UNIVERSIDADE FEDERAL DO PARÁ
INSTITUTO DE CIÊNCIAS EXATAS E NATURAIS
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

Silvério Sirotheau Corrêa Neto

AVALIAÇÃO AUTOMÁTICA DE TEXTOS NA LÍNGUA PORTUGUESA BASEADA EM ATRIBUTOS LINGUÍSTICOS EM QUATRO DIMENSÕES

Belém - PA

2020

Silvério Sirotheau Corrêa Neto

AVALIAÇÃO AUTOMÁTICA DE TEXTOS NA LÍNGUA PORTUGUESA BASEADA EM ATRIBUTOS LINGUÍSTICOS EM QUATRO DIMENSÕES

Tese de Doutorado apresentada ao Programa de Pós-Graduação em Ciência da Computação. Instituto de Ciências Exatas e Naturais. Universidade Federal do Pará.

Orientador: Prof. Dr. Elói Luiz Favero

Belém - PA

2020

Dados Internacionais de Catalogação na Publicação (CIP) de acordo com ISBD
Sistema de Bibliotecas da Universidade Federal do Pará
Gerada automaticamente pelo módulo Ficat, mediante os dados fornecidos pelo(a) autor(a)

C824a Corrêa Neto, Silvério Sirotheau
Avaliação automática de textos na língua portuguesa baseada em atributos linguísticos em quatro dimensões / Silvério Sirotheau Corrêa Neto. — 2020.
150 f. : il. color.

Orientador(a): Prof. Dr. Elói Luiz Favero
Tese (Doutorado) - Programa de Pós-Graduação em Ciência da Computação, Instituto de Ciências Exatas e Naturais, Universidade Federal do Pará, Belém, 2020.

1. avaliação automática. 2. questões discursivas. 3. atributos linguísticos. 4. análise semântica latente . 5. n-gramas. I. Título.

CDD 006.35

UNIVERSIDADE FEDERAL DO PARÁ
INSTITUTO DE CIÊNCIAS EXATAS E NATURAIS
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

SILVÉRIO SIROTHEAU CORRÊA NETO

AVALIAÇÃO AUTOMÁTICA DE TEXTOS NA LÍNGUA PORTUGUESA BASEADA EM ATRIBUTOS LINGUÍSTICOS EM QUATRO DIMENSÕES

Tese de Doutorado apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal do Pará como requisito para obtenção do título de Doutor em Ciência da Computação, defendida e aprovada em 14/02/2020, pela banca examinadora constituída pelos seguintes membros:

Prof. Dr. Eloi Luiz Favero
Orientador – PPGCC/UFPA

Prof. Dr. Jefferson Magalhães de Morais
Membro Interno - PPGCC/UFPA

Prof. Dr. Nelson Cruz Sampaio Neto
Membro Interno - PPGCC/UFPA

Prof. Dr. Tássio Costa de Carvalho
Membro Externo - UFPA

Prof. Dr. Leonel Figueiredo de Alencar
Membro Externo – UFC

Visto:
Prof. Dr. Nelson Cruz Sampaio Neto
Coordenador do PPGCC/UFPA

Nelson Cruz Sampaio Neto
Coordenador do PPGCC
SIAPE: 2659210

Dedico este trabalho à minha família.

Agradecimentos

Agradeço a Deus.

Agradeço ao meu filho João Claudio pela inspiração de lutar todos os dias.

A minha esposa Ana Paula por todo apoio e incentivo nesse período de estudo.

Agradeço aos meus amados pais Maria de Fátima e Valdomiro por terem me incentivado a estudar sempre.

Agradeço a toda minha família, em especial Glayce, Glauce, Gabrielly, Alessandro, Marco, Fredson e aos meus sogros José Claudio e Janete pelo apoio e incentivo.

Minha eterna gratidão ao Prof. Elói Luiz Favero (Mestre Eloi) por mostrar o caminho da luz e do conhecimento.

Agradeço ao Prof. João Carlos pelas orientações, amizade e conhecimento adquirido.

Agradeço a Profa. Simone Negrão de Freitas por todas as revisões e dicas sobre o trabalho.

Agradeço a todos os professores da Pós-Graduação, em especial Prof. Jefferson Magalhães de Morais, Prof. Dr. Nelson Cruz Sampaio Neto e Prof. Dr. Bianchi Serique Meiguins pelo apoio e orientações.

Minha gratidão aos meus amigos do Campus Universitário de Salinópolis em especial a Profa. Midori Makino e ao Prof. Adilson Oliveira do Espírito Santo.

“Tenho em mim todos os sonhos do mundo."

Fernando Pessoa

Resumo

Juntamente com o desenvolvimento da educação à distância surgem ambientes virtuais com métodos de avaliação automática de exercícios. Neste contexto, dada à relevância das questões discursivas no processo avaliativo, emerge a demanda para avaliadores automáticos para esse tipo de questão. Na literatura foram encontrados resultados promissores, bem como sistemas comerciais, de avaliadores automáticos de questões discursivas para língua inglesa. No entanto, na língua portuguesa só existem estudos preliminares. Esta pesquisa foca na avaliação automática de respostas discursivas: respostas curtas (até um parágrafo) e redações (mais de um parágrafo) na língua portuguesa. Nos experimentos foram utilizados: 192 (cento e noventa e duas) respostas curtas de Filosofia oriundas de uma plataforma virtual de ensino; 131 (cento e trinta e um) e 229 (duzentos e vinte e nove) respostas curtas de Biologia e de Geografia de um processo seletivo para ingresso no ensino superior, respectivamente; 1.000 (mil) redações de um processo seletivo para ingresso em concurso público, de nível técnico. A abordagem baseou-se em técnicas de aprendizagem de máquina que segue uma arquitetura *pipeline* com cinco passos: preparação de *corpus*, pré-processamento, coleta de atributos, modelo preditivo e avaliação. No pré-processamento foram explorados técnicas de limpeza/normalização, *stemmer* e remoção de *stopword*. Na coleta de atributos foram exploradas características linguísticas em quatro dimensões: léxica, sintática, semântica e de coerência. Foram coletados e avaliados mais de 140 (cento e quarenta) atributos. Para gerar o escore de cada resposta foi utilizado o classificador *Random Forest.* Como resultado para a prova de Biologia e de Geografia, respectivamente, o sistema alcançou um índice Kappa quadrático (KQ) de 0.72 e 0.76 (Sistema *versus* Humano - *SxH*), contra 0.89 e 0.58 (Humano *versus* Humano - *HxH*). Para as redações foi obtido um índice KQ de 0.68 *SxH* contra *HxH* de 0.56. Assim tanto para respostas curtas (em parte) como para redações superou o índice de concordância obtido entre dois avaliadores humanos. Este trabalho mostrou a influência de mais de uma centena de atributos na avaliação automática das respostas textuais em português. Além disso, este estudo mostra que esta tecnologia está alcançando maturidade para ser utilizada com grandes vantagens nos ambientes virtuais de ensino: baixo custo, *feedback* imediato, libera o professor do trabalho de correção e atende grandes turmas.

Palavras-chave: avaliação automática, questões discursivas, redações, atributos linguísticos, *n*-gramas, análise semântica latente

Abstract

Along with the development of distance education, virtual environments appear with automatic exercise evaluation methods. In this context, given the relevance of discursive questions in the evaluation process, the demand for automatic evaluators for this type of question emerges. Promising results were found in the literature, as well as commercial systems, of automatic evaluators of discursive questions for the English language. However, in the Portuguese language, there are only preliminary studies. This research focuses on the automatic evaluation of discursive responses: short responses (up to one paragraph) and essays (more than one paragraph) in Portuguese. The experiments used: 192 (one hundred ninety-two) short Philosophy responses from a virtual teaching platform; 131 (one hundred thirty-one) and 229 (two hundred twenty-nine) short answers from Biology and Geography from a selection process to enter higher education, respectively; 1.000 (one thousand) essays from a selection process for entering a public tender, at a technical level. The approach based on machine learning techniques follows a pipeline architecture with five steps: corpus preparation, pre-processing, attribute collection, predictive model and evaluation. In the pre-processing, we explored cleaning/normalization, stemmer and stopword removal techniques. In the collection of attributes, linguistic characteristics were explored in four dimensions: lexical, syntactic, semantic and coherence. More than 140 (one hundred forty) attributes were collected and evaluated. To generate the escore for each answer, we used the Random Forest classifier. As a result for the Biology and Geography test, respectively, the system reached a quadratic Kappa index (KQ) of 0.72 and 0.76 (System versus Human - SxH), against 0.89 and 0.58 (Human versus Human - HxH). For the essays, we obtained a KQ index of 0.68 SxH against HxH of 0.56. Thus, both for short answers (in part) and for essays, we managed to overcome the agreement index obtained between two human evaluators. This work showed the influence of more than a hundred attributes in the automatic evaluation of textual responses in Portuguese. In addition, this study shows that this technology is reaching maturity to be used with great advantages in virtual teaching environments: low cost, immediate feedback, frees the teacher from correction work and serves large classes.

