UNIVERSIDADE FEDERAL RURAL DE PERNAMBUCO

Curso de Graduação em Ciência da Computação

Lucas de Holanda Vieira Rocha

**Mineração de Texto Aplicada na Análise de Redações do ENEM**

Recife

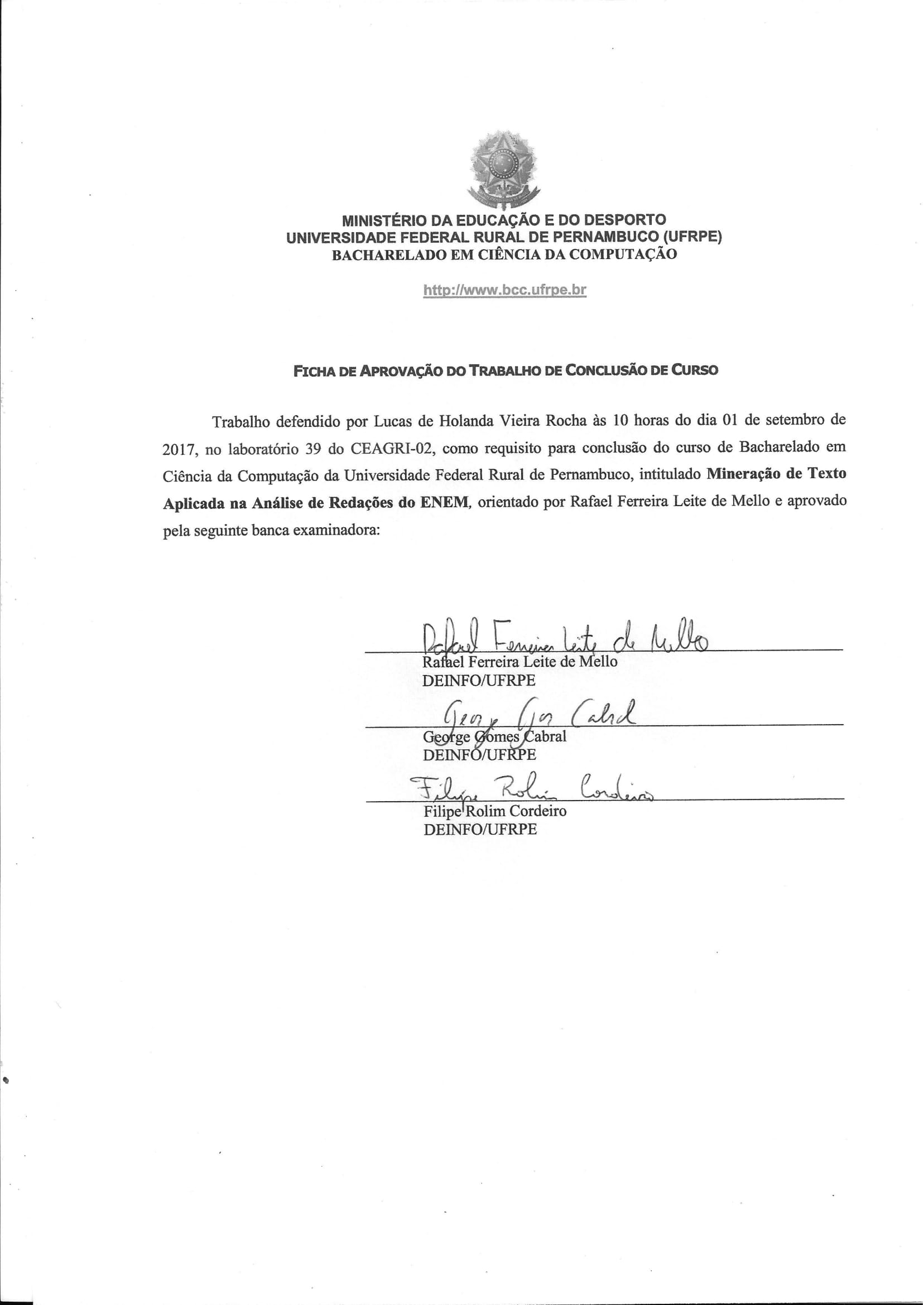
2017

Lucas de Holanda Vieira Rocha

Mineração de Texto Aplicada na Análise de Redações do ENEM

Trabalho apresentado à banca examinadora da Universidade Federal Rural de Pernambuco, como requisito para a obtenção do título de bacharel em Ciência da Computação, sob a orientação do professor Rafael Ferreira Leite de Mello.

Recife, AGOSTO de 2017



Agradecimentos

Agradeço ao meu orientador, Rafael Ferreira, por toda ajuda na criação deste trabalho.

À professora Jeisa Oliveira, que ajudou na elaboração inicial.

E aos meus colegas de curso, que tornaram esses anos todos de curso uma experiência mais agradável.

Resumo

Com a constante produção de textos no meio acadêmico, analisar a todos manualmente torna-se uma tarefa cara e desgastante. Entre estes textos destacam-se as redações, por serem maiores e mais subjetivas que questões dissertativas comuns. Portanto, torna-se necessária a utilização de ferramentas automáticas para auxiliar no processo de correção e produção de redações. As redações do ENEM são corrigidas de acordo com 5 Critérios.

Este trabalho tem como objetivo propor uma ferramenta que utilize diferentes técnicas de mineração de texto para tratar diferentes problemas relacionados à produção e análise textual, bem como um *web crawler* para varrer o Banco de Redações da UOL, criando um banco de redações local para a análise das diferentes técnicas. Ele contém 194 redações, divididas em 10 temas. Cada tema é composto de um título e um texto base para contextualizá-lo.

Ele visa analisar duas medidas de similaridade de texto para calcular a adesão da redação do aluno ao tema proposto, uma delas é provida pelo Swoogle e a outra é um algoritmo baseado em matriz de similaridade proposto por Pinheiro et al. (2017). Para a análise de coesão textual ele analisa seis diferentes medidas de legibilidade, Flesch Reading Ease, Flesch-Kincaid Grade Level, Gunning-Fog Index, Coleman-Liau Index, SMOG Index e Automated Readability Index.

Para calcular a adesão ao tema foram calculadas as similaridades entre as redações e os textos base de todos os temas, então foram propostos limiares entre 0,05 e 0,35, em incrementos de 0,05, para indicar que a similaridade era alta o bastante para considera-la dentro do tema. Foi calculada a acurácia das duas medidas para cada um dos limiares, onde a medida do Swoogle com um limiar de 0,25 demonstrou-se a melhor combinação, com uma acurácia de 0,67.

Para a análise de coesão textual calculou-se o grau de correlação entre os valores de legibilidade e o Critério 4 das redações, que trata da coesão, foi utilizado o Coeficiente Pearson. A medida Flesch Reading Ease apresentou o melhor coeficiente com um valor de 0,2385.

Palavras-chave: Redação, Mineração de Texto, Coesão, Similaridade de Texto

Abstract

With the constant production of texts in the academic world, analyzing them all manually becomes an expensive and exhausting task. Among these texts stand out the essays, because they are bigger and more subjective than common open questions. Therefore, it is necessary to use automatic tools to aid in the process of correction and production of essays. The ENEM essays are graded according to 5 Criteria.

This work aims to propose a tool that uses different text mining techniques to handle different problems related to textual production and analysis, as well as a web crawler to scan the UOL Essay Base, creating a local essay database for the analysis of different techniques. It contains 194 essays, divided into 10 themes. Each theme is composed of a title and a base text to contextualize it.

It aims to analyze two measures of text similarity to calculate the student's essay adherence to the proposed theme, one is provided by Swoogle and the other is an algorithm based on similarity matrix proposed by Pinheiro et al. (2017). For the textual cohesion analysis it analyzes six different readability measures, Flesch Reading Ease, Flesch-Kincaid Grade Level, Gunning-Fog Index, Coleman-Liau Index, SMOG Index, and Automated Readability Index.

In order to calculate adherence to the theme, the similarities between the essays and the base texts of all themes were calculated, then thresholds between 0.05 and 0.35 were proposed, in increments of 0.05, to indicate that the similarity was high enough to consider it within the theme. The accuracy of the two measures was calculated for each of the thresholds, where Swoogle’s measure with a threshold of 0.25 was the best combination, with an accuracy of 0.67.

For the textual cohesion analysis, the degree of correlation between readability values ​​and the 4th Criteria of the essays, which deals with cohesion, was calculated, using the Pearson Coefficient. The Flesch Reading Ease showed the best coefficient with a value of 0.2385.

Keywords: Essay, Text Mining, Cohesion, Text Similarity

Lista de Figuras

[Figura 1 – Interface Online do CoGrOO 19](#_Toc491415251)

[Figura 2 – Estrutura da Ferramenta 25](#_Toc491415252)

[Figura 3 – Tela Inicial 26](#_Toc491415253)

[Figura 4 – Tela do Aluno 27](#_Toc491415254)

[Figura 5 – Tela do Professor 27](#_Toc491415255)

[Figura 6 – Título e Texto Base 28](#_Toc491415256)

[Figura 7 – Calculando a Similaridade 30](#_Toc491415257)

[Figura 8 – Similaridade Calculada 30](#_Toc491415258)

[Figura 9 – Detecção de Erros Gramaticais 31](#_Toc491415259)

[Figura 10 – Distribuição das Notas 36](#_Toc491415260)

Lista de Tabelas

[Tabela 1 – Exemplos de Similaridade Entre Palavras 22](#_Toc491415289)

[Tabela 2 – As palavras e são removidas da matriz 23](#_Toc491415290)

[Tabela 3 – Relação entre resultado do teste de Flesch Reading Ease e série escolar americana 32](#_Toc491415291)

[Tabela 4 – Relação entre resultado do teste de Automated Readability Index, idade e série escolar americana 34](#_Toc491415292)

[Tabela 5 – Quantidade de Redações, Data e Média de Notas por Tema 35](#_Toc491415293)

[Tabela 6 – Acurácia por Medida com Diferentes Limiares 38](#_Toc491415294)

[Tabela 7 – Relação entre Técnica de Coesão e o Coeficiente Pearson 39](#_Toc491415295)

Sumário

[1 Introdução 9](#_Toc491414396)

[1.1 Justificativa 10](#_Toc491414397)

[1.2 Objetivos 11](#_Toc491414398)

[1.2.1 Objetivo Geral 11](#_Toc491414399)

[1.2.2 Objetivos Específicos 11](#_Toc491414400)

[1.3 Organização do Trabalho 11](#_Toc491414401)

[2 Trabalhos Relacionados 13](#_Toc491414402)

[3 Embasamento Teórico 16](#_Toc491414403)

[3.1 Critérios de Avaliação do ENEM 16](#_Toc491414404)

[3.2 Mineração de Texto 17](#_Toc491414405)

[3.3 CoGrOO 18](#_Toc491414406)

[3.4 Medidas de Similaridade 19](#_Toc491414407)

[3.4.1 Swoogle 19](#_Toc491414408)

[3.4.2 Medida Baseada em Matriz de Similaridade 20](#_Toc491414409)

[3.4.2.1 TF-IDF 20](#_Toc491414410)

[3.4.2.2 Word2Vec 21](#_Toc491414411)

[3.4.2.3 Método Baseado em Matriz 21](#_Toc491414412)

[3.4.2.4 Método Utilizado 23](#_Toc491414413)

[4 Desenvolvimento 25](#_Toc491414414)

[4.1 Persistência 27](#_Toc491414415)

[4.2 Similaridade 29](#_Toc491414416)

[4.3 Detecção de Erros 30](#_Toc491414417)

[4.4 Coesão 31](#_Toc491414418)

[4.4.1 Flesch Reading Ease 31](#_Toc491414419)

[4.4.2 Flesch-Kincaid Grade Level 32](#_Toc491414420)

[4.4.3 Gunning-Fog Index 32](#_Toc491414421)

[4.4.4 Coleman-Liau Index 33](#_Toc491414422)

[4.4.5 SMOG Index 33](#_Toc491414423)

[4.4.6 Automated Readability Index 33](#_Toc491414424)

[5 Experimento 35](#_Toc491414425)

[5.1 Banco de Dados 35](#_Toc491414426)

[5.2 Métricas de Avaliação 36](#_Toc491414427)

[5.2.1 Coeficiente de Correlação Pearson 36](#_Toc491414428)

[5.2.2 Acurácia 36](#_Toc491414429)

[5.3 Avaliação do Módulo de Similaridade 37](#_Toc491414430)

[5.4 Avaliação do Módulo de Coesão 38](#_Toc491414431)

[6 Conclusões 40](#_Toc491414432)

[6.1 Limitações 40](#_Toc491414433)

[6.2 Trabalhos Futuros 41](#_Toc491414434)

[REFERÊNCIAS 42](#_Toc491414435)

1. Introdução

Um grande volume de textos vem sendo gerado constantemente no meio acadêmico como, por exemplo, questões dissertativas e redações. Estes, por sua vez, precisam ser analisados e avaliados (WARSCHAUER e WARE, 2006). As redações, além de maiores, são mais subjetivas que a resposta de uma questão dissertativa, tornando sua leitura e avaliação um processo muito custoso.

No Brasil, um dos exames mais importantes é o Exame Nacional do Ensino Médio (ENEM), inicialmente proposto como uma ferramenta de auxílio ao Ministério da Educação para elaborar políticas de melhoria do ensino escolar, através do cruzamento de dados e pesquisas nos resultados no ENEM com os Parâmetros Curriculares Nacionais do Ensino Médio (BRASIL, 1998). Criado em 1998, foi reformado em 2009 para tentar unificar as avaliações de ingresso das faculdades federais nacionais através do Sistema de Seleção Unificada (BRASIL, 2010).

O ENEM é composto de 4 avaliações das diferentes áreas de conhecimento: Ciências Humanas, Ciências da Natureza, Matemática e Linguagens, com questões objetivas, e uma redação[[1]](#footnote-1). Logo, todos que pretendem ingressar no ensino superior precisam escrever uma redação. De acordo com Villalon e Calvo (2009) redações “são consideradas um excelente reflexo dos conhecimentos do aluno”. Eles também afirmam que “pesquisadores educacionais já estabeleceram que escrever é uma tarefa onde funções cognitivas mais elevadas, como análise e síntese, são completamente desenvolvidas”.

A redação do ENEM é avaliada de acordo com 5 competências[[2]](#footnote-2):

1. Demonstrar Domínio da Norma Padrão da Língua Escrita;
2. Compreender a Proposta de Redação e Aplicar Conceitos das Várias Áreas de Conhecimento para Desenvolver o Tema, Dentro dos Limites Estruturais do Texto Dissertativo-Argumentativo;
3. Selecionar, relacionar, organizar e interpretar informações, fatos, opiniões e argumentos em defesa de um ponto de vista;
4. Demonstrar conhecimento dos mecanismos linguísticos necessários para a construção da argumentação;
5. Elaborar proposta de solução para o problema abordado, respeitando os valores humanos e considerando a diversidade sociocultural.

Cada competência trata de um aspecto distinto da produção de uma redação. Dentre essas, a competência *i* avalia os erros gramaticais; a competência *ii*, entre outras coisas, avalia se a redação foi escrita dentro do tema solicitado (PERSING e NG, 2014); e a competência *iv* avalia a coesão textual (NOBRE e PELLEGRINO, 2010).

Diante da dificuldade de correção do grande número de redações (G1, 2016), este trabalho propõe a criação de uma ferramenta que utiliza de técnicas de mineração de texto para auxiliar professores e alunos na correção de redações. O trabalho foca principalmente nos 3 aspectos apresentados acima: erros gramaticais através da ferramenta CoGrOO, adesão ao tema através da análise de similaridade e coesão textual através da análise de legibilidade.

* 1. Justificativa

Devido à grande quantidade de redações que o ENEM recebe e deve avaliar todo ano (G1, 2016), faz-se necessário a criação de uma ferramenta automática para auxiliar na correção das redações. Contudo, existem diferentes aspectos que precisam ser analisados, como por exemplo: erros gramaticais, coesão do texto e pertinência ao tema.

“Erros Gramaticais ocorrem quando a estrutura gramatical da sentença não segue as regras gramaticais vigentes” (KINOSHITA, SALVADOR e MENEZES, 2005). Eles tipos de erros consistem na inconsistência e concordância de um diverso conjunto, incluindo artigos, preposições, substantivos, verbos, concordância verbal, pronomes, escolha de palavras, estrutura de sentença, pontuação, entre outros. (NG et al., 2014)

Coesão textual se refere à presença ou ausência de sugestões explícitas no texto que permitem ao leitor fazer conexões entre ideias no texto (CROSSLEY, KYLE e MCNAMARA, 2016). Legibilidade é definida como “a soma de todos os elementos em um material textual que afetam a compreensão, velocidade de leitura e nível de interesse do material” (DALE e CHALL, 1949).

Por último, a adesão ao tema se refere ao quão relacionado está o conteúdo da redação com o tema proposto. Uma redação com uma adesão alta permanece constantemente no tópico proposto, não inserindo informação irrelevante (PERSING e NG, 2014).

Diante disto, este trabalho visa fornecer estatísticas relacionadas à avaliação automática de cada um desses aspectos, principalmente focando em coesão do texto e pertinência ao tema, auxiliando assim tanto na correção como na criação de redações.

* 1. Objetivos

Essa seção apresenta os objetivos do trabalho.

* + 1. Objetivo Geral

Utilizar mineração de texto para analisar automaticamente textos de redações quanto à coesão e pertinência ao tema.

* + 1. Objetivos Específicos

1. Propor um *crawler* para extração e indexação das redações já escritas e corrigidas disponíveis no Banco de Redações da UOL[[3]](#footnote-3).
2. Integração de ferramentas para análise léxica e sintática de textos de redações.
3. Utilização de diferentes medidas de similaridade para análise da adesão ao tema.
4. Analisar diferentes algoritmos para tratar coesão de texto em português.
   1. Organização do Trabalho

Este trabalho encontra-se organizado da seguinte forma:

* O Capítulo 2 aborda outros trabalhos na área, apresentando uma revisão da literatura e nos trazendo a par do estado da arte;
* O Capítulo 3 apresenta um conjunto de definições importantes para o trabalho, abordando técnicas comumente utilizadas na área, ferramentas utilizadas no trabalho e algoritmos utilizados no processo de desenvolvimento do trabalho;
* O Capítulo 4 apresenta as técnicas e o banco de dados utilizados no desenvolvimento da ferramenta bem como sua estrutura;
* O Capítulo 5 apresenta os resultados obtidos com o uso da ferramenta;
* O Capítulo 6 apresenta as considerações finais sobre o resultado deste trabalho.

1. Trabalhos Relacionados

Existem diversos trabalhos que lidam com a avaliação automática de redações (NEWMAN *et al.*, 2010) (SHERMIS *et al.*, 2010) (VILLALON e CALVO, 2009), mas devido à dificuldade do desenvolvimento de técnicas de mineração e análise de texto em português, os trabalhos neste idioma ainda são poucos (BAZELATO e AMORIM, 2010) (EPSTEIN e REATEGUI, 2015) (NOBRE e PELLEGRINO, 2010).

Para a extração de tópicos de um texto, Newman *et al.* (2010) sugerem a avaliação de coerência de tópicos, onde um conjunto de palavras, geradas por um modelo de tópicos, onde este modelo utiliza 4 tópicos, é classificada por coerência ou interpretabilidade. Para tal avaliação, eles se utilizam de um modelo de tópicos e um conjunto de tópicos, baseando-se no método de *Latent Dirichlet Allocation* (BLEI, NG e JORDAN, 2003) para a geração de tópicos.

Em um modelo de tópicos, cada documento, em um conjunto de *d* documentos, é modelado como uma distribuição multinominal de *t* tópicos, e estes são representados como uma distribuição multinominal de *p* palavras. Este artigo segue a convenção de usar apenas as 10 palavras mais significativas para representar cada tópico. O artigo então compara diferentes métodos de avaliação de tópicos, baseados em 3 sistemas de busca, WordNet, Wikipédia e Google.

Villalon e Calvo (2009) utilizam-se de *parsers* gramaticais e análise semântica para a extração de Mapas de Conceito, que são pequenos resumos que devem conter os principais temas da redação.

A extração de conceito é feita através da análise das palavras ou frases em potencial que podem fazer parte do Mapa de Conceito, como Conceitos, Relacionamentos e Topologia. Ele utiliza primeiro a identificação de conceitos usando uma árvore gramatical, então aplica a *Latent Semantic Analysis*, para a remoção de redundâncias.

Em português, Epstein e Reategui (2015) e Klemman, Reategui e Lorenzati (2009) utilizaram a ferramenta de mineração de textos Sobek, que identifica os conceitos mais relevantes de um texto, a relação entre eles e utiliza um grafo das ideias apresentadas no texto para exibir os resultados.

No experimento de Epstein e Reategui dois textos foram utilizados para a avaliação, onde o conjunto de tópicos *gold standard* foi criado pela avaliação de três especialistas na área de educação. *Gold standard* é um termo usado para descrever algo como sendo o objetivo ideal. Neste caso, isto quer dizer que quanto mais próximo à lista de tópicos *gold standard* estiver a lista que o Sobek gerar, melhor ele será.

O resultado de sensibilidade obtido por eles, que se dá pelo número de termos que o Sobek considerou corretamente em relação ao *gold standard*, foi de 62,5% e 69%. O resultado de precisão, que se dá pelo número de termos que o Sobek considerou corretamente em relação ao número total de termos que há sua própria lista, foi de 100%.

Klemman, Reategui e Lorenzati também utilizaram a ferramenta Sobek, porém para servir de apoio na criação textual, ao invés de usá-la para classificação, como Epstein e Reategui. Um estudante de 20 anos, no segundo semestre de Engenharia Mecânica inicialmente inseriu seu texto no Sobek, que extraiu os conceitos e os mostrou em uma lista. Ele então foi modificando a lista e sua redação, removendo da lista conceitos que foram identificados erroneamente e adicionando conceitos que não haviam sido identificados.

Observou-se que a alternância do aluno entre a ferramenta e o texto foi construtivo para promover uma melhor compreensão do texto e do tema tratado. A produção do texto de maneira bastante fluida foi atribuída às várias etapas adotadas no uso da ferramenta. Ao final, foi mostrado um grafo dos conceitos e como eles se relacionam ao aluno, que imediatamente os identificou como corretos.

Bazelato e Amorim (2010) propõem um avaliador baseado em uma rede bayesiana composta de 21 classificadores, um para cada nota de 0 a 10 em incrementos de 0,5, sobre um conjunto de redações extraídas do Banco de Redações do site da UOL.[[4]](#footnote-4) Este classificador já é baseado em um desenvolvido por Larkey (1998), que classificava um conjunto de redações em 3 categorias, Soc, sobre estudos sociais, Phys, sobre física, e Law, sobre assuntos legais. Usando uma rede com 3 classificadores bayesianos e aplicando uma regressão linear, este foi capaz de alcançar uma precisão de 54% a 62% em Soc, 44% a 55% em Phys e 24% a 42% em Law.

Nobre e Pellegrino (2010) utilizam-se de um método desenvolvido a partir da Teoria da Centragem e do Foco e aplicação da Lógica Difusa. A partir de um conjunto de conhecimentos morfossintáticos e semânticos, obtidos através do *parser* Palavras, o avaliador automático de coesão desenvolvido processa os dados. Comparando às notas dadas por uma comissão avaliadora, o artigo afirma ter conseguido uma taxa de 70% das redações classificadas na mesma faixa de pontuação dos humanos. Com 15% apresentando uma diferença inferior a 4 décimos, e os 15% restante, uma diferença maior.

Este trabalho visa integrar diversas abordagens para a análise de redações. Utilizando medidas de similaridade para calcular o grau de adesão ao tema, uma baseada em matriz de similaridade e a outra baseada na distribuição de similaridade e da Análise Semântica Latente; análise léxica e sintática para a análise gramatical; e seis medidas de legibilidade para garantir a coesão da redação: Flesch Reading Ease, Flesch-Kincaid Grade Level, Gunning-Fog Index, Coleman-Liau Index, SMOG Index e Automated Readability Index.

1. Embasamento Teórico

Este capítulo irá explorar em mais detalhes os principais conceitos, técnicas e ferramentas utilizados na elaboração deste trabalho.

* 1. Critérios de Avaliação do ENEM

A redação do ENEM é avaliada de acordo com 5 competências abordadas na seção 1.

O item *i* envolve ter a distinção entre as modalidades oral e escrita da língua portuguesa, bem como a constituição de frases. O aluno deve demonstrar também a perícia no uso da pontuação para expressar ideias da oralidade, como pausas e entoação. Além disso, ele deve também mostrar domínio gramatical e precisão vocabular da língua.

O item *ii* envolve a compreensão do tema proposto para a redação, através de um texto dissertativo-argumentativo. Ele deve elaborar um texto relacionado ao tema, que defenda uma tese. Dois pontos importantes a se evitar neste item são o tangenciamento, que é uma abordagem apenas parcial do tema, colocando-o em segundo plano, e a fuga ao tema, que envolve redações onde nem o tema, nem outros assuntos amplos relacionados ao tema foram desenvolvidos.

O item *iii* envolve a forma como se seleciona, relaciona e organiza suas informações, fatos, opiniões e argumentos em defesa do ponto de vista da sua tese. O texto apresentado deve ser claro e conciso, e deve defender bem a posição assumida.

O item *iv* envolve a organização textual, onde o aluno deve demonstrar-se capaz de desenvolver uma estrutura lógica e formal entre as partes da redação, garantindo a coesão textual e a interdependência das ideias.

O item *v* envolve a apresentação da solução proposta ao problema abordado, apresentando uma tese sobre o tema, suportada por argumentos consistentes. A solução proposta na tese deve ter um vínculo direto com o tema abordado e deve demonstrar-se coerente com os argumentos utilizados.

Destas cinco competências, este trabalho irá propor soluções de análises para o item *i*, através do uso da ferramenta CoGrOO, item *ii*, a partir do uso de medidas de similaridade textual e do item *iv*, através das medidas de coesão textual.

* 1. Mineração de Texto

Mineração de Texto pode ser definida como um método de extração de informações relevantes em bases de dados não estruturadas, ou semi-estruturadas (FELDMAN e SANGER, 2007).

Klemann, Reategui e Rapkiewicz (2011) afirmam que:

Trata-se de um campo multidisciplinar que inclui conhecimentos de áreas como Informática, Estatística, Linguística e Ciência Cognitiva. A mineração de textos busca extrair regularidades, padrões ou tendências de textos em linguagem natural, normalmente, para objetivos específicos.

Outra definição também é a de que mineração de textos é uma aplicação de sistemas de computação que envolve hardware e software dedicados à análise textual de documentos (KLEMANN, REATEGUI e RAPKIEWICZ apud MATTISON, 1999).

A mineração de textos permite recuperar informações, extrair dados, resumir documentos, descobrir padrões, dentre outras análises possíveis de se realizar em documentos de texto. Pode ser utilizada com muitos propósitos, como por exemplo identificar documentos similares entre si, buscar dados relevantes dentro do documento, entre outras (KLEMANN, REATEGUI e RAPKIEWICZ, 2011).

Visto que o uso da mineração de texto é aplicado em documentos não estruturados, podem ser extraídos dados não importantes para seu objetivo, logo, é necessário realizar um pré-processamento, para adaptar os dados do texto a um formato que seja útil ao seu objetivo.

Meyer, Hornik e Feinerer (2008) definem o pré-processamento como “a aplicação de métodos para a limpeza e estruturação do texto de entrada para análises adicionais”, além disso, também frisam que ele “é um componente principal no estudo de mineração de texto prática”.

Hotho, Nürnberger e Paaß (2005) descrevem alguns dos métodos de pré-processamento mais comuns, como:

* Remoção de *Stopwords:* Seu objetivo é limpar o texto removendo palavras que acrescentam pouco ou nenhuma informação, como artigos, conjunções, preposições etc.
* *Stemming*: Envolve tentar reconstruir a forma básica da palavra, por exemplo, removendo o “s” de substantivos no plural, ou “ei”, “ou”, “ar”, “er” etc, de verbos conjugados ou no infinitivo, entre outros afixos.
* Lematização: Seu objetivo é tentar mapear verbos conjugados com suas formas no infinitivo.
  1. CoGrOO

O *OpenOffice*[[5]](#footnote-5) é uma ferramenta conhecida como “pacote de escritório”, composto de processador de textos, planilhas de cálculos, editor HTML e editor de apresentação.

Apesar de suas funcionalidades serem comparáveis a outros pacotes de escritório, como o *software* proprietário *Microsoft Office*, ele ainda não possui um corretor gramatical, apenas ortográfico.

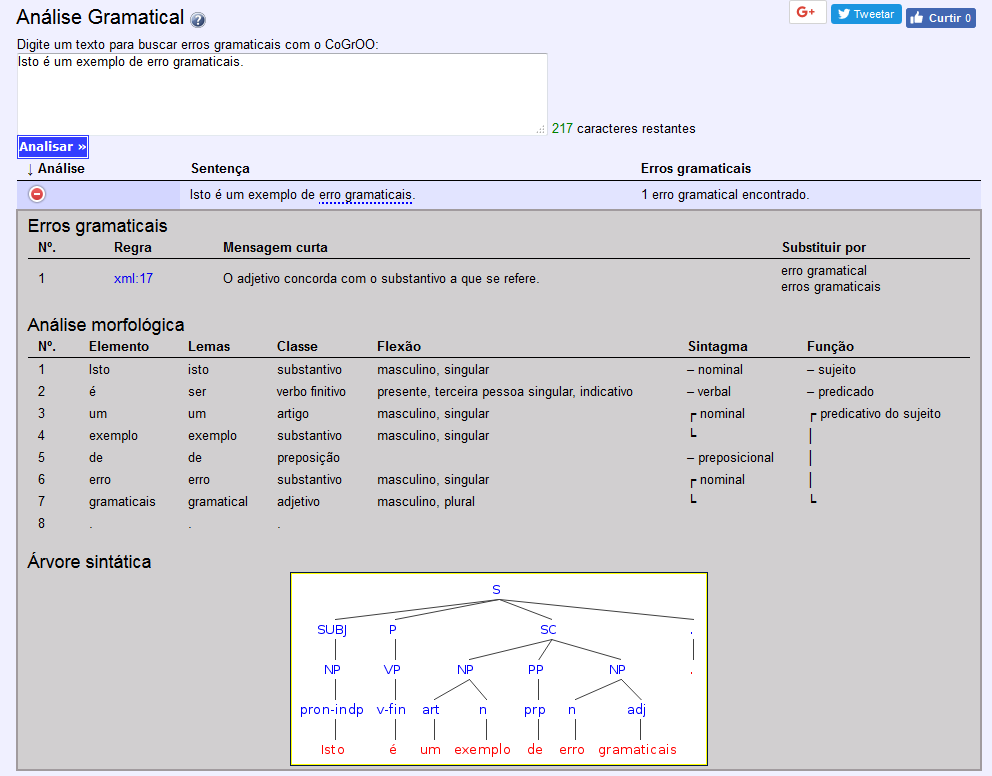
Erros Ortográficos ocorrem quando o usuário digita uma palavra erroneamente. Pode ser por um erro de digitação, onde ele inverte letras, ou aperta a tecla errada próxima à correta, ou ainda pode ser por desconhecer a forma correta da palavra. Exemplos respectivamente são: “gahnar”, “viyória” e “adezivo”.

Erros Gramaticais ocorrem quando a estrutura da sentença não está de acordo com as normas gramaticais. Por exemplo: “Nós gosta de pão” Teria um erro gramatical de concordância nominal.

O CoGrOO então foi proposto como um corretor gramatical que pode ser acoplado ao OpenOffice (KINOSHITA, SALVADOR e MENZES, 2005). Tendo como objetivo identificar erros gramaticais, tais como erros de concordância nominal e verbal, crase (a contração da preposição “a” com o artigo “a”, resultando em “à”), regência nominal e verbal, e outros erros comuns que são encontrados na língua portuguesa falada no Brasil (KINOSHITA, SALVADOR e MENZES, 2006).

A API do CoGrOO[[6]](#footnote-6) está disponibilizada online para ser usada como uma biblioteca Java. Além desta, ele também possui uma interface[[7]](#footnote-7) online onde o usuário pode interagir diretamente, colocando seu texto, como na Figura 1. Nela você ainda pode expandir a análise para ver mais detalhes sobre os erros, como as regras quebradas, soluções propostas para resolver os erros e a árvore da análise morfológica.

Figura – Interface Online do CoGrOO



Fonte: CoGrOO Comunidade

* 1. Medidas de Similaridade

Esta seção detalhará os dois métodos de cálculo de similaridade analisados durante a criação da ferramenta, o Serviço de Similaridade Semântica da UMBC através da API do Swoogle, e o modelo de Pinheiro et al. (2017).

* + 1. Swoogle

A Web Semântica é uma web paralela à Web convencional. Os Documentos da Web Semântica (DWS) são caracterizados por anotações semânticas e referências a outros DWS. O Swoogle é um sistema de indexação e busca para a Web Semântica. (DING et al., 2004)

Han et al. (2013) propuseram um modelo de similaridade de palavras, desenvolvido inicialmente para o Projeto de Grafo de Relações (UMBC, 2013). Este modelo utiliza uma abordagem híbrida, combinando os dois métodos mais populares de similaridade de palavras, um baseado no uso de um tesauro, que é uma lista de palavras com significados semelhantes, dentro de um domínio específico de conhecimento, e o outro num conjunto de estatísticas geradas a partir de uma larga coleção de documentos. O método estatístico é baseado na distribuição de similaridade e da Análise Semântica Latente (ASL), complementado com relações semânticas extraídas do WordNet.

O Swoogle provê um *webservice* com uma API de acesso a este serviço que recebe as duas frases a serem comparadas e retorna um número entre 0 e 1 que representa seu grau de similaridade.

* + 1. Medida Baseada em Matriz de Similaridade

Pinheiro et al. (2017) propõem um modelo de cálculo de similaridade que extrai quatro características de um par de sentenças para calcular sua similaridade.

As técnicas utilizadas estão descritas abaixo, seguidas de como sua integração no modelo ocorrem.

* + - 1. TF-IDF

Esta abordagem visa medir o grau de importância de uma palavra em relação a um conjunto de documentos, utilizando sua frequência (TF) e relevância no conjunto de textos (IDF). O TF-IDF é calculado pelas equações abaixo.

Isto gera uma matriz de sentenças x palavras, e o valor de TF-IDF de cada palavra para cada sentença. A similaridade das sentenças então é calculada através do cosseno da distância entre os vetores TF-IDF dos pares de sentenças (SALTON e YANG, 1973).

* + - 1. Word2Vec

Word2Vec é um modelo para a geração de um vetor para cada palavra em um conjunto, com o objetivo de medir a similaridade semântica entre as palavras (XUE, FU e SHAOBIN, 2014). Ele utiliza o modelo de predição *skip-gram*, que prevê as palavras próximas, dada uma palavra. O vetor da palavra corresponde aos pesos entre a entrada e a primeira camada escondida da rede neural *feedforward* utilizada. O tamanho final deste vetor é um parâmetro de entrada.

O vetor então é simplificado, e a similaridade entre as palavras é calculada através do cosseno da distância entre os vetores de cada palavra.

* + - 1. Método Baseado em Matriz

Ferreira et al. (2016) propõe uma representação em 3 camadas para calcular a similaridade entre um par de sentenças:

1. A Camada Rasa, composta pela análise léxica, *stopwords* e um Reconhecedor de Entidades Nomeadas (REN)
2. A Camada Sintática, composta pela análise sintática, REN e relações de correferência.
3. A Camada Semântica, que descreve a anotação semântica.

Ele utiliza um método baseado em matriz para calcular a similaridade entre as sentenças.

Tomando e , onde é uma palavra da frase A, e é uma palavra da frase B, n é o número de palavras em A, e m o número de palavras em B.

Tomando duas sentenças A e B com seis palavras cada, para calcular a similaridade entre elas, o primeiro passo seria calcular a similaridade entre as palavras das duas sentenças, gerando uma tabela-matriz (Tabela 1).

Tabela – Exemplos de Similaridade Entre Palavras

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |
|  | 0,3 | 0,2 | 0,56 | 0,88 | 0,25 | 0,13 |
|  | 0,12 | 0,5 | 0,31 | 0,22 | 0,87 | 0,65 |
|  | 0,56 | 0,23 | 0,5 | 0,28 | 0,6 | 0,63 |
|  | 0,7 | 0,62 | 0,6 | 0,38 | 0,12 | 0,1 |
|  | 0,84 | 0,21 | 0,54 | 0,78 | 0,29 | 0,56 |
|  | 0,4 | 0,35 | 0,47 | 1,0 | 0,23 | 0,33 |

Fonte: Pinheiro et al., 2017

O segundo passo é remover as palavras que tiveram o maior grau de similaridade, neste caso e como mostrado na Tabela 2.

Os passos 1 e 2 então são repetidos até não haver mais palavras para terem sua similaridade calculada.

O último passo então é calcular a média entre os maiores valores de similaridade obtidos, a partir da fórmula abaixo:

O resultado obtido é a similaridade entre as sentenças.

Tabela – As palavras e são removidas da matriz

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |
|  | 0,3 | 0,2 | 0,56 | 0,88 | 0,25 | 0,13 |
|  | 0,12 | 0,5 | 0,31 | 0,22 | 0,87 | 0,65 |
|  | 0,56 | 0,23 | 0,5 | 0,28 | 0,6 | 0,63 |
|  | 0,7 | 0,62 | 0,6 | 0,38 | 0,12 | 0,1 |
|  | 0,84 | 0,21 | 0,54 | 0,78 | 0,29 | 0,56 |
|  | 0,4 | 0,35 | 0,47 | 1,0 | 0,23 | 0,33 |

Fonte: Pinheiro et al., 2017

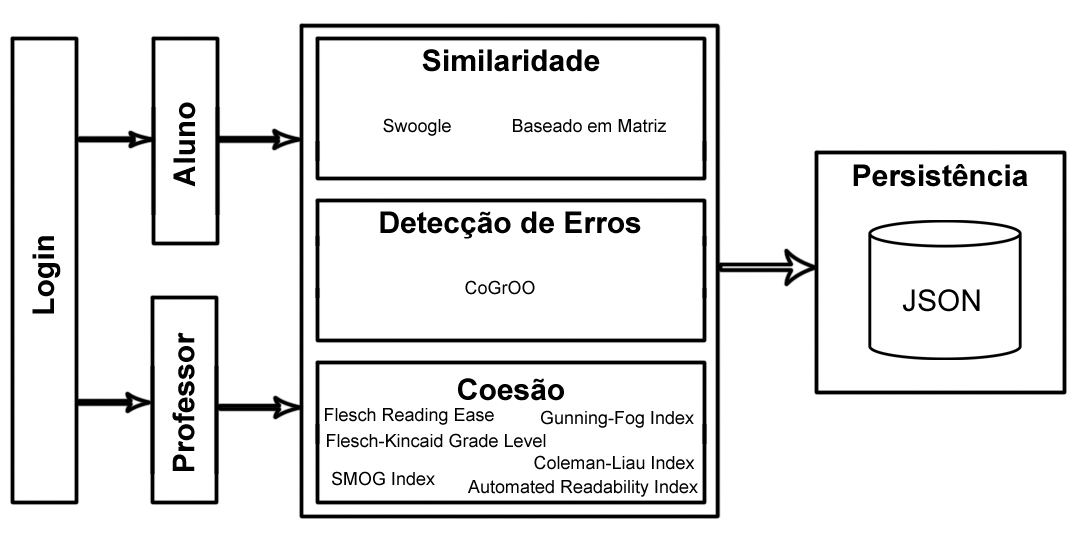
* + - 1. Método Proposto

O método proposto por Pinheiro et al. (2017) consiste na extração de 4 características e o uso de um classificador para calcular a similaridade

1. A primeira característica utiliza o cosseno da distância entre o TF-IDF de cada sentença para calcular a similaridade. Antes do cálculo do TF-IDF, dois métodos de pré-processamento foram utilizados: *Stemming*, para reduzir a variação dos dados, e um método de expansão de palavras para incluir até dois sinônimos para cada palavra de acordo com o TEP (Um Thesaurus paras Português do Brasil).
2. A segunda característica é obtida através do uso em conjunto de um método baseado em matriz e Word2Vec. Alguns métodos padrão de pré-processamento foram utilizados, como transformação em minúsculas e remoção de pontuação, bem como a remoção de *stopwords* e lematização O modelo foi construído utilizando os seguintes parâmetros:
   1. Dimensão: 250
   2. *Window*: 10
   3. Menor frequência de palavras: 5
   4. Número de Iterações: 10
3. Outro método baseado em matriz também é utilizado, o método de matriz binária, para amenizar o fato de que o método anterior também gerava resultados altos para sentenças com pouca similaridade. Ele atribui os valores de acordo com:
4. A última característica é o tamanho das sentenças. Para calcular um valor que represente este tamanho, a quantidade de palavras da sentença menor é dividida pela quantidade de palavras da sentença maior. Para este método, as *stopwords* de ambas as frases foram removidas.
5. Por fim, um algoritmo de classificação de Regressão Linear é aplicado para combinar as características num valor final. Ele consiste na execução de uma análise estatística para verificar a existência de uma relação entre uma variável dependente com uma ou mais variáveis independentes.
6. Desenvolvimento

Esta seção detalha os processos de construção da ferramenta, incluindo os diferentes métodos utilizados durante sua produção e a criação da base de dados usando um *web* *crawler*.

Figura – Estrutura da Ferramenta



Fonte: O autor

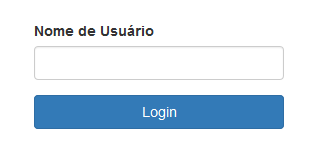
A Figura 2 apresenta uma visão geral da proposta, dividida em 7 módulos, onde:

* **Login**: Trata das credenciais do usuário, para determinar se é um professor ou aluno e direcioná-lo à tela correta.
* **Aluno:** Trata da redação do aluno, oferecendo ajuda conforme ele escreve sobre sua aderência ao tema e coesão, e ele também pode ativamente requisitar a detecção de erros.
* **Professor**: Permite que ele selecione um aluno de uma lista e veja a redação do mesmo, contando com os mesmos auxílios do módulo aluno
* **Similaridade**: Utiliza técnicas de similaridade entre textos para analisar se as redações estão dentro do tema proposto.
* **Detecção de erros**: Identifica erros gramaticais na redação.
* **Coesão**: Analisa a legibilidade da redação.
* **Persistência**: Guarda os dados gerados pelos alunos, de modo que possam ser acessados posteriormente pelos professores.

A ferramenta a ser desenvolvida visa atender a alunos e professores, ajudando tanto na escrita como na correção das redações, automatizando vários processos. Ela terá uma interface onde alunos e professores poderão realizar seu *login* (Figura 3) e entrarem em suas respectivas páginas, onde terão acesso a diferentes módulos da ferramenta, como mostrado na Figura 4 para os alunos, e na Figura 5 para os professores.

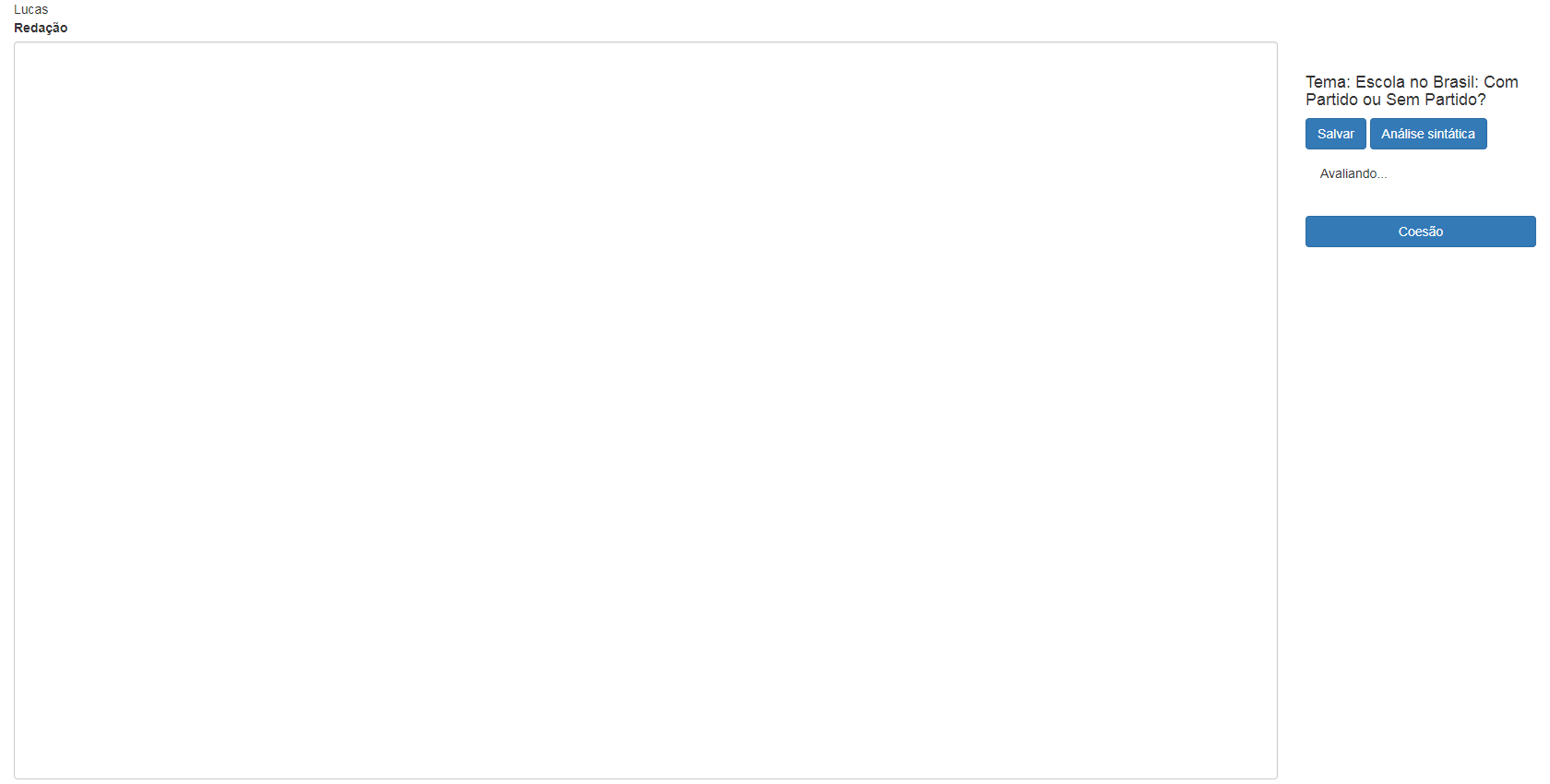
A Figura 3 apresenta a tela inicial da ferramenta desenvolvida.

Figura – Tela Inicial



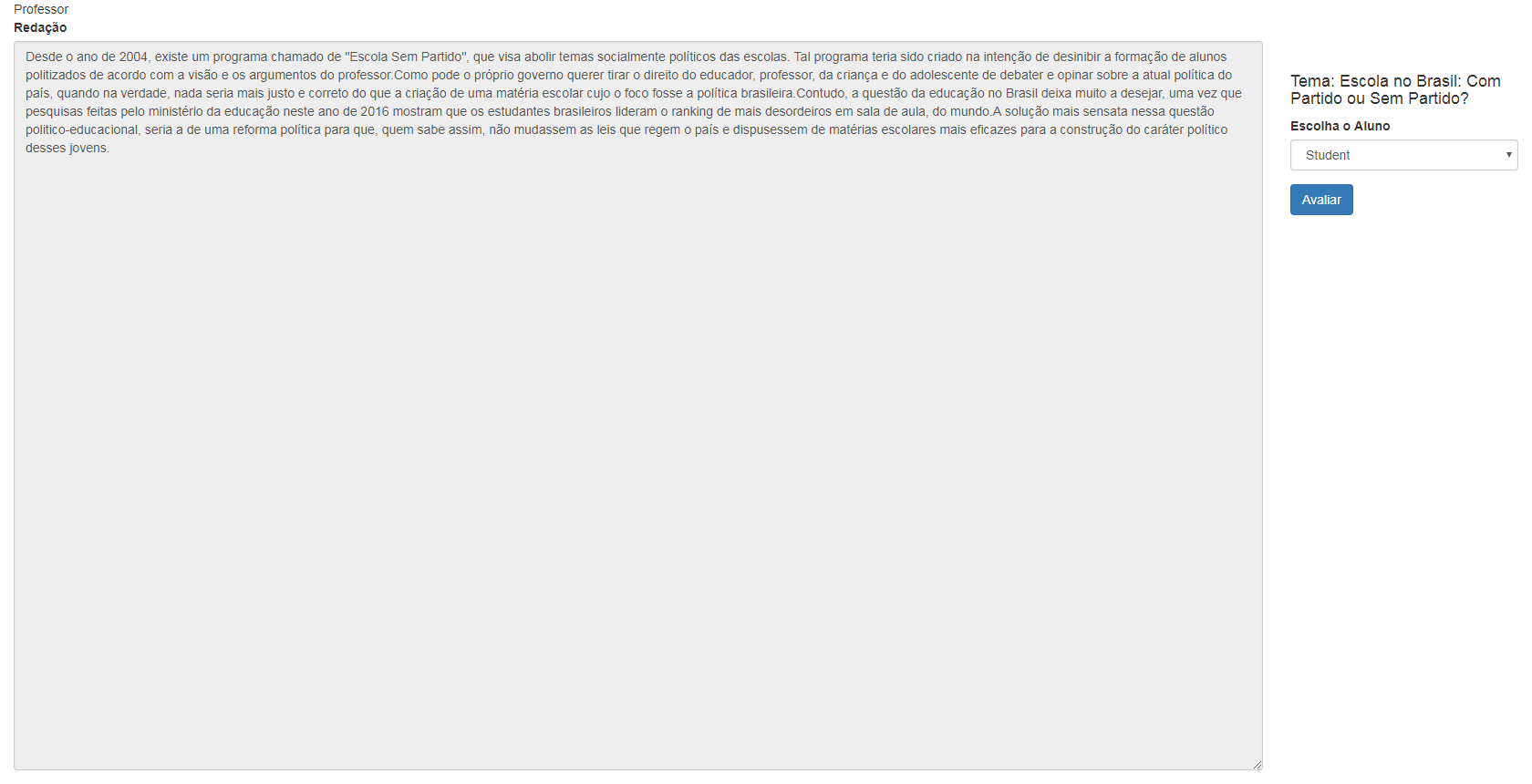
Fonte: O autor

Figura – Tela do Aluno



Fonte: O Autor

Figura – Tela do Professor



As próximas seções vão apresentar em detalhes os principais módulos do sistema: Persistência, Similaridade, Detecção de Erros e Coesão.

* 1. Persistência

O Banco de Redações da UOL é uma base de redações públicas disponível gratuitamente no site da UOL. Um novo tema é proposto mensalmente pela UOL, e os usuários submetem suas próprias redações sobre ele. Uma banca de professores associados então corrige e comenta 20 textos sorteados, e estes são publicados no site. A avaliação consiste em uma nota e comentários baseados nos critérios de correção do ENEM[[8]](#footnote-8). O tema proposto consiste de um título e um texto base para contextualizar o tema, conforme demonstrado na Figura 66.

Figura – Título e Texto Base



Fonte: Banco de Redações da UOL[[9]](#footnote-9)

Foi criado um *web* *crawler*, utilizando a biblioteca jsoup[[10]](#footnote-10) para a conexão com a página de redações e para realizar as buscas das informações no documento HTML. Ele utiliza uma lista com os *links* das páginas web dos temas desejados, e busca o link para a página de cada redação daquele tema, recuperando as informações desejadas, criando uma pasta para cada tema, e um arquivo XML para cada redação, que será salvo dentro da pasta de seu respectivo tema.

No futuro, caso se deseje aumentar o banco de redações, basta adicionar os *links* das páginas dos temas desejados do Banco de Dados da UOL e executar o *web crawler*.

O arquivo gerado é um XML que guarda as informações de título da redação, nota, texto original, erros e suas correções, comentário geral, aspectos pontuais, e uma lista de cada competência das redações do ENEM e sua nota associada. Os erros, por sua vez, possuem atributos que descrevem um id único, o ponto onde começam e seu comprimento

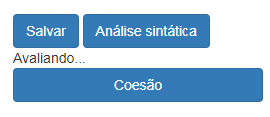
A persistência de dados, neste caso o texto da redação e o nome do aluno ao qual ela está associada, é feita através de um arquivo JSON, onde está guardado um *vetor* cujas chaves são os nomes dos alunos e seus valores são as redações dos respectivos alunos.

* 1. Similaridade

Este módulo é responsável por comparar a similaridade da redação do aluno com o texto introdutório do tema proposto no Banco de Redações da UOL para detecção de possível fuga ao tema.

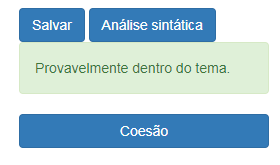
Inicialmente foi usado o serviço de similaridade do Swoogle[[11]](#footnote-11), onde o texto da redação do aluno será constantemente comparado ao texto introdutório, para verificar se ele está ou não dentro do tema. A Figura 7 e a Figura 8 mostram as telas que aparecem no momento da avaliação do texto. O processo é repetido automaticamente a cada 20 segundos para que o aluno possa perceber se a sua redação está começando a fugir do tema.

Figura – Calculando a Similaridade



Fonte: O autor

Figura – Similaridade Calculada



Fonte: O Autor

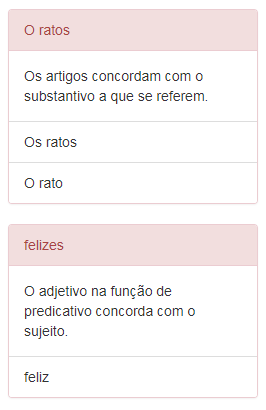
O sistema também tem a opção de utilizar a medida de similaridade baseada em matriz apresentada na seção 3.3.2. Contudo, a medida do Swoogle alcançou melhores resultados (como apresentado na seção de experimento mais a frente) e possui um custo computacional menor.

* 1. Detecção de Erros

Este módulo é responsável por analisar o texto, identificar os erros gramaticais cometido pelo aluno e propor possíveis formas de corrigi-los.

A detecção de erros é feita a partir de um WebService Java, utilizando a biblioteca CoGrOO para fazer a análise e retornar dados sobre o texto e seus possíveis erros. A Figura 99 mostra como os erros são identificados e exibidos para o usuário, para o exemplo de frase “O ratos entra na casa. Ele está felizes.”

Figura – Detecção de Erros Gramaticais



Fonte: O autor

* 1. Coesão

Este módulo trata de avaliar o grau de legibilidade da redação, de acordo com 6 medidas: Flesch Reading Ease, Flesch-Kincaid Grade Level, Gunning-Fog Index, Coleman-Liau Index, SMOG Index e Automated Readability Index.

Estes testes, em geral, se utilizam do número de palavras, sílabas, frases ou letras para calcular um número que pode ser um valor arbitrário que é associado a uma série escolar dos Estados Unidos através de uma tabela ou pode ser diretamente o número da série. Para a análise de coesão de uma redação, quanto maior o valor das medidas maior a coesão da redação

* + 1. Flesch Reading Ease

Este teste usa a média de palavras por sentença e de sílabas por palavra para calcular o grau de legibilidade de um texto, através da fórmula abaixo[[12]](#footnote-12). O resultado máximo é em torno de 120, enquanto podem haver resultados abaixo de 0.

O resultado desta fórmula é usado para comparar à seguinte tabela. As séries são séries escolares do Estados Unidos.

Tabela – Relação entre resultado do teste de Flesch Reading Ease e série escolar americana

|  |  |
| --- | --- |
| **Resultado** | **Série Escolar** |
| 100.00-90.00 | 5ª Série |
| 90.0–80.0 | 6ª Série |
| 80.0–70.0 | 7ª Série |
| 70.0–60.0 | 8ª e 9 ª Séries |
| 60.0–50.0 | 10ª a 12 ª Séries |
| 50.0–30.0 | Faculdade |
| 30.0–0.0 | Graduado |

Fonte: Flesch, 2017

* + 1. Flesch-Kincaid Grade Level

Este teste usa as mesmas variáveis do teste anterior, porém seu resultado é um número que índica a série para a qual o texto é indicado, conforme a fórmula abaixo (KINCAID, 1975). Para números maiores que 12, ele também pode ser visto como o número de anos de educação que a pessoa precisa ter para compreender o texto.

* + 1. Gunning-Fog Index

Este teste usa a média de palavras por sentença e o percentual de “palavras complexas”, neste caso palavras com 3 ou mais sílabas que não incluam nomes próprios, jargões familiares ou palavras compostas (PEARSON CANADA, 2009). A fórmula completa é apresentada a seguir.

Assim como o Flesch-Kincaid Grade Level, o Gunning-Fog Index representa a série dos Estados Unidos correspondente à dificuldade de leitura, e acima de 12 representa o número de anos de educação que a pessoa precisa ter para compreender o texto.

* + 1. Coleman-Liau Index

Este teste usa a média de letras e a média de sentenças por 100 palavras para calcular o número que representa a série do Estados Unidos indicada para o texto, conforme a fórmula abaixo[[13]](#footnote-13).

* + 1. SMOG Index

Este teste utiliza o número de palavras polissílabas e o número de sentenças para chegar a um resultado que representa a série do Estados unidos em que a pessoa deve estar para conseguir entender o texto. O texto deve ter pelo menos 30 frases. A fórmula completa está apresentada abaixo (MCLAUGHLIN, 1969).

* + 1. Automated Readability Index

Este teste utiliza a quantidade média de letras por palavra e palavras por sentenças para a obtenção de um valor, que então deve ser comparado à uma tabela. A seguir está apresentada a fórmula completa (SENTER e SMITH, 1967).

O valor resultante da aplicação desta fórmula deve ser comparado à seguinte tabela para obter-se a série do Estados Unidos indicada para sua leitura.

Tabela – Relação entre resultado do teste de Automated Readability Index, idade e série escolar americana

|  |  |  |
| --- | --- | --- |
| **Valor** | **Idade** | **Série** |
| 1 | 5-6 | Jardim de Infância |
| 2 | 6-7 | 1ª Série |
| 3 | 7-8 | 2 ª Série |
| 4 | 8-9 | 3 ª Série |
| 5 | 9-10 | 4 ª Série |
| 6 | 10-11 | 5 ª Série |
| 7 | 11-12 | 6 ª Série |
| 8 | 12-13 | 7 ª Série |
| 9 | 13-14 | 8 ª Série |
| 10 | 14-15 | 9 ª Série |
| 11 | 15-16 | 10 ª Série |
| 12 | 16-17 | 11 ª Série |
| 13 | 17-18 | 12 ª Série |
| 14 | 18-22 | Faculdade |

Fonte: Senter e Smith, 1967

1. Experimento

Essa seção apresenta detalhes sobre a avaliação dos módulos de coesão e adesão ao tema.

* 1. Banco de Dados

Um total de 194 redações de 10 diferentes temas foram extraídas com o *web crawler* proposto para serem usadas neste experimento. A distribuição das mesmas, bem como a data que o tema foi proposto e a média de notas das redações pode ser vista na Tabela 5.

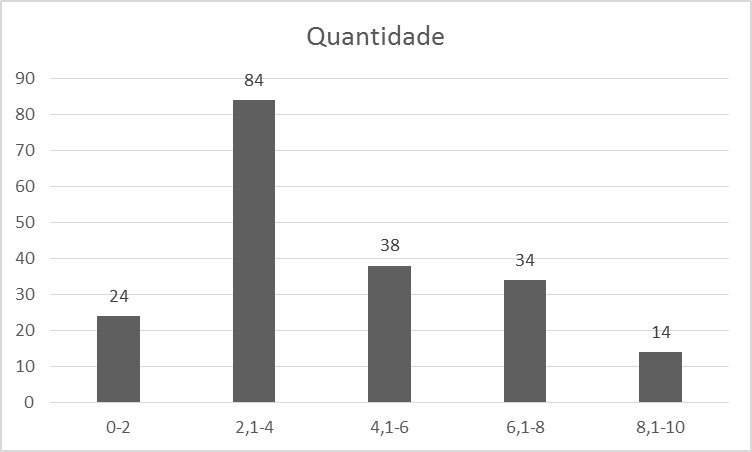
Tabela – Quantidade de Redações, Data e Média de Notas por Tema

|  |  |  |  |
| --- | --- | --- | --- |
| **Título** | **Quantidade** | **Data** | **Média** |
| A tecnologia e a eliminação de empregos | 20 | Fev/16 | 4,3 |
| Bandido bom é bandido morto? | 20 | Nov/15 | 3,9 |
| Carta-convite: discutir discriminação na escola | 19 | Mar/16 | 3,7 |
| Disciplina, ordem e autoridade favorecem a educação? | 18 | Set/15 | 4,8 |
| Escola no Brasil: com partido ou sem partido? | 20 | Ago/16 | 4,7 |
| Forma física, corpo perfeito e consumismo | 20 | Ago/15 | 6,4 |
| Impeachment: a presidente deve perder o mandato? | 20 | Abr/16 | 4,6 |
| Mariana: fatalidade ou negligência? | 19 | Dez/15 | 4,4 |
| O sucesso vem da escola ou do esforço individual? | 18 | Out/15 | 3,7 |
| Por que o Brasil não consegue vencer o *Aedes aegypti*? | 20 | Jan/16 | 4,0 |

Fonte: O autor

Além disto, na Figura 10 está disposta a distribuição das redações de acordo com suas notas em 5 intervalos diferentes. É possível perceber que a grande maioria das redações estão concentradas no intervalo entre 2,1-4.

Figura – Distribuição das Notas



Fonte: O autor

* 1. Métricas de Avaliação

Nesta seção serão descritos em detalhes as métricas utilizadas para analisar os resultados obtido.

* + 1. Coeficiente de Correlação Pearson

Moore (2007) define o Coeficiente de Correlação Pearson (CP) como “A correlação mensura a direção e o grau da relação linear entre duas variáveis quantitativas”. Dalson e José (2010) também o definem mais simplificadamente como “uma medida de associação linear entre variáveis”. Este coeficiente é calculado pela fórmula:

Onde é o valor de similaridade obtido, e é o valor desejado.

O CP varia entre -1 e 1, onde 0 indica que não há correlação alguma entre as variáveis, e os valor entre 0 e os extremos (-1 e 1), indicam a força da relação das variáveis, onde o sinal indica a direção positiva ou negativa. (DALSON e JOSÉ, 2010)

* + 1. Acurácia

Monico et al. (2009) define acurácia como “o termo utilizado para indicar a qualidade de uma grandeza observada ou parâmetro estimado”.

Mikhail e Ackermann (1976) apresentam acurácia como sendo o grau de proximidade de uma estimativa com seu parâmetro (ou valor verdadeiro). E acrescentam que acurácia reflete a proximidade de uma grandeza estatística ao valor do parâmetro para o qual ela foi estimada.

A Acurácia é calculada através da seguinte formula:

Onde VP e VN significam, respectivamente, Verdadeiros Positivos e Verdadeiros Negativos, ou seja, a quantidade de Positivos e Negativos que foram classificados corretamente, e FP e FN significam, respectivamente, Falso Positivo e Falso Negativo, ou seja, a quantidade de Positivos e Negativos que foram classificados incorretamente.

* 1. Avaliação do Módulo de Similaridade

Este trabalho propõe o uso de medidas de similaridade para a análise de fuga ao tema.

A avaliação seguiu o seguinte protocolo: para cada tema, seu texto base foi comparado ao texto de cada redação. Idealmente as redações associadas ao tema teriam um alto grau de similaridade, enquanto as outras não.

Para a análise do método baseado em matriz, foi comparado o grau de similaridade do texto base com a redação, e da redação com o texto base. Já usando o serviço de similaridade da UMBC, foi feita apenas uma comparação do texto base com a redação, visto que a ordem da comparação não altera o resultado.

Nos 3 métodos a saída é um número entre 0 e 1, onde 1 indica textos iguais, e quanto mais próximo a zero, maior a diferença entre os textos. Logo, um limiar precisa ser definido para indicar que a partir dele os textos são parecidos o bastante para indicarem que tratam do mesmo tema.

O maior valor gerado entre os textos da redação e o texto base foi 0,39, então foram testados limiares de 0,05 a 0,35, em incrementos de 0,05 para determinar com qual deles se obtém uma maior acurácia. Os resultados destes testes com os diferentes limiares para cada medida estão dispostos na Tabela 6.

Tabela – Acurácia por Medida com Diferentes Limiares

|  |  |  |
| --- | --- | --- |
| **Medida** | **Limiar de Similaridade** | **Acurácia** |
| Swoogle | 0,05 | 0,20 |
| 0,10 | 0,23 |
| 0,15 | 0,44 |
| 0,20 | 0,54 |
| **0,25** | **0,67** |
| 0,30 | 0,17 |
| 0,35 | 0,10 |
| Matriz | 0,05 | 0,08 |
| 0,10 | 0,15 |
| 0,15 | 0,24 |
| 0,20 | 0,30 |
| **0,25** | **0,44** |
| 0,30 | 0,14 |
| 0,35 | 0,04 |

Fonte: O autor

Como pode ser visto, em ambas as medidas o limiar de 0,25 foi o que produziu a melhor acurácia ao determinar se a redação se encontrava ou não dentro do tema. Porém o Swoogle apresentou um melhor resultado e desempenho, com uma acurácia de 0,67, 33% maior que o melhor resultado da Matriz, que foi 0,44.

* 1. Avaliação do Módulo de Coesão

Na avaliação do Módulo de Coesão, cada redação foi submetida às 6 técnicas de medida de coesão: Flesch Reading Ease, Flesch-Kincaid Grade Level, Gunning-Fog Index, Coleman-Liau Index, SMOG Index e Automated Readability Index; seus valores foram associados ao valor do critério 4, que é o critério associado com a coesão textual.

Então utilizamos o CP para analisar o nível de correlação entre os valores de cada medida e o valor do Critério 4 da avaliação do ENEM. Seu valor para cada medida de legibilidade está disposto na Tabela 7.

Tabela – Relação entre Técnica de Coesão e o Coeficiente Pearson

|  |  |
| --- | --- |
| **Técnica de Coesão** | **CP** |
| Flesch Reading Ease | **0,2385** |
| Flesch-Kincaid Grade Level | 0,1363 |
| Gunning-Fog Index | 0,1738 |
| Coleman-Liau Index | 0,1930 |
| SMOG Index | 0,0290 |
| Automated Readability Index | 0,0304 |

Fonte: O autor

De acordo com o CP, a melhor técnica para calcular a coesão textual é o Flesch Reading Ease, pois é a que tem não só maior valor absoluto, como positivo, com um valor de 0,2385. Já com o valor mais próximo a 0, mostrando uma quase completa falta de correlação, o SMOG Index foi a pior técnica.

Os resultados ainda não possuem uma forte correlação com o critério 4 pois além da coesão, ele também leva em consideração relações de discurso e estruturação do texto. Além disso, inicialmente estas técnicas foram propostas para avaliar textos em inglês, logo o fato delas estarem sendo aplicadas sobre textos em português pode ter influenciado.

1. Conclusões

Uma grande quantidade de novas redações vem sendo produzida anualmente no Brasil devido ao ENEM, pois ele requer o desenvolvimento de uma redação para compor a nota final do aluno. Diante disto mostra-se necessário o desenvolvimento de uma ferramenta que possa auxiliar não só aos professores durante a correção, reduzindo custos, tempo gasto na correção e variações subjetivas de pontuação, como também aos alunos durante o desenvolvimento da mesma.

A ferramenta que este trabalho propõe visa disponibilizar um ambiente onde o aluno possa ir criando sua redação enquanto recebe *feedback* sobre várias partes que integram a nota da mesma, como a coesão, gramática e adesão ao tema. A ferramenta está disponível para download[[14]](#footnote-14).

As medidas para análise de fuga ao tema apresentaram uma taxa de acerto de 67%. Por outro lado, as medidas de coesão atingiram, no máximo, uma correlação de 0,2385 com as notas atribuídas às notas do ENEM.

* 1. Limitações

Algumas limitações deste trabalho envolvem o Banco de Dados da UOL, visto que eles possuem poucas redações para cada tema, 20, onde seria necessário um volume maior para uma análise mais precisa sobre as medidas de adesão ao tema. Ele também não possui uma padronização para a exibição dos erros, usando apenas *tags* HTML sem significado semântico e sem estrutura no texto.

O serviço de similaridade do Swoogle, mesmo apresentando bons resultados na análise dos textos em português, foi inicialmente desenvolvido para textos em inglês, onde ele possui uma base própria de palavras e sinônimos para aumentar ainda mais sua precisão. Similar ao serviço do Swoogle, as técnicas de coesão usadas também foram inicialmente desenvolvidas para a análise de textos em inglês, onde várias delas usam a quantidade de sílabas em seus cálculos, cuja distribuição varia bastante entre os idiomas.

* 1. Trabalhos Futuros

Este trabalho teve como foco uma abordagem parcial de algumas das competências exigidas pela redação do ENEM.

Em trabalhos futuros, pode-se ampliar esta abordagem, seja focando também nas outras competências parcialmente ou tentar propor uma análise integral das mesmas. Pode-se também propor ou utilizar outras técnicas para a análise de coesão para textos em português, podendo até mesmo adaptar as já existentes em inglês. Também é possível a análise de mais medidas de similaridade, com diferentes limiares e um banco de dados mais extenso. Por fim, pode-se também utilizar técnicas de extração de tópicos para avaliar a fuga ao tema.

REFERÊNCIAS

BAZELATO, B. S.; DE AMORIM, E. C. F. A Bayesian Classifier to Automatic Correction of Portuguese Essays. 2010.

BLEI, D. M.; NG, A. Y.; JORDAN, M. I. Latent dirichlet allocation. **Journal of machine Learning research**, v. 3, n. Jan, p. 993-1022, 2003.

BRASIL. Portaria MEC Nº 438, 28 de maio de 1998.

BRASIL. Portaria Normativa Nº 2, 26 de janeiro de 2010.

CROSSLEY, S. A.; KYLE, K.; MCNAMARA, D. S. The tool for the automatic analysis of text cohesion (TAACO): Automatic assessment of local, global, and text cohesion. **Behavior research methods**, v. 48, n. 4, p. 1227-1237, 2016.

DALE, E.; CHALL, J. S. The concept of readability. **Elementary English**, v. 26, n. 1, p. 19-26, 1949.

DALSON, B.; JOSÉ, A. Desvendando os Mistérios do Coeficiente de Correlação de Pearson (r). **Revista Política Hoje-ISSN: 0104-7094**, v. 18, n. 1, 2010.

DING, L.; FININ, T.; JOSHI, A.; PAN, R.; COST, R. S.; PENG, Y.; REDDIVARI, P.; DOSHI, V.; SACHS, J. Swoogle: A semantic web search and metadata engine. In: **Proc. 13th ACM Conf. on Information and Knowledge Management**. 2004. p. 10.1145.

EPSTEIN, D.; REATEGUI, E. Uso de mineração de textos no apoio à compreensão textual. **RENOTE: revista novas tecnologias na educação [recurso eletrônico]. Porto Alegre, RS**, 2015.

FELDMAN, R.; SANGER, J. **The text mining handbook: advanced approaches in analyzing unstructured data**. Cambridge university press, 2007.

FERREIRA, R.; LINS, R.D.; SIMSKE, S.J.; FREITAS, F.; RISS, M. Assessing sentence similarity through lexical, syntactic and semantic analysis. **Computer Speech & Language**, v. 39, p. 1-28, 2016.

HAN, L.; KASHYAP, A.; FININ, T.; MAYFIELD, J.; WEESE, J. UMBC\_EBIQUITY-CORE: Semantic Textual Similarity Systems. In: **\* SEM@ NAACL-HLT**. 2013. p. 44-52.

HOTHO, A.; NÜRNBERGER, A.; PAAß, G. A brief survey of text mining. In: **Ldv Forum**. 2005. p. 19-62.

KINCAID, J. P.; FISHBURNE JR, R. P.; ROGERS, R. L.; CHISSOM, B. S **Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel**. Naval Technical Training Command Millington TN Research Branch, 1975.

KINOSHITA, J.; SALVADOR, L. N.; MENEZES, C. E. D. CoGrOO–Um Corretor Gramatical para a língua portuguesa, acoplável ao OpenOffice. In: **Proc. of Latin American Informatics Conf., Cali, Colombia**. 2005.

KINOSHITA, J.; SALVADOR, L. N.; MENEZES, C. E. D. CoGrOO: a Brazilian-Portuguese Grammar Checker based on the CETENFOLHA Corpus. In: **Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC’2006), Genoa, Italy**. 2006.

KLEMANN, M.; REATEGUI, E.; RAPKIEWICZ, C. Análise de ferramentas de mineração de textos para apoio a produção textual. In: **Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)**. 2011.

LARKEY, L. S. Automatic essay grading using text categorization techniques. In: **Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval**. ACM, 1998. p. 90-95.

MCLAUGHLIN, G. H. SMOG grading-a new readability formula. **Journal of reading**, v. 12, n. 8, p. 639-646, 1969.

MEYER, D.; HORNIK, K.; FEINERER, I. Text mining infrastructure in R. **Journal of statistical software**, v. 25, n. 5, p. 1-54, 2008.

MIKHAIL, E.; ACKERMAN, F. Observations and least squares. 1976.

MONICO, J. F. G.; PÓZ, A. P. D.; GALO, M.; SANTOS, M. C. D.; OLIVEIRA, L. C. D. Acurácia e precisão: revendo os conceitos de forma acurada. **Boletim de Ciências Geodésicas**, v. 15, n. 3, 2009.

MOORE, D. S. **The basic practice of statistics**. New York: WH Freeman, 2007.

NEWMAN, D.; LAU, J. H.; GRIESER, K.; BALDWIN, T. Automatic evaluation of topic coherence. In: **Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics**. Association for Computational Linguistics, 2010. p. 100-108.

NG, H. T.; WU, S. M.; BRISCOE, T.; HADIWINOTO, C.; SUSANTO, R. H.; BRYANT, C. The CoNLL-2014 Shared Task on Grammatical Error Correction. In: **CoNLL Shared Task**. 2014. p. 1-14.

NOBRE, J. C. S.; PELLEGRINO, S. R. M. Avaliador Automático de Coesão Textual em Redação Dissertativa - AVAC. In: **Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)**. 2010.

PERSING, I.; NG, V. Modeling Prompt Adherence in Student Essays. In: **ACL (1)**. 2014. p. 1534-1543.

PERSON CANADA. **In Search of Clear Writing: A Use and Assessment of the Fog Index**. 2009.

PINHEIRO, A.; FERREIRA, R.; FERREIRA, M. A. D.; ROLIM, V. B.; TENÓRIO, J. V. S. Statistical and Semantic Features to Measure Sentence Similarity in Portuguese. 2017.

Salton G.; Yang, C.S. On the specification of term values in automatic indexing. **Journal of documentation**, v. 29, n. 4, p. 351-372, 1973.

SENTER, R. J.; SMITH, E. A. **Automated readability index**. CINCINNATI UNIV OH, 1967.

SHERMIS, M. D.; BURSTEIN, J.; HIGGINS, D.; & ZECHNER, K. Automated essay scoring: Writing assessment and instruction. **International encyclopedia of education**, v. 4, n. 1, p. 20-26, 2010.

UMBC**.** Graph of relations project. 2013.

VILLALON, J. J.; CALVO, R. A. Concept extraction from student essays, towards concept map mining. In: **Advanced Learning Technologies, 2009. ICALT 2009. Ninth IEEE International Conference on**. IEEE, 2009. p. 221-225.

WARSCHAUER, M.; WARE, P. Automated writing evaluation: Defining the classroom research agenda. **Language teaching research**, v. 10, n. 2, p. 157-180, 2006.

XUE, B.; FU, C.; SHAOBIN, Z. A study on sentiment computing and classification of sina weibo with word2vec. In: **Big Data (BigData Congress), 2014 IEEE International Congress on**. IEEE, 2014. p. 358-363.

Anexo I ­– Redação com um Alto Grau de coesão textual

Hoje, mais do que nunca, podemos dizer que o Brasil está dividido. É comum em todos os países democráticos, eleitores que têm posições divergentes, mas o que se vê no Brasil são grupos com determinada ideologia em "pé de guerra", defendendo o que pensam ser "o melhor para o Brasil". O fracasso de um é comemorado pelo outro. Sair vestido de vermelho é um chamado à guerra e o que salva o verde e amarelo é a nossa bandeira, caso contrário, seria perigoso usar tais cores. Deixamos de valorizar a essência para valorizar a ideologia: se é convergente somos amigos, se divergente, inimigos.

Isso, claro, chega às escolas e o que fazer? Os professores - e tal profissional deve ser muito respeitado - não podem se calarem enquanto o Brasil porta afora está a beira de uma guerra civil. Não é só nos jornais que podemos ter informações do nosso país, hoje, o Brasil inteiro discute sobre política: em praças, restaurantes, bibliotecas e escolas! O professor não está na escola apenas para ler um livro de Geografia - sem desprezo a disciplina, o professor deve e, sem duvidas, pode desenvolver um senso crítico em seu aluno. Obviamente os pais vão se queixar se a opinião do educador for diferente e sairão pedindo leis que blindem a ideologia de seus filhos, porém, o que se faz em casa? Os pais que acusam professores de fazerem uma "lavagem cerebral" na cabeça dos jovens não o fazem dentro de casa? Sim. O que ocorre dentro das escolas não é pior do que acontece dentro de casa: pais com determinada ideologia política impõe ao seus filhos que sigam a mesma ideia. Para exemplo, se um pai petista tivesse como professor de seu filho o Suplicy, acharia essa lei um absurdo.

Portanto, esse PL só irá jogar sal em uma ferida que está tentando cicatrizar: a divisão do Brasil. Com respeito ao Senador Malta, o Congresso deveria se preocupar em unificar o povo brasileiro e não o contrário. Os professores já são desprezados pelo salário, se ficarem impedidos de opinar o Brasil deixa de ser do povo.[[15]](#footnote-15)

Anexo II ­– Redação com um BAIXO Grau de coesão textual

Atualmente é muito discutido um problema na questão política que pode ter tido início em meados 1930 com o governo provisório de Getulio Vargas que induzia a população de baixa renda através de propostas educacionais.

A situação abordada refere-se a politicagem introduzida nas instituições de ensino, e com isso podemos citar um fato que ocorre principalmente nas escolas públicas que acontece a partir de um programa chamado Bolsa Família e segundo o MEC 95% dos alunos participantes só frequentam a escola para conseguirem desembolsar os benefícios que o governo oferece as famílias carentes.

Outro fato a ser abordado é a questão dos professores introduzirem sua opinião pessoal em questões políticas para favorecer algo ou alguém.

Portanto medidas são necessárias para resolver o impasse e com isso foi criado uma proposta que tem como nome "Escola Sem Partido" onde existe deveres a serem cumpridas pelos educadores. Como cita o filósofo Immanuel Kant "O homem não é nada além daquilo que a educação faz dele" e assim não devemos em hipótese alguma aceitar esses abusos cometidos.[[16]](#footnote-16)

1. http://portal.inep.gov.br/web/enem/conteudo-das-provas [↑](#footnote-ref-1)
2. https://www.infoenem.com.br/wp-content/uploads/2011/12/Redação\_Enem1.pdf [↑](#footnote-ref-2)
3. http://educacao.uol.com.br/bancoderedacoes [↑](#footnote-ref-3)
4. http://educacao.uol.com.br/bancoderedacoes [↑](#footnote-ref-4)
5. https://www.openoffice.org/ [↑](#footnote-ref-5)
6. http://cogroo.sourceforge.net/download/current.html [↑](#footnote-ref-6)
7. http://comunidade.cogroo.org/grammar;jsessionid=3FA0321B4AC366172395B3A6F5C8AFA3 [↑](#footnote-ref-7)
8. https://educacao.uol.com.br/bancoderedacoes/como-participar.jhtm [↑](#footnote-ref-8)
9. https://educacao.uol.com.br/bancoderedacoes/propostas/forma-fisica-corpo-perfeito-e-consumismo.htm [↑](#footnote-ref-9)
10. https://jsoup.org/ [↑](#footnote-ref-10)
11. http://swoogle.umbc.edu/SimService/api.html [↑](#footnote-ref-11)
12. http://www.mang.canterbury.ac.nz/writing\_guide/writing/flesch.shtml [↑](#footnote-ref-12)
13. http://www.readabilityformulas.com/coleman-liau-readability-formula.php [↑](#footnote-ref-13)
14. <https://1drv.ms/f/s!AgaoZPB3qN0FgQ21cyUFcfaS_Ejk> [↑](#footnote-ref-14)
15. https://educacao.uol.com.br/bancoderedacoes/redacoes/um-brasil-democratico-ainda-com-marcas-da-ditadura.htm [↑](#footnote-ref-15)
16. https://educacao.uol.com.br/bancoderedacoes/redacoes/sem-titulo-097.htm [↑](#footnote-ref-16)