Proposta de um Sistema de Avaliação Automática de Redações do ENEM Utilizando Técnicas de Aprendizagem de Máquina e Processamento de Linguagem Natural

Celso R. C. A. Júnior[1], **Marcos A. Spalenza**[1] **Elias de Oliveira**[1]

[1] Programa de Pós-Graduação em Informática - Universidade Federal do Espírito Santo (UFES) - 29.075-910 - Vitória - ES - Brasil

celsoromao@gmail.com, marcos.spalenza@gmail.com, elias@lcad.inf.ufes.br

Abstract. *The ENEM essay evaluation process involves around 10,000 evaluators, correcting an average of 70 essays per day. The number of corrected essays rose by 3.15 million from 2009 to 6.54 million in 2015, as the cost of each evaluation increased from R$ 3.15 to R$ 15.88. This work proposes the use of machine learning techniques and natural language processing for the construction of a system of automatic evaluation of ENEM essays, in Competence 1, mastering of written language standard. The results show that the system is able to imitate grades given by the evaluators for a dataset of the UOL site with a mean absolute error of 0.25.*
Keywords: Natural Language Processing, Machine Learning, Essay, ENEM.

Resumo. *O processo de avaliação de redações do ENEM envolve cerca de 10.000 avaliadores, corrigindo uma média de 70 redações por dia. O número de redações corrigidas subiu de 3,15 milhões em 2009 para 6,54 milhões em 2015, assim como o custo de cada avaliação subiu de R$ 3,15 para R$ 15,88. Esse trabalho propõe o uso de técnicas de aprendizagem de máquina e processamento de linguagem natural para a construção de um sistema de avaliação automática de redações do ENEM, na Competência 1, domínio da norma culta da língua escrita. Os resultados nos mostram que o sistema é capaz de imitar a avaliação de avaliadores do banco de redação do site UOL com um erro médio absoluto de 0.25.*
Palavras-chave: Processamento de Linguagem Natural, Aprendizagem de Máquina, Redações, ENEM.

1. Introdução

A redação é a única prova do ENEM que não é submetida ao método Teoria de Resposta ao Item (TRI) [Pasquali 2004], cujo valor de cada questão varia de acordo com o percentual de acertos e erros de cada estudante. Escrever bem, seguindo as normas cultas da língua portuguesa é essencial para quem deseja ingressar em uma Instituição de Ensino Superior. Nesse cenário, escolas têm dedicado uma atenção especial na preparação dos alunos para a prova de redação. No entanto sua correção e avaliação, no modelo do ENEM, é um processo custoso para quem avalia, que além da correção e avaliação precisa indicar os erros e dar um *feedback* para o aluno. No processo de avaliação do ENEM, cada avaliador atribui uma nota de 0 a 200 em cada uma das 5 competências abaixo:

1. Domínio da norma padrão da língua portuguesa;

2. Compreensão da proposta de redação;
3. Seleção e organização das informações;
4. Demonstração de conhecimento da língua necessária para argumentação do texto
5. Elaboração de uma proposta de solução para os problemas abordados, respeitando os valores e considerando as diversidades socioculturais.

Na Competência 1, foco desse trabalho, são observados os desvios gramaticais como sintaxe de concordância, regência e colocação, pontuação, flexão, entre outros. Em relação as convenções da escrita, espera-se do participante o domínio e o respeito às particularidades da modalidade escrita. São avaliados também a ortografia, acentuação e o uso adequado de letras maiúsculas e minúsculas. Características comuns ao *internetês* e uso de gírias são consideradas desvio da norma culta. Para a atribuição da nota na Competência 1 , cada avaliador atribui uma nota de acordo com os níveis de conhecimento da Tabela 1, estabelecidos pelo anexo IV do Edital ENEM 2016 [1].

Nível	Proficiência	Pontuação
0	Muito baixa ou ausente	0%
1	Baixa	20%
2	Mediana	40%
3	Boa	60%
4	Muito boa	80%
5	Excelente	100%

Tabela 1. Níveis de conceitos na Competência 1.