Keyword: automatic evaluation, discursive answers, essays, linguistic attributes, n-grams, latent semantic analysis

Lista de ilustração

Lista de Tabelas

Lista de Abreviaturas e Siglas

AAT	Avaliação Automática de Texto
ASAG	*Automatic Short Answer Grading*
AES	*Automated Essay Scoring*
AEG	*Automated Essay Grading*
AWE	*Automated Writing Evaluation*
AEE	*Automated Essay Evaluation*
PLN	Processamento de Linguagem Natural
PEG	*Project Essay Grade*
LSA	*Latent Semantic Analysis*

Sumário

Introdução

1.1 Contextualização

Durante seu percurso escolar, o aluno passa por um processo de avaliação de ensino e aprendizagem ininterrupto, cumulativo e sistemático. Neste tipo de processo, Mohler e Mihalcea (2009) afirmam que um dos aspectos mais importantes do processo de aprendizagem do conhecimento adquirido pelo aluno é a avaliação. Também sustentam a mesma ideia Rodrigues e Araújo (2012) quando dizem que a avaliação desempenha um papel central em qualquer processo educacional, porque é uma maneira de avaliar o conhecimento dos alunos em relação aos conceitos de aprendizagem. A tarefa de avaliação é utilizada como uma ferramenta de *feedback* para os professores, para medir e melhorar a qualidade do processo de aprendizagem quando os resultados não estão sendo alcançados (AMÁLIA et al., 2019). Geralmente, esse tipo de verificação é feito manualmente pelo professor. Com o avanço da tecnologia de informação (TI), vários processos de avaliação podem ser automatizados, ou pelo menos assistidos pela computação, onde o computador passa a ser um auxiliar ao professor no processo avaliativo (BULL; MCKENNA, 2001).

Nos tempos atuais, instituições acadêmicas vêm promovendo cursos abertos na modalidade de ensino à distância (EaD). Cursos esses, acessíveis a qualquer pessoa com acesso à *internet*, tais como *Coursera*, *Udacity*, *OpenClass*, *Edx, Lúmina,* entre outros. Neste contexto, a aplicação de avaliações compostas por questões discursivas tem forte relevância, pois avaliam resultados de aprendizagem do aluno, em particular, a capacidade de escrita e na estruturação do discurso argumentativo (PAGE, 1966; YANG, 2012; LEE, 2014; ZUPANC; BOSNIC, 2015).

No entanto, a tarefa de correção manual de respostas discursivas para um número considerável de alunos é uma atividade demorada, intensiva e dispendiosa (HE; HUI; QUAN, 2009; ZEN; ISKANDAR; LINANG, 2011; ISLAN; HOQUE, 2012; ZUPANC; BOSNIC, 2017). Em muitos casos, essa sobrecarga de trabalho compromete o desempenho e a qualidade do ensino (CARLOTTO, 2002). De modo geral, esse tipo de problema poderia ser resolvido com o aumento do número de professores, uma vez que suas atividades poderiam ser divididas. Porém, esse aumento pode gerar mais custos para as instituições de ensino. Este fato acentua-se ainda mais em um contexto de um processo seletivo de grande abrangência

composto também por questões discursivas. Por exemplo, a prova do Exame Nacional do Ensino Médio (ENEM) que é um processo seletivo nacional para o acesso a cursos de nível superior, com aproximadamente 7 (sete) milhões de candidatos, com provas objetivas e uma prova discursiva para avaliar a habilidade de escrita dos estudantes (INEP, 2018). Quanto custa a correção manual destas provas? Quantos recursos humanos deverão ser necessários para correção? Quanto tempo se gasta para correção destas provas? Neste contexto, o desenvolvimento de abordagens para automatizar a correção de questões discursivas é muito relevante, permitindo que o computador auxilie os avaliadores humanos na tarefa de correção das questões (PEREZ *et al*., 2005a).

Diante destes, cria-se uma relevância no estudo e desenvolvimento de sistemas de **Avaliação Automática de Texto** (AAT). A área AAT há bastante tempo atrai o interesse das comunidades científicas como da área de linguística, da filosofia, da teoria da informação, da matemática e da computação (HATZVASSILOGLOU *et al*., 1999).

No contexto dos Ambientes Virtuais de Aprendizagem (AVA), abordagens sobre avaliação automática de respostas discursivas torna-se uma ferramenta essencial no processo de aprendizagem do aluno, pois é um recurso que possibilita um *feedback* imediato – está sempre disponível, permite múltiplas re-submissões e tem um baixo custo.

Quanto ao tipo de respostas discursivas na área avaliação automática de texto, existem duas principais categorias: curtas e ensaios. Na literatura, existem várias definições para diferenciar uma resposta do tipo curta de uma resposta do tipo ensaio: Siddiqi, Harrison e Siddiqi (2010) definem resposta curta composta "*de algumas frases para três ou quatro sentenças*"; Sukkarieh e Stoyanchev (2009) definem respostas curtas contendo "*algumas palavras até aproximadamente 100 palavras*". Segundo Burrows, Gurevych e Stein (2015) ensaios são definidos por "*dois ou mais parágrafos até várias páginas*". Para este trabalho a palavra ensaio equivale à redação, porém, especificamente o termo redação é usado para ensaios argumentativos com aproximadamente uma página.

Além do comprimento do texto, as respostas da categoria curtas podem ser mais bem definidas com o auxílio de outros critérios. Para Burrows, Gurevych e Stein (2015), o termo "*Automatic Short Answer Grading - ASAG*" expressa a avaliação automática de respostas curtas que compreendem questões que satisfazem cinco critérios:

1. A questão deve exigir uma resposta tipo "*recall*" (relembrar);
2. O estudante deve expor a sua compreensão em linguagem natural;

3. O texto é delimitado entre uma frase até um parágrafo;
4. O foco da avaliação deve ser o conteúdo e não o estilo;
5. E a questão deve ser do tipo aberta, porém objetiva.

O termo AAT para este trabalho associasse a termologia em inglês "*AEE - Automated Essay Evaluation*" (HEARST, 2000), que generaliza o termo *Automatic Essay Assesment* (AEA). O termo AEA foca somente na atribuição de um escore final enquanto que o termo AEE permite a geração de um escore associado a um *feedback* orientador para os estudantes (ZUPANC; BOSNIC, 2015).

Pesquisas sobre avaliação automática de texto ocorrem desde a década de 1960 (PAGE, 1966; HEARST, 2000; NOORBEHBAHANI; KARDAN, 2011). O professor Page (1966) iniciou um dos primeiros trabalhos nesta área denominado *Project Essay Grade* (PEG) com foco em avaliar as habilidades do estilo de escrita. Porém, somente a partir dos anos 1990, com o uso de técnicas de Processamento de Linguagem Natural (PLN) houve um avanço considerável em pesquisas nesta área. Apesar dos avanços, esse tipo de tecnologia ainda não é utilizado em larga escala sobre aplicações reais, devido à incipiência da tecnologia: sistemas ainda não são robustos – podem ser enganados por alunos fraudulentos (ATTALI; BURSTEIN, 2006); educadores desconfiam da tecnologia; e o desempenho do sistema ainda é inferior ao desempenho dos especialistas humanos (HALEY *et al*., 2007).

