Predição da Complexidade Textual de Recursos Educacionais Abertos em Português

Murilo Gazzola[1] **, Sidney Evaldo Leal**[1] **, Sandra Maria Aluisio**[1]

[1]Instituto de Ciências Matemáticas e de Computação – Universidade de São Paulo (USP)
Caixa Postal 668 – 13560-970 – Sao Carlos – SP

mgazzola@icmc.usp.br, sidleal@gmail.com, sandra@icmc.usp.br

Abstract. *In 2016, UNESCO stated the priorities for the use of Open Educational Resources (OER), highlighting the main research challenges. The lack of quality of OER is a challenge to be overcome. In a data analysis of the Integrated Platform of the Ministry of Education (MEC-RED) in May 2018, 41% of the resources did not have the teaching stage metadata, making it difficult to search OER, use and edit them. The Textual Complexity task can help identify texts that have linguistic complexity appropriate to specific series, allowing to complete the teaching stage in MEC-RED. In this article, we evaluate the impact of the textual genre in the evaluation of textual complexity, using a model trained in a large corpus of didactic texts and tested in 2 OER datasets of MEC-RED. The best trained model (F-measure 0.804) had an F-measure of 0.518 in a set of OER of the same genre and 0.389 of F-measure for the animation/simulation and practical experiment, two genres of interest in this research.*

Resumo. *Em 2016, a UNESCO escreveu em seu relatório as prioridades para o uso de Recursos Educacionais Abertos (REA), destacando os principais desafios de pesquisas. A falta de qualidade dos REA é um desafio a ser superado. Em uma recolha na Plataforma Integrada do Ministério da Educação (MEC-RED) de maio de 2018, 41% dos recursos não possuíam classificação da etapa de ensino, dificultando sua busca, uso e edição. A tarefa Complexidade Textual pode ajudar a identificar textos que tem complexidade linguística adequada a séries específicas, permitindo completar a etapa de ensino. Neste artigo, avaliamos o impacto do gênero textual na avaliação da complexidade textual, com um modelo treinado em um grande córpus de textos didáticos e testado em 2 conjuntos de REA da MEC-RED. O melhor modelo treinado (0.804 de F-measure) teve uma F-measure de 0.518 em um conjunto de REA de mesmo gênero e 0.389 de F-measure para os recursos do tipo animação/simulação e experimento prático, dois gêneros de interesse nesta pesquisa.*

1. Introdução

O termo Recursos Educacionais Abertos (REA) foi cunhado em 2002 pela Unesco [UNESCO. 2002] no fórum sobre o impacto de cursos de ensino superior aberto em países em desenvolvimento. Os REA podem ser caracterizados como materiais de ensino, aprendizagem e pesquisa em qualquer meio de armazenamento, que estão disponíveis através da licença que permite quatro liberdades mínimas conhecidas como 4R: **R**evisar, **R**eusar, **R**ecombinar e **R**edistribuir [Wiley et al. 2014]. Portanto, devido à importância do tema

[Miao et al. 2016] [Wiley et al. 2014], a Organização das Nações Unidas (ONU) definiu os principais problemas relacionados ao desenvolvimento e uso de REA: i) *o problema de qualidade dos REA; ii) o problema da descoberta, ou seja, como encontrar REA*; iii) *o problema da sustentabilidade*, isto é, como financiá-los; iv) *o problema de localização e recontextualização* dos REA; e v) *o problema do remix*, isto é, a dificuldade de identificar a granularidade da mudança de conteúdo por outras pessoas e o nível de mudança. Consideramos que a qualidade é um fator importante para garantir uma educação e ensino de qualidade, utilizando esses materiais. Assim, buscamos critérios e métodos da área de Processamento de Línguas Naturais (PLN) para avaliar a qualidade dos REA. Os principais trabalhos que avaliaram a qualidade de REA ou materiais semelhantes aos REA não trazem uma análise linguística, nem avaliam o conteúdo de REA, apenas tratam os seus metadados para julgar a qualidade ou usam indicadores subjetivos [Bethard et al. 2009], [Dalip et al. 2011], [Leary et al. 2011], [Cechinel et al. 2011], [Ahmed and Fuge 2017].