A redação é uma das avaliações do Enem que mais deixam os estudantes temerosos, pois muitos ainda possuem dúvidas sobre a elaboração dessa, qual seu objetivo, como se estrutura seguindo as normas cultas da língua portuguesa. Como consequência disto, surgem plataformas privadas para correção e avaliação de redações no modelo ENEM como Redação *online* [2], Mais Correções [3] e *Imaginie* [4], que disponibilizam pacotes mensais para estudantes submeterem suas redações. O *site* UOL, permite que gratuitamente seus usuários submetam suas redações de um tema proposto para avaliação. Essas redações recebem uma nota entre 0.0 e 2.0, com passos de 0.5, em cada uma das 5 competências. Depois, o *site* disponibiliza algumas das redações corrigidas, as notas e os comentários dos avaliadores. Em cima desse conjunto de dados [5] elaboramos nossas estratégias para construção do sistema.

O objetivo desse trabalho é apresentar uma estratégia para redução do esforço no processo de correção e avaliação, auxiliando os avaliadores. Para isto propomos a construção de um sistema de avaliação automática de redações do ENEM, na Competência 1, domínio da norma culta da língua escrita. Técnicas de Processamento de Linguagem Natural e Inteligência Artificial foram estudas e implementadas.

Esse trabalho está estruturado em 5 seções. Na Seção 2 estão listados trabalhos relacionados com avaliação automáticas de redações. A Seção 3 apresenta a metodologia usada no desenvolvimento desse trabalho. Na Seção 4 são apresentados os resultados dos

[1] http://download.inep.gov.br/educacaobasica/enem/edital/2016/editalenem2016.pdf
[2] http://www.redacaonline.com.br/redacao.php
[3] https://maiscorrecoes.com.br/
[4] http://www.imaginie.com/
[5] http://vestibular.brasilescola.uol.com.br/banco-de-redacoes/

experimentos realizados com o sistema. A Seção 5 apresenta as considerações finais e trabalhos futuros.

2. Fundamentação Teórica

Para diminuir a sobrecarga de trabalho dos professores e proporcionar maior agilidade no ensino da língua escrita, pesquisadores tem procurado desenvolver aplicações para correção e avaliação automática de atividades escritas. Com um trabalho que iniciou a avaliação de redações, [Page and Paulus 1968] propôs um sistema para que baseia-se na criação de modelos estatísticos segundo uma base de dados de redações, capturando o estilo de escrita associado à cada nota. Porém, ao observar o formato de escrita de cada aluno, desconsiderava seu conteúdo. Hoje, validar a qualidade de escrita dos estudantes não é só mais um requisito na avaliação do ensino, sendo necessário o acompanhamento desde o início da escolaridade. Segundo [Mota et al. 2000], o processo de formação de uma grafia competente passa por inúmeros estágios desde a infância, sendo aprimorada até que o estudante domine a estrutura léxico-sintática da norma culta.

A partir dos anos 90 mais ferramentas lidaram com a avaliação de redações, iniciando a análise da verificação dos argumentos, como o *e-rater* [Burstein et al. 1998]. Partindo do pressuposto que é essencial para uma boa escrita o domínio da grafia correta, muitos trabalhos nessa área seguiram com estudos na linguística computacional. O avanço do *Processamento de Linguagem Natural - PLN* teve como consequência o início da identificação da sintaxe atribuída à linguagem humana pelo computador. Integrando tal metodologia nos trabalhos [Burstein et al. 2003] apresentou a ferramenta *Criterion*, atentando para o contexto de aplicação das palavras nos textos além da análise morfossintática.

Com a evolução das técnicas de PLN aplicadas à correção de textos, a análise textual passou à ser revisada conforme diferentes competências linguísticas. Como [Burrows et al. 2015] divide-as, a escrita pode ser interpretada de forma léxica, morfológica, semântica, sintática e estrutural. As verificações léxicas observam a distribuição das palavras no texto e sua grafia. De forma morfológica, podemos definir equivalências intrínsecas à construção das palavras e suas derivações. Com verificação semântica a interpretação do significado ou valor atribuído à palavra e sua associação. Analisando a sintaxe, encontramos a dependência entre as palavras e a função à ela atribuída na frase. E no formato estrutural, são definidos os padrões de escrita conforme pontuação, tamanho das sentenças, número de caracteres, dando embasamento estatístico para os sistemas.

O estudo da grafia correta e suas divisões ganhou foco na criação de ferramentas de tradução e dos bancos de dados de redações provenientes dos testes de proficiência, como o GMAT (*Graduate Management Admissions Test*) e o TWE *Test of Written English*. Nesses casos, assim como durante a formação do estudante, há grande necessidade de validar a escrita conforme os padrões linguísticos pois, *à priori*, esse é considerado um desconhecedor da sua estrutura. A avaliação de redações para estudantes de língua estrangeira, como a utilizada pelo *e-rater* [Burstein and Chodorow 1999], serviu de referencial para o trabalho de [Wang et al. 2008] para textos em chinês, [Ishioka and Kameda 2004] no desenvolvimento do *Jess* para o japonês e a ferramenta produzida por [Pérez et al. 2004] para o espanhol. Essa última, foi atrelada à um sistema de traduções com grande eficácia ao encontrar erros no uso das palavras, o *Bleu*

[Papineni et al. 2002], com reconhecimento de termos sinônimos e desordenados.

Para o português, [Santos et al. 2015] propôs um analisador ortográfico-gramatical para avaliação da escrita. Em sua abordagem, o autor utilizou Algoritmos Genéticos (GA) e técnicas de PLN. O módulo que atua na verificação ortográfica utiliza distância de Levenshtein durante a otimização de sugestões de correção geradas pelo GA. Enquanto isso, o módulo responsável pela verificação gramatical aplica técnicas de PLN com o uso da ferramenta OpenNLP[6] para marcação de classes de cada palavra e o CoGrOO[7] na identificação de erros. Outra abordagem foi apresentada por [Bazelato and Amorim 2013], onde o autor utilizou a técnica de aprendizado de máquina de probabilidades independentes *Naïve Bayes*. As redações foram avaliadas de 0 à 10 com passos de 0.5 pontos. Como base de dados esse trabalho também aplicou a técnica no banco de redações do site UOL, onde obteve acurácia de 52% considerando notas com 1.5 pontos de distância da atribuída pelo especialista.

Trabalho similar foi desenvolvido por [Júnior and Oliveira 2016], onde foi proposto um sistema para avaliação automática de redações do ENEM, na Competência 1, utilizando técnicas de aprendizagem de máquina e PLN. Nessa proposta, o sistema utiliza o revisor gramatical ReGra [Pinheiro 2007] para identificação dos erros gramaticais. Com 954 redações do site UOL, obteve 0.313 pontos de erro médio absoluto (MAE), 35% de *precision* e 47% de *recall*.

3. Metodologia

Essa seção apresenta as etapas realizadas e as técnicas utilizadas nesse trabalho para avaliação e predição de notas de redações, na Competência 1 do ENEM.

3.1. Pré-processamento das Redações

Na mineração de dados, o procedimento padrão é, primeiro submeter o texto a uma fase de pré-processamento. Nessa fase elimina-se o conjunto de termos que eventualmente não trazem significado ao texto, pelo contrário pode causar alguma perturbação de ruído para a classificação automática. No nosso caso, retiramos das redações, números, datas, *tags html*, entre outros. Depois utilizamos o *Apache OpenNLP, framework* para atribuir marcas (*tags*) morfológicas e de inflexão a cada *token* da redação.

3.2. Extração de Características

Representamos cada redação por um vetor com as seguintes características: quantidades de parágrafos, frases, palavras, caracteres, erros ortográficos, 124 erros gramaticais identificados pelo CoGrOO [Silva 2014], vírgulas, pontos, pontos de interrogação e exclamação. Também definimos como características as 18 principais classes gramaticais da língua portuguesa como substantivos, verbos, adjetivos, entre outros.

Para identificar os erros ortográficos o sistema utiliza *Hunspell*, corretor ortográfico e analisador morfológico, além do algoritmo que implementamos baseando-se na probabilidade *bayesiana* para analisar o contexto da palavra, isto é, a sua colocação na frase. Uma importante parte dos corretores ortográficos são os dicionários utilizados por

[6]https://opennlp.apache.org/

[7]http://ccsl.ime.usp.br/cogroo/

eles. Nesse trabalho, além de complementar o dicionário utilizado pelo *Hunspell*, com palavras extraídas dos textos tema do *site*, utilizamos um dicionário auxiliar composto por palavras retiradas de textos jornalísticos de 10 anos do jornal A Tribuna. Esses textos jornalísticos também foram utilizados na composição de nossos dicionários de *bigrams* e *trigrams*, ou seja, contabilizando a frequência das palavras adjacentes com dois ou três termos.

Na identificação dos erros gramaticais optamos por utilizar o corretor gramatical CoGrOO, um corretor de código aberto, capaz de identificar erros como: colocação pronominal, concordância nominal, concordância sujeito-verbo, uso de crase, concordância nominal e verbal e outros erros comuns de escrita em Português do Brasil.

3.3. Seleção de Características

Como mencionado em [Baeza-Yates and Ribeiro-Neto 2011], um grande espaço de recursos (ou termos) pode tornar o problema de classificação de documentos impraticável, devido ao tempo do processo. A solução clássica para esse problema é reduzir o tamanho do espaço de recursos selecionando um subconjunto de todos os termos para a representação de documentos. Chamamos essa etapa de seleção de características.

O algoritmo evolucionário *Particle Swarm Optimization (PSO)* ajusta as trajetórias de uma população de "partículas" através de um espaço de problema com base em informações sobre o melhor desempenho anterior de cada partícula e de seus vizinhos. As versões anteriores do enxame de partículas operaram em espaço contínuo, onde as trajetórias são definidas como mudanças de posição em algumas dimensões[Kennedy and Eberhart 1997].

Usamos o algoritmo evolucionário *Particle Swarm Optimization (PSO)* para reduzir a dimensionalidade. Através da utilização desse algoritmo selecionamos as características que mais influenciam na nota da norma culta. Nossa função de qualidade selecionou a configuração de características que obteve a melhor separação entre as classes durante o treinamento.

3.4. Classificação

Para as etapas de treinamento e testes utilizamos o método de validação cruzada estratificada. Normalmente os conjuntos são estratificado para garantir que cada parte seja uma boa representação do conjunto original de dados. Dividimos os dados em 10 partes (*folds*) de tamanho iguais, com amostras mutuamente exclusivas. Depois, 10 iterações de treinamento são realizados, tal que em cada iteração uma diferente parte é usada para teste e as outras 9 para treinamento.

Conforme já mencionamos, uma redação pode ser pontuada com uma nota entre 0 e 2 com passos de 0.5. Sendo essa uma característica dos problemas de natureza multiclasses, utilizamos a estratégia um contra todos, que consiste na montagem de um classificador por classe. Onde para cada classificador, a classe está equipada contra todas as outras classes. Escolhemos o *Support Vector Machine (SVM)* como classificador devido a sua boa capacidade de generalização, robustez em grandes dimensões e por possuir uma teoria bem definida[Cristianini and Shawe-Taylor 2000].

3.5. Métricas

A avaliação é um importante passo para o desenvolvimento de qualquer método de classificação. Sem esse não há forma de determinar o quão bom é o novo método proposto [Baeza-Yates and Ribeiro-Neto 2011]. Nesse artigo foram utilizadas para avaliação as métricas: *precision*, *recall* e erro médio absoluto.

Precision, representada na Equação 1, para classificação binária, mede a proporção de amostras classificadas como positivas que realmente são positivas. *Recall*, representada na Equação 2, mede a proporção de amostras positivas que foram classificadas como positivas.

$$Precision(C_p) = \frac{TP(C_p)}{TP(C_p) + FP(C_p)} \tag{1}$$

$$Recall(C_p) = \frac{TP(C_p)}{TP(C_p) + FN(C_p)} \tag{2}$$

Para o nosso problema p corresponde a uma das $classes = \{0.0, 0.5, 1.0, 1.5, 2.0\}$ que corresponde as notas na Competência 1 do ENEM. *True Positive*$(TP(Cp))$ é a quantidade de redações atribuídas corretamente a classe Cp pelo classificador. *False Positive*$(FP(Cp))$ é a quantidade de redações atribuídas incorretamente a classe Cp pelo classificador. *False Negative* $(FN(Cp))$ é a quantidade de redações que pertencem a classe Cp, mas que foram classificadas incorretamente em outra classe.

Em estatística, o Erro Médio Absoluto - MAE, dado pela Equação 3, é a média das diferenças entre os valores reais e preditos. Onde n é número de amostras, y_i é a classe real e y_i^t é classe predita pelo classificador

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |y_i - y_i^t| \tag{3}$$

[Page 1994] destacou que notas atribuídas pelos avaliadores humanos são geralmente diferentes. Em seus experimentos, constatou que a correlação de Pearson entre dois humanos avaliadores foi entorno de 0.564. [Bazelato and Amorim 2013] afirma que é razoável avaliar não exatamente o grau atribuído automaticamente, mas também notas adjacentes. Embasado nessas afirmações e no relato de avaliadores do ENEM, entrevistado por [Luna 2009], da dificuldade de diferenciar os níveis de uma competência, decidimos avaliar também as notas adjacentes, classificadas como 0.5,1.0 e 1.5 pontos longe da nota atribuída pelo avaliador do *site* UOL, considerada aqui como nota verdadeira.

4. Experimentos e Resultados

Nosso conjunto de dados possui 4547 redações do *site* UOL. Cada redação pertence a uma classe, que é dada pela nota atribuída pelo avaliador do *site* na Competência 1 do ENEM. A Tabela 2 mostra a distribuição de redações por nota.

Durante a etapa de pré-processamento incluímos as *POS-tags* em cada redação, para categorização de cada termo do documento, como no trecho abaixo.

Nota	Quantidade
0.0	76
0.5	479
1.0	2181
1.5	1556
2.0	255

Tabela 2. Distribuição das classes (notas) da base de dados.

> ”A_art legislação_n brasileira_adj atribui_v-fin a_prp maioridade_n a_prp pessoas_n apartir_v-inf ._punc Nos_pron-det Estados_n Unidos_prop e_conj-c na_n Inglaterra_prop não_adv existe_v-fin idade_n miníma_adj”

Durante a identificação de erros ortográficos, encontramos palavras sintaticamente corretas, mas fora do contexto da frase. Palavras como agencia e transito foram utilizadas no lugar de agência e trânsito. Essas não foram identificadas como erros pelo sistema.

Para solucionar o problema implementamos um algoritmo para analisar o contexto da palavra na frase. Esse algoritmo gera uma lista de palavras modificadas, incluindo a palavra em análise. Para cada uma, juntamente com as adjacentes, consulta as frequências nos dicionários de *bigrams* e *trigrams*. A palavra com maior probabilidade é comparada com a palavra em análise.

4.0.1. Ajustando o *Threshold*

Conforme dito anteriormente, podemos configurar o sistema para considerar notas adjacentes, isso seria feito relaxando nosso limite (*threshold*) do que queremos que nosso sistema considere como nota correta. Esse limite pode ser movido tanto para cima, como também para baixo, em cada posição do limiar, obteríamos um valor de *precision* diferente, um valor de *recall* e um erro absoluto médio. A Tabela 3 mostra os resultados iniciais com nosso conjunto de dados para experimentações, em que K indica o quanto variamos esse limite para longe da nota tomada como a correta.

K	Precision	Recall	MAE
0.0	41%	45%	0.3
0.5	85%	89%	0.116
1.0	99%	99%	0.048
1.5	100%	100%	0.0

Tabela 3. Resultados com a base de dados inicial nos limiares de 0.0 a 1.5

A fim de melhorar a separação entre as classes decidimos usar uma estratégia para selecionar e ponderar as características. Utilizamos uma função qualidade que aumenta a densidade de cada classe individual, proporcionado uma melhor separação entre as classes. A Equação 4 proposta por [Saude et al. 2014] foi adaptada e utilizada como função de qualidade do Algoritmo *PSO*.

$$Density(C_p) = \frac{\sum_{i=1}^{|N_{C_p}|} euclidean(d_i, cent_{C_p})}{|N_{C_p}|} \tag{4}$$

Na Equação 4 C_p é a classe cuja as características serão selecionadas, $|N_{C_p}|$ é o número de redações da classe C_p, $euclidean(d_i, cent_{C_p})$ é a distância euclidiana entre

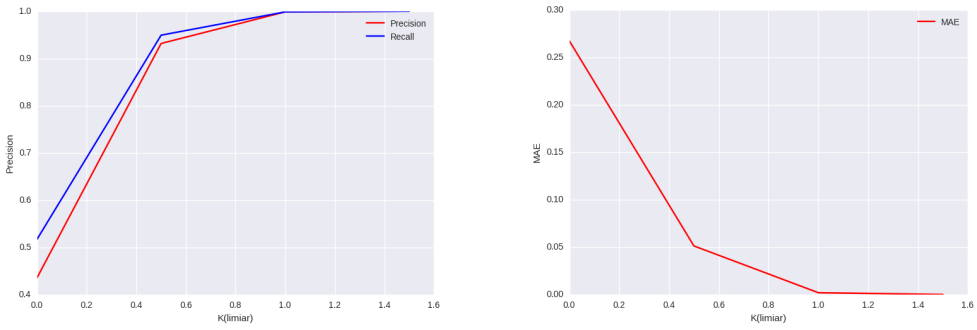
uma redação e seu centroide e $cent_{c_p}$ é o centroide da classe C_p. A função de qualidade seleciona as melhores características de cada classe que maximize o resultado da Equação 4. Nossa estratégia selecionou as características com maior poder discriminatório entre as classes. No total foram selecionadas 13 características, entre elas: erros gramaticais, repetição de palavras ou símbolos, regência verbal, acentuação gráfica e máximo de palavras por frase. Com isto aumentamos a separação entre as classes, melhoramos os resultados como mostra Tabela 4.

Nossa estratégia selecionou as características com maior poder discriminatório entre as classes. No total foram selecionadas 13 características, entre elas: erros gramaticais, repetição de palavras ou símbolos, regência verbal, acentuação gráfica e máximo de palavras por frase. Com isto aumentamos a densidade intra-classe e reduzimos a extra-classe, melhorando os resultados, como mostra Tabela 4.

K	Precision	Recall	MAE
0.0	44%	52%	0.26
0.5	93%	95%	0.05
1.0	100%	100%	0.0
1.5	100%	100%	0.0

Tabela 4. Resultados por classe após a seleção de características.

Conforme citamos, [Bazelato and Amorim 2013] avaliou a redação como um todo, atribuindo notas que variam de 0 a 10 continuamente, conseguindo 52% de acerto (*accuracy*) considerando notas adjacentes aquelas a 1.5 pontos de distância da nota do avaliador humano, 15% da nota total. Nosso sistema avaliou somente a Competência 1, conseguindo 52% de acertos e 93% considerando notas adjacentes a 0.5 pontos de distância da nota do avaliador na nota total da Competência 1. Quanto ao Erro Médio Absoluto obtivemos 0.26 e 0.05 considerado notas adjacentes a 0.5 de distância da nota do avaliador humano. Conforme mostram as Figuras 1(a) e 1(b).



(a) Precision e Recall (b) Erro Médio Absoluto

Figura 1. Resultados finais dos experimentos.

5. Considerações Finais e Trabalhos Futuros

Podemos observar muitos vícios e erros da língua portuguesa nos meios comuns de escrita como as redes sociais, plataformas educacionais, serviços de mensagens, portais de notícia e entretenimento. A avaliação do domínio linguístico, durante o ensino, é um trabalho recorrente do professor na correção de atividades. Sua importância é ainda maior

em redações, inclusive pela sua relevância para o ENEM, assim como para a formação desse profissional. Desse modo, os sistemas de verificação da Competência 1, além de realizarem a análise das redações dos estudantes, podem servir como ferramentas para alertar os erros de escrita nos meio de comunicação.

Com resultados de 93% de *precision*, considerando apenas um nível de nota para avaliadores humanos, o sistema têm possibilidades de aplicação diária, dada a grande necessidade de correção de erros gramaticais. Contudo, além dos ganhos em sala de aula, podemos listar como relevantes a redução de custo, gasto de tempo e o apoio aos estudantes na melhoria da escrita. Esses resultados indicam uma economia imediata de cerca de 20% no custo total de realização do ENEM, ou seja, a redução de mais de 90 milhões de reais.

Como trabalhos futuros planejamos: a expansão dos dicionários do sistema, a melhoria da análise gramatical por contexto e o uso de outros algoritmos sofisticados na predição das notas como regressão linear e redes neurais. Com tais alterações esperamos a aproximação dos resultados do sistema aos critérios do especialista para melhores estudos de adequação do sistema ao avaliador humano.

Referências

Baeza-Yates, R. and Ribeiro-Neto, B. (2011). *Modern Information Retrieval: The Concepts and Technology Behind Search*. Addison Wesley.

Bazelato, B. S. and Amorim, E. C. F. (2013). A Bayesian Classifier to Automatic Correction of Portuguese Essays. In *Conferência Internacional sobre Informática na Educação (TISE, 2013)*, volume 18, pages 779–782, Porto Alegre, RS, Brasil.

Burrows, S., Gurevych, I., and Stein, B. (2015). The Eras and Trends of Automatic Short Answer Grading. *International Journal of Artificial Intelligence in Education*, 25(1):60–117.

Burstein, J. and Chodorow, M. (1999). Automated essay scoring for nonnative english speakers. In *Proceedings of a Symposium on Computer Mediated Language Assessment and Evaluation in Natural Language Processing*, ASSESSEVALNLP '99, pages 68–75.

Burstein, J., Chodorow, M., and Leacock, C. (2003). CriterionSM Online Essay Evaluation: An Application for Automated Evaluation of Student Essays. In *Innovative Applications of Artificial Intelligence Conference (IAAI)*, volume 15, pages 3–10, Acapulco, Mexico.

Burstein, J., Kukich, K., Wolff, S., Lu, C., Chodorow, M., Braden-Harder, L., and Harris, M. D. (1998). Automated Scoring Using a Hybrid Feature Identification Technique. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, ACL '98, pages 206–210.

Cristianini, N. and Shawe-Taylor, J. (2000). *An Introduction to Support Vector Machines: And Other Kernel-based Learning Methods*. Cambridge University Press, New York, NY, USA.

Ishioka, T. and Kameda, M. (2004). Automated Japanese Essay Scoring System: Jess. In *15th International Workshop on Database and Expert Systems Applications, 2004.*, pages 4–8.

Júnior, C. R. C. A. and Oliveira, E. (2016). Proposta de um Sistema de Avaliação Automática de Redações do ENEM. In *Workshop de Pesquisa e Desenvolvimento em Inteligência Artificial, Inteligência Coletiva e Ciências de Dados - (WORKPEDIA, 2016)*, volume 2, Niterói, RJ, Brasil.

Kennedy, J. and Eberhart, R. C. (1997). A Discrete Binary Version of the Particle Swarm Algorithm. In *1997 IEEE International Conference on Systems, Man, and Cybernetics. Computational Cybernetics and Simulation*, volume 5, pages 4104–4108.

Luna, E. A. A. (2009). Avaliação Da Produção Escrita no ENEM: Como se Faz e o que Pensam os Avaliadores. Master's thesis, Programa de Pós-Graduação em Letras - Universidade Federal de Pernambuco (UFPE), Recife, PE.

Mota, M. d., Moussatchê, A. H., Castro, C. R., Moura, M. L. S., and D'Angelis, T. (2000). Erros de Escrita no Contexto: Uma Análise na Abordagem do Processamento da Informação. *Psicologia: Reflexão e Crítica*, 13:01 – 06.

Page, E. B. (1994). Computer grading of student prose, using modern concepts and software. *The Journal of Experimental Education*, 62(2):127–142.

Page, E. B. and Paulus, D. H. (1968). *The Analysis of Essays by Computer. Final Report.* Office of Education (DHEW), Washington, DC.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: A Method for Automatic Evaluation of Machine Translation. In *40th Annual Meeting on Association for Computational Linguistics*, pages 311–318.

Pasquali, L. (2004). *Psicometria - Teoria dos Testes na Psicologia e na Educação*. Vozes.

Pérez, D., Alfonseca, E., and Rodríguez, P. (2004). Upper Bounds and Extension of the Bleu Algorithm Applied to Assessing Student Essays. In *Conference of International Association for Educational Assessment (IAEA)*, Philadelphia, PA, USA.

Pinheiro, G. M. (2007). Redações do ENEM: Estudo dos Desvios da Norma Padrão sob a Perspectiva de Corpus. Master's thesis, Faculdade de Filosofia, Letras e Ciências Humanas - Universidade de São Paulo (USP), São Paulo, SP.

Santos, J. J., Paiva, R. O. A., and Bittencourt, I. I. S. P. (2015). Avaliação Automática de Atividades Escritas Baseadas em Algoritmo Genético e Processamento de Linguagem Natural. *Anais dos Workshops do IV Congresso Brasileiro de Informática na Educação (CBIE 2015)*, 4:95–104.

Saude, M. R., de Medeiros Soares, M., Basoni, H. G., Ciarelli, P. M., and Oliveira, E. (2014). A Strategy for Automatic Moderation of a Large Data Set of Users Comments. In *2014 XL Latin American Computing Conference (CLEI)*, pages 1–7.

Silva, W. D. C. d. M. (2014). Aprimorando o Corretor Gramatical CoGrOO. Master's thesis, Instituto de Matemática e Estatística - Universidade de São Paulo (USP), São Paulo, SP.

Wang, H.-C., Chang, C.-Y., and Li, T.-Y. (2008). Assessing Creative Problem-Solving with Automated Text Grading. *Computers & Education*, 51(4):1450–1466.