Como se pode medir a acurácia de um sistema AAT? Muitas respostas de provas discursivas são avaliadas por dois especialistas humanos de forma independente, considerando que a escrita é um tema subjetivo. Tipicamente os avaliadores humanos divergem em muitos dos seus escores. Por exemplo, nas redações do nosso *corpus* uma divergência maior que um ponto se classifica como discrepância e um terceiro avaliador entra para julgar e decidir o escore final. Numa escala de 0 (zero) a 10 (dez), se os avaliadores divergem em média 2 (dois) pontos, se diz que a acurácia Humano *versus* Humano ($H \times H$) entre eles é 0.8 (percentual de concordância entre 0 e 1). De forma similar, se uma mesma prova é avaliada por um sistema e por um especialista humano também se pode medir a acurácia Sistema *versus* Humano ($S \times H$). Um sistema possui bom desempenho quando a acurácia $S \times H$ é próxima de $H \times H$. Na literatura, vários estudos relatam boa aproximação entre $S \times H$ e $H \times H$; em alguns casos os sistemas até superam os humanos, isto é, a acurácia $S \times H$ supera a $H \times H$, como se pode ver nos exemplos da Tabela 1 (um).

Tabela 1 - Acurácias entre *S*×*H* e *H*×*H* na literatura.

Autores	Metodologia	Respostas	*S*×*H*	*H*×*H*	*Métrica**
Landauer et al., 1997	LSA	Ensaio	0.77	0.70	*corr*
Larkey (1998)	Bayesiana, knn e regressão	Curtas	0.88	0.87	*corr*
Burstein e Chodorow (1999)	PLN	Ensaio	0.92	0.75	*agree*
Burstein et al., (2003)	PLN e Regressão	Ensaio	0.86	0.90	*prec*
Mohler e Mihalcea (2009)	LSA	Curtas	0.50	0.64	*corr*
Santos e Favero (2015)	LSA	Curtas	0.83	0.84	*acc*
Palma e Atkinson (2018)	*Random Forest*	Ensaio	0.87	0.75**	*kq*
Zupanc e Bosnic (2017)	*Random Forest*	Ensaio	0.93	0.75**	*kq*

*Correlação (corr), porcentagem de concordância entre as notas produzidas pelos sistemas e as notas atribuídas por especialistas humanos (agree), *precision* (prec), acurácia de erro médio (*acc*) e kappa quadrático (kq).

**resultado da média de 8 (oito) conjuntos de dados dos valores referente a competição *kaggle*.

Fonte: próprio autor (2020).

A Tabela 1 mostra sistemas com avanços significativos, nela a maioria dos sistemas tem acurácia *SxH* superior a *HxH*. Apesar destes avanços significativos, ainda não se tem métodos confiáveis e robustos para serem utilizados em larga escala na correção automática de textos discursivos (BURROWS; GUREVYCH; STEIN, 2015; ZUPANC; BOSNIC, 2016). Em especial se pode falar da falta de estudos de AAT na língua portuguesa.

1.2 Motivação

A avaliação humana de textos toma decisões holísticas sob a influência de várias dimensões, por exemplo, foco no tema, organização do discurso, argumentos, vocabulário, entre outros. Pelos avanços relatados na literatura, pode-se iniciar um trabalho de pesquisa

que resultará num futuro próximo em ferramentas que venham auxiliar o professor na tarefa de avaliação de questões discursivas. Por exemplo, numa prova onde se tem dois avaliadores humanos, um deles pode ser substituído pelo computador. Ou, num ambiente de ensino virtual o computador pode ser o principal avaliador, sendo supervisionado pelo avaliador humano. Assim, o computador pode liberar uma parte do tempo do professor, dando a ele mais disponibilidade para focar em outras questões do processo de ensino e aprendizagem.

Acredita-se apostar também na tecnologia de AAT na evolução e amadurecimento do processo de ensino, onde se apresenta várias vantagens no uso desta tecnologia:

a) *Feedback* imediato para o aluno, mesmo com um grande grupo de estudantes;
b) Estar sempre disponível, permite repetidas re-submissões do estudante (ZUPANC; BOSNIC, 2015);
c) Ser de baixo custo, permite avaliação ainda da resposta parcial e/ou *feedback* da resposta parcial (ZUPANC; BOSNIC, 2015);
d) Permitir avaliação independe do estresse do avaliador, o método de avaliação é mais consistente;
e) Liberar o professor da correção manual, permitindo que o mesmo direcione maior atenção para estudantes com baixa pontuação.

Considerando a necessidade do desenvolvimento de tecnologia no campo da avaliação automática de texto para a língua portuguesa, com uma acurácia próxima a dos avaliadores humanos, define-se a hipótese de pesquisa.

1.3 Hipótese

Em virtude da pequena quantidade de pesquisa em AAT já existente para a língua portuguesa, onde se busca adaptar e ajustar a tecnologia já desenvolvida para sistemas de AAT para a língua inglesa, esta tese tem a seguinte hipótese: *Os resultados sobre AAT da língua inglesa se reproduzem para língua portuguesa, com adaptações da tecnologia. Essa tecnologia de AAT para a língua portuguesa supera a acurácia alcançada entre dois especialistas humanos.*

Os primeiros estudos foram focados nas questões do tipo ensaios, estudos posteriores enfatizaram respostas do tipo curtas (HEARST, 2000). Conforme a literatura (GOMAA;

FAHMY, 2014) a tecnologia para respostas curtas e ensaios é diferente, pois as respostas curtas tem pouco texto, por esse motivo, muitos dos métodos sobre ensaios não se aplicam a respostas curtas, por exemplo, o método *Latent Semantic Analysis* (LSA) (DEERWESTER *et al*., 1990).

Considerando essa hipótese e o estudo feito na revisão da literatura, foram definidas 17 (dezessete) questões de pesquisa, 10 (dez) para respostas do tipo curtas e 7 (sete) para respostas do tipo ensaios.

As questões para respostas curtas são: **(QP1):** Como conseguir um *corpus* (para questões curtas e ensaios) para se iniciar os primeiros experimentos desta pesquisa? **(QP2):** Quais as melhores técnicas de pré-processamento para abordagem de similaridade de texto? O pré-processamento influencia na acurácia final para abordagem de similaridade de texto? **(QP3):** O que é melhor, ter uma única resposta de referência dada por um especialista humano ou compor uma resposta de referência a partir da concatenação das melhores respostas avaliadas? **(QP4):** Sobre medidas de similaridade entre textos, é melhor trabalhar com interseção de conjuntos ou frequência dos termos? Qual é a melhor medida de similaridade de teoria dos conjuntos? **(QP5):** O uso de bigramas minimiza o problema? Combinando bigramas com unigramas se tem uma boa acurácia? **(QP6):** O método de avaliação centrado em atributos de conteúdo alcança a acurácia dos avaliadores humanos? **(QP7):** O pré-processamento influencia na acurácia final para abordagem em dimensões linguísticas sobre respostas do tipo curta? **(QP8):** Quais os melhores atributos preditores para a língua portuguesa em questões de respostas discursivas curtas**? (QP9):** A importância de contribuição dos atributos se repete em diferentes conjuntos de dados nas respostas curtas? **(QP10):** O método de avaliação centrado em atributos entre três dimensões alcança a acurácia dos avaliadores humanos?

Por outro lado, se tem algumas questões de pesquisas com foco na categoria de ensaios: **(QP11-12-13-14, Léxica, Sintática, Semântica e Coerência)** qual a contribuição de cada uma das quatro dimensões? Quais os melhores atributos para cada uma das dimensões? Se explorar também o cruzamento entre as dimensões: **(QP15)** qual a acurácia da combinação das dimensões duas a duas: léxica *x* sintática, léxica *x* semântica, léxica *x* coerência, sintática *x* semântica, sintática *x* coerência e semântica *x* coerência? Além disso, como semântica tem maior influência sendo explorado semântica + léxica + coerência, semântica + sintática + coerência e semântica + léxica + sintática. Nas 1.000 (mil) redações o escore humano é

gerado a partir de três variáveis relacionadas às competências que foram avaliadas nas redações (tema, regras e conteúdo). **(QP16)** por meio da correlação de *Pearson* foi medida a influência de cada uma das dimensões em relação a três variáveis das competências. **(QP17)** por fim, se tem a questão de pesquisa sobre acurácia. Com base nas quatros dimensões, a acurácia final do método alcança a acurácia dos avaliadores humanos?