A plataforma de Recursos Educacionais do Ministério da Educação (MEC), conhecida como Plataforma Integrada MEC[1] (MEC-RED), faz parte de um dos compromissos do Brasil na *Open Government Partnership* (OGP) para fortalecer práticas que envolvem a transparência dos atos governamentais e promovem a participação social e o acesso à informação pública [MEC 2019b]. Além disso, a MEC-RED centralizou todos os materiais do Portal do Professor[2], Banco Internacional de Objetos Educacionais[3], Domínio Público[4] e TV Escola[5] [MEC 2019a]. Na MEC-RED, há quatro filtros de busca por recursos: componentes curriculares, tipos de recurso, etapas de ensino e palavra-chave, embora a plataforma tenha outros metadados como título, pessoas que favoritam o recurso, escala de estrelas, URL para download associado ao material, descrição do recurso, autor do envio, autor do material, *tags* associadas ao material e tipo de recurso. No início de maio de 2019, a MEC-RED contava com 31.488 recursos, mas as avaliações feitas neste trabalho foram realizadas com informações de uma recolha realizada em maio de 2018, quando a plataforma possuía 28.026 recursos. Destes últimos, 6.966 recursos (41%) não possuíam a informação sobre etapas de ensino preenchido. Particularmente o metadado etapas de ensino permite a busca por material relacionado com a sua complexidade textual e conceitual, para a recuperação de material adequado a uma das etapas do Sistema Educacional Brasileiro. Assim, espera-se o seu preenchimento correto, sendo um item importante para se avaliar a qualidade de um recurso.

Sabemos que certos conteúdos são estudados em séries específicas do Ensino Fundamental I e II, Ensino Médio e Ensino Superior e que a cada nova etapa novos gêneros textuais (por exemplo, romances, crônicas, fábulas, ensaios, anúncios, editoriais e reportagens de jornal, cartas, relatórios, anedotas, dentre outros) são trabalhados e que também os próprios componentes curriculares (ou disciplinas) trazem características de complexidade textual variadas, como, por exemplo, os textos de ciência trazem uma terminologia técnica, textos de história trazem ideologias e interpretação de eventos com grande número de personagens e locais geográficos, os de matemática símbolos e organização textual novos e conceitos mais abstratos [Fang 2016]. Neste trabalho, propomos iden-

[1] https://plataformaintegrada.mec.gov.br/
[2] http://portaldoprofessor.mec.gov.br
[3] http://objetoseducacionais2.mec.gov.br
[4] http://www.dominiopublico.gov.br/)
[5] http://tvescola.org.br/

tificar automaticamente a etapa de ensino, via uma tarefa do PLN chamada predição automática da complexidade textual (*text readability*, em inglês) que estuda as características lexicais, sintáticas, semânticas e discursivas [Fang 2016] que podem impactar na dificuldade/facilidade de um texto ser lido por um aluno, que tem um conjunto de conhecimentos prévios, no contexto de uma dada atividade escolar. Como a MEC-REC possui 14 tipos de recursos, excluímos a análise dos áudios, imagens, infográfico, mapas e vídeos, por não se apresentarem no formato textual. De especial interesse neste trabalho são os tipos de recursos animação/simulação, aplicativo móvel, jogos, experimento prático e software educacional. Entretanto, 100% dos aplicativos móveis e 99,47% dos jogos não apresentam informação sobre a etapa de ensino, inviabilizando a compilação de um grande córpus de REA para o treinamento para um preditor de complexidade textual.

Vários trabalhos da literatura de predição de complexidade textual utilizam materiais de séries escolares ([Vajjala and Meurers 2014], ([Hartmann et al. 2016, Wagner Filho et al. 2016b]), considerando a série na qual o texto é usado como substituto (*proxy*) para a sua complexidade linguística; neste artigo também adotamos essa abordagem. Neste trabalho, primeiro a investigar a predição automática da complexidade de textos para REA, buscamos avaliar as *features* importantes para a tarefa, usando um arcabouço de análise multinível, assim como faz o ambiente Coh-Metrix [Graesser and McNamara 2011], envolvendo métricas que tratam de: (i) palavras, (ii) sintaxe, (iii) conexão entre sentenças no discurso. Essas *features* são discutidas na Seção 3.2. Como não há nenhum grande córpus disponível publicamente para predizer a complexidade textual para as etapas do ensino no Brasil, compilamos um córpus anotado com as quatro etapas (Seção 3.1).

As seções a seguir são organizadas da seguinte maneira: os trabalhos relacionados são apresentados na Seção 2; na Seção 3 são apresentados os detalhes do córpus, as *features* e os métodos de aprendizado de máquina e seleção de *features* utilizados no trabalho; e na Seção 4 os resultados da avaliação intrínseca e extrínseca, usando o melhor modelo que foi avaliado no grande córpus em dois conjuntos de REA de gêneros diferentes.

2. Trabalhos Relacionados

[Graesser and McNamara 2011, Graesser et al. 2011] desenvolveram a ferramenta Coh-Metrix[6] para língua inglesa que analisa textos usando métricas dos vários níveis da língua e que estão alinhadas com um arcabouço teórico de compreensão discursiva multinível. Em seu trabalho, utilizaram extratos de textos com média de 288,6 palavras para extraírem suas métricas, porém não informam a quantia mínima de palavras em cada extrato, enquanto nossa proposta traz uma quantia mínima de 300 palavras e uma média de 448 palavras, para viabilizar o cálculo de métricas como o Índice Flesch, por exemplo. Utilizam 53 métricas textuais, enquanto nossa proposta extrai 79 métricas dos extratos de textos. Usaram um grande córpus da língua inglesa com 37.520 extratos fornecido pelo *Touchstone Applied Science Associates* (TASA), já neste artigo tratamos a língua portuguesa para avaliação intrínseca de um preditor de complexidade textual, além da avaliação extrínseca na MEC-RED, considerando gêneros textuais semelhantes e diferentes do preditor criado.

[Scarton and Aluísio 2010] classificaram de forma binária os textos (simples versus complexos), usando o Coh-Metrix-Port com 40 métricas textuais. Usaram 4 córpus

[6] http://tea.cohmetrix.com/

para treinamento/teste: textos jornalísticos do jornal Zero Hora (ZH) dos anos de 2006 e 2007, textos reescritos para crianças da seção "Para o seu filho ler" (PSFL) do ZH e textos do gênero científico do Ciência Hoje (CH) e Ciência Hoje das Crianças (CHC). O trabalho apresenta resultados da classificação binária usando o SVM do Weka com precisão do melhor classificador treinado de 97%, porém sabemos que classificadores binários que medem uma grande distância de idade (crianças versus adultos) são mais simples do que classificadores com mais classes. Neste trabalho, usamos 4 classes e avaliarmos o preditor de forma intrínseca também extrínseca, usando córpus de gêneros de texto distintos e semelhantes ao preditor.

[Hartmann et al. 2016] reportam a classificação da complexidade de textos do gênero didático em português para cinco anos do Ensino Fundamental (3º, 4º, 5º, 6º e 7º anos). O córpus compilado é formado por textos de diversas fontes e possui 7.645 textos compilados de Livros Didáticos, NILC Corpus, Testes do SARESP, CHC, FSP, PSFL e Mundo Estranho; sem indicação da quantidade mínima de palavras que esses textos possuem. Apresentam uma classificação mais fina do que a tratada neste artigo, e utilizam 108 métricas para a criação do modelo preditivo. Já a nossa proposta utiliza 79 métricas e trata a divisão por etapas escolares. Utilizam apenas o classificador SVM implementado no libsvm que obteve 56% de acurácia.

[Wagner Filho et al. 2016b] reportam a predição automática do nível escolar da Wikilivros, considerando 3 níveis escolares (nível 1, 2 e 3). O córpus possui 77 textos e usaram 7 *features* para avaliação da inteligibilidade, com Regressão Logística[7]. Nossa proposta se diferencia, trazendo uma avaliação com *Logistic Regression, SVM, RandomForest e Multilayer Perception*, além de usar 79 métricas textuais. Também realizamos uma avaliação extrínseca nos dados da plataforma MEC-RED, dada a motivação inicial do trabalho de avaliar a qualidade de REA no Brasil. Também utilizamos para o treinamento do preditor a Wikilivros (Wikibook em português) e realizamos uma avaliação intrínseca usando *cross-validation*. [Wagner Filho et al. 2016a] trazem uma continuação do trabalho de [Wagner Filho et al. 2016b], considerando a língua portuguesa do Brasil e a língua inglesa para a criação do córpus de trabalho. Usaram 9.829 textos da língua portuguesa com níveis de inteligibilidade mistas de 2 e 3 níveis compostos por Wikilivros, É Só o Começo[8] (ESOC), PSFL, ZH, BrEscola[9]. Para a língua inglesa, usaram Wikibooks, *Simple* Wikipedia (SW) [Coster and Kauchak 2011] e Biografias Britânicas (BB); os níveis de inteligibilidade também são mistos, variando de 2,3 e 4 níveis. Usaram o Weka para geração dos modelos de classificação, usando os métodos SVM, Regressão Logística, DecisionStump, RandomForest, com *cross-validation*. A quantidade de *features* totaliza 134 para o idioma português e 89 para o inglês. Os melhores resultados foram com o SVM e a Regressão Logística. O trabalho [Wagner Filho et al. 2016a] avalia a generalização do modelo de classificação para inteligibilidade em diferentes níveis e em dois idiomas. Porém, os resultados são ruins quando utilizam 3 classes, sendo melhores para classes binárias e do mesmo nível de complexidade, por exemplo, textos apenas para crianças. Este trabalho diferencia do nosso, pois usamos 4 classes de complexidade textual e avaliamos gêneros textuais diferentes.

[7]Modelo SimpleLogistic da ferramenta Weka.
[8]Contrasta obras da literatura clássica brasileira com versões adaptadas.
[9]Córpus de materiais educativos para crianças e adolescentes.

Tabela 1. Córpus de livros-textos da Língua Portuguesa compilado: Marcha criança, Tudo É Linguagem, Projeto Porta Aberta, Projeto Ápis, Português, Buriti, Porta Aberta, Mundo Amigo, Nos Dias de Hoje, Projeto Teláris, CNEC Educação.

	Ensino Fund I	Ensino Fund II	Ensino Médio	Ensino Superior	Total	Dataset 1 (D1)	Dataset 2 (D2)
Fontes de textos	Livros -texto + PSFL	Livros-texto, SAEB, E-Book CNEC Educação	Wikilivros, ENEM 2015, 2016 e 2017	Wikilivros			
Docs	296	325	627	819	2.067	60	40
Sents	5.258	5.598	9.316	10.416	30.588	720	540
MTSP	20.58	24.31	29.81	39.15	31.35	31.17	46.27
Type	63.081	75.698	134.788	177.054	450.621	10870	9281
Token	101.911	127.705	241.267	342.534	813.417	20040	16216
TTR	0.618	0.592	0.558	0.516	0.553	0.542	0.572
D1	10	10	10	10	40		
D2	10	10	20	20	60		

3. Materiais e Métodos

3.1. Córpus dos Quatro Estágios Escolares do Sistema Educacional Brasileiro

Compilamos um grande córpus que abrange textos utilizados em diferentes etapas de ensino (ou níveis escolares) do Sistema Educacional Brasileiro, organizado nas seguintes etapas: Ensino Fundamental I (1º ao 5º ano), Ensino Fundamental II (6º ao 9º ano), Ensino Médio e Ensino superior. Essas quatro etapas de ensino são as mesmas utilizadas na MEC-RED para classificar os REA nos Estágios Escolares.

O córpus[10] inclui: livros-texto, notícias da Seção *Para Seu Filho Ler* (PSFL) do jornal Zero Hora que apresenta algumas notícias sobre os mesmos tópicos do Zero Hora, mas escritas para crianças de 8 a 11 anos de idade , Exames do SAEB , Livros Digitais do Wikilivros em Português , Exames do Enem dos anos 2015, 2016 e 2017. Nosso córpus de trabalho compreende 2.067 extratos (min = 300 palavras, max = 596 palavras, média = 448) dos textos do córpus compilado. Como pode ser visto na Tabela 1, nosso córpus não é balanceado, pois o número de textos do Ensino Médio possui aproximadamente o dobro da quantidade do Ensino Fundamental I e do Ensino Fundamental II, por exemplo. Para resolver esse problema, foi utilizado o método ClassBalancer do Weka[11] antes da execução dos métodos de aprendizado de máquina (cf. mais detalhes na Seção 4).

3.2. Métricas de Complexidade Textual

A seleção inicial das métricas para a avaliação da complexidade textual baseou-se no estudo de [Graesser and McNamara 2011] que utilizou 53 métricas do Coh-Metrix agrupadas nas relacionadas às palavras, sentenças e conexões entre sentenças. Para selecionar métricas similares para o português, escolhemos duas ferramentas disponíveis publicamente: Coh-Metrix-Port [Scarton et al. 2010], Coh-Metrix-Dementia [da Cunha 2015] e o trabalho [dos Santos et al. 2017]. A Figura 1 mostra um recorte[12] das 79 métricas

[10]Disponível em http://www.nilc.icmc.usp.br/nilc/index.php/tools-and-resources

[11]https://www.cs.waikato.ac.nz/ ml/weka/

[12]A lista completa está disponível em https://github.com/gazzola/corpus_readability_nlp_portuguese

Palavras	Sentenças	Conexões entre Sentenças
1.adjective_ratio: proporção de Adjetivos em relação à quantidade de palavras. **2.adverbs:** proporção de Advérbios em relação à quantidade de palavras. **3.pronoun_ratio:** proporção de pronomes em relação à quantidade de palavras. **4.first_person_pronouns:** proporção de pronomes pessoais nas primeiras pessoas em relação à quantidade de pronomes pessoais. **5.third_person_pronouns**: proporção de pronomes pessoais nas terceiras pessoas em relação à quantidade de pronomes pessoais do texto. **6.content_density**: proporção de palavras de conteúdo em relação à quantidade de palavras funcionais do texto. **7.conn_ratio**: proporção de conectivos em relação à quantidade de palavras do texto. **8.add_neg_conn_ratio**: proporção de conectivos aditivos negativos em relação à quantidade de palavras. **9.add_pos_conn_ratio**: proporção de conectivos aditivos positivos em relação à quantidade de palavras. **10.cau_pos_conn_ratio**: proporção de conectivos causais positivos em relação à quantidade de palavras. **11.concretude_mean:** média dos valores de concretude das palavras de conteúdo. **12.familiaridade_mean:** média dos valores de familiaridade das palavras de conteúdo. **13.imageabilidade_mean:** média dos valores de imageabilidade das palavras de conteúdo.	**1.words_per_sentence**: média de palavras por sentença. **2.sentence_length_min**: quantidade Mínima de palavras por sentença. **3.sentence_length_max**: quantidade Máxima de palavras por sentença. **4.sentence_length_standard_dev**: desvio Padrão da quantidade de palavras por sentença. **5.mean_noun_phrase**: média dos tamanhos médios dos sintagmas nominais nas sentenças. **6.min_noun_phrase**: mínimo entre os tamanhos de sintagmas nominais do texto. **7.max_noun_phrase**: máximo entre os tamanhos de sintagmas nominais do texto. **8.std_noun_phrase**: desvio-padrão do tamanho dos sintagmas nominais do texto. **9.words_before_main_verb**: quantidade Média de palavras antes dos verbos principais das orações principais das sentenças. **10.passive_ratio**: proporção de orações na voz passiva analítica em relação à quantidade de orações do texto. **11.yngve**: Complexidade Sintática de Yngve. **12.frazier**: Complexidade Sintática de Frazier. **13.dep_distance**: distância na árvore de dependências.	**1.adj_cw_ovl**: Quantidade média de palavras de conteúdo que se repetem nos pares de sentenças adjacentes. **2.adj_arg_ovl**: Quantidade média de referentes que se repetem nos pares de sentenças. **3.arg_ovl**: Quantidade média de referentes que se repetem nos pares de sentenças adjacentes. **4.adj_stem_ovl**: Quantidade média de radicais de palavras de conteúdo que se repetem nos pares de sentenças. **5.stem_ovl**: Quantidade média de radicais de palavras de conteúdo que se repetem nos pares de sentenças adjacentes. **6.ttr**: Proporção de types (despreza repetições de palavras) em relação à quantidade de tokens (computa repetições de palavras). **7.content_word_diversity**: Proporção de types de palavras de conteúdo em relação à quantidade de tokens de palavras de conteúdo. **8.verb_diversity**: proporção de types de verbos em relação à quantidade de tokens de verbos. **9.lsa_adj_mean:** similaridade média entre pares de sentenças adjacentes. **10.lsa_adj_std:** desvio padrão de similaridade entre pares de sentenças adjacentes. **11.lsa_all_mean:** média de similaridade entre todos os pares de frases. **12.lsa_all_std:** Desvio padrão de similaridade entre palavras **13.lsa_givennes_mean:** média de *givenness* de cada sentença do texto, a partir da segunda sentença

Figura 1. Recorte de 39 das 79 métricas usadas nesta pesquisa

incluídas; agrupadas naquelas relacionadas às palavras, sentenças e conexões entre sentenças. Entretanto, 17 métricas do estudo de [Graesser and McNamara 2011] não foram adaptadas para o português, seja por falta de recursos linguísticos ou ferramentas precisas de PLN. Elas são listadas aqui para futuras pesquisas: conectivos adversativos, *meaningfulness*, verbos causais, ações intencionais, eventos e partículas, similaridade sintática (sentenças no parágrafo), sobreposição de palavras de conteúdo em todas as sentenças, dissimilaridade de PoS entre sentenças e dissimilaridade de palavras entre sentenças, coesão causal, temporal e intencional, repetição de tempo e de aspecto verbal, log da frequência de palavras, sobreposição de verbo adjacente e sobreposição de verbo no modelo LSA em sentenças adjacentes. Para suprir essa falta, novas foram anexadas, como, por exemplo, Complexidade de Yngve e de Frasier, Distância de Dependência, dentre outras.

3.3. Métodos de Seleção de Features Avaliados

Foram avaliados o *o Correlation-based Feature Selection* (CFS) e o *Least Absolute Shrinkage and Selection Operator* (Lasso); o CFS resultou em 34 *features* (Tabela 2). Foram realizados experimentos de predição usando as 34 features selecionadas pelo CFS com os classificadores Logistic Regression (LR), *Random Forest* (RF), *Support Vector Machines* (SVM) e *Multilayer Perceptron* (MLP) (Tabela 3).

Tabela 2. Features selecionadas pelo método CFS

1	noun_ratio	12	verbs_ambiguity	23	idade_aquisicao_1_25_ratio
2	pronoun_ratio	13	yngve	24	idade_aquisicao_55_7_ratio
3	verbs	14	std_noun_phrase	25	idade_aquisicao_25_4_ratio
4	negation_ratio	15	passive_ratio	26	imageabilidade_std
5	min_cw_freq	16	concretude_25_4_ratio	29	imageabilidade_55_7_ratio
6	first_person_pronouns	17	concretude_4_55_ratio	30	sentence_length_std_deviation
7	conn_ratio	18	familiaridade_std	31	verb_diversity
8	tmp_neg_conn_ratio	19	familiaridade_4_55_ratio	32	adj_mean
9	tmp_pos_conn_ratio	20	familiaridade_55_7_ratio	33	span_mean
10	adjectives_ambiguity	21	idade_aquisicao_mean	34	content_density
11	adverbs_ambiguity	22	idade_aquisicao_std		

Tabela 3. Resultados da classificação com as features selecionadas pelo CFS

Classificador	Ensino Fundamental I	Ensino Fundamental II	Ensino Médio	Ensino Superior	F-Measure (*Weighted Avg.*)
SVM	85,30%	60,90%	74,80%	83,90%	0.777
MLP	80,30%	63,10%	73,30%	83,30%	0.767
Logistic Regression	86,00%	63,10%	75,10%	84,10%	0.783
RandomForest	**87,10%**	**67,60%**	**76,00%**	**85,00%**	**0.798**

O experimento com o Lasso teve como entrada os dados normalizados e o parâmetro alpha ajustado com 0.2. O método selecionou 8 *features* consideradas mais representativas para o conjunto de dados: proporção de pronomes, pronomes de primeira pessoa, proporção de palavras de conteúdo do texto com familiaridade entre 4 e 5.5, desvio padrão da imageabilidade, proporção de palavras de conteúdo do texto com imageabilidade entre 4 e 5.5, proporção de palavras de conteúdo do texto com imageabilidade entre 5.5 e 7, desvio padrão do comprimento da sentença e densidade de conteúdo. Nos experimentos, os melhores resultados foram com o classificador RandomForest, considerando a seleção das *features* do Lasso, com média ponderada de *F-Measure* de 69.6.

4. Avaliação Intrínseca e Extrínseca

4.1. Avaliação da Complexidade Textual no Córpus de Textos Didáticos

Para avaliar a tarefa de classificação da complexidade textual em nosso córpus, anotado com quatro etapas de ensino e com 79 métricas, quatro classificadores foram escolhidos, com base nos trabalhos relacionados, que foram revisados na Seção 2. Os algoritmos selecionados do Weka foram: SVM, MLP, LR e RF. Para a avaliação do melhor modelo, a validação cruzada foi usada com valor *10-folds*. Os melhores resultados foram do SVM, que alcançou uma média ponderada de F-Measure de **0.804**; o resultado do RF foi 0.794, da MLP foi de 0.698 e da LR foi de 0.802, caracterizando um empate técnico com o SVM. Comparando as predições dos modelos com seleção de via Lasso e CFS e treinado com as 79 *features*, foi possível observar que o desempenho dos modelos com seleção de features é inferior ao do modelo com todas as features. Os resultados por nível escolar do classificador SVM com todas as *features* pode ser visto na Tabela 4. Como havia desbalanceamento da classe nível escolar, usamos o ClassBalancer [Jain et al. 2018]. Esse método reutiliza instâncias para que a soma total de pesos em todas as instâncias seja equilibrada. Desta forma, ficamos com 516,8 instâncias em cada classe.

Tabela 4. Resultados da Classificação do SVM com todas as features

	Precisão	Precisão c/ Balanceamento	Recall	Recall c/ Balanceamento	F-Measure	F-Measure c/ Balanceamento
Ensino Fund I	81.60%	85.0%	91.2%	92.6%	0.861	**0.886**
Ensino Fund II	69.8%	75.0%	65.50%	75.7%	0.676	**0.754**
Ensino Médio	80.10%	80.5%	71.80%	69.1%	**0.757**	0.743
Ensino Superior	83.40%	81.6%	88.50%	85.2%	**0.859**	0.834

4.2. Avaliação da Complexidade Textual com REA da MEC-RED

Para avaliar a utilidade e robustez do melhor modelo treinado no grande córpus descrito na Seção 3.1 em predizer a complexidade textual de REA, realizamos uma avaliação extrínseca em dois conjuntos de REA: com gêneros diferentes do modelo treinado e com mesmo gênero textual (cf. Seção 3). Utilizamos o melhor modelo de classificação que foi o SVM com 79 *features*. O *dataset 1* é composto por 60 REA de experimentos práticos e animações/simulações. Na avaliação de robustez, a média poderada da *F-Measure* foi de 0.389. O *dataset 2* é composto por 40 textos do gênero textual didático; na avaliação de robustez, a média ponderada da *F-Measure* foi de 0.518. Em uma análise detalhada dos textos disponíveis, verificamos que eles possuem muitos erros ortográficos e a anotação da Etapa Escolar estipulada pelos autores dos materiais que disponibilizaram no MEC-RED parecia equivocada para alguns REA. Sendo este o primeiro trabalho a avaliar a Etapa de Ensino de REA, antevemos novas pesquisas para validar a predição automática. Por exemplo, fazer uma correção gramatical nos textos e utilizar REA para os quais os metadados *pessoas que favoritam o recurso*, e *escala de estrelas sejam usados* nos indique que os recursos são usados nas escolas. Há também algumas melhorias para a tarefa como inclusão de novas métricas de complexidade, como as 17 métricas do trabalho de [Graesser and McNamara 2011], citadas na Seção 3.2, que não foram incluídas neste estudo atual.

5. Conclusões e Trabalhos Futuros

Em resumo, este artigo explorou métodos automáticos para predizer o metadado etapa de ensino da plataforma MEC-RED, embora este trabalho possa ser utilizado para outras plataformas. Foi criado um grande córpus para modelar a tarefa de complexidade textual e assim avaliar o modelo com textos de gêneros textuais didáticos e outros como animação/simulação e experimento prático. A avaliação intrínseca mostrou um ótimo desempenho para o modelo treinado (F-measure de 0.804). Foi feita uma seleção de *features* usando 2 métodos de redução de dimensionalidade que foram comparados com o modelo treinado com todas as 79 *features*, além de usarmos um método de balanceamento de classes. Por fim, foi possível observar o impacto dos gêneros textuais na complexidade textual para predizer a Etapa Escolar. A avaliação extrínseca usando recursos da MEC-RED mostrou que a tarefa é difícil e merece ser melhor explorada. Os trabalhos futuros consistem em (i) estudar métricas linguísticas que distinguem os tipos de recursos animação/simulação, aplicativo móvel, jogos, experimento prático e software educacional, que são de especial interesse para essa pesquisa, para explorar novas *features* para os modelos e (ii) explorar novos métodos para a tarefa de complexidade textual como as arquiteturas neurais avaliadas em [Nadeem and Ostendorf 2018] para tentar mitigar os problemas de desempenho dos modelos treinados com engenharia de *features*.

Referências

[Ahmed and Fuge 2017] Ahmed, F. and Fuge, M. (2017). Capturing winning ideas in online design communities. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, CSCW '17, pages 1675–1687, New York, NY, USA. ACM.

[Bethard et al. 2009] Bethard, S., Wetzer, P., Butcher, K., Martin, J. H., and Sumner, T. (2009). Automatically characterizing resource quality for educational digital libraries. In *Proceedings of the 9th ACM/IEEE-CS joint conference on Digital libraries*, pages 221–230. ACM.

[Cechinel et al. 2011] Cechinel, C., Sanchez-Alonso, S., and Garcia-Barriocanal, E. (2011). Statistical profiles of highly-rated learning objects. *Comput. Educ.*, 57(1):1255–1269.

[Coster and Kauchak 2011] Coster, W. and Kauchak, D. (2011). Simple english wikipedia: a new text simplification task. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, pages 665–669. Association for Computational Linguistics.

[da Cunha 2015] da Cunha, A. L. V. (2015). Coh-metrix-dementia: análise automática de distúrbios de linguagem nas demências utilizando processamento de línguas naturais. Master's thesis, Universidade de São Paulo, ICMC - USP São Carlos.

[Dalip et al. 2011] Dalip, D. H., Gonçalves, M. A., Cristo, M., and Calado, P. (2011). Automatic assessment of document quality in web collaborative digital libraries. *Journal of Data and Information Quality (JDIQ)*, 2(3):14.

[dos Santos et al. 2017] dos Santos, L. B., Duran, M. S., Hartmann, N. S., Jr., A. C., Paetzold, G. H., and Aluísio, S. M. (2017). A lightweight regression method to infer psycholinguistic properties for brazilian portuguese. *CoRR*, abs/1705.07008.

[Fang 2016] Fang, Z. (2016). Text complexity in the us common core state standards: A linguistic critique. *Australian Journal of Language and Literacy*, 39(3):195–206.

[Graesser and McNamara 2011] Graesser, A. C. and McNamara, D. S. (2011). Computational analyses of multilevel discourse comprehension. *Topics in cognitive science*, 3(2):371–398.

[Graesser et al. 2011] Graesser, A. C., McNamara, D. S., and Kulikowich, J. M. (2011). Coh-metrix: Providing multilevel analyses of text characteristics. *Educational researcher*, 40(5):223–234.

[Hartmann et al. 2016] Hartmann, N., Cucatto, L., Brants, D., and Aluísio, S. (2016). Automatic classification of the complexity of nonfiction texts in portuguese for early school years. In *International Conference on Computational Processing of the Portuguese Language*, pages 12–24. Springer.

[Jain et al. 2018] Jain, S., Kotsampasakou, E., and Ecker, G. F. (2018). Comparing the performance of meta-classifiers—a case study on selected imbalanced data sets relevant for prediction of liver toxicity. *Journal of computer-aided molecular design*, pages 1–8.

[Leary et al. 2011] Leary, H., Recker, M., Walker, A., Wetzler, P., Sumner, T., and Martin, J. (2011). Automating open educational resources assessments: a machine learning

generalization study. In *Proceedings of the 11th annual international ACM/IEEE joint conference on Digital libraries*, pages 283–286. ACM.

[MEC 2019a] MEC (2019a). Sobre a plataforma MEC-RED. https://plataformaintegrada.mec.gov.br/sobre. Acessado: 2019-01-24.

[MEC 2019b] MEC (2019b). Termos de serviços - plataforma mec-red. https://plataformaintegrada.mec.gov.br/termos-de-uso. Acessado em: 2019-01-24.

[Miao et al. 2016] Miao, F., Mishra, S., and McGreal, R. (2016). *Open educational resources: policy, costs, transformation*. UNESCO Publishing.

[Nadeem and Ostendorf 2018] Nadeem, F. and Ostendorf, M. (2018). Estimating linguistic complexity for science texts. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 45–55, New Orleans, Louisiana. Association for Computational Linguistics.

[Scarton et al. 2010] Scarton, C., Gasperin, C., and Aluísio, S. (2010). Revisiting the readability assessment of texts in portuguese. *Advances in Artificial Intelligence – IBERAMIA - Volume 6433 of Lecture Notes in Computer Science*, pages 306–315.

[Scarton and Aluísio 2010] Scarton, C. E. and Aluísio, S. M. (2010). Análise da inteligibilidade de textos via ferramentas de processamento de língua natural: adaptando as métricas do coh-metrix para o português. *Linguamática*, 2(1):45–61.

[UNESCO. 2002] UNESCO. (2002). Forum on the impact of open courseware for higher education in developing countries:: final report.

[Vajjala and Meurers 2014] Vajjala, S. and Meurers, D. (2014). Assessing the relative reading level of sentence pairs for text simplification. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 288–297.

[Wagner Filho et al. 2016a] Wagner Filho, J. A., Wilkens, R., and Villavicencio, A. (2016a). Automatic construction of large readability corpora. In *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC)*, pages 164–173.

[Wagner Filho et al. 2016b] Wagner Filho, J. A., Wilkens, R., Zilio, L., Idiart, M., and Villavicencio, A. (2016b). Crawling by readability level. In *International Conference on Computational Processing of the Portuguese Language*, pages 306–318. Springer.

[Wiley et al. 2014] Wiley, D., Bliss, T., and McEwen, M. (2014). Open educational resources: A review of the literature. pages 781–789.