analisa a representatividade de cada vetor de respostas, realiza a amostragem e coleta as notas do professor no papel de especialista na avaliação. A próxima etapa recebe os subconjuntos de respostas, uma parte com requisição da avaliação e outra separada para avaliação do próprio sistema. Com isso, o terceiro módulo compreende o reconhecimento do padrão de correções para as amostras e a reprodução do critério avaliativo. A reprodução do processo observado nas amostras selecionadas é dada através de técnicas de classificação e regressão, de acordo com o padrão de notas. Ao fim desta etapa, todas as respostas contém notas atribuídas, sejam dadas pelo professor ou pelo sistema de forma colaborativa. Por fim, com o conjunto de informações utilizadas durante os processos, a quarta etapa, produz históricos, relatórios e *feedbacks* para descrever com detalhes cada *dataset*.

Antes da execução do sistema, a criação de bases de dados compreende organização dos dados e o cumprimento dos padrões de leitura. Cada base de dados deve apresentar uma série de respostas discursivas curtas e um índice como referência a cada aluno. A origem destes dados podem ser arquivos estruturados ou Ambientes Virtuais de Aprendizagem (AVA). Os arquivos estruturados, são conjuntos de amostras de resposta delimitados de forma organizada em colunas para descrever uma comunicação entre sistema e professor, incluindo índice, resposta, nota e *feedback*. Por outro lado, os AVA são plataformas utilizadas

pelos professores para interação direta com o aluno. Podemos citar como exemplos de AVA o Moodle e o *Google Classroom*. O uso deste tipo de sistema ganha ainda mais notoriedade com o Ensina a Distância (EaD), entretato, não está restrito ao mesmo. Para isso, utilizamos de um *framework* de coleta, transferência e controle das atividades da sala virtual para processamento externo. Portanto, é responsabilidade da aplicação a coleta as atividades no ambiente virtual, a transferência para um servidor de processamento e o envio de resultados para o professor. A Figura X apresenta o funcionamento do método de coleta de dados em diferentes plataformas de ensino.

A Figura X apresenta os métodos de extração de informação dos sistema com os AVA. Inicialmente, o módulo de transferência de dados é configurado em ambas as partes, no cliente AVA e no servidor de processamento. Com a configuração, o módulo acessa cada cliente e transfere as atividades ao qual os professores marcaram para análise. O *p*Nota avalia o conjunto de atividades e na primeira etapa realiza a requisição de anotação (avaliação) de amostras para treinamento do algoritmo de avaliação. O módulo de transferência envia marcações nas respostas requisitadas e o professor as avalia em seu ambiente de ensino. O sistema mantém sincronizada a versão da avaliação do professor e a versão no servidor. Em um segundo momento, com todas as respostas requisitadas já avaliadas, o sistema avaliador é treinado e recria o método de avaliação em modelos de classificação / regressão conforme as notas atribuídas. Os resultados, somando as notas atribuídas e os *feedbacks* gerados, são enviados para a plataforma de ensino novamente.

O professor, em qualquer momento fica aberto para finalizar/cancelar o processamento ao liberar a chave de buscas em sua plataforma. Da mesma forma, a nota atribuída pelo professor é considerada a correta, sendo objetivo do sistema atender seu modelo avaliativo. Assim, é livre ao professor a alteração e controle de qualquer nota mesmo que ainda em processo de análise do sistema. Portanto, o professor a todo momento fica responsável por monitorar o processo avaliativo e ajustar os resultados propostos pelo sistema. A análise textual, seleção de amostras, modelos avaliativos e materiais explicativos aplicados pelo *p*Nota em cada atividade são apresentadas detalhadamente em quatro etapas do processo de correção automática.

3.1 Extração das Componentes Textuais

A primeira etapa, denominada de extração de componentes textuais, compreende a análise do conteúdo textual para a extração de conhecimento. Com os documentos organizados, o primeiro processo é a leitura do modelo de dados. Foram observados 3 diferentes modelos de dados: um único arquivo para todo o *dataset* da atividade, um arquivo de resposta textual para cada estudante em uma coleção para a atividade e, por fim, uma estrutura por aluno que contém os arquivos enviados para a com o conteúdo textual

do aluno para a atividade. Com o carregamento dos arquivos, cada aluno é representado pelo seu identificador, seja ele uma referência da plataforma de origem ou sua respectiva ordem na estrutura do *dataset*.

Após a leitura do conjunto de dados, o sistema realiza uma série de processos de padronização, segmentação, filtragem, transformação e vetorização dos documentos. As etapas são sequenciais e encadeadas para análise do conteúdo em diferentes níveis. A padronização é composto pela coleta do conteúdo da resposta, remoção de conteúdos extras que permeiam o texto e a garantia da equivalência na ocorrência de cada termo. A segmentação visa a construção de vetores de resposta, termo-a-termo, identificando séries de *tokens* através de uma heurística cada palavra que a compõe. A filtragem, a partir dos vetores de resposta, seleciona as palavras com potencial relação com o conteúdo. A transformação compreende extrair as estruturas das componentes textuais e modificar cada palavra para um token representante. Por fim, com os documentos padronizados, ocorre a vetorização, analisando a frequência de *tokens* ou séries de *tokens* para representarem o conteúdo da submissão do aluno.

3.1.1 Padronização

Com a extração do texto do documento enviado pelo aluno, o conteúdo, neste primeiro momento, está em estado bruto. No estado bruto o documento precisa ser normalizado seguindo um padrão para os diferentes espaçamentos, acentuação e pontuação. Além disso, é fundamental remover conteúdos não interpretáveis incluindo caracteres não alfanuméricos e *tags* (marcações). Portanto, esta etapa é composta pelos seguintes processos:

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

Após cada um dos processos o conteúdo do aluno está normalizado para as próximas etapas. Os sinais gráficos auxiliam na identificação, pronúncia e leitura dos termos. Porém, computacionalmente, os sinais gráficos não é relevante para identificação de cada termo. O inverso ocorre com marcações de arquivos estruturados. As marcações, apesar da interpretação computacional, não fazem parte do conteúdo produzido pelo estudante. Portanto, ambos os casos não adicionam semanticamente ao conteúdo das respostas.

3.1.2 Segmentação

A partir dos documentos em formato padrão, com o texto normalizado, é possível partir para análise detalhada do conteúdo. Segmentações comuns podem particionar o conteúdo por palavras, por caracteres, por frases ou por parágrafos. Cada particionamento tem um aspecto específico alinhado com as tarefas realizadas na sequência. Neste caso, a análise detalhada do conteúdo depende do particionamento do que foi obtido em segmentos de palavras. Cada segmento é denominado *token*. Os *tokens*, neste caso, representam as palavras separadas por uma heurística, que delimita de cada segmento. Uma heurística simples é a *tokenização* por espaçamento, porém, como é um método simples é sujeito a muitas falhas. Apesar disso, métodos com melhor desempenho compreendem a aplicação da linguagem e consideram formas específicas de pontuação ou divisões do estilo textual.

A segmentação é o método que transforma o conteúdo em uma lista de palavras. A sequência de palavras permite que os próximos níveis trabalhem a perspectiva de cada *token* desta lista ou sua vizinhança. É muito comum que, durante o processo, o documento seja manipulado de diferentes formas, inclusive passando várias vezes pela transformação de texto em lista de *tokens* e vice-versa. Deste modo, os *tokens* permitem que cada palavra seja trabalhada de forma independente, sem impactar no conteúdo adjacente.

3.1.3 Filtragem

A filtragem de conteúdo é uma etapa muito importante desse processo. Uma dificuldade da filtragem é o balanceamento para alcançar níveis desejáveis de aquisição de informação. Portanto, como esperado, a filtragem estabelece uma grande perda de informação e redução do conteúdo dos documentos. Entretanto, vale ressaltar que a perda de informação inerente ao processo caracteriza uma melhoria na consistência e na equivalência dos documentos. Como é uma limpeza conduzida pelo sistema, as características removidas representam detalhes com baixa correlação com a essência de cada resposta.

Em geral, nem todos os termos de uma sentença fazem parte do núcleo de interesse para análise das respostas. Algumas palavras independem do contexto ao qual são empregadas e não são aderentes ao tema. Esse é o caso das *stopwords*. As *stopwords* são palavras que são empregadas na linguagem como conectivos e não representam o conteúdo passado. Assim, são extremamente importantes para a leitura e interpretação do contexto, mas não adicionam informação quando empregadas. Assim, a lista de *stopwords* é um método que restringe a frase a palavras com maior potencial de relação com o contexto e o tema da resposta do estudante. Outros métodos de filtragem também incluem a remoção de palavras com poucos ou muitos caracteres e com tendência a serem muito específicos, quando não se enquadram como *stopwords*. Assim, podemos incluir como parte deste processo as seguintes etapas:

- Remover palavras pequenas (menores que 3 caracteres);
- Remover palavras grades (maiores que 15 caracteres);
- Remover *stopwords*;
- Remover números.