1.4 Objetivos

1.4.1 Objetivo Geral

Desenvolver um método computacional para a língua portuguesa que permita avaliar automaticamente textos discursivos para dois tipos de questões, respostas curtas e ensaios, focando principalmente em aspectos de atributos linguísticos, considerando quatro dimensões: léxica, sintática, semântica e de coerência.

1.4.2 Objetivos específicos

1. Fazer uma revisão da literatura sobre AAT levantando a tecnologia utilizada na língua inglesa;
2. Criar um *corpus* mínimo para poder fazer os experimentos;
3. Analisar estratégias para avaliação automática de textos discursivos para respostas do tipo curta e ensaio;
4. Desenvolver um método de avaliação automática de texto discursivo considerando aspectos de quatro dimensões de atributos (Léxica, Sintática, Semântica e Coerência); propor uma arquitetura baseada num modelo *pipeline*, tanto para respostas curtas quanto para ensaios;
5. Criar experimentos para testar a tecnologia propostas para respostas curtas, rodando experimentos para clarificar e responder as questões de pesquisa **QP2** até **QP10**;
6. Criar experimentos para testar a tecnologia proposta para ensaios, rodando experimentos para clarificar e responder as questões de pesquisa **QP11** até **QP17**.
7. Avaliar o método proposto, buscando uma acurácia próxima dos avaliadores humanos.

1.5 Metodologia

A metodologia adotada nesse trabalho pode ser dividida em quatro fases:

Fase 1: Revisão da literatura

Inicialmente foi realizada a fundamentação teórica sobre o tema Avaliação Automática de Texto (AAT) para respostas discursivas. Em seguida foram levantadas as áreas de conhecimento neste tema, compreendendo dois tipos de respostas: resposta curta (até um parágrafo) e ensaio (mais de um parágrafo). Também, uma revisão complementar foi realizada para verificar se houve novos trabalhos sobre esse tema, que foram publicados após a conclusão da etapa de revisão da literatura.

Fase 2: Conjunto de dados e arquitetura

A partir do estudo inicial foram definidos os objetivos e estudos relacionados aos conjuntos de dados e a arquitetura da pesquisa. O conjunto de dados foi constituído por duas bases, um voltado para resposta de tipo curta e outro para ensaio. Um dos conjuntos de dados da pesquisa foi constituído em três bases de textos escritos para respostas curtas: uma questão discursiva conceitual de Biologia, uma questão discursiva argumentativa de Geografia e uma questão discursiva argumentativa de Filosofia. E o outro composto por um conjunto de redações.

Partir da revisão da literatura, foi definida uma arquitetura *pipeline* para desenvolvimento que constitui 5 (cinco) etapas: (1) Etapa de preparação do *corpus*, (2) Pré-processamento, (3) Coleta de atributos, (4) Modelo de predição e (5) Avaliação.

Fase 3: Aplicação do processo

Logo após a definição do conjunto de dados e da arquitetura de desenvolvimento, iniciou-se o desenvolvimento de algoritmos para cada fase da arquitetura. Na etapa de **preparação do *corpus*** as respostas das questões foram digitadas e organizadas numa coleção. Cada resposta deverá ter pelo menos um escore de um avaliador humano. Esta etapa faz a leitura de um arquivo no formato *comma-separated-values* (csv) das quais as colunas representam os escores atribuídos por dois avaliadores humanos e os textos.

Na etapa de **pré-processamento** foram implementadas algumas tarefas, tais como: retirada de acentuação, converter maiúscula em minúscula, remoção de pontuação e caracteres especiais, entre outros. Foram utilizadas combinações de técnicas de pré-processamento, como por exemplo, remoção de *stopwords* com remoção de sufixos (*stemming*) e etc.

Na etapa de **coleta de atributos,** foram coletados atributos (características) em quatro dimensões linguísticas. Foram usadas tecnologias de PLN e inúmeras bibliotecas em *Python* específicos para coleta de atributos.

Na etapa **modelo de predição** foi utilizado um algoritmo em aprendizagem de máquina de classificação supervisionado *Random Forest*, que é um algoritmo de fácil implementação e utilização, e que produz bons resultados sem a necessidade de ajuste de parâmetros (FERNÁNDEZ-DELGADO *et al*., 2014). Este algoritmo também mede a importância de cada atributo utilizado. O modelo divide os dados em duas partições: uma de treinamento e outra de teste. Em seguida, o algoritmo constrói várias árvores de decisão a partir da partição de treinamento para gerar o classificador, que executando sobre o conjunto de teste gera os vetores de escore*s* da avaliação do sistema.

Na etapa de **avaliação**, esses vetores de escores do sistema são contrastados com os vetores de escores dos humanos. Nos experimentos foi utilizado o erro médio e o coeficiente *Weighted Cohen's Kappa* para avaliar a acurácia final.

Fase 4: Publicação de artigos e escrita da tese

Após os experimentos realizados, os resultados obtidos foram divulgados em quatro eventos nacionais e um periódico. Assim, proporcionando o desenvolvimento da escrita da tese.

1.6 Contribuição

1.6.1 As principais contribuições desta tese são:

i. Uma contribuição científica relevante dessa pesquisa é a disponibilidade do *corpora* para área AEE (*Automated Evaluation Essay*) e AES (*Automated essay scoring*), na língua portuguesa, devido à grande dificuldade de encontrar um conjunto de respostas já digitalizadas e com a avaliação dos especialistas humanos.

ii. A principal contribuição do estudo foi a coleta e classificação de mais de cento e quarenta atributos em quatro dimensões linguísticas. Relacionando esses atributos com os escores parciais das competências usadas na avaliação humana.

iii. Outra contribuição positiva são os números encontrados na acurácia que mostram que esta área está amadurecendo e já se podem ter avaliadores para auxiliar na tarefa do professor na correção de textos escritos.

1.6.2 Publicações relacionadas com a tese.

i. Sirotheau, S., Santos, J., e Favero, E. **Avaliação automática de respostas textuais curtas por similaridades de n-gramas: refinamentos por regressão linear.** In: Simpósio Brasileiro de Informática na Educação - SBIE, 2018, Fortaleza-CE. XXIX Simpósio Brasileiro De Informática Na Educação. Fortaleza, CE, Brasil. Sociedade Brasileira de Computação. SBC, 2018. p. 1-1973. DOI: 10.5753/cbie.sbie.2018.1433

ii. Sirotheau, S., Favero, E., Santos, J. e Balieiro, R. **LabPy: Laboratório virtual de ensino em Python.** In: Workshop de Ensino em Pensamento Computacional, Algoritmos e Programação, 2018, Fortaleza. Workshops do Congresso Brasileiro de Informática na Educação. Fortaleza, CE: Sociedade Brasileira de Computação. SBC, 2018. p. 1-1147. DOI: 10.5753/cbie.wcbie.2018.749

iii. Sirotheau, S., Santos, J., e Favero, E. **Avaliação Automática de Ensaios, em português, centrada em atributos linguísticos de superfície e de conteúdo.** In: Workshop sobre educação em computação da SBC (WEI-SBC), 2019, BELEM. XXVII Workshop sobre Educação em Computação, 2019. p. 255-265.

iv. Sirotheau, S., Santos, J., Favero, E. e Freitas, S. **Avaliação Automática de respostas discursivas curtas baseado em três dimensões linguísticas**. In: Simpósio Brasileiro de Informática na Educação - SBIE, 2019, Brasília-DF. XXX Simpósio Brasileiro de Informática na Educação. Brasília, DF, Brasil: Sociedade Brasileira de Computação. SBC, 2019. p. 1-1973. DOI: 10.5753/cbie.sbie.2019.1551

v. Sirotheau, S., Santos, J., Favero, E. e Freitas, S. (2019) ***Automated evaluation of short answers using text similarity for the Portuguese language***. Journal of Computer Science. DOI: 10.3844/jcssp.2019.1669.1677

1.6.3 Publicação em andamento:

i. Sirotheau, S., Santos, J., Favero, E. e Freitas, S. (2020). ***Automatic evaluation of essays in Portuguese based on semantic and shallow features.***

1.7 Organização do trabalho

Além deste capítulo de introdução, o texto possui mais seis capítulos e um apêndice.

O **capítulo 2** apresenta a revisão da literatura, listando os principais trabalhos e resultados publicados sobre o tema de pesquisa.

O **capítulo 3** apresenta os conjuntos de dados (corpora) utilizados nos experimentos desta pesquisa.

O **capítulo 4** apresenta a proposta de um método para a pesquisa. Nele são abordadas as etapas utilizadas no desenvolvimento dos experimentos buscando para maximizar a acurácia final dos resultados.

O **capítulo 5** relata os procedimentos, resultados e discussão dos experimentos realizados para as questões de respostas do tipo curtas.

O **capítulo 6** relata os procedimentos, resultados e discussão dos experimentos realizados para as questões de respostas do tipo ensaio.

O **capítulo 7** apresenta as considerações finais, comparando e analisando os resultados obtidos, a importância da pesquisa e trabalhos futuros.

O **Apêndice** apresenta alguns fragmentos de códigos em Python, para ilustrar para o leitor como foi implementada a arquitetura do método e como foram rodados os experimentos.

2 Revisão da literatura para avaliação automática de textos

Avaliação automática de textos (AAT) é um campo específico voltado para avaliação de respostas discursivas em linguagem natural. Estas são classificadas em dois tipos: resposta curta (até um parágrafo) e ensaio (mais de um parágrafo) (BURROWS; GUREVYCH; STEIN, 2015).

2.1 Revisão da literatura

A revisão da literatura se refere a uma análise que sintetiza dados orientados por meio de um protocolo, centrada em questões-chave relacionadas à questão principal de pesquisa (RUSSELL *et al*., 2009; KITCHENHAM, 2004). De acordo com Kitchenham (2004) uma revisão envolve algumas atividades em etapas bem distintas:

- A identificação da pesquisa;
- Os estudos primários;
- A identificação da necessidade de Revisão;
- O protocolo de Revisão; e
- A análise da revisão da literatura.

2.2 Identificação da pesquisa

Este trabalho foca o campo de pesquisa denominado Avaliação Automática de Texto (AAT) de respostas discursivas. AAT com métodos computacionais é um campo bastante específico voltado para avaliação de respostas escritas em linguagem natural. O desenvolvimento desse tipo de campo (sistemas e algoritmos) envolve trabalho sobre questões discursivas e/ou abertas de grande relevância para avaliação do aluno exigindo compreensão do processo de aprendizagem, enfatizando o desempenho dos alunos na escrita, incluindo habilidades de ordem de pensamento, como síntese e análise (MAGNINI et al., 2005; ZUPANC; BOSNIC, 2017; SHERMIS et al., 2002). Dentro das questões com respostas textuais, as técnicas de avaliação que surgiram, ramificaram-se em duas principais subáreas

dependendo da estrutura da pergunta: resposta curta (até um parágrafo) e ensaio (mais de um parágrafo).

2.3 Estudos primários

Avaliação Automática de Texto (AAT) é uma área de pesquisa em andamento desde a década de 1960, quando o professor Ellis Batten Page propôs o primeiro sistema em uma escola de ensino médio (PAGE, 1966). O sistema usava tecnologia que buscava medir indiretamente dois tipos de variáveis, *trins* (variáveis intrínsecas) e *proxes* (proximidade das variáveis entre si) para avaliar a qualidade de um ensaio. Este tipo de abordagem teve fraco desempenho. A partir da década de 1990, o progresso no campo do processamento de linguagem natural (PLN) incentivou os pesquisadores a aplicarem novas técnicas computacionais para extrair outros atributos mais reais da qualidade dos textos, tais como, léxicos (qualidade do vocabulário), de sintaxe (as etiquetas) e semânticos de conteúdo, por exemplo, com *Latent Semantic Analysis* (LSA).

Ao longo do seu desenvolvimento, para a área de AAT várias denominações foram usadas de forma intercambiável, tanto para respostas do tipo curta como para ensaio. Para respostas do tipo curta (*short answer*) o termo mais utilizado é *Automatic Short Answer Grading* (ASAG) (BURROWS; GUREVYCH; STEIN, 2015). Para questão do tipo ensaio (*essay*), os termos *Automated Essay Scoring* (AES) e *Automated Essay Grading* (AEG) gradativamente foram substituídos pelos termos *Automated Writing Evaluation* (AWE) ou *Automated Essay Evaluation* (AEE). O termo "avaliação" dentro das siglas (AWE e AEE) surgiu porque o processo automatizado permite que os alunos recebam *feedback* construtivo sobre seus ensaios (ZUPANC; BOSNIC, 2016).

2.4 Identificação da necessidade de revisão

Boa parte das pesquisas sobre AAT foi conduzida por organizações comerciais que protegem seus investimentos, restringindo o acesso aos detalhes tecnológicos e aos métodos utilizados. Segundo Landauer *et al*., (2003) isto pode ser um estado compreensível no mercado de tecnologia de software altamente competitivo na atualidade, porém cria dificuldades para a confiança pública, para o debate profissional e para a pesquisa acadêmica. Para Zupanc e Bosnic (2017) um dos principais obstáculos para obter progresso nessa área era

a falta de sistemas de avaliação automática com código aberto, o que permitiria uma compreensão de suas abordagens de classificação. Apesar de tudo, alguns sistemas tiveram suas abordagens técnicas publicadas pela comunidade acadêmica, tais como, *Bayesian Essay Test Scoring* (BETSY) (RUDNER; LIANG, 2002) e LightSIDE (MAYFIELD; PENSTEIN-ROSÉ, 2010). Recentemente, a publicação de um livro sobre o tema tornou o campo mais transparente para os pesquisadores: *Handbook of Automated Essay Evaluation* (SHERMIS; BURSTEIN, 2013).

Para a língua inglesa existem sistemas de AAT que são usados em combinação com a avaliação humana em provas que possuem grande credibilidade no mercado, tais como o *Test of English as a Foreign Language* (TOEFL) e o *Graduate Management Admissions Test* (GMAT) (ZUPANC; BOSNIC, 2016).

Para a língua portuguesa os estudos ainda são incipientes, pois foram encontrados apenas 3 (três) trabalhos: Passero; Haendchen e Dazzi (2016), Galhardi *et al*., (2018) e Fonseca *et al*., (2018).

Quanto à acurácia dos sistemas, atualmente, pesquisas demonstram que a AAT possui forte correlação com a avaliação humana (KEITH, 2003; RUDNER; LIANG, 2002; ZUPANC, 2017; AZMI; AL-JOUIE; HUSSAIN, 2019). No entanto, é ainda um campo bem ativo de pesquisa mostrando um crescimento expressivo nos últimos 3 (três) anos. As pesquisas focam em tornar os sistemas mais robustos e com acurácia próxima das dos avaliadores humanos ou até superar a acurácia dos avaliadores humanos (ZUPANC; BOSNIC, 2017; ZUPANC, 2018; SHEHAB; FAROUN; RASHAD, 2018; AZMI; AL-JOUIE; HUSSAIN, 2019).

2.5 Protocolo de revisão

O protocolo de revisão determina um roteiro para apresentar os resultados da análise da revisão, conforme os itens estabelecidos abaixo:

- Especificar os objetivos;
- Definir as questões de pesquisa;
- Definir os critérios de seleção dos artigos;
- Compilar os dados.

Os objetivos desta etapa de revisão de literatura são:

1. Apresentar o estado da arte em AAT focando na metodologia e abordagem dos sistemas, gerando elementos para fundamentar a pesquisa em AAT buscando adaptar a tecnologia para a língua portuguesa;
2. Apresentar uma panorâmica de estado da arte na área AAT, coletando a acurácia dos principais sistemas.

Outra etapa da pesquisa bibliográfica refere-se à definição das **questões para pesquisa (QPP)**, que deverão guiar a leitura dos artigos; no final elas deverão ser respondidas por meio dos resultados levantados. As QPP são as seguintes:

1. Apresentar uma panorâmica de estado da arte na área AAT, considerando publicações por ano, subáreas de pesquisa, país que mais pública, entre outros;
2. Analisar a existência de conjunto de dados público para testar os algoritmos, se existe algum em Português **(QPP1);**
3. Coletar informações sobre número de publicações por ano **(QPP2)**
4. Coletar informações sobre o país do qual o trabalho é relacionado **(QPP3)**;
5. Coletar informações sobre pesquisadores com o maior número de publicações **(QPP4)**;
6. Coletar informações sobre o tipo de pré-processamento **(QPP5)**;
7. Coletar informações sobre o tipo de *feedback* dos trabalhos levantados **(QPP6)**;
8. Coletar informações sobre as métricas de avaliação para acurácia dos trabalhos **(QPP7)**;

Mecanismo de Pesquisa: As palavras usadas na busca de trabalhos foram: "*automated essay evaluation*", "*automatic essay grading*", "*natural language processing*", "*free-text answer*". Para melhorar os resultados da busca foram trabalhadas algumas combinações lógicas entre as palavras utilizadas: (("*automated essay evaluation*" OR "*automatic essay grading*" OR "*free-text answer*") AND ("*natural language processing*")) como exemplo. A busca procurou qualquer parte do trabalho que continha as palavras usando motores de busca de pesquisas científicas como o *Scholar* Google[1] e o *Semantic Scholar*[2];

[1] Ferramenta de pesquisa do Google que permite pesquisar em trabalhos acadêmicos, literatura escolar, jornais de universidades e artigos variados.

[2] Projeto desenvolvido no Instituto Allen para Inteligência Artificial para ser um serviço de pesquisa "inteligente" para artigos de revistas.

Critérios de seleção: Após a busca pelos motores de pesquisas científicas, uma verificação rápida de cada trabalho é efetuada com a leitura do título e resumo de cada um: os falsos positivos foram removidos.

Critérios de Extração dos dados: Após a busca e seleção, os trabalhos coletados foram catalogados; definimos algumas variáveis para extração de dados principais, como, título, autores, onde foi publicado, ano da publicação, resumo, palavras chaves, proposta, observações, área de concentração, ferramentas, metodologia, análise dos resultados, trabalhos futuros e conclusões.

A condução da revisão de literatura levou aproximadamente um ano e meio para sua conclusão. Nesta pesquisa foram obtidos 110 (cento e dez) trabalhos, entre artigos (66), dissertações (18), teses (5) e livros (21). Destes trabalhos foram retirados estudos repetidos, temas não alinhados e estudos não relacionados à área de AAT. Assim, apenas 66 (sessenta e seis) tratam diretamente de experimentos, onde o foco é avaliação automática de textos.

A última fase deste protocolo consistiu em compilar os dados extraídos dos trabalhos relacionados e apresentá-los em forma de tabelas, gráficos ou texto descritivo. Os resultados obtidos foram dados quantitativos e qualitativos que serão apresentados na subseção de análise.

2.6 Análise da revisão

O estudo dos trabalhos relacionados possibilitou uma análise ampla da área de AAT. Em relação a **QPP1** é sobre a existência de conjuntos de dados público, em Português, para testar os algoritmos.

Foi encontrado um *corpus* de redações com avaliação no *link* https://github.com/gpassero/uol-redacoes-xml, porém para uso no estudo surgiram algumas dificuldades, pois para cada tema de redação existia menos de uma centena de redações. Esse número baixo de redações não permitia o treinamento adequado no modelo de arquitetura. Por isso, decidiu-se utilizar um *corpus* de mil redações todas sobre o mesmo tema, que estavam disponíveis no formato manuscrito em *Portable Document Format* (pdf).

Em relação a **QPP2** onde se refere a coletar dados sobre número de publicações por ano, a Figura 1 (um) apresenta o número de publicações desde 1966 até 2018.

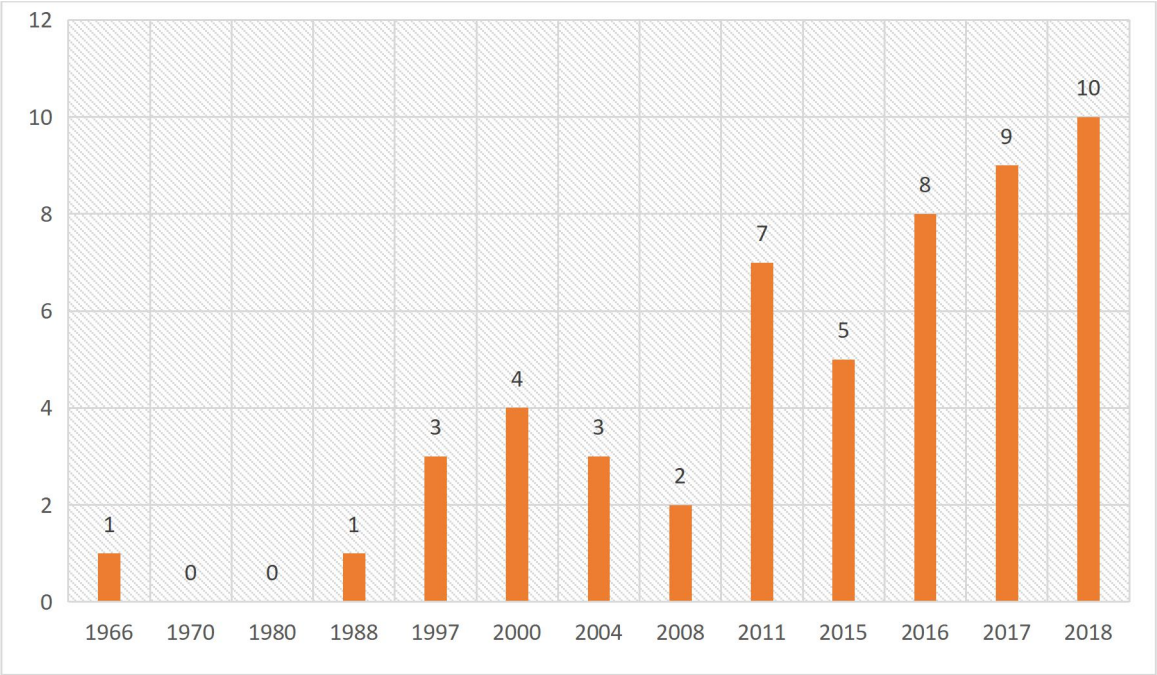


Figura 1 - Publicações anuais sobre AAT.

Percebe-se uma ausência nas publicações desde a primeira publicação de PAGE, (1966) até o início dos anos 1980. Como foi comentado somente com as técnicas de PLN é que a pesquisa foi retomada com certa intensidade. Por outro lado, ocorreu um aumento no número de publicações nos três últimos anos.

A maioria dos trabalhos pesquisados apresentam sistemas (bem como a publicação) para a língua inglesa. Essa análise relaciona-se com a **QPP3** (Coletar dados sobre o país do qual o trabalho está relacionado) conforme apresentado na Figura 2 (dois). Dos países destacam-se os Estados Unidos com 36 (trinta e seis) e a Espanha com 12 (doze) publicações. No Brasil foram encontradas 3 (três) publicações.

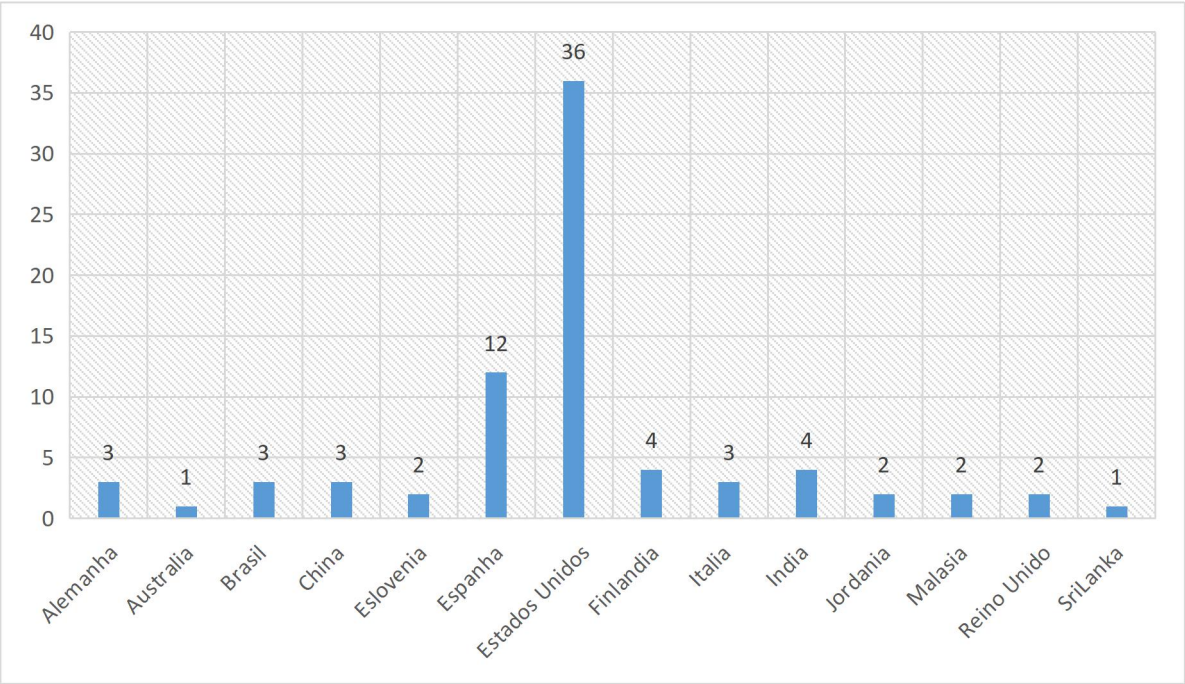


Figura 2 - Publicações por país relacionadas à Avaliação Automática de Texto

Dentro das subáreas de AAT, destacam-se 6 (seis) nomenclaturas: *Automated Essay Evaluation* (AEE), *Automated Essay* Escore (AES), *Automated Essay Grading* (AEG), *Automated Writing Evaluation* (AWE), *Automated Summary Evaluation* (ASE) e *Automated Short Answer Grading* (ASAG). Destaca-se a subárea *Automated Essay Evaluation* (AEE) com 23 (vinte e três) publicações (Figura 3). Acredita-se que esse resultado pode estar relacionado com a mudança gradativa das expressões utilizadas em Avaliações Automáticas, como mencionadas nos trabalhos de (ZUPANC; BOSNIC, 2015; BRENT; ATKISSON; GREEN, 2010).

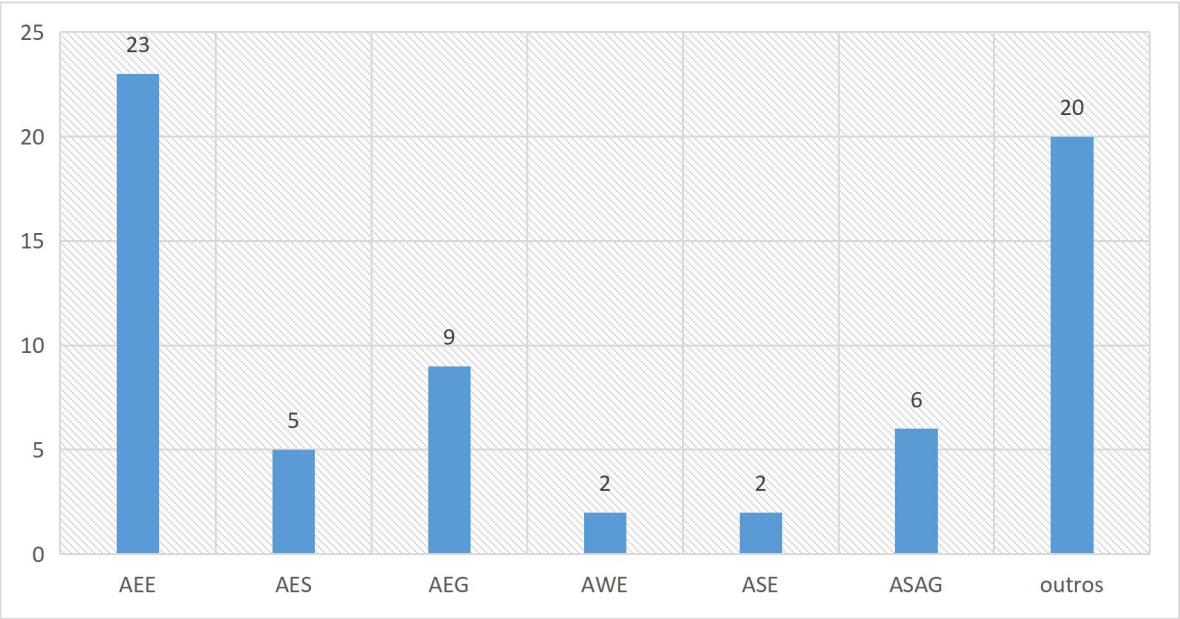


Figura 3 – Nomenclaturas na área de Avaliação Automática de Textos

A Tabela 2 (dois) mostra os pesquisadores com o maior número de publicações (**QPP4**). De 66 (sessenta e seis) trabalhos levantados nesta pesquisa sobre AAT foi verificado que oito pesquisadores concentraram mais de 53% do total destes trabalhos.

Tabela 2 - Pesquisadores mais produtivos em Avaliação Automática de Texto

Pesquisador	Universidade/companhia	Publicações	País
Jill Burstein	*Educational Testing Service* - ETS	10	USA
Martin Chodorow	*Hunter College and The Graduate Center of CUNY*	5	USA
Tuomo Kakkonen	Universidade de Eastern Finland	5	Finlândia
Diana Pérez	Universidade Rey Juan Carlos	3	Espanha
Kaja Zupanc	Universidade de Ljubljana	3	Eslovênia
Zoran Bosnic	Universidade de Ljubljana	3	Eslovênia
Ellis Batten Page	Universidade de Connecticut	3	USA

Fonte: Do próprio autor (2020)

2.7 Técnicas de pré-processamento

A etapa de pré-processamento de textos tem como objetivo melhorar os índices de similaridade, classificação e de eficiência no comparativo dos textos. Dentre os 66 (sessenta e seis) trabalhos, 13 (treze) informaram o tipo de pré-processamento, conforme a Figura 4 (quatro). A utilização de *Stemmer* e a remoção de *stopword* foram às técnicas mais utilizadas, obtendo mais de 53% entre os trabalhos pesquisados.

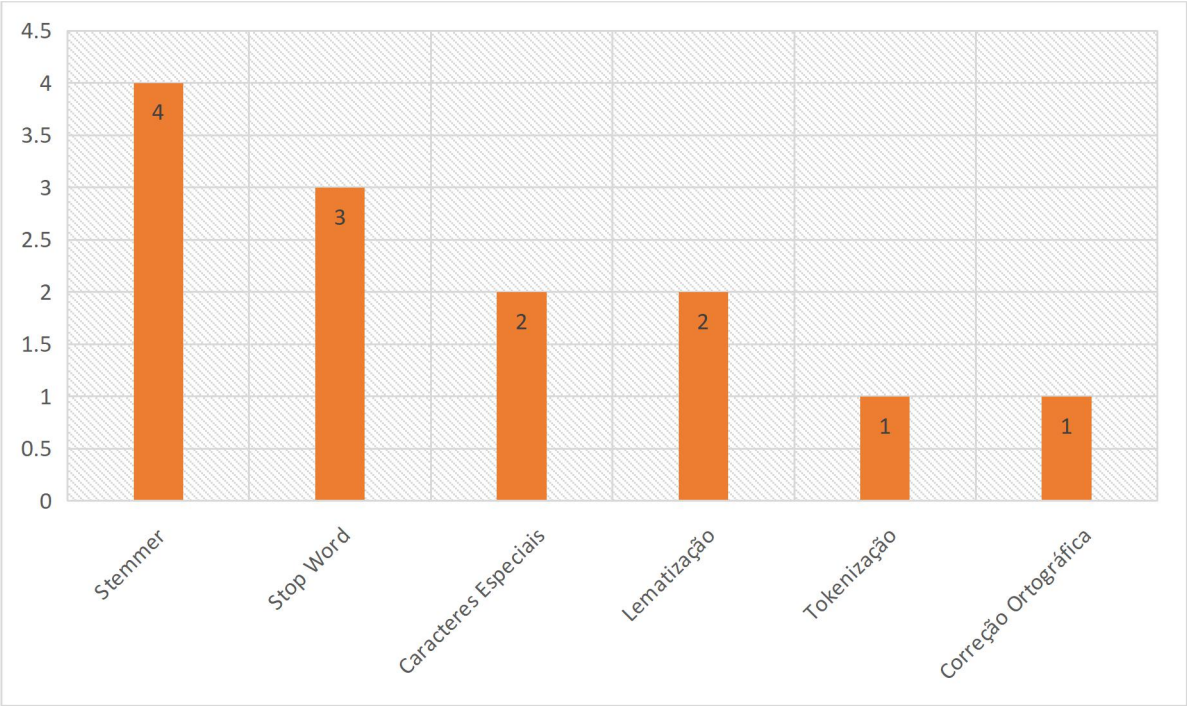


Figura 4 - Métodos de pré-processamento utilizados em AEE

2.8 Tipo de *feedback* dos sistemas

AAT pode dar dois tipos de *feedback*: avaliativo (dar um escore) ou formativo (indicar uma direção) (BRENT; ATKISSON; GREEN, 2010). O processo avaliativo mede o desempenho ao final de um curso com o objetivo de classificação ou certificação; enquanto que a avaliação formativa é conduzida durante todo o processo de ensino com a finalidade de identificar as dificuldades dos alunos e monitorar o progresso da instrução (BROWN; KNIGHT, 1994; TORRANCE; PRYOR, 1998). Os sistemas de avaliação automatizada podem gerar instruções individualizadas para apoiar abordagens do processo avaliativo enfatizando, assim o valor de um ensaio (BURSTEIN *et al.,* 2012). Também os sistemas de AAT podem motivar e orientar o aluno, promovendo uma autonomia do aluno por meio de um *feedback* automatizado (ATTALI; POWERS, 2008).

O *feedback* pode ser bastante complexo e pode ser fornecido com um grande nível de detalhes e explicações. Os sistemas retornam o rascunho do aluno com comentários destacando questões que exigem atenção, que podem estar relacionadas a erros de linguagem, complexidade sintática, variação no tipo de frase, estilo, organização, desenvolvimento de ideias, conteúdo conceitual e outros traços de escrita, conforme apresentado na Tabela 3 (três).

Tabela 3 - Sistemas de avaliação automática com *feedback*

Sistemas	*Feedback*	*Características*
E-rater	*Criterion*	Sintaxe, discurso, conteúdo tópico, complexidade lexical, gramática, uso, mecânica, estilo
IEA	*WriteToLearn*	Ideias, organização, convenções, fluência de frase, escolha de palavras, voz do escritor, ortografia, cópia, redundância, irrelevância
IntelliMetric	*MY Access!*	Foco e significado, organização, conteúdo e desenvolvimento, uso e estilo de linguagem, mecânica
Bookette	*Bookette*	Gramática, ortografia, convenções no nível da sentença
CRASE	*True score and CRASE*	Ideias, organização, voz, escolha de palavras, fluência de frase, convenções

Fonte: do próprio autor (2020)

2.9 Avaliação de acurácia

Existe uma discussão quanto à forma de quantificar a acurácia dos modelos de avaliação automática. No trabalho de Burrows, Gurevych e Stein (2015) coletou-se o uso de três principais medidas: kappa, correlação e erro médio. Os trabalhos de Pribadi *et al*., (2017) e Gomaa e Fahmy (2012) realizam a avaliação da acurácia do método com a medida de correlação, para escore*s* contínuos.

Dentre os 66 (sessenta e seis) trabalhos pesquisados, 20 (vinte) relatam sobre o teste de acurácia. A Figura 5 (cinco) mostra que Correlação de *Pearson* foi à medida de acurácia mais utilizada pelos trabalhos relacionados nesta revisão.

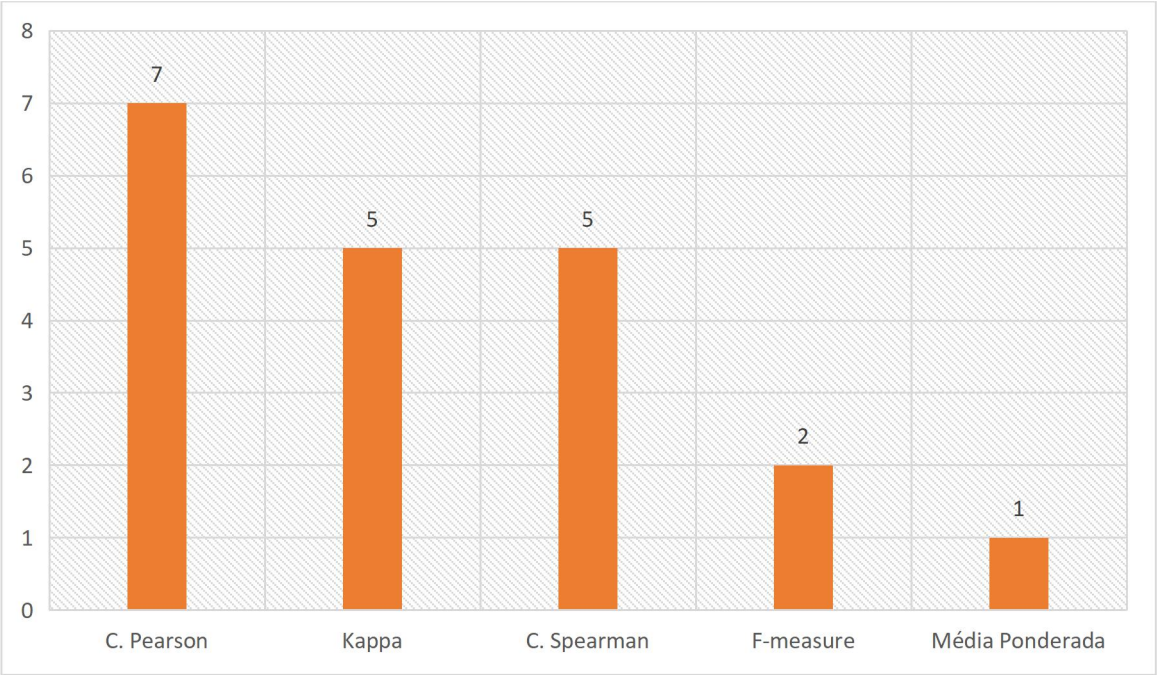


Figura 5 - Medidas de acurácia mais utilizados em AEE.

2.10 Avaliação automática de ensaios

Nesta sessão, detalha-se a revisão do estado da arte sobre avaliação do tipo ensaio. Apresentam-se as metodologias de aplicação, as abordagens, os resultados alcançados e eventuais peculiaridades focando nos objetivos específicos listados no capítulo 1 (um), que são:

- Analisar estratégias para avaliação automática de textos discursivos para respostas do tipo curta e ensaio;
- Desenvolver um método de avaliação automática de texto discursivo considerando aspectos de 4 (quatro) dimensões de atributos linguísticos (Léxica, Sintática, Semântica e de Coerência);
- