CBIE 2023
CONGRESSO BRASILEIRO DE INFORMÁTICA NA EDUCAÇÃO
Uma escola para o futuro: Tecnologia e conectividade a serviço da educação

Aplicando ChatGPT para Recomendação de Tags para Auxiliar Professores na Correção de Atividades Abertas

Rodrigues Neto, Gabriel Alves, Rafael Ferreira Mello

[1] Universidade Federal Rural de Pernambuco (UFRPE)

rodriguesnt01@gmail.com, {gabriel.alves, rafael.mello}@ufrpe.br

Abstract. *Effective tag selection is crucial in automating the open-ended activity correction process, supporting teachers to provide valuable feedback to students. This research addresses the growing demand for tools that assist educators in assessing written responses, saving time, and enhancing efficiency. We evaluate the performance of ChatGPT in comparison to traditional Natural Language Processing (NLP) approaches, considering metrics such as precision, recall, and F1-score. The ChatGPT presented worse results than some traditional algorithms besides having additional costs.*

Resumo. *A seleção eficaz de tags é crucial na automatização do processo de correção de atividades abertas, auxiliando professores no fornecimento de feedbacks aos estudantes. Nesse contexto, essa pesquisa aborda a crescente demanda por ferramentas que auxiliem os educadores na avaliação de respostas escritas, economizando tempo e melhorando a eficiência. Foi avaliado o desempenho do ChatGPT em comparação a abordagens de Processamento de Linguagem Natural (PLN) tradicionais, considerando métricas como precisão, revocação e a medida F1. O ChatGPT apresentou um desempenho inferior ao de alguns algoritmos tradicionais, possuindo ainda custos adicionais.*

1. Introdução

O avanço das tecnologias de processamento de linguagem natural (PLN) tem revolucionado a forma como se lida com a correção de atividades abertas em ambientes educacionais [José et al. 2015, de Lima Dias and Pazoti 2023, Mello et al. 2022]. A avaliação de respostas escritas de alunos é uma tarefa fundamental para os educadores [Wiggins 1998], mas também pode ser intensiva em termos de tempo e recursos [Boud and Molloy 2013]. Nesse contexto, a automação desse processo tornou-se uma prioridade, visando não apenas economizar tempo, mas também melhorar a eficiência e a qualidade das avaliações.

A recomendação de tags, etiquetas, ou marcadores, é uma estratégia que vem sendo empregada para categorizar e organizar respostas de alunos em sistemas de correção automática [Mello et al. 2022]. Essas tags são cruciais, pois ajudam a identificar os principais conceitos e elementos presentes nas respostas dos alunos, facilitando a análise e o *feedback* posterior. Portanto, a escolha de tags eficazes é de suma importância para o sucesso da automação do processo de correção.

Nesse contexto, este artigo se propõe a comparar o desempenho do ChatGPT [Rahman and Watanobe 2023], um modelo de linguagem grande (do inglês *large language model*), com abordagens tradicionais de PLN na recomendação de tags para auxiliar os professores na correção de atividades abertas. O ChatGPT, notório por sua ca-

pacidade de entender e gerar texto de alta qualidade, oferece a promessa de aprimorar significativamente a seleção de tags em respostas de alunos.

A pesquisa aborda uma demanda crescente por ferramentas que apoiem os educadores na avaliação de respostas escritas, permitindo que eles se concentrem em tarefas mais estratégicas e de valor agregado. A automação da seleção de tags também pode contribuir para a uniformidade e consistência nas avaliações, garantindo que todos os alunos sejam tratados de forma equitativa. Nesse estudo, será avaliado o desempenho do ChatGPT em relação às técnicas tradicionais de PLN na seleção de tags, empregando métricas objetivas, como precisão, revocação e a medida F1. Além disso, também será considerada a capacidade do ChatGPT de compreender e inferir conceitos subjacentes nas respostas dos alunos [Kasneci et al. 2023].

