Aplicação de reconhecimento de entidades nomeadas para análise automática de textos narrativos em produções textuais do ensino fundamental

Aluno: Erverson Bruno Gomes de Sousa[1]
Orientador: Rafael Ferreira Mello[1]

[1]Centro de Estudos e Sistemas Avançados do Recife (CESAR SCHOOL)
Recife – PE – Brasil

{ebgs,rflm}@cesar.school

Resumo. *A produção de texto é uma tarefa imprescindível para os estudantes do Ensino Fundamental (EF). O texto narrativo é abordado nessa etapa de ensino. Os professores, na correção dos textos, precisam identificar manualmente os elementos dos textos narrativos, o que pode gerar uma sobrecarga e dificultar um feedback personalizado. Prover a identificação automática dos elementos do texto narrativo, no idioma português do Brasil, é um desafio na área de reconhecimento de entidades nomeadas, bem como apoiar os docentes, por meio de um dashboard, a acompanharem individualmente as dificuldades dos alunos. O objetivo da pesquisa é desenvolver uma ferramenta para apoiar os professores na correção das produções textuais de estudantes do EF.*

1. Introdução

As escolas têm enfrentado novos desafios devido à pandemia mundial de Covid 19, o que vem causando impactos, principalmente, no que diz respeito ao processo de ensino e aprendizagem [Onyema et al. 2020]. Muitas instituições escolares, mesmo sem o tempo e recursos devidos, tiveram que migrar às aulas para o mundo digital, por meio de aplicativos e plataformas educacionais [Peres 2020]. A adoção abrupta do ensino remoto, evidenciou a precariedade socioeducacional de vários países, dentre eles o Brasil. As estratégias adotadas para dar continuidade ao ensino, não alcançaram alguns alunos e professores, devido ao contexto de vulnerabilidade social [Giordano 2021, Peres 2020].

No cenário da educação básica, tais dificuldades são potencializadas, pois compreende um dos períodos mais importantes para os estudantes, onde é realizado o processo de alfabetização e a construção de habilidades matemáticas, que servem de base durante todo o seu percurso escolar [de Queiroz et al. 2021, Ferreira et al. 2020]. Essas experiências e conhecimentos foram afetados completamente devido a adaptação ao novo cenário de ensino remoto, aumentando a lacuna de aprendizagem que já existia na educação brasileira [Oliveira et al. 2008, Buselli et al. 2020].

Diante desse contexto, foi publicado no diário oficial da união o decreto nº 11.079[1] que estabelece a Política Nacional para Recuperação das Aprendizagens na Educação Básica, a qual tem o intuito de implementar estratégias, programas e ações para a recuperação das aprendizagens e o enfrentamento da evasão e do abandono escolar na

[1]Decreto nº 11.079, de 23 de maio de 2022: https://www.in.gov.br/web/dou/-/decreto-n-11.079-de-23-de-maio-de-2022-402040949

DOI: 10.5753/cbie_estendido.2022.226747

educação básica. Tal ação reforça a importância do desenvolvimento de estudos e projetos que colaborem para sanar as lacunas de aprendizagem que foram geradas e potencializadas no ensino remoto.

O ensino de produção textual, elemento imprescindível para o desenvolvimento da capacidade de escrita de estudantes da educação básica, foi uma das etapas do ensino da língua portuguesa afetada pelo contexto pandêmico e que está contemplada pela Política Nacional para Recuperação das Aprendizagens. O desenvolvimento da produção de texto inicia desde o processo de alfabetização e acompanha o estudante durante todo o seu percurso acadêmico e profissional, logo assim é essencial que tal capacidade seja desenvolvida de forma satisfatória desde os anos iniciais e finais do ensino fundamental[Rocha 2020].

A tipologia textual narrativa é utilizada nas produções de diversos gêneros textuais nos do ensino fundamental, diferentemente do que ocorre no ensino médio, onde os alunos desenvolvem produções de textos dissertativos-argumentativos [Rezende and de Souza 2018]. O tipo de texto narrativo contempla elementos como narrador, enredo, personagens, tempo e espaço. E propicia ao estudante a capacidade de "construção de conhecimento do mundo, do próprio sujeito produtor de textos e, consequentemente, da linguagem verbal, o que pode ser de grande valor educativo" [Rezende and de Souza 2018]. Tais elementos da narrativa são imprescindíveis para caracterizar o texto como narrativo e isso requer do professor uma análise para a correção de produções textuais que identifique-os, e consiga fornecer um feedback útil para os alunos do que faltou e o que precisa ser aprimorado, proporcionando um ensino personalizado.

Por um lado, o ensino personalizado tem ganhado força nos últimos anos, principalmente no período de ensino remoto, que propiciou a geração de dados no contexto educacional, devido a adoção massiva de ferramentas digitais nos ambientes escolares [Gaftandzhieva et al. 2021]. Os dados gerados das interações dos alunos com ferramentas digitais podem ser utilizados para monitorar, analisar, prever, intervir, recomendar e, principalmente, melhorar a qualidade no processo de ensino e aprendizagem por meio de técnincas de Learning Analytics (LA)[Sousa et al. 2021]. LA é um campo emergente que aborda esse contexto de análise de dados educacionais, cujo objetivo é a coleta, análise e relatório de dados sobre os estudantes e os contextos onde eles ocorrem [Siemens and Gasevic 2012]. Essas são funcionalidades preconizadas na prática docente, o qual tem o intuito de manter um processo de ensino e aprendizagem personalizado, para atender as necessidades específicas de cada estudante.

Por outro lado, é necessário utilizar técnicas para a análise automática da produção textual. Neste contexto, o Processamento de Linguagem Natural (PLN) é uma subárea da inteligência artificial que utiliza técnicas computacionais para análise de uma língua, seja ela falada ou escrita, que esteja no formato digital [Alhawiti 2014]. Têm havido aplicações de técnicas de PLN para correção automática de textos/redações no contexto da educação básica, as quais auxiliam os professores a fornecer um feedback mais assertivo, eficaz e em tempo hábil para aprimoramento da escrita, possibilitando ao docente o uso do seu tempo pedagógico focado em elaborar estratégias para sanar as dificuldades de escritas mais recorrentes dos discentes [Ramesh and Sanampudi 2021].

Uma das tarefas de PLN é o reconhecimento de entidades nomeadas (do inglês,

Named Entity Recognition - NER), que se propõe a aplicar técnicas de extração de informação em textos escritos com o intuito de identificar automaticamente classes pré determinadas, tais como pessoas, organizações e locais [Goyal et al. 2018]. O NER tem sido aplicado em diversos domínios, principalmente no idioma inglês, como no contexto de extração de informações de redes sociais, do setor jurídico e educacional[Nasar et al. 2021]. No âmbito educacional, em textos da Língua Portuguesa, tem havido soluções de NER para gerar automaticamente questões a partir de textos didáticos [Nasar et al. 2021], na área da geologia [Amaral et al. 2017] e para análise de diferentes tipos de gêneros textuais [Pirovani and Oliveira 2021].

A aplicação de técnicas de NER, em português do Brasil, no contexto educacional, ainda são incipientes e carecem de mais estudos para o desenvolvimento de soluções que forneçam suporte para os professores na tarefa de correções de produções textuais [Ferreira-Mello et al. 2019]. Principalmente na educação básica, onde a base para a capacidade de escrita é iniciada, e os discentes carecem de um acompanhamento e um feedback mais personalizado e individualizado.

Levando em consideração as informações apresentadas, uma das características dos modelos de análise textual de NER é que propostas aplicadas em um idioma específico, dentro de um domínio particular, como o da educação básica, focado em uma tipologia textual específica, geralmente tem um melhor desempenho do que os que são desenvolvidos dentro de um contexto mais genérico [Nasar et al. 2021]. Diante do exposto é relevante avançar o estado da arte no que diz respeito a pesquisas do uso de NER para análise de produções textuais, no idioma português do Brasil, no domínio do ensino fundamental - anos finais, da educação básica brasileira.

Partindo da contextualização apresentada, este projeto de tese, tem-se o intuito de responder a seguinte questão de pesquisa: *Como apoiar docentes do ensino fundamental a corrigir produção de textos dos alunos, na perspectiva de automatizar o processo de análise de enredo na narrativa, com o intuito de prover um feedback mais personalizado para os estudantes?*

Com o propósito de responder a questão de pesquisa, foram definidos o seguinte objetivo geral e objetivos específicos:

Objetivo Geral:

- Desenvolver uma ferramenta para apoiar os professores na correção de produções de texto de estudantes dos anos finais do ensino fundamental por meio de técnicas de Reconhecimento de Entidades Nomeadas.

Objetivos Específicos:

- Coletar estudos, por meio de uma revisão sistemática da literatura, que utilizaram técnicas de PLN para correção automática de produção textual no contexto da educação básica do Brasil;
- Desenvolvimento de algoritmo para análise de tipologia textual;
- Utilização de NER para identificação de elementos da narrativa;
- Utilização de modelos de redes neurais profundas para análise de enredo na narrativa;
- Aplicação dos modelos desenvolvidos em provas reais realizadas em escolas de educação básica, nos anos finais do ensino fundamental;

- Desenvolvimento de um dashboard para apoiar os professores na análise das produções textuais dos estudantes.

2. Metodologia

O presente trabalho utilizará a pesquisa aplicada como tipo de estudo, que de acordo com [Gil 2010] são "pesquisas voltadas à aquisição de conhecimentos com vistas à aplicação numa situação específica", onde a necessidade de produzir conhecimento para aplicação de seus resultados é a motivação para "contribuir para fins práticos, visando à solução imediata do problema encontrado na realidade" [BARROS and SILVEIRA BARROS 2007].

Para fins da pesquisa, ela se caracteriza como descritiva, pois tem o objetivo de descrever as características de determinadas populações ou fenômenos e "podem ser elaboradas também com a finalidade de identificar relações entre variáveis" [Gil 2010]. A sua abordagem será qualitativa para análise dos dados da pesquisa, de acordo com [Gil et al. 2002] "a análise qualitativa depende de muitos fatores, tais como a natureza dos dados coletados, a extensão da amostra, os instrumentos de pesquisa e os pressupostos teóricos que nortearam a investigação".

Quanto ao procedimento técnico, o estudo se classifica como pesquisa-ação, que é definida como um tipo de pesquisa com base empírica e tem uma "estreita associação com uma ação ou com a resolução de um problema coletivo e no qual os pesquisadores e participantes representativos da situação ou do problema estão envolvidos de modo cooperativo ou participativo" [Thiollent 2011].

Com o intuito de responder às questões de pesquisa e alcançar o objetivo deste estudo, as ferramentas de coleta de dados serão, a utilização de formulários, observações e entrevistas com os professores da disciplina de Língua Portuguesa, dos anos finais do ensino fundamental, da Escola Professora Olindina Roriz Dantas, bem como serão realizados testes com os modelos desenvolvidos em produções textuais reais, do corpus de redações, com aproximadamente 2000 textos narrativos, de escolas que aderiram ao programa Brasil na Escola do Ministério da Educação (MEC).

Para alcançar o objetivo geral desta pesquisa, pretende-se desenvolver as seguintes etapas: Levantamento do estado da arte por meio de uma Revisão Sistemática na aplicação de PLN à produção textual de alunos do ensino fundamental; Seleção de modelos de redes neurais profundas aplicados ao contexto de NER; Modelagem de um banco de dados para tipologia textual; Desenvolvimento de algoritmo para análise de textos no gênero narrativo; Aplicação dos modelos desenvolvidos em provas reais de alunos do ensino fundamental brasileiro; Avaliação dos resultados; Desenvolvimento de dashboard para o professor acompanhar os resultados dos alunos.

3. Resultados

Com o desenvolvimento desta pesquisa pretende-se alcançar os seguintes resultados:

- Construção de uma revisão sistemática da literatura no uso de técnicas de PLN para correção automática de produções textuais na educação básica;
- Disponibilização de um algoritmo para análise de tipologia textual, elementos da narrativa e enredo por meio de técnicas de NER;

- Desenvolvimento de artigo com os resultados obtidos a partir da avaliação do modelo construído com produções textuais reais de alunos do ensino fundamental;
- Desenvolvimento de *Dashboard* para o professor acompanhar os resultados das produções de texto e fornecer um feedback personalizado para os estudantes.

Com tais resultados, é esperado a otimização do tempo pedagógico dos professores, no que diz respeito a correção de produções textuais, para haver a possibilidade de um feedback mais assertivo e individualizado para cada aluno, levando em consideração o seu histórico de produções textuais, identificando os pontos de melhorias automaticamente à medida que a ferramenta, por meio do algoritmo desenvolvido, é alimentada com mais produções de textos no decorrer dos bimestres e anos letivos no ensino fundamental - anos finais.

4. Considerações finais

Este projeto está sendo desenvolvido no programa de Doutorado Profissional em Engenharia de Software da CESAR School. O ingresso no curso foi em setembro de 2020 na área de pesquisa de Learning Analytics. Em março de 2022, integramos a equipe de pesquisadores do programa Brasil na Escola do Ministério de Educação em parceria com a Universidade Federal de Alagoas (UFAL). A defesa da tese está prevista para agosto de 2024.

Vale ressaltar que dos resultados propostos para esse trabalho, a revisão sistemática da literatura, que proverá o estado da arte do uso de técnicas de PLN para correção automática de produções textuais na educação básica, já foi iniciada e está na etapa final da redação do texto, que resultará em um artigo. Ao ser concluído será submetido a um periódico para possível publicação.

Referências

Alhawiti, K. M. (2014). Natural language processing and its use in education. *International Journal of Advanced Computer Science and Applications*, 5(12).

Amaral, D. O. F. d. et al. (2017). Reconhecimento de entidades nomeadas na área da geologia: bacias sedimentares brasileiras.

BARROS, A. J. d. S. and SILVEIRA BARROS, N. A. d. S. (2007). Lehfeld. fundamentos de metodologia científica. *Makron*.

Buselli, M., Estevão, K. P., and Sambugari, M. F. (2020). A importância da alfabetização matemática no ciclo i do ensino fundamental. *Revista Eletrônica de Ciências Humanas*, 3(2).

de Queiroz, M., de Sousa, F. G. A., and de Paula, G. Q. (2021). Educação e pandemia: impactos na aprendizagem de alunos em alfabetização. *Ensino em Perspectivas*, 2(4):1–9.

Ferreira, L. G., Ferreira, L. G., and Zen, G. C. (2020). Alfabetização em tempos de pandemia: perspectivas para o ensino da língua materna. *Fólio-Revista De Letras*, 12(2).

Ferreira-Mello, R., André, M., Pinheiro, A., Costa, E., and Romero, C. (2019). Text mining in education. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 9(6):e1332.

Gaftandzhieva, S., Docheva, M., and Doneva, R. (2021). A comprehensive approach to learning analytics in bulgarian school education. *Education and Information Technologies*, 26(1):145–163.

Gil, A. C. (2010). Como elaborar projetos de pesquisa/–12. reimpressão.–são paulo: Atlas, 2009. _. *Como elabora projetos de pesquisa./5. Ed.–São Paulo: Atlas*.

Gil, A. C. et al. (2002). *Como elaborar projetos de pesquisa*, volume 4. Atlas São Paulo.

Giordano, D. X. F. (2021). A pandemia e as consequências no setor educacional: Desafios para os gestores escolares. *Colóquios-Geplage-PPGED-CNPq*, (2):28–35.

Goyal, A., Gupta, V., and Kumar, M. (2018). Recent named entity recognition and classification techniques: a systematic review. *Computer Science Review*, 29:21–43.

Nasar, Z., Jaffry, S. W., and Malik, M. K. (2021). Named entity recognition and relation extraction: State-of-the-art. *ACM Computing Surveys (CSUR)*, 54(1):1–39.

Oliveira, K. L. d., Boruchovitch, E., and Santos, A. A. A. d. (2008). Reading and school performance in portuguese and mathematics in elementary school. *Paidéia (Ribeirão Preto)*, 18:531–540.

Onyema, E. M., Eucheria, N. C., Obafemi, F. A., Sen, S., Atonye, F. G., Sharma, A., and Alsayed, A. O. (2020). Impact of coronavirus pandemic on education. *Journal of Education and Practice*, 11(13):108–121.

Peres, M. R. (2020). Novos desafios da gestão escolar e de sala de aula em tempos de pandemia. *Revista de Administração Educacional*, 11(1):20–31.

Pirovani, J. P. and Oliveira, E. (2021). Studying the adaptation of portuguese ner for different textual genres. *The Journal of Supercomputing*, 77(11):13532–13548.

Ramesh, D. and Sanampudi, S. K. (2021). An automated essay scoring systems: a systematic literature review. *Artificial Intelligence Review*, pages 1–33.

Rezende, N. L. and de Souza, M. C. (2018). Do ensino escolar da escrita de textos narrativos. *Linha D'Água*, pages 143–158.

Rocha, A. G. A. (2020). A importância dos gêneros textuais no processo de ensino-aprendizagem de língua portuguesa. *Revista Científica Multidisciplinar Núcleo do Conhecimento. Ano*, 5:18–32.

Siemens, G. and Gasevic, D. (2012). Guest editorial-learning and knowledge analytics. *Journal of Educational Technology & Society*, 15(3):1–2.

Sousa, E. B. d., Alexandre, B., Ferreira Mello, R., Pontual Falcão, T., Vesin, B., and Gašević, D. (2021). Applications of learning analytics in high schools: a systematic literature review. *Frontiers in Artificial Intelligence*, 4:737891.

Thiollent, M. (2011). Metodologia da pesquisa-ação. In *Metodologia da pesquisa-ação*, pages 136–136.