A remoção de partes do conteúdo devem cautelosamente observadas para não impactarem na capacidade de análise do conteúdo. Um bom exemplo é a extração dos números de uma resposta. Este método nem sempre é utilizado pelo sistema, dada sua influência no conteúdo. De modo prático, a aplicação deste método impacta diretamente na capacidade de análise de respostas compostas por números ou datas.

No entanto, podemos destacar a relevância da filtragem de conteúdo através de exemplos de aplicação. Por exemplo, uma situação ao qual encontramos uma fórmula em meio ao texto e, após as padronizações, se tornam comuns os caracteres soltos em meio ao texto. Neste caso, a função que elimina os *tokens* com menos de 3 caracteres remove tal conteúdo descontextualizado e direciona o sistema para atentar-se no contexto descrito pelo aluno. Por outro lado, a função que elimina com mais de 15 caracteres também tem papel fundamental. Podemos tomar como exemplo respostas que o estudante insere uma série de *links* como fontes do que foi descrito na resposta. Sob a perspectiva do sistema os *links* são grandes palavras únicas e não-interpretáveis. Então, a remoção de palavras com uma extensão incomum visa eliminar resquícios de conteúdo como *links* que não foram retirados nos níveis inciais de tratamento por conterem caracteres alfanuméricos.

Nesta etapa, portanto, os filtros de conteúdos são métodos de redução de ruído, responsáveis por discernir quais termos podem ser extraídos de cada item de resposta. O ruído em meio ao texto pode ser um grande problema para o desempenho do classificador em relação a interpretação do conhecimento. Identificando apenas a essência de cada resposta, esperamos que o sistema tenha maior capacidade de interpretativa da relação entre respostas.

3.1.4 Transformação

Em meio aos métodos de padronização, uma importante etapa é a transformação da série de *tokens*. As transformações envolvem métodos complexos de NLP, treinados para classificação de cada *token* segundo sua função na linguagem. Neste nível de trabalho do texto, o texto original é particionado, sendo observado em diferentes perpectivas. Os diferentes níveis analizados nesta etapa é apresentado à seguir:

- Análise gramatical: *Part-of-Speech Tags* (POS-Tags);

- Análise semântica:*Named Entity Recognition* (NER);
- Análise morfológica;
- Modificação: *Stemming*;
- Modificação: *Lemmatization*;
- Modificação: Tipografia.

Cada uma das atividades aplica uma diferente transformação no texto. Os primeiros três níveis, analíticos, acrescentam diversidade na informação de cada *token* do texto. O primeiro, usando *POS-Tags* realiza a análise gramatical de toda sequência. O método *POS-Tag* classifica cada palavra segundo sua função no âmbito gramatical, dentre verbos, adjetivos, pronomes, dentre 17 categorias (MARNEFFE et al., 2021). Em nível semântico, o NER é uma atividade de identificação de entidades nomeadas em meio ao texto livre (PIROVANI et al., 2019). Através do NER, categorias de nomes são definidas conforme a instância ao qual ele representa. Dentre as categorias reconhecidas neste trabalho estão *pessoa* (PER), *local* (LOC), *organização* (ORG) e *diversos* (MISC). Por último, o analisador morfológico identifica detalhes na construção de cada palavra. Pela análise morfológica certas palavras são identificadas segundo sua flexão. Dentre as flexões classificadas por cada termo estão as nominais (como gênero, número e definição) e verbais (pessoa, modo, tempo, voz). Adicionalmente, esse módulo também realiza algumas classificações léxicas de pronomes, adjetivos e advérbios (MARNEFFE et al., 2021).

Com análises linguísticas complexas, cada *token* é observado em diferentes perspetivas. Adicionalmente, para lidar com o texto em si, aplicamos três tipos de modificadores. Os modificadores alteram o texto original para adicionar mais uma padronização. Aqui, no entanto, a padronização torna a linguagem mais próxima da compreensão do sistema do que da linguagem humana. Inicialmente, o processo de modificação de tipografia (*case*) torna todo o texto em letras maiúsculas ou minúsculas. Ao realizar essa mudança o sistema define palavras equivalentes para uma mesma forma. Do mesmo modo, os processos de *stemming* e *lemmatization* extraem das palavras flexionadas suas formas básicas. A forma simplificada extraída através do processo de *stemming* é a raiz da palavra. Enquanto isso, a simplificação resultante do processo de *lemmatization* é o *lemma* da palavra, ou seja, sua forma sem flexões. Em ambos os casos os *tokens* são modificados e todas as palavras de uma mesma base são dadas como coincidentes.

A resultante desses processos é uma forte análise das componentes textuais de cada documento de forma a compreender a construção do texto (SPALENZA et al., 2020). Através dessas verificações textuais, o texto recebe adição de diversos modelos que, em conjunto, caracterizam a construção termo-a-termo de cada frase. Os módulos analíticos, ampliam a informação de cada documento, tornando as nuances da escrita uma variável

de interesse. Enquanto isso, os demais módulos visam aumentar a compatibilidade entre documentos para que escritas similares sejam trabalhadas de modo uniforme. Assim, o sistema como avaliador é responsável por compreender que o *corpus* foi trabalhado em diferentes perspectivas. Os termos são uma referência que buscam alinhamento com a resposta esperada na avaliação. Por outro lado, a estrutura frasal representa diferentes níveis linguísticos da escrita do estudante.

3.1.5 Vetorização

A vetorização, como última etapa do pré-processamento, é responsável por extrair o modelo numérico de cada documento, permitindo mensurar a diferença ou equivalência para os demais itens da coleção. Deste modo, os documentos são representados por vetores numéricos segundo seu padrão de características. Cada uma das características é analisada conforme sua frequência de ocorrência em cada documento do *dataset*. A representação vetorial numérica de cada documento pela frequência é denominada *Term Frequency* (TF). Sendo a coleção de documentos $D = d_0, d_1, d_2, \ldots d_i$ e as características (*features*) encontradas nos documentos $F = f_0, f_1, f_2, \ldots f_j$. Portanto, para cada documento d na coleção D, contamos a frequência de cada *feature* f_j do vocabulário F. Deste modo, a forma vetorial do documento de índice i é dada por d_i, sendo o vetor composto pela frequência n de cada *feature* no documento $n_{i,j}$. Então, podemos representar cada documento em D por sua forma vetorial $d_i = n_{i,0}, n_{i,1}, n_{i,2} \ldots n_{i,j}$, usando TF.

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

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

Portanto, o IDF é uma ponderação na frequência de cada *feature* no vetor $d_{i,j}$, segundo o total de documentos n_{d_j} que contém f_j em relação ao total de documentos da coleção D. Essa ponderação reduz a diferença numérica entre uma característica encontrada em todos os documentos para as características que estão em grupos de documentos. Assim, o uso deste modelo potencialmente delimita melhor características relevantes em avaliações com mais gradações de notas. Então, a aplicação deste modelo está diretamente associada à capacidade de identificação de características com alta correlação a grupos específicos de nota.

No método de vetorização, durante a verificação de frequência de cada característica,

existe a preocupação de manter a relação de vizinhança entre os termos e sua construção sequencial. Assim, para preservar o aspecto textual em sequências e identificar características adjacentes com alta correlação, utilizamos a análise por *n-grams*. Através dos *n-grams*, em vez de cada documento ser representado por um vetor simples da frequência de cada característica, essa frequência é calculada segundo uma sequências de n termos. Sendo aplicado valores n de 1 a 5-*grams*, utilizamos sequencias de 1 até 5 termos para analisar comportamento de cada documento em cada uma de suas perspectivas textuais. Portanto, as diferentes componentes textuais identificadas através de *n-grams* em busca de padrões mais complexos e associações de termos fortemente correlatos ao método avaliativo (SPALENZA et al., 2020).

3.2 Particionamento do Conjunto de Respostas

A partir dos vetores de documentos, o sistema *p*Nota torna-se capaz de comparar itens de resposta segundo suas componentes textuais. Com as características em formato numérico, começamos a interagir com o professor em busca da criação dos modelos que relacionem os documentos com as notas atribuídas. Entretanto, para criação destes modelos, o sistema precisa receber avaliação de alguns documentos para estabelecer a relação entre o que é o conteúdo de cada documento e a nota ao qual cada um recebe. Apesar de que muitos sistemas realizam uma amostragem aleatória, o *p*Nota realiza uma amostragem baseada na distribuição dos vetores. A análise da distribuição dos vetores e suas características é dada por meio de métodos de *clusterização*.

3.2.1 Clusterização

A *clusterização* é realizada com a otimização segundo o *elbow method*. Esse método é designado por testar sequência de valores de parâmetros para identificar a melhor combinação de *clusters* segundo uma métrica de qualidade. Em geral, a métrica de qualidade é diretamente relacionada com o propósito de uso dos *clusters*. O algoritmo de *clusterização* utilizado é o *Agglomerative Clustering* (SPALENZA; PIROVANI; OLIVEIRA, 2019), um método hierárquico de agrupamento por proximidade. O *Agglomerative* compreende formar clusters agrupando item a item até que um limiar de proximidade seja alcançado dado um k número de *clusters* (EVERITT et al., 2011).

Dentre as métricas estudadas estão o *Calinski-Harabasz Score* (CHS) (CALIŃSKI; J., 1974), *Davies-Bouldin Score* (DBS) (DAVIES; BOULDIN, 1979), *Silhouette Score* (SS) (ROUSSEEUW, 1987) e *Sum of Squared Errors* (SSE) (MAIMON; ROKACH, 2005). Essas métricas são denominadas índices de validação interna e avaliam os agrupamentos sem considerar a anotação de cada amostra, ou seja, de modo não-supervisionado. Cada índice é uma heurística utilizada para mensurar, sob diferentes perspectivas, a qualidade dos

clusters gerados em relação a outros agrupamentos em um mesmo *dataset*. CHS mensura a razão entre a dispersão dos itens intra-*cluster* e a dispersão extra-*cluster*. DBS é o indice que estabelece a relação entre a média de similaridade entre as amostras do *cluster* para a média de similaridade entre-*clusters*. O SS é a média entre as distâncias das amostras pertencentes a um *cluster* em relação às amostras do *cluster* mais próximo. Por fim, SSE é uma métrica que avalia o erro de cada amostra que compõe um *cluster* em relação ao seu centróide. O centróide é o ponto médio dos itens que constituem cada *cluster*. Portanto, o centróide é uma instância representante da dispersão dos itens no *cluster*, porém é um ponto artificial e não necessáriamente uma amostra que o compõe.

Para a avaliação de respostas abertas, consideramos que o ideal são as análises que balanceam os itens de cada *cluster* em relação aos *cluters* adjacentes. Por padrão escolhemos a análise de *silhouette*, para identificar os resultados de *clusterização* com maior separabilidade entre os *clusters*. A separabilidade indica se os *clusters* formados são bem definidos, consistentes e sem sobreposição. Nesse índice, valores próximos a 1,0 representam agrupamentos consistentes, com distância para o *cluster* mais próximo. Valores negativos, aproximando-se de -1,0, indicam aleatoriedade na associação entre *clusters* e amostras, com confusão entre as os agrupamentos. Por outro lado, valores próximos a 0,0 indicam sobreposição entre *clusters*, com itens no limiar de pertencer diferentes grupos.

Em relação a verificação dos coeficientes de SS, a otimização com *elbow-method* identifica no intervalo de busca qual maximiza o resultado do índice. A otimização utiliza *Gaussian Process* para redução das possibilidades de busca. Esse método analisa cada teste pela distribuição dos valores da métrica de qualidade como uma *gaussiana*, buscando pontos de máxima da função. A resultante é dada pelo melhor valor encontrado (SPALENZA; PIROVANI; OLIVEIRA, 2019). O atributo de controle é o k, número de *clusters*. O intervalo de k é definido por valores de 2 até $raizn$, sendo n o número de amostras do *dataset* (HAN; PEI; KAMBER, 2011). Simultaneamente, para cada combinação de k são realizados testes com vinte métricas de distância.

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

A resultante da otimização em clustering é escolhida como o teste que apresenta a melhor coeficiente de SS e com maior número de *clusters* formados. Enquanto o SS avalia a separabilidade dos agrupamentos, o maior número de *clusters* formados indica, na perspectiva do modelo avaliativo, uma possível coincidência entre conteúdo e notas. O agrupamento selecionado é utilizado para amostragem em um percentual do conjunto de respostas disponíveis.

3.2.2 Seleção de Amostras

Com a formação dos *clusters*, buscamos identificar as principais respostas de cada agrupamento para coleta do modelo avaliativo do professor. A amostragem é realizada com a coleta de um percentual dos itens que compõe o *dataset*. Essa coleta analisa padrões de documentos de cada grupo, a fim de compreender como é dada a avaliação do especialista para cada um dos diferentes itens. As amostras são selecionadas conforme critérios específicos, descrevendo um padrão específico do *cluster*. Os sete critérios de seleção de amostras utilizados estão listados abaixo:

- Par de amostras de menor similaridade no *cluster*;
- Par de amostras de maior similaridade no *cluster*;
- Amostra com mais características do *cluster*;
- Amostra com menos características do *cluster*;
- Amostra com menor índice de *silhouette* do *dataset*;
- Amostra aleatória do maior *cluster*;
- Amostra aleatória do *dataset*.

As sete instâncias, seguem uma ordem de prioridade conforme o percentual de itens coletados. O par de maior e o par de menor similaridade compreendem os itens mais convergentes e mais divergentes que compõe cada *cluster*. Em uma diferente perspectiva, a coleta de amostras dos itens de maior e menor número de características indicam as respostas que foram mais extensas e mais concisas, respectivamente. A coleta destes itens indica a consistência do padrão reconhecido na atribuição dos *clusters*. Em *clusters* com apenas um par de itens, por exemplo, todos os quatro modelos de amostragem são dados para as duas únicas amostras. Portanto, após essa seleção, não necessariamente foram identificadas 6 instâncias por *cluster*. A composição de até o percentual de amostragem por *dataset* é dado de três formas, aplicada conforme a demanda do sistema. A forma padrão, e mais robusta, é a análise de dispersão de cada item.

A análise de dispersão calcula o coeficiente de *silhouette* da amostra. Tal qual o SS, esse índice determina a razão entre a distância da amostra para os demais itens do grupo em relação aos itens do *cluster* mais próximo. Desta forma, esse método incrementa as amostras por dispersão até alcançar o percentual de amostragem selecionado. Uma outra opção de seleção é a escolha de amostras por balanceamento do tamanho dos *clusters*. Tal método determina que um item seja aleatoriamente selecionado, sendo a seleção ponderada de acordo com a quantidade de itens que compõe cada grupo. Por fim, caso seja descartada

a análise para amostragem dos demais itens, a seleção é realizada aleatoriamente, coletando itens sem uso dos agrupamentos.

Terminando este procedimento de seleção de amostras, as representantes de cada grupo são enviadas para avaliação do professor no papel de especialista. O especialista é responsável por atribuir notas de acordo com seu método avaliativo. Fica a cargo do próprio sistema identificar como os padrões de nota estão alinhados com os padrões textuais das respostas. Para isso, as notas coletadas devem ser estudadas pelo sistema para identificar o alinhamento do modelo avaliativo com o conteúdo das respostas.

3.3 Modelo Avaliativo

O desenvolvimento do modelo SAG acontece depois que o professor analisa todos as requisições de anotação, avaliando-as. Para a criação de um modelo SAG, é fundamental o desenvolvimento de um padrão de associação entre termos e notas. Entretanto, identificar os detalhes observados pelo professor na avaliação não é trivial. Através do conjunto de respostas, o *p*Nota compara respostas que receberam a mesma avaliação para identificar padrões correspondentes. Os padrões encontrados em uma mesma nota, indicam detalhes que provavelmente foram levados em conta pelo professor na hora da avaliação. Deste modo, o sistema visa recuperar a expectativa de resposta por nota diante do alinhamento entre as componentes textuais.

3.3.1 Classificação

3.3.2 Regressão

Em notas atribuídas em intervalos contínuos, são aplicados métodos de regressão. A regressão estima valores segundo o intervalo conhecido através das amostras. Os métodos buscam compreender o ajuste dos dados segundo a distribuição, elaborando o modelo minimizando o erro preditivo. Os cinco métodos de regressão aplicados são a *Regressão Linear*, *Lasso*, *K-Nearest Neighboors*, *Decision Tree* e *WiSARD*.

A Regressão Linear (LNREG) é um algoritmo que avalia a tendência linear das amostras segundo sua distribuição. Essa tendência linear busca, no espaço n-dimensional das características, definir os coeficientes de reta que minimizam o resíduo entre as amostras. É importante para o algoritmo determinar uma função de tendência dos dados. Minimizar o erro através dos coeficientes da função reflete na simplificação do conjunto de dados. Entretanto, é determinante que o modelo não apresente *overfitting* e um baixo desempenho com o viés dos dados de treinamento. Por outro lado, como espera-se do algoritmo, a aquisição de informação deve extrair um modelo que minimamente descreva os dados conhecidos, evitando a ocorrência de *underfitting*. Assim, o modelo simplificado deve ser

direcionado ao desempenho linear e não apenas à associação forte com o conjunto de treinamento. Também é utilizada uma variante do LNREG tradicional, denominada *Least Absolute Shrinkage and Selection Operator - Lasso* (LSREG), que utiliza a normalização dos dados com a função $L1$, reduzindo a complexidade do modelo de dados e previnindo o *overfitting*.

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

Segundo o modelo de notas contínuo, o método de avaliação dos métodos de regressão são dados através do erro da predição em relação a nota esperada. Assim, para mensurar a diferença entre a expectativa do professor e a nota resultante do sistema utilizamos o *Mean Absolute Error* (MAE), o *Mean Squared Error* (MSE) e o *Maximum Error* (MaxE). O MAE, erro médio absoluto, mensura a resíduo absoluto entre a nota predita e a nota dada pelo professor. Em outras palavras, o MAE avalia as diferênças em módulo entre os valores obtidos, segundo o alinhamento de cada predição com a expectativa do professor. Enquanto isso, MSE ou erro médio quadrático, é uma medida do resíduo entre os valores com penalização dos erros absolutos. Assim, através do MSE erros maiores têm maior impacto no sistema quando comparados com erros de menor grau. Por fim, o MaxE ou erro máximo, é a extração do maior erro obtido pelo modelo em relação à avaliação do professor. A Equação X apresenta a fórmula de cada uma das métricas utilizadas para avaliação dos métodos de regressão citados.

Apesar de serem comuns os erros entre modelos computacionais e a expectativa do especialista, é crucial para um bom avaliador automático a proximidade entre os modelos. Assim, é esperado ao sistema que o erro seja minimizado e, como descrito, seja capaz de lidar com diferentes situações. Portanto, a capacidade avaliativa do sistema e seu nível de interpretação da linguagem podem ser mais relevantes a longo prazo do que o erro apresentado em situações incomuns (outliers). Destacamos ainda que, devido o nível de subjetividade inato ao processo avaliativo, os erros são comuns em qualquer correção, inclusive entre dois especialistas humanos. Nos sistemas SAG, foram observados durante a

correção entre professores até x pontos de diferença em uma escala de 0 a 10.

Para seleção do regressor mais adequado utilizamos a correlação de *Pearson*. Independente do nível de erro o método selecionado deve cumprir gradações similares ao método de atribuição de notas do professor. Considerando todos os algoritmos aplicados como capazes de realizar uma avaliação, mesmo que de forma básica, o uso da correlação nos permite verificar se a gradação é equivalente ao que foi atribuído pelo professor no treinamento. Para mensurar isso antes da avaliação, particionamos as amostras anotadas com KFold. O algoritmo é treinado com dois terços das amostras coletadas e validado em um terço. Após a predição no *subset* de avaliação, o modelo com maior índice de correlação observado é dado como padrão. Os resultados são gerados para todos os regressores, mas é encaminhado ao professor como resultado o de melhor correlação conhecida.

3.4 Relatórios

Após a classificação, com a atribuição de notas de todos os alunos, os relatórios são ferramentas importantes para aplicação direta em sala de aula. Os relatórios visam descrever para professores, estudantes e até desenvolvedores como é efetuada a análise das respostas, os métodos de reconhecimento de padrões e a coerência entre avaliação e modelos de resposta. Sabendo que cada modelo de resposta é até o momento uma associação entre termos e notas identificada pelo sistema, é determinante no processo de relatórios a apresentação deste modelo como descritor do método avaliativo.

Os relatórios básicos são apresentação de conteúdos e dados do sistema. Dentre eles estão a lista de respostas dos alunos, lista de amostras requisitadas pelo sistema, lista da frequência das principais características, descrição do particionamento de treino e teste e a descrição das *features* extraídas dos documentos. Em geral, esse *feedback* é um conteúdo explicativo, apresentando o que o sistema interpretou em cada documento e algumas ações básicas tomadas durante os processos como a vetorização e contagem da frequência das características.

Outro relatório inclui a descrição do modelo avaliativo de acordo com cada um dos algoritmos testados. Independente do formato aplicado, o uso das métricas, matrizes de confusão e a comparação do desempenho de cada algoritmo ilustra a entrega de resultados do sistema. Assim, conforme os resultados finais apresentados ao fim do processo avaliativo, podemos acompanhar também a capacidade do modelo de avaliação automática gerado pelo sistema. Em diferentes perspectivas, podemos vislumbrar a capacidade de cada algoritmo avaliador. Entretanto, como principal interessado no ajuste avaliativo, destacamos a relevância do *p*Nota ilustrar a efetividade do processo avaliativo diretamente ao professor. Assim, para além de mensurar a qualidade como classificador, alinhamos os objetivos do sistema com a expectativa de resultados do professor (NASCIMENTO; KAUARK;

MOURA, 2020). Para auxiliar a interpretação, dividimos em três categorias a avaliação automática sob a ótica do professor:

1. Intervalo de 75 - 100%: Nível Avançado;
2. Intervalo de 36 - 75%: Nível Adequado;
3. Intervalo de 0 - 35%: Nível Insuficiente.

Em nível *Insuficiente* a relação entre as notas finais divulgadas pelo professor e a predição do teste para o conjunto tem desempenho baixo. Em nível *Adequado* as notas apresentam aprendizado das notas e modelos avaliativos alinhados com o professor. Por fim, em nível *Avançado*, o desempenho do classificador automático para as demais notas foi similar ao humano, identificando bem o método avaliativo do professor. É importante também descrevermos o detalhe que cada uma das métricas utilizadas pelo sistema sob a ótica do professor (NASCIMENTO; KAUARK; MOURA, 2020):

- ACC: apresenta ao professor a quantidade percentual de respostas que foi avaliada de forma equivalente em relação a sua expectativa avaliativa. Através da acurácia identificamos o percentual de verdadeiros positivos, ou seja, alunos que receberam notas iguais para ambos os avaliadores, em relação ao total de respostas avaliadas. Em uma outra perspectiva podemos também considerar que, com o uso do sistema, é o percentual de notas não alteradas pelo professor após a avaliação automática.
- PRE: apresenta ao professor o percentual de acertos do total de amostras da avaliação automática em relação a sua expectativa avaliativa. A precisão exibe a quantidade de notas avaliadas de forma equivalente (verdadeiros positivos) em relação ao índice de notas que incorretamente avaliadas com o mesmo nível de nota (falsos positivos). Em geral, essa métrica determina ao professor o desempenho médio do sistema na avaliação de cada uma das notas atribuídas.
- REC: apresenta ao professor, em percentual, a capacidade do sistema em discernir cada um dos níveis de nota. A revocação é um indicativo da quantidade de notas avaliadas corretamente (verdadeiros positivos) em relação aos itens que foram categorizados de forma incoerente em outras categorias (falsos negativos). Portanto, é uma métrica que determina ao professor a capacidade média do modelo avaliativo segundo os modelos de cada nota.
- F1: é uma ponderação entre as métricas PRE e REC. Indica ao professor, em âmbito geral, a capacidade dos modelos avaliativos na identificação dos padrões de nota. Deste modo, é uma métrica que leva em conta os erros (falsos positivos e falsos negativos) do sistema entre as categorias em relação a coerência do modelo adquirido para os níveis de nota (verdadeiros positivos e verdadeiros negativos).

- MAE: apresenta ao professor o nível de erro do sistema. O erro médio absoluto determina para o conjunto de notas qual foi a diferença entre as notas atribuídas pelo professor e pelo avaliador automático. A resultante é a variação média entre as notas do professor e do sistema.

- MSE: apresenta ao professor o nível de erro do sistema com penalização. O erro médio quadrático determina a diferença entre as notas do professor e do sistema. Entretanto, a diferença quadrática entre notas atribui maior peso aos modelos avaliativos que apresentam discrepâncias maiores entre notas. A resultante, portanto, é a variação média entre notas do professor e do sistema após a penalização.

Em um contexto geral, os níveis e as métricas representam uma associação entre a consistência de cada avaliação dada pelo sistema automático e a representatividade do modelo de notas. Assim, podemos interpretar os erros sob duas perspectivas determinantes para a melhoria avaliativa e o bom uso do sistema. A primeira é a concordância entre a correção e as componentes textuais das respostas analisadas pelo *p*Nota. E a segunda é a capacidade de montar modelos com a quantidade de respostas por cada nível de nota. O modelo formado, portanto, precisa de conteúdos coincidentes em relação aos padrões linguísticos e os modelos de nota para demonstrar o aprendizado e realizar uma avaliação efetiva. Assim, os níveis de aprendizado, *Avançado*, *Adequado* e *Insuficiente*, descrevem como cada um dos algoritmos de classificação ou regresssão foi capaz de compreender o vínculo entre as componentes textuais e o método avaliativo.

3.4.1 Identificação de Respostas Candidatas

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

Para criação do quadro de *rubrics*, ou rúbricas, os exemplos que receberam mesma nota são considerados alinhados com uma expectativa de resposta. Sabendo que o método não utiliza ou compõe respostas candidatas, a ideia é que este processo elenque uma série de respostas para representar cada grau de nota. Para isso, utilizamos o Latent (LDA). O LDA é um método que identifica o grau de correlação do grupo de respostas de mesma avaliação e cada uma das características. A resultante é uma seleção dos 10 principais termos que compõe as respostas selecionadas. Assim, formamos o quadro de *rubrics* organizando as respostas conforme a proximidade de cada uma com as palavras

selecionadas. Um exemplo de *rubrics* resultante do processo de identificação da simetria entre termos e notas é dado abaixo.

Como podemos observar, ... A ideia, portanto, é criar uma visualização do processo avaliativo de acordo com a simetria entre respostas e a avaliação observada pelo sistema. Deste modo, o uso em sala de aula relaciona diretamente cada grau de nota, as principais termos observados nas respostas e os exemplos de resposta coletados dos próprios estudantes. Por fim, como modelo descritivo da avaliação diretamente ao estudante, utilizamos a visualização de correlação entre conteúdo textual e a classificação através do Lime [1]. O Lime é uma ferramenta de visualização que ilustra a categorização de acordo com os padrões do conteúdo. Para isso, é apresentado em cada resposta a correlação da resposta e suas principais componentes para cada um dos grupo de nota da avaliação. A Figura X apresenta essa estrutura de *feedback* diretamente ao estudante descrevendo sua avaliação.

Como podemos observar na Figura X ...

[1] Lime

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 ordinais, discretas e contínua (MORETTIN; BUSSAB, 2010).

Em bases de dados com notas *ordinais* o método avaliativo do tutor é dado de forma textual e categórica. A representação do rótulo não estabelece escalas para o sistema, não sendo possível mensurar a diferenças na escala *a priori*. O modelo formado deve compreender as estruturas textuais de forma simbólica, caracterizando a essência de cada nível. Portanto, o classificador deve ser robusto para aprender a relevância das respostas pela equivalência de palavras-chave. Basicamente, é fundamental para o classificador produzir um modelo com as informações essenciais para a resposta receber tal categoria e reproduzir o modelo.

Por outro lado, outra situação acontece com bases de dados 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 *discreta* 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	discreto	Inglês
Kaggle PTASAG	15	7473	discreto	Português
Projeto Feira Literária	10	700	discreto	Português
SEMEVAL2013 Beetle	47	3941	ordinal	Inglês
SEMEVAL2013 SciEntBank	143	5251	ordinal	Inglês
UK Open University	20	23790	discreto	Inglês
University of North Texas	87	2610	contínuo	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 et al., 2018) 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* do *SEMEVAL'2013 : Task 7 (Inglês)*

Beetle (DZIKOVSKA et al., 2013) 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 et al., 2013) 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 graaduadas em cinco notas ordinais: *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* (JORDAN, 2012). 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 discreta 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 discretas, de 5 equivalente a resposta perfeita até 0 completamente incorreta. Foram avaliadas por dois avaliadores independentes, estudantes do curso de Ciência da Computação. Para os autores, o modelo seguido pelo sistema deve ser a resultante da média ente os avaliadores, em intervalo contínuo. 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

4.3 Discussão de Resultados

É importante destacar que, apesar de serem comuns os modelos rígidos, direcionados a domínios específicos ou dependentes de regras, o modelo proposto neste trabalho foi o mesmo aplicado para todas as questões. Assim, são passadas ao sistema apenas alguns poucos parâmetros sobre as questões, como sua linguagem. Os parâmetros do modelo estão todos descritos em ANEXO.

Beetle (5 Categorias)							
	Métricas						
	ACC	PRE	REC	F1(m)	F1(w)	Kappa(ln)	Kappa(qu)
DTR	55,68%	28,27%	28,80%	28,01%	54,89%	0,1161	**0,1123**
GBC	59,25%	32,26%	33,12%	31,68%	**57,34%**	0,1247	0,0852
KNN	59,65%	32,31%	33,59%	31,59%	56,76%	0,1217	0,1049
RDF	60,24%	33,22%	34,57%	32,54%	56,73%	0,1239	0,0851
SVM	58,00%	29,24%	32,60%	28,99%	51,96%	0,0592	0,0639
WSD	**60,42%**	**33,30%**	**34,97%**	**32,67%**	56,59%	**0,1262**	0,0909

Tabela 2

SciEntsBank (5 Categorias)							
	Métricas						
	ACC	PRE	REC	F1(m)	F1(w)	Kappa(ln)	Kappa(qu)
DTR	41,85%	28,73%	29,91%	28,06%	40,88%	0,1431	0,1306
GBC	**43,29%**	**29,53%**	**31,62%**	**29,28%**	**41,30%**	**0,1525**	**0,1514**
KNN	40,59%	26,05%	29,72%	25,85%	36,80%	0,0626	0,0519
RDF	41,18%	25,43%	30,56%	25,86%	36,07%	0,0685	0,0605
SVM	38,34%	20,96%	29,65%	22,98%	31,10%	0,0263	0,0235
WSD	40,20%	25,31%	28,70%	25,11%	36,57%	0,0641	0,0528

Tabela 3

North Texas University (Notas 0 - 5)			
	Métricas		
	Avaliador1		
	MAE	MSE	RMSE
LINR	1,0047	**2,5064**	**1,1930**
LSSR	1,3249	3,1709	1,4688
KNRG	0,9366	2,9032	1,2557
DTRG	**0,9233**	3,7338	1,4483
WSRG	1,2828	3,0113	1,4236
	Avaliador2		
	MAE	MSE	RMSE
LINR	**0,4735**	0,6096	0,6099
LSSR	0,6484	0,8605	0,7621
KNRG	0,4914	**0,0000**	**0,0000**
DTRG	0,5112	1,1997	0,7856
WSRG	0,6518	**0,0000**	**0,0000**
	Média		
	MAE	MSE	RMSE
LINR	**0,5043**	**0,5473**	**0,6179**
LSSR	0,7278	0,8461	0,8150
KNRG	0,5055	0,6804	0,6765
DTRG	0,5811	1,1244	0,8372
WSRG	0,7019	0,8087	0,7915

Tabela 4

Projeto Feira Literária (Notas 0 - 3)							
	Métricas						
	ACC	PRE	REC	F1(m)	F1(w)	Kappa(ln)	Kappa(qu)
DTR	58,59%	46,20%	43,85%	42,49%	56,75%	0,3273	0,3791
GBC	64,14%	46,21%	46,32%	43,76%	60,54%	0,3702	**0,4324**
KNN	50,00%	35,54%	34,25%	32,81%	49,64%	0,1363	0,1346
RDF	**68,18%**	**51,60%**	**51,51%**	**48,98%**	**63,56%**	**0,3842**	0,4073
SVM	57,07%	29,28%	38,52%	31,05%	47,70%	0,1121	0,1212
WSD	65,15%	48,30%	48,10%	45,25%	60,60%	0,3282	0,3543

Tabela 5

5 Considerações Finais

5.1 Trabalhos Futuros

5.2 Conclusões

Referências

ALMEIDA-JÚNIOR, C. R. C.; SPALENZA, M. A.; OLIVEIRA, E. de. Proposta de um Sistema de Avaliação Automática de Redações do ENEM Utilizando Técnicas de Aprendizagem de Máquina e Processamento de Linguagem Natural. In: *Computer on the Beach*. Florianópolis (SC), Brasil: Sociedade Brasileira de Computação, 2017. v. 8, p. 474–483. Citado na página 29.

ARTER, J. A.; CHAPPUIS, J. *Creating & Recognizing Quality Rubrics*. 1st. ed. New York (NY), USA: Pearson Education, 2006. (Assessment Training Institute, Inc Series). Citado na página 19.

ARTSTEIN, R.; POESIO, M. Inter-Coder Agreement for Computational Linguistics. *Computational Linguistics*, MIT Press, v. 34, n. 4, p. 555–596, 2008. Citado 2 vezes nas páginas 23 e 24.

AZAD, S. et al. Strategies for Deploying Unreliable AI Graders in High-Transparency High-Stakes Exams. In: *Proceedings of the 21st International Conference on Artificial Intelligence in Education*. Ifrane, Morocco: Springer International Publishing, 2020. (AIED' 2020, v. 21), p. 16–28. Citado 3 vezes nas páginas 23, 32 e 35.

BAEZA-YATES, R.; RIBEIRO-NETO, B. *Modern Information Retrieval: The Concepts and Technology behind Search*. 2nd. ed. Boston (MA), USA: Addison-Wesley Publishing Company, 2011. Citado 4 vezes nas páginas 24, 31, 33 e 43.

BAILEY, S.; MEURERS, D. Diagnosing Meaning Errors in Short Answers to Reading Comprehension Questions. In: *Proceedings of the Third Workshop on Innovative Use of NLP for Building Educational Applications*. Columbus (OH), USA: Association for Computational Linguistics, 2008. (EANL '08, v. 3), p. 107–115. Citado 3 vezes nas páginas 21, 28 e 29.

BANJADE, R. et al. Evaluation Dataset (DT-Grade) and Word Weighting Approach Towards Constructed Short Answers Assessment in Tutorial Dialogue Context. In: *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*. San Diego (CA), USA: Association for Computational Linguistics, 2016. v. 11, p. 182–187. Citado na página 33.

BANJADE, R.; RUS, V.; NIRAULA, N. B. Using an Implicit Method for Coreference Resolution and Ellipsis Handling in Automatic Student Answer Assessment. In: *The Twenty-Eighth International Flairs Conference*. Hollywood (FL), USA: AAAI Press, 2015. v. 28, p. 150–155. Citado na página 30.

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.

BASU, S.; JACOBS, C.; VANDERWENDE, L. Powergrading: A Clustering Approach to Amplify Human Effort for Short Answer Grading. *Transactions of the Association for Computational Linguistics*, MIT Press, v. 1, n. 1, p. 391–402, 2013. Citado 2 vezes nas páginas 31 e 33.

BEZERRA, M. A. Questões Discursivas para Avaliação Escolar. *Acta Scientiarum. Language and Culture*, Universidade Estadual de Maringá, v. 30, n. 2, p. 149–157, 2008. Citado na página 28.

BIGGS, J. Assessment and Classroom Learning: A Role for Summative Assessment? *Assessment in Education: Principles, Policy & Practice*, Routledge, v. 5, n. 1, p. 103–110, 1998. Citado na página 19.

BILGIN, A. A.; ROWE, A. D.; CLARK, L. Academic Workload Implications of Assessing Student Learning in Work-Integrated Learning. *Asia-Pacific Journal of Cooperative Education*, ERIC, v. 18, n. 2, p. 167–183, 2017. Citado na página 27.

BURROWS, S.; GUREVYCH, I.; STEIN, B. The Eras and Trends of Automatic Short Answer Grading. *International Journal of Artificial Intelligence in Education*, Springer, v. 25, n. 1, p. 60–117, 2015. Citado 9 vezes nas páginas 20, 23, 25, 27, 28, 30, 35, 36 e 37.

BUTCHER, P. G.; JORDAN, S. E. A Comparison of Human and Computer Marking of Short Free-Text Student Responses. *Computers & Education*, Elsevier, v. 55, n. 2, p. 489–499, 2010. Citado 5 vezes nas páginas 21, 25, 28, 31 e 35.

CALIŃSKI, T.; J., H. A Dendrite Method for Cluster Analysis. *Communications in Statistics*, Taylor & Francis, v. 3, n. 1, p. 1–27, 1974. Citado na página 44.

CAMUS, L.; FILIGHERA, A. Investigating Transformers for Automatic Short Answer Grading. In: *Proceedings of the 21st International Conference on Artificial Intelligence in Education*. Ifrane, Morocco: Springer International Publishing, 2020. (AIED' 2020, v. 21), p. 43–48. Citado na página 36.

CHAKRABORTY, U. K.; ROY, S.; CHOUDHURY, S. A Fuzzy Indiscernibility Based Measure of Distance between Semantic Spaces Towards Automatic Evaluation of Free Text Answers. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, World Scientific, v. 25, n. 6, p. 987–1004, 2017. Citado 2 vezes nas páginas 22 e 35.

CONDOR, A. Exploring Automatic Short Answer Grading as a Tool to Assist in Human Rating. In: *Proceedings of the 21st International Conference on Artificial Intelligence in Education*. Ifrane, Morocco: Springer International Publishing, 2020. (AIED' 2020, v. 21), p. 74–79. Citado 3 vezes nas páginas 23, 26 e 32.

DAVIES, D. L.; BOULDIN, D. W. A Cluster Separation Measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, IEEE, v. 1, n. 2, p. 224–227, 1979. Citado na página 44.

DING, Y. et al. Don't Take "nswvtnvakgxpm" for an Answer - The Surprising Vulnerability of Automatic Content Scoring Systems to Adversarial Input. In: *Proceedings of the 28th International Conference on Computational Linguistics*. Barcelona, Spain (Virtual Event): International Committee on Computational Linguistics, 2020. v. 28, p. 882–892. Citado 5 vezes nas páginas 22, 25, 26, 29 e 35.

DZIKOVSKA, M. et al. SemEval-2013 Task 7: The Joint Student Response Analysis and 8th Recognizing Textual Entailment Challenge. In: *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*. Atlanta (GE), USA:

Association for Computational Linguistics, 2013. v. 7, p. 263–274. Citado 2 vezes nas páginas 22 e 58.

DZIKOVSKA, M. O.; NIELSEN, R. D.; BREW, C. Towards Effective Tutorial Feedback for Explanation Questions: A Dataset and Baselines. In: *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Montréal, Canada: Association for Computational Linguistics, 2012. v. 11, p. 200–210. Citado na página 22.

EVERITT, B. S. et al. *Cluster Analysis*. 5th. ed. Chichester, United Kingdom: John Wiley, 2011. Citado 3 vezes nas páginas 24, 31 e 44.

FERREIRA-MELLO, R. et al. Text Mining in Education. *WIREs Data Mining and Knowledge Discovery*, Wiley Online Library, v. 9, n. 6, p. e1332.1–e1332.49, 2019. Citado na página 27.

FILIGHERA, A.; STEUER, T.; RENSING, C. Fooling Automatic Short Answer Grading Systems. In: *Proceedings of the 21st International Conference on Artificial Intelligence in Education*. Ifrane, Morocco: Springer International Publishing, 2020. (AIED' 2020, v. 21), p. 177–190. Citado 6 vezes nas páginas 20, 22, 26, 29, 35 e 36.

FOWLER, M. et al. Autograding "Explain in Plain English" Questions Using NLP. In: *Proceedings of the 52nd ACM Technical Symposium on Computer Science Education*. New York (NY), USA (Virtual Event): Association for Computing Machinery, 2021. (SIGCSE'21, v. 52), p. 1163–1169. Citado 2 vezes nas páginas 29 e 35.

FUNAYAMA, H. et al. Preventing critical scoring errors in short answer scoring with confidence estimation. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*. Online Event: Association for Computational Linguistics, 2020. v. 58, p. 237–243. Citado 5 vezes nas páginas 21, 22, 26, 29 e 34.

GALHARDI, L. B.; BRANCHER, J. D. Machine Learning Approach for Automatic Short Answer Grading: A Systematic Review. In: *Proceedings of the 16th Ibero-American Conference on Artificial Intelligence - IBERAMIA 2018*. Trujillo, Peru: Springer International Publishing, 2018. (IBERAMIA 2018, v. 16), p. 380–391. Citado 2 vezes nas páginas 25 e 32.

GALHARDI, L. B. et al. Exploring Distinct Features for Automatic Short Answer Grading. In: *Anais do XV Encontro Nacional de Inteligência Artificial e Computacional*. São Paulo (SP), Brazil: Sociedade Brasileira de Computação, 2018. (XV ENIAC, v. 15), p. 1–12. Citado 3 vezes nas páginas 32, 36 e 57.

GHAVIDEL, H.; ZOUAQ, A.; DESMARAIS, M. Using BERT and XLNET for the Automatic Short Answer Grading Task. In: *Proceedings of the 12th International Conference on Computer Supported Education (CSEDU)*. Prague, Czechia (Virtual Event): SciTePress, 2020. (CSEDU 2020, v. 12), p. 58–67. Citado na página 34.

GOLDBERG, Y.; HIRST, G. *Neural Network Methods in Natural Language Processing*. 1st. ed. San Rafael (CA), USA: Morgan & Claypool Publishers, 2017. Citado na página 34.

GUNTHER, H.; LOPES-JÚNIOR, J. Perguntas Abertas Versus Perguntas Fechadas: Uma Comparação Empírica. *Psicologia: Teoria e Pesquisa*, Universidade de Brasília, v. 6, n. 2, p. 203–213, 2012. Citado na página 20.

HAN, J.; PEI, J.; KAMBER, M. *Data Mining: Concepts and Techniques*. 3rd. ed. Waltham (MA), USA: Elsevier, 2011. Citado 2 vezes nas páginas 24 e 45.

HEILMAN, M.; MADNANI, N. The Impact of Training Data on Automated Short Answer Scoring Performance. In: *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*. Denver (CO), USA: Association for Computational Linguistics, 2015. v. 10, p. 81–85. Citado 2 vezes nas páginas 25 e 30.

HIGGINS, D. et al. *Is Getting the Right Answer Just About Choosing the Right Words? The Role of Syntactically-Informed Features in Short Answer Scoring*. Princeton (NJ), USA, 2014. Citado 3 vezes nas páginas 21, 26 e 34.

HORBACH, A.; PINKAL, M. Semi-supervised clustering for short answer scoring. In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki, Japan: European Language Resources Association (ELRA), 2018. (LREC 2018, v. 11), p. 4065–4071. Citado 4 vezes nas páginas 21, 22, 30 e 34.

JIMENEZ, S.; BECERRA, C.; GELBUKH, A. SOFTCARDINALITY: Hierarchical Text Overlap for Student Response Analysis. In: *Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*. Atlanta (GA), USA: Association for Computational Linguistics, 2013. v. 7, p. 280–284. Citado 3 vezes nas páginas 30, 34 e 35.

JOHNSTONE, K. M.; ASHBAUGH, H.; WARFIELD, T. D. Effects of Repeated Practice and Contextual-Writing Experiences on College Students' Writing Skills. *Journal of Educational Psychology*, American Psychological Association, v. 94, n. 2, p. 305–315, 2002. Citado na página 27.

JORDAN, S. Student Engagement with Assessment and Feedback: Some Lessons from Short-Answer Free-Text e-Assessment Questions. *Computers & Education*, Elsevier, v. 58, n. 2, p. 818–834, 2012. Citado 5 vezes nas páginas 21, 25, 35, 36 e 59.

JURAFSKY, D.; MARTIN, J. H. *Speech and Language Processing*. 2nd. ed. Upper Saddle River (NJ), USA: Prentice-Hall, Inc., 2009. Citado 3 vezes nas páginas 24, 32 e 34.

KAR, S. P.; CHATTERJEE, R.; MANDAL, J. K. A novel automated assessment technique in e-learning using short answer type questions. In: *Proceedings of the 1st International Conference on Computational Intelligence, Communications, and Business Analytics*. Kolkata, India: Springer Singapore, 2017. (CICBA 2017, v. 1), p. 141–149. Citado 2 vezes nas páginas 30 e 35.

KRITHIKA, R.; NARAYANAN, J. Learning to grade short answers using machine learning techniques. In: *Proceedings of the Third International Symposium on Women in Computing and Informatics*. Kochi, India: Association for Computing Machinery, 2015. (WCI '15, v. 3), p. 262–271. Citado 2 vezes nas páginas 21 e 34.

KUMAR, S.; CHAKRABARTI, S.; ROY, S. Earth Mover's Distance Pooling over Siamese

LSTMs for Automatic Short Answer Grading. In: *Proceedings of the 26th International Joint Conference on Artificial Intelligence*. Melbourne, Australia: AAAI Press, 2017. (IJCAI'17, v. 26), p. 2046–2052. Citado na página 33.

KUMAR, Y. et al. Get it Scored Using AutoSAS - An Automated System for Scoring Short Answers. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Honolulu (HI), USA: AAAI Press, 2019. v. 33, p. 9662–9669. Citado 6 vezes nas páginas 22, 23, 25, 31, 33 e 35.

LANDAUER, T. K.; FOLTZ, P. W.; LAHAM, D. An Introduction to Latent Semantic Analysis. *Discourse Processes*, Routledge, v. 25, n. 2, p. 259–284, 1998. Citado na página 33.

LEFFA, V. J. Análise Automática da Resposta do Aluno em Ambiente Virtual. *Revista Brasileira de Linguística Aplicada*, SciELO, v. 3, n. 2, p. 25–40, 2003. Citado na página 20.

LUN, J. et al. Multiple Data Augmentation Strategies for Improving Performance on Automatic Short Answer Scoring. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. New York (NY), USA: AAAI Press, 2020. v. 34, n. 09, p. 13389–13396. Citado na página 31.

MADERO, C. Secondary Teacher's Dissatisfaction with the Teaching Profession in Latin America: The Case of Brazil, Chile, and Mexico. *Teachers and Teaching*, Routledge, v. 25, n. 3, p. 358–378, 2019. Citado na página 27.

MAIMON, O.; ROKACH, L. *Data Mining and Knowledge Discovery Handbook*. 1st. ed. New York (NY), USA: Springer, 2005. Citado na página 44.

MANNING, C.; SCHUTZE, H. *Foundations of Statistical Natural Language Processing*. 1st. ed. Cambridge (MA), USA: MIT Press, 1999. Citado na página 33.

MAO, L. et al. Validation of Automated Scoring for a Formative Assessment that Employs Scientific Argumentation. *Educational Assessment*, Routledge, v. 23, n. 2, p. 121–138, 2018. Citado na página 35.

MAQUINÉ, G. Recursos para Avaliação da Aprendizagem: Estudo Comparativo entre Ambientes Virtuais de Aprendizagem. In: *Anais do XXVI Workshop de Informática na Escola*. Natal (RN) (Online), Brasil: Sociedade Brasileira de Computação, 2020. v. 26, p. 299–308. Citado na página 19.

MARNEFFE, M.-C. et al. Universal Dependencies. *Computational Linguistics*, MIT Press, v. 47, n. 2, p. 255–308, 2021. Citado na página 42.

MARVANIYA, S. et al. Creating Scoring Rubric from Representative Student Answers for Improved Short Answer Grading. In: *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*. Torino, Italy: Association for Computing Machinery, 2018. (CIKM '18, v. 27), p. 993–1002. Citado 5 vezes nas páginas 22, 23, 26, 31 e 34.

MENINI, S. et al. Automated Short Answer Grading: A Simple Solution for a Difficult Task. In: *Proceedings of the Sixth Italian Conference on Computational Linguistics*. Bari, Italy: CEUR-WS, 2019. (CLiC-it, v. 6), p. 48.1–48.7. Citado na página 35.

MING, L. S. Reduction of Teacher Workload in a Formative Assessment Environment through use of Online Technology. In: *6th International Conference on Information Technology Based Higher Education and Training*. Santo Domingo, Dominican Republic: IEEE, 2005. v. 6, p. 18–21. Citado na página 27.

MIZUMOTO, T. et al. Analytic Score Prediction and Justification Identification in Automated Short Answer Scoring. In: *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*. Florence, Italy: Association for Computational Linguistics, 2019. v. 14, p. 316–325. Citado 2 vezes nas páginas 26 e 31.

MOHLER, M.; BUNESCU, R.; MIHALCEA, R. Learning to Grade Short Answer Questions using Semantic Similarity Measures and Dependency Graph Alignments. In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Portland (OR), USA: Association for Computational Linguistics, 2011. v. 10, p. 752–762. Citado 4 vezes nas páginas 21, 32, 34 e 59.

MOHLER, M.; MIHALCEA, R. Text-to-text semantic similarity for automatic short answer grading. In: *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*. Athens, Greece: Association for Computational Linguistics, 2009. v. 12, p. 567–575. Citado 2 vezes nas páginas 32 e 33.

MORETTIN, P. A.; BUSSAB, W. O. *Estatística Básica*. 6. ed. Pinheiros (SP), Brasil: Editora Saraiva, 2010. Citado 2 vezes nas páginas 24 e 55.

NASCIMENTO, P. V.; KAUARK, F. S.; MOURA, P. R. G. *Construindo uma Atividade Experimental Problematizada (AEP) e Avaliando Seu Nível Cognitivo de Aprendizagem Através do Software pNota no Contexto do Ensino Fundamental*. 9. ed. Vila Velha (ES), Brasil: Instituto Federal do Espírito Santo, 2020. (Série Guia Didático de Ciências/Química). Citado 2 vezes nas páginas 51 e 59.

OLIVEIRA, E. et al. Combining Clustering and Classification Approaches for Reducing the Effort of Automatic Tweets Classification. In: *Proceedings of the International Conference on Knowledge Discovery and Information Retrieval - KDIR, (IC3K 2014)*. Rome, Italy: SciTePress, 2014. (KDIR '14, v. 6), p. 465–472. Citado na página 31.

OLIVEIRA, K. L. d.; SANTOS, A. A. A. Compreensão em Leitura e Avaliação da Aprendizagem em Universitários. *Psicologia: Reflexão e Crítica*, SciELO, v. 18, n. 1, p. 118–124, 2005. Citado na página 19.

OLIVEIRA, M. G.; CIARELLI, P. M.; OLIVEIRA, E. Recommendation of Programming Activities by Multi-label Classification for a Formative Assessment of Students. *Expert Systems with Applications*, Elsevier, v. 40, n. 16, p. 6641–6651, 2013. Citado na página 27.

PADÓ, U.; PADÓ, S. Determinants of Grader Agreement: An Analysis of Multiple Short Answer Corpora. *Language Resources and Evaluation*, Springer, v. 55, n. 2, p. 1–30, 2021. Citado 5 vezes nas páginas 20, 23, 24, 26 e 32.

PAIVA, R. et al. Mineração de Dados e a Gestão Inteligente da Aprendizagem: Desafios e Direcionamentos. In: *I Workshop de Desafios da Computação Aplicada á Educação (DesafIE!2012)*. Curitiba (PR), Brasil: Sociedade Brasileira de Computação, 2012. v. 1. Citado 2 vezes nas páginas 19 e 32.

PÉREZ-MARÍN, D.; PASCUAL-NIETO, I.; RODRÍGUEZ, P. Computer-Assisted Assessment of Free-Text Answers. *The Knowledge Engineering Review*, Cambridge University Press, v. 24, n. 4, p. 353–374, 2009. Citado 2 vezes nas páginas 20 e 27.

PIROVANI, J. P. C. et al. Adapting NER (CRF+LG) for Many Textual Genres. In: *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019)*. Bilbao, Spain: CEUR-WS, 2019. (IberLEF - SEPLN 2019, v. 35), p. 421–433. Citado na página 42.

PISSINATI, E. M. *Uma Proposta de Correção Semi-Automática de Questões Discursivas e de Visualização de Atividades para Apoio à Atuação do Docente*. Dissertação (Mestrado) — PPGI - Universidade Federal do Espírito Santo, Vitória (ES), Brasil, Set 2014. Citado na página 60.

PRIBADI, F. S. et al. Automatic Short Answer Scoring Using Words Overlapping Methods. In: *AIP Conference Proceedings*. Bandung, Indonesia: AIP Publishing LLC, 2017. v. 1818, p. 020042:1–020042:6. Citado na página 31.

RAES, A. et al. A Systematic Literature Review on Synchronous Hybrid Learning: Gaps Identified. *Learning Environments Research*, Springer, v. 23, n. 3, p. 269–290, 2020. Citado na página 19.

RAMACHANDRAN, L.; CHENG, J.; FOLTZ, P. Identifying Patterns For Short Answer Scoring Using Graph-based Lexico-Semantic Text Matching. In: *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*. Denver (CO), USA: Association for Computational Linguistics, 2015. v. 10, p. 97–106. Citado 4 vezes nas páginas 29, 30, 33 e 35.

RAMACHANDRAN, L.; FOLTZ, P. Generating Reference Texts for Short Answer Scoring Using Graph-based Summarization. In: *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*. Denver (CO), USA: Association for Computational Linguistics, 2015. v. 10, p. 207–212. Citado 3 vezes nas páginas 21, 22 e 30.

RIORDAN, B.; FLOR, M.; PUGH, R. How to Account for Mispellings: Quantifying the Benefit of Character Representations in Neural Content Scoring Models. In: *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*. Florence, Italy: Association for Computational Linguistics, 2019. v. 14, p. 116–126. Citado 3 vezes nas páginas 22, 32 e 35.

RIORDAN, B. et al. Investigating Neural Architectures for Short Answer Scoring. In: *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*. Copenhagen, Denmark: Association for Computational Linguistics, 2017. v. 12, p. 159–168. Citado na página 34.

ROMERO, C. et al. *Handbook of Educational Data Mining*. 1st. ed. Boca Raton (FL), USA: CRC Press, 2010. Citado 3 vezes nas páginas 23, 24 e 27.

ROUSSEEUW, P. J. Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis. *Journal of Computational and Applied Mathematics*, Elsevier, v. 20, n. 1, p. 53–65, 1987. Citado na página 44.

ROY, S. et al. Wisdom of Students: A Consistent Automatic Short Answer Grading

Technique. In: *Proceedings of the 13th International Conference on Natural Language Processing*. Varanasi, India: NLP Association of India, 2016. v. 13, p. 178–187. Citado 2 vezes nas páginas 30 e 33.

ROY, S.; RAJKUMAR, A.; NARAHARI, Y. Selection of Automatic Short Answer Grading Techniques Using Contextual Bandits for Different Evaluation Measures. *International Journal of Advances in Engineering Sciences and Applied Mathematics*, Springer, v. 10, n. 1, p. 105–113, 2018. Citado na página 36.

SAHA, S. et al. Sentence Level or Token Level Features for Automatic Short Answer Grading?: Use Both. In: *Proceedings of the 19th International Conference on Artificial Intelligence in Education*. London, United Kingdom: Springer International Publishing, 2018. (AIED' 2018, v. 19), p. 503–517. Citado 4 vezes nas páginas 21, 22, 25 e 33.

SAHA, S. et al. *Joint Multi-Domain Learning for Automatic Short Answer Grading*. New Delhi, India, 2019. Citado 2 vezes nas páginas 22 e 36.

SAHU, A.; BHOWMICK, P. K. Feature Engineering and Ensemble-Based Approach for Improving Automatic Short-Answer Grading Performance. *IEEE Transactions on Learning Technologies*, IEEE, v. 13, n. 1, p. 77–90, 2020. Citado 3 vezes nas páginas 22, 25 e 33.

SAKAGUCHI, K.; HEILMAN, M.; MADNANI, N. Effective Feature Integration for Automated Short Answer Scoring. In: *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Denver (CO), USA: Association for Computational Linguistics, 2015. v. 14, p. 1049–1054. Citado 2 vezes nas páginas 22 e 33.

SIDDIQI, R.; HARRISON, C. J. On the Automated Assessment of Short Free-Text Responses. In: *Proceedings of the 34th International Association for Educational Assessment Annual Conference*. Cambridge, United Kingdom: IAEA, 2008. (IAEA Conference, v. 34), p. 1–11. Citado na página 29.

SIEMENS, G.; BAKER, R. S. J. d. Learning Analytics and Educational Data Mining: Towards Communication and Collaboration. In: *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge*. Vancouver, Canada: Association for Computing Machinery, 2012. (LAK '12, v. 2), p. 252–254. Citado na página 27.

SPALENZA, M. A. et al. Construção de mapas de características em classes de respostas discursivas. In: *Conferência Internacional sobre Informática na Educação (TISE 2016)*. Santiago, Chile: Centro de Computación y Comunicación para la Construcción del Conocimiento (C5), 2016. (TISE 2016, v. 12), p. 630–635. Citado na página 34.

SPALENZA, M. A. et al. Uso de Mapa de Características na Avaliação de Textos Curtos nos Ambientes Virtuais de Aprendizagem. In: *XXVII Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação*. Uberlândia (MG), Brazil: Sociedade Brasileira de Computação, 2016. (SBIE 2016, v. 27), p. 1165–1174. Citado 3 vezes nas páginas 20, 22 e 34.

SPALENZA, M. A. et al. Using NER + ML to Automatically Detect Fake News. In: *Proceedings of the 20th International Conference on Intelligent Systems Design and Applications*. Online Event: Springer International Publishing, 2020. (ISDA 2020, v. 20), p. 1176–1187. Citado 2 vezes nas páginas 42 e 44.

SPALENZA, M. A.; PIROVANI, J. P. C.; OLIVEIRA, E. de. Structures Discovering for Optimizing External Clustering Validation Metrics. In: *Proceedings of the 19th International Conference on Intelligent Systems Design and Applications*. Auburn (WA), USA: Springer International Publishing, 2019. (ISDA 2019, v. 19), p. 150–161. Citado 2 vezes nas páginas 44 e 45.

SULTAN, M. A.; SALAZAR, C.; SUMNER, T. Fast and Easy Short Answer Grading with High Accuracy. In: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. San Diego (CA), USA: Association for Computational Linguistics, 2016. v. 15, p. 1070–1075. Citado na página 33.

SUNG, C. et al. Pre-Training BERT on Domain Resources for Short Answer Grading. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, 2019. v. 9, p. 6071–6075. Citado 2 vezes nas páginas 30 e 36.

SUNG, C.; DHAMECHA, T. I.; MUKHI, N. Improving Short Answer Grading Using Transformer-Based Pre-training. In: *Proceedings of the 20th International Conference on Artificial Intelligence in Education*. Chicago (IL), USA: Springer, 2019. (AIED' 2019, v. 20), p. 469–481. Citado na página 34.

SÜZEN, N. et al. Automatic Short Answer Grading and Feedback Using Text Mining Methods. *Procedia Computer Science*, Elsevier, v. 169, n. 1, p. 726–743, 2020. Citado 3 vezes nas páginas 26, 29 e 36.

TAN, H. et al. Automatic Short Answer Grading by Encoding Student Responses via a Graph Convolutional Network. *Interactive Learning Environments*, Taylor & Francis, v. 28, n. 1, p. 1–15, 2020. Citado 2 vezes nas páginas 33 e 35.

WANG, T. et al. Inject Rubrics into Short Answer Grading System. In: *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*. Hong Kong, China: Association for Computational Linguistics, 2019. (DeepLo 2019, v. 2), p. 175–182. Citado na página 30.

YANG, S. J. H. et al. Human-Centered Artificial Intelligence in Education: Seeing the Invisible through the Visible. *Computers and Education: Artificial Intelligence*, Elsevier, v. 2, n. 1, p. 100008, 2021. Citado 2 vezes nas páginas 32 e 35.

ZESCH, T.; HEILMAN, M.; CAHILL, A. Reducing Annotation Efforts in Supervised Short Answer Scoring. In: *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*. Denver (CO), USA: Association for Computational Linguistics, 2015. v. 10, p. 124–132. Citado 3 vezes nas páginas 21, 26 e 31.

ZESCH, T.; HORBACH, A. ESCRITO - An NLP-Enhanced Educational Scoring Toolkit. In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki, Japan: European Language Resources Association (ELRA), 2018. v. 11, p. 2310–2316. Citado 2 vezes nas páginas 33 e 36.

ZHANG, Y.; LIN, C.; CHI, M. Going Deeper: Automatic Short-Answer Grading by Combining Student and Question Models. *User Modeling and User-Adapted Interaction*, Springer, v. 30, n. 1, p. 51–80, 2020. Citado na página 34.

ZHANG, Y.; SHAH, R.; CHI, M. Deep Learning+ Student Modeling+ Clustering: A Recipe for Effective Automatic Short Answer Grading. In: *Proceedings of the 9th International Conference on Educational Data Mining*. Raleigh (NC), USA: ERIC, 2016. (EDM 2016, v. 09), p. 562–567. Citado 2 vezes nas páginas 31 e 36.

ZIAI, R.; OTT, N.; MEURERS, D. Short Answer Assessment: Establishing Links Between Research Strands. In: *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*. Montreal, Canada: Association for Computational Linguistics, 2012. (NAACL HLT '12, v. 7), p. 190–200. Citado na página 35.

Apêndices