2. Trabalhos Relacionados

Ferramentas computacionais como o Gradescope [Singh et al. 2017] e o ArTEMiS [Krusche and Seitz 2018] atuam com o objetivo de automatizar ou sistematizar o processo de avaliação. Outros trabalhos como os apresentados em Cutrone e Chang [Cutrone and Chang 2010] discutem a automação da avaliação de questões abertas, enquanto Siddiqi et al. [Siddiqi et al. 2010] exploram a melhoria do ensino e da aprendizagem por meio da marcação automática de respostas curtas. Assim como o Tutoria, essas propostas buscam fornecer soluções computacionais para dar o suporte à avaliação educacional. No caso do Tutoria, esse suporte é focado especialmente no *feedback* relacionado a atividades abertas, com o uso de PLN.

Diversas medidas de similaridade relacionadas ao PLN foram avaliadas e utilizadas pelo Tutoria. Yujian e Bo [Yujian and Bo 2007] apresentam uma métrica de distância de Levenshtein normalizada para comparação de texto. O trabalho de José et al. [José et al. 2015] apresenta um Avaliador Ortográfico-Gramatical baseado em algoritmos genéticos e PLN. Esse estudo destaca a importância de abordagens automatizadas para a correção de atividades escritas. De Lima Dias e Pazoti [de Lima Dias and Pazoti 2023] também contribuem para essa área, apresentando um método de correção automatizada de questões dissertativas utilizando medidas de similaridade e PLN. Por fim, Marin [Marın 2004] discute a avaliação automática de ensaios curtos por meio de técnicas estatísticas e PLN. Wang et al. [Wang et al. 2017] propõem uma métrica de distância Jaro-Winkler eficiente para comparação de entidades.

Modelos de linguagem avançados como os propostos por Manning e Schutze [Manning and Schutze 1999] oferecem uma base sólida em processamento de linguagem natural. O trabalho de Devlin et al. [Devlin et al. 2019] introduz o modelo BERT, que tem sido influente na compreensão de linguagem natural. Esses e outros modelos tradicionais da área de PLN foram comparados com os resultados obtidos com a utilização do ChatGPT, com o objetivo de aprimorar o modelo utilizado no Tutoria.

3. Plataforma Tutoria

A plataforma Tutoria[1] oferece suporte para correção de tarefas escritas e elaboração de *feedback* direcionado para o usuário [Falcão et al. 2023, Neto et al. 2022,

[1]https://tutor-ia.com/

Falcao et al. 2022b, Falcão et al. 2020]. Para isso, os professores importam as atividades do ambiente virtual de aprendizagem (por exemplo, Google Sala de Aula ou Moodle). Após importar as respostas, o professor pode optar por realizar a correção por pergunta ou aluno. A Figura 1 apresenta uma visão geral da plataforma de correção.

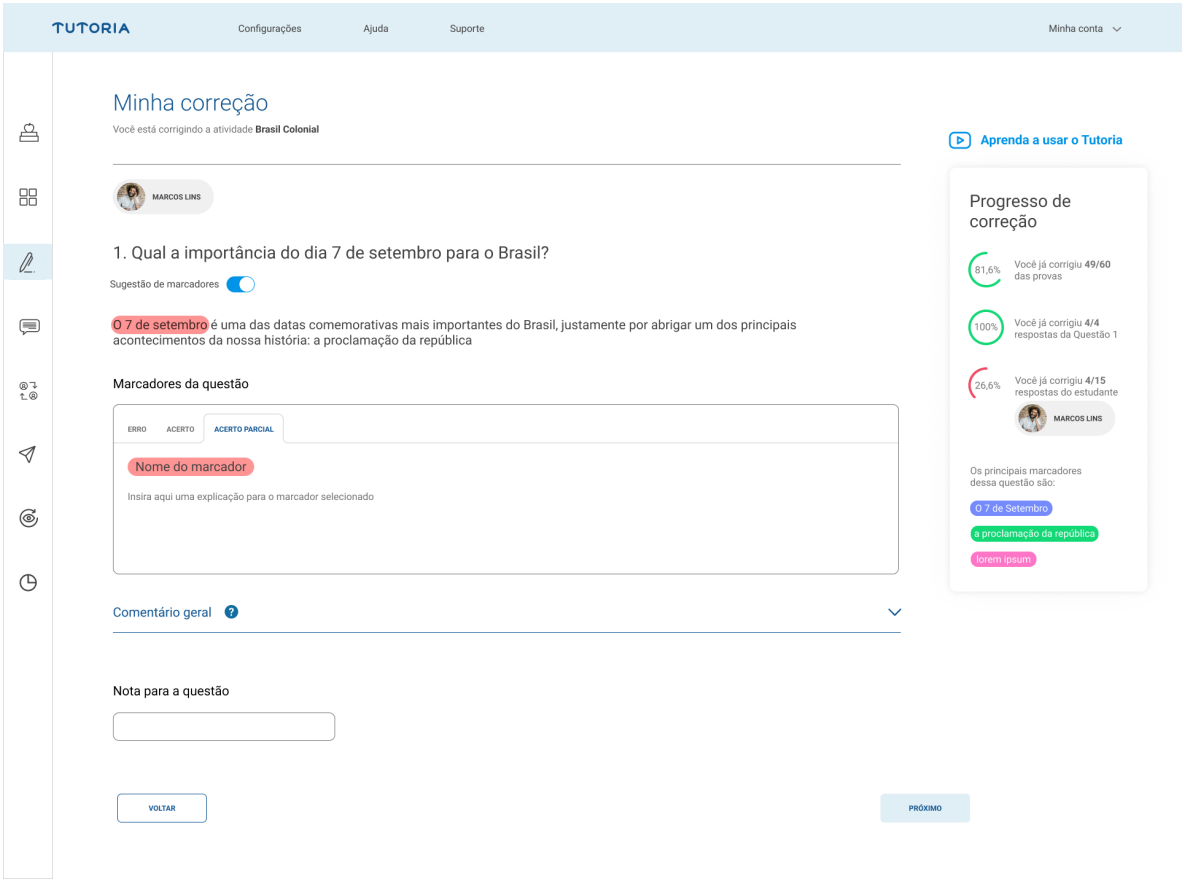

Figura 1. Tela de correção de questão aberta.

Atualmente a plataforma Tutoria oferece suporte à correção de questões abertas. Para corrigir uma questão aberta o professor pode criar um conjunto de tags, relacionadas a erros e acertos, que contém um *feedback* próprio. Essas tags podem ser reaproveitadas para corrigir atividades de diferentes alunos, assim, o professor consegue facilmente reutilizar a mensagem de *feedback* escrita [Falcao et al. 2022a].

A principal contribuição desta pesquisa está na aplicação de algoritmos de inteligência artificial durante o fluxo de correção, como ilustrado na Figura 2. No primeiro estágio, o professor recebe a resposta de um estudante (Estudante 1) e procede à seleção dos trechos relevantes, identificando erros ou acertos na resposta do aluno. Em seguida, o professor gera *feedbacks* correspondentes a cada tag, conforme exemplificado no passo 2.

Durante a correção subsequente, destinada a um novo estudante (Estudante 2), conforme representado no passo 3, um algoritmo de inteligência artificial entra em ação. Ele analisa as tags previamente atribuídas na correção do Estudante 1 e recomenda possíveis erros ou acertos com base nessas tags. Importante destacar que a recomendação do algoritmo pode ser aceita ou rejeitada pelo professor.

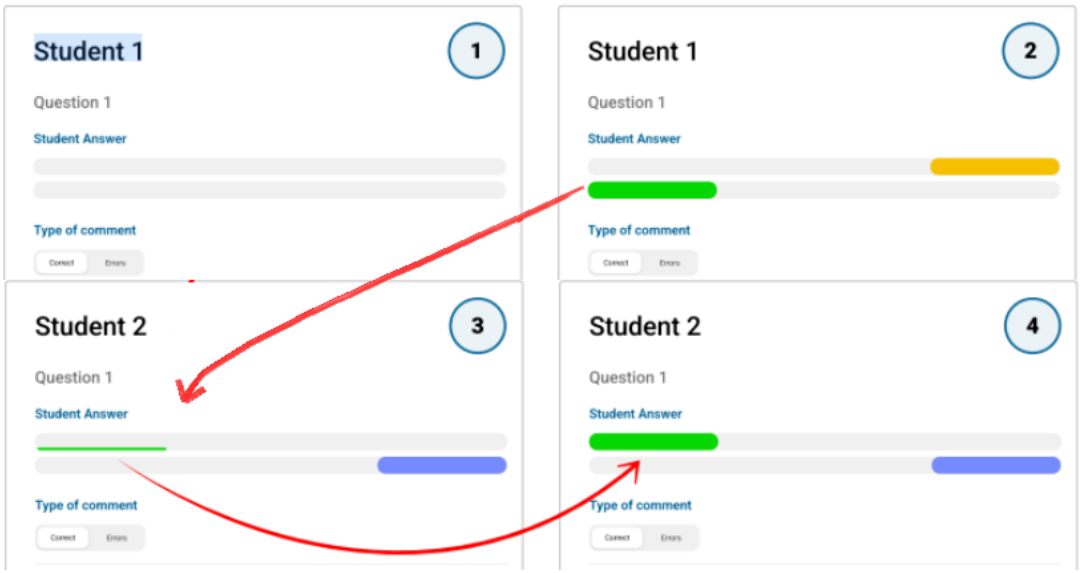

Figura 2. Fluxo de sistema de recomendação de tag proposto.

Quando o professor aceita a recomendação, como indicado no passo 4, o conteúdo do *feedback* já elaborado anteriormente é automaticamente aplicado à correção do estudante atual. Essa abordagem demonstra a integração eficaz de algoritmos de IA no processo de correção da plataforma Tutoria.

Nesse contexto, esse estudo apresenta o resultado de vários algoritmos na recomendação de tags para esse processo de correção. Especificamente a contribuição desse artigo vai na direção de avaliar o ChatGPT nesse problema.

4. Método

Esta seção apresenta o método utilizado para avaliar a recomendação de tags. Esse método parte da **coleta** das respostas de questões abertas, relacionadas a um curso de computação. Em seguida, é realizado o **processamento** destas respostas para então ser realizada a **extração** das características que serão utilizadas para **comparar** os diferentes algoritmos de PLN. Esta comparação é realizada a partir de um conjunto de referência com as tags selecionadas pelo instrutor da disciplina.

Os dados educacionais utilizados para a validação da recomendação de tag correspondem a um exercício extraído de um curso de graduação totalmente online sobre Informática Básica, que explora temas relacionados a hardware, software, redes, sistema operacional, entre outros. Esse curso incluiu uma série de vídeos instrutivos sobre diferentes tópicos usados em combinação com tarefas online contendo perguntas de múltipla escolha e abertas. A cada duas semanas, novos vídeos e tarefas eram fornecidos aos alunos. Esses trabalhos representaram 50% da nota final. Na oferta do curso analisado, um total de 47 alunos responderam à primeira tarefa, contendo cinco questões abertas.

O instrutor desse curso, com formação em ciência da computação, concordou em usar a plataforma Tutoria para avaliar as respostas abertas sem o sistema de recomendação de tags, a fim de gerar as tags para cada resposta manualmente e produzir o padrão ouro neste estudo. Assim, esse estudo avalia o sistema de recomendação com base nas tags in-

cluídas por um instrutor. A Tabela 1 apresenta os detalhes do número de tags dividido por cada questão. Mostra o (1) número de respostas dos alunos; (2) número de tags exclusivas que o instrutor incluiu; (3) número total de tags, incluindo a repetição da mesma tag para respostas diferentes; e (4) número máximo de tags exclusivas que podem ser sugeridas se o sistema recomendar todas as tags exclusivas para todas as respostas dos alunos. Esse experimento avalia respostas com 100 a 200 palavras.

Tabela 1. Distribuição de tags por questão.

Questão	Respostas	Tags únicas	Número total de tags	Recomendação potencial
Q1	33	4	35	132
Q2	33	4	42	132
Q3	47	10	98	470
Q4	47	3	63	141
Q5	47	4	37	188
Total	207	25	275	1063

Após a coleta destas respostas, foram utilizadas técnicas de Processamento de Linguagem Natural (PLN) para o processamento e extração das características dos textos destas respostas. As medidas de similaridade adotadas nesse estudo precisam ser aplicadas apenas às palavras. Portanto, remove-se pontuação, espaços duplicados e caracteres especiais.

Os modelos elaborados para a identificação da similaridade são baseados no *TF-IDF*, no *BERT* e em *Modelo de Linguagem Grande*. O TF-IDF é uma das abordagens mais utilizadas em modelos de mineração de texto para extrair características de textos [Manning and Schutze 1999]. Esse algoritmo converte documentos textuais (por exemplo, respostas dos alunos) em um vetor que consiste na contagem de termos, nesse caso os valores TF-IDF. O presente estudo adotou a técnica tradicional de TF-IDF [Manning and Schutze 1999].

O BERT é uma abordagem de aprendizado profundo estado da arte na comparação de similaridades entre textos. Esta abordagem incorpora o contexto de cada palavra, o que tende a melhorar o desempenho em várias aplicações de PLN [Devlin et al. 2019]. Estudos anteriores mostraram o potencial do uso do BERT em sistemas automatizados de classificação de respostas curtas [Bonthu et al. 2021]. É importante mencionar que não foi realizada nenhuma preparação de dados em nosso conjunto de dados antes de usar o BERT, como sugerido pelos estudos anteriores [Devlin et al. 2019, Bonthu et al. 2021].

Além das técnicas tradicionais de Processamento de Linguagem Natural (PLN) e BERT também foi explorado o uso do ChatGPT como parte da análise. O ChatGPT é um modelo de linguagem largo (LLM), conhecido por sua capacidade de compreender e gerar texto de alta qualidade [Topal et al. 2021]. O ChatGPT foi integrado a fim de avaliar sua eficácia na atividade de recomendação de tags. Ele foi utilizado para para analisar e interpretar respostas de alunos, identificando padrões e conceitos subjacentes. Essa inclusão permite uma perspectiva única e avançada na extração de características, considerando o contexto e a semântica das respostas dos alunos.

As medidas de similaridade avaliadas neste artigo são compostas por métodos

estatísticos para realizar correspondência de *strings*, correspondência de palavras, a abordagem de aprendizado profundo usando BERT[Devlin et al. 2019] e a utilização do modelo geracional ChatGPT. O resultado de cada medida de similaridade é um número entre 0 e 1, onde 1 é a similaridade máxima. Com base nos estudos anteriores, as tags são recomendadas caso o coeficiente de similaridade seja superior a $0,7$ [Marın 2004, Cutrone and Chang 2010, Siddiqi et al. 2010, Bonthu et al. 2021]. Esse trabalho utiliza uma abordagem baseada em outros trabalhos da literatura [Mello et al. 2022] com a adição da avaliação do ChatGPT.

Métodos como a distância de Levenshtein, que conta as modificações necessárias para converter uma *string* em outra [Yujian and Bo 2007] foram utilizados na comparação. Também foram considerados os algoritmos *Partial Ratio*, *Token Sort Ratio*, *Partial Token Sort Ratio*, *Token Set Ratio* e *Partial Token Set Ratio*, que incorporam processos de tokenização e tratamento de *stopwords* para melhorar as comparações. Os algoritmos *Fuzzy Search*, *Edit Distance* e *Rapidfuzz*, cada um com suas particularidades [Wang et al. 2017], também foram incluídos na análise. Esses algoritmos têm em comum o fato de buscarem aferir a similaridade com base na sobreposição de *strings* a nível de caracteres.

Além dos algoritmos anteriores, também foram analisados algoritmos que utilizam o TF-IDF para vetorizar os textos analisados, computando as semelhanças usando:

1-gram: Compara a similaridade de cada palavra nos dois textos.
2-gram: Compara a similaridade de cada par de palavras nos dois textos.
3-gram: Compara a similaridade de cada segmento de três palavras em ambos os textos.
4-gram: Compara a similaridade de cada segmento de quatro palavras nos dois textos.
n-gram: Usa as similaridades de 1, 2, 3 e 4-gram para calcular a pontuação final.

Diferentemente dos algoritmos mencionados anteriormente, o BERT encapsula um vetor para cada palavra, em vez de considerar a frase como um todo. Essa abordagem permite a comparação de duas matrizes, o que possibilita a captura de informações semânticas detalhadas sobre as palavras em comparação. Isso amplia nossa capacidade de compreender a semelhança entre os textos analisados em um nível mais profundo e contextual. Esse modelo também foi incluído no experimento.

Por fim, o modelo geracional ChatGPT foi utilizado a fim de observar se teria um desempenho melhor que os demais algoritmos analisados. Esse modelo foi utilizado através de sua API chamada remotamente a partir de um servidor. Foram realizados dois *prompts* distintos. No primeiro prompt, foi informado ao ChatGPT que ele deveria realizar o trabalho de recomendar tags baseado nas correções já feitas por um professor para um aluno novo. O ChatGPT retornou sua confiança (um valor entre 0 e 1) nas tags que deveriam ou não ser sugeridas na resposta nova, com base nas correções anteriores. No segundo prompt, além de informar ao ChatGPT sobre o trabalho de recomendação de tags com base nas correções do professor, também foi revelado o tipo de cada uma das tags que o professor já havia marcado como acerto ou erro.

Para avaliar o desempenho dos algoritmos, foram utilizadas as medidas tradicionais de aprendizado de máquina: *precisão*, *revocação* e a *medida F1* [Manning and Schutze 1999]. No contexto da recomendação de tags, priorizar a precisão pode fazer com que o sistema não recomende todas as tags esperadas, enquanto

a revocação pode fazer com que muitas tags sejam recomendadas de forma indevida. Portanto, apesar de observar estas duas métricas, esse trabalho tem o foco na medida F1, que é a média harmônica entre a precisão e a revocação.

5. Resultados

Os resultados apresentados nesta seção mostram o desempenho de cada medida de similaridade para a recomendação de 25 tags para respostas de 207 alunos. A Tabela 2 apresenta os resultados de cada algoritmo de similaridade que foi avaliado usando precisão, revocação, medida F1 e o tempo para executar a recomendação para todas as respostas em segundos.

Em geral, todos os algoritmos alcançaram bons resultados em termos de precisão. Isso significa que os algoritmos conseguiram recomendar tags para os instrutores corretamente. Em contraste, algoritmos múltiplos obtiveram uma revocação inferior a 0,50. Em outras palavras significa que o sistema não recomendou tags para a maioria das questões e o instrutor teve que fazê-lo manualmente, mantendo uma carga de trabalho semelhante à não utilização da abordagem proposta.

Três abordagens alcançaram a medida F1 superior a 0,85: Partial Token Set Ratio, ChatGPT (COM revelação de tipos de tag) e ChatGPT (SEM revelação de tipos de tag). Esses algoritmos alcançaram um equilíbrio entre precisão e revocação, e foram os algoritmos mais adequados para esta tarefa. No entanto, dentre esses, o ChatGPT foi mais lento, levando pouco mais de 1 segundo para realizar as previsões de tags[2] e com maior custo total[3] enquanto o *Partial Token Set Ratio* fez as mesmas recomendações em 0,01 segundos.

Tabela 2. Resultados de cada algoritmo na recomendação de tag.

#	Algoritmo de Similaridade	Precisão	Revocação	F1	Tempo (s)	Custo (R$)[4]
1	Levenshtein	0.80	0.01	0.01	00.02	0.00
2	Partial Ratio	0.98	0.32	0.48	00.06	0.00
3	Token Sort Ratio	0.94	0.02	0.03	00.05	0.00
4	Token Set Ratio	0.97	0.43	0.59	00.05	0.00
5	**Partial Token Set Ratio**	**0.90**	**0.92**	**0.91**	**00.01**	**0.00**
6	Partial Token Sort Ratio	0.96	0.25	0.39	00.07	0.00
7	Fuzzy Search	0.90	0.42	0.57	01.22	0.00
8	Edit Distance	0.93	0.62	0.74	01.03	0.00
9	Rapidfuzz	0.93	0.60	0.72	00.86	0.00
10	TF-IDF 1-gram	0.90	0.74	0.81	05.10	0.00
11	TF-IDF 2-gram	0.98	0.06	0.10	04.84	0.00
12	TF-IDF 3-gram	0.98	0.02	0.03	04.78	0.00
13	TF-IDF 4-gram	0.96	0.01	0.02	04.71	0.00
14	TF-IDF n-gram	0.94	0.19	0.31	05.27	0.00
15	BERT	0.89	0.79	0.83	93.28	0.00
16	**ChatGPT (SEM tipos de tag)**	**0.88**	**0.84**	**0.86**	**01.11[3]**	**6.44**
17	**ChatGPT (COM tipos de tag)**	**0.89**	**0.88**	**0.88**	**01.07[3]**	**6.44**

[2]Tempo referente a uma chamada via API

[3]Custos para chamada de uma API privada, considerando 300 palavras em média para cada resposta no dataset de validação 1. Algoritmos que rodam localmente já têm seu custo atrelado ao custo do servidor da aplicação e foram desconsiderados.

6. Discussões e implicações

O principal objetivo desse artigo foi destacar a utilização do ChatGPT no contexto de recomendação de tags para avaliação de questões abertas. Como apresentado nos resultados o desempenho do ChatGPT foi bem próximo ao melhor algoritmo apresentado em estudos anteriores [Mello et al. 2022]. Apesar de não ter sido o melhor modelo é importante destacar alguns aspectos relacionados a utilização do ChatGPT.

Primeiro, o domínio da atividade era bastante restrito, por isso a medida Partial Token teve um bom desempenho. É possível que em um teste com atividades mais elaboradas e de disciplinas diversas o ChatGPT consiga um resultado mais consistente [Kasneci et al. 2023].

Segundo, o prompt é uma questão que influencia bastante o resultado final em qualquer modelo de linguagem. Não foi o foco desse trabalho avaliar diferentes prompts, mas é possível que o ChatGPT melhore seus resultados com a utilização de técnicas avançadas de engenharia de prompt [White et al. 2023]. Este tipo de análise pode vir a ser realizada em trabalhos futuros.

Por fim, as limitações da utilização do ChatGPT é que ele tem um custo alto associado e ainda existem muitas considerações éticas que devem ser estudadas antes da aplicação massiva desses modelos na educação [Yan et al. 2023].

7. Agradecimentos

Esse trabalho foi desenvolvido a partir do programa de computador *FEEDBACKBOT* (ou *Tutoria*), registrado no processo de número *BR512023000376-0* com a titularidade da *Rede Nacional de Ensino e Pesquisa* (RNP) e da *Universidade Federal Rural de Pernambuco* (UFRPE). A plataforma *Tutoria*, bem como uma startup homônima criada a fim de evoluir essa plataforma, foram resultados do *Programa de Pesquisa e Desenvolvimento em Serviços Avançados* da RNP. Os autores agradecem à RNP, à UFRPE e à startup Tutoria pelo apoio no desenvolvimento desse trabalho.

Referências

Bonthu, S., Rama Sree, S., and Krishna Prasad, M. (2021). Automated short answer grading using deep learning: A survey. In *International Cross-Domain Conference for Machine Learning and Knowledge Extraction*, pages 61–78. Springer.

Boud, D. and Molloy, E. (2013). Rethinking models of feedback for learning: the challenge of design. *Assessment & Evaluation in higher education*, 38(6):698–712.

Cutrone, L. A. and Chang, M. (2010). Automarking: automatic assessment of open questions. In *2010 10th IEEE International Conference on Advanced Learning Technologies*, pages 143–147. IEEE.

de Lima Dias, A. N. and Pazoti, M. A. (2023). Correção automatizada de questões dissertativas utilizando medidas de similaridade e processamento de linguagem natural. In *Colloquium Exactarum. ISSN: 2178-8332*, volume 15, pages e234595–e234595.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186. doi: 10.18653/v1/N19-1423.

Falcão, T. P., Arêdes, V., de Souza, S. B. J., Fiorentino, G., Neto, J. R., Alves, G., and Mello, R. F. (2023). Tutoria: a software platform to improve feedback in education. *Journal on Interactive Systems*, 14(1):383–393.

Falcão, T. P., Arêdes, V., Souza, S., Luisi, V., Neto, G. F., Neto, R., Morais, D., Miranda, P. B., and Mello, R. F. (2020). Tutoria: uma plataforma para apoiar boas práticas de feedback no processo de ensino e aprendizagem. In *Anais dos Workshops do X Congresso Brasileiro de Informática na Educação*, pages 213–220. SBC.

Falcao, T. P., Arêdes, V., Wagner, S. S., Uchoa, J. P. C., Luisi, V., and Mello, R. F. (2022a). What did i get wrong? supporting the feedback process in computer science education. In *Anais do XXX Workshop sobre Educação em Computação*, pages 239–250. SBC.

Falcao, T. P., Oliveira, V., Souza, S., Fiorentino, G., Neto, J. R., Galdino, J. V., Alves, G., and Mello, R. F. (2022b). Tutoria: Supporting good practices for providing written educational feedback. In *Anais do XXXIII Simpósio Brasileiro de Informática na Educação*, pages 668–679. SBC.

José, J., Paiva, R., and Bittencourt, I. I. (2015). Avaliação automática de atividades escritas baseada em algoritmo genético e processamento de linguagem natural: Avaliador ortográfico-gramatical. In *Anais dos Workshops do Congresso Brasileiro de Informática na Educação*, volume 4, page 95.

Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., et al. (2023). Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and individual differences*, 103:102274.

Krusche, S. and Seitz, A. (2018). Artemis: An automatic assessment management system for interactive learning. In *Proceedings of the 49th ACM Technical Symposium on Computer Science Education*, pages 284–289. ACM.

Manning, C. and Schutze, H. (1999). *Foundations of statistical natural language processing*. MIT press.

Marın, D. R. P. (2004). Automatic evaluation of users' short essays by using statistical and shallow natural language processing techniques. *Advanced Studies Diploma Work, University of Madrid*.

Mello, R. F., Neto, R., Fiorentino, G., Alves, G., Arêdes, V., Silva, J. V. G. F., Falcão, T. P., and Gašević, D. (2022). Enhancing instructors' capability to assess open-response using natural language processing and learning analytics. In *European Conference on Technology Enhanced Learning*, pages 102–115. Springer.

Neto, J. R., Falcao, T. P., Oliveira, V., Souza, S., Fiorentino, G., Galdino, J. V., Alves, G., and Mello, R. F. (2022). Tutoria: Plataforma para suporte à correção de atividades e envio de feedback personalizado. In *Anais do I Workshop de Aplicações Práticas de Learning Analytics em Instituições de Ensino no Brasil*, pages 21–29. SBC.

Rahman, M. M. and Watanobe, Y. (2023). Chatgpt for education and research: Opportunities, threats, and strategies. *Applied Sciences*, 13(9):5783.

Siddiqi, R., Harrison, C. J., and Siddiqi, R. (2010). Improving teaching and learning through automated short-answer marking. *IEEE Transactions on Learning Technologies*, 3(3):237–249.

Singh, A., Karayev, S., Gutowski, K., and Abbeel, P. (2017). Gradescope: a fast, flexible, and fair system for scalable assessment of handwritten work. In *Proceedings of the fourth (2017) acm conference on learning@ scale*, pages 81–88.

Topal, M. O., Bas, A., and van Heerden, I. (2021). Exploring transformers in natural language generation: Gpt, bert, and xlnet. *arXiv preprint arXiv:2102.08036*.

Wang, Y., Qin, J., and Wang, W. (2017). Efficient approximate entity matching using jaro-winkler distance. In *International Conference on Web Information Systems Engineering*, pages 231–239. Springer.

White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., Elnashar, A., Spencer-Smith, J., and Schmidt, D. C. (2023). A prompt pattern catalog to enhance prompt engineering with chatgpt. *arXiv preprint arXiv:2302.11382*.

Wiggins, G. (1998). *Educative Assessment. Designing Assessments To Inform and Improve Student Performance.* ERIC.

Yan, L., Sha, L., Zhao, L., Li, Y., Martinez-Maldonado, R., Chen, G., Li, X., Jin, Y., and Gašević, D. (2023). Practical and ethical challenges of large language models in education: A systematic scoping review. *British Journal of Educational Technology*.

Yujian, L. and Bo, L. (2007). A normalized levenshtein distance metric. *IEEE transactions on pattern analysis and machine intelligence*, 29(6):1091–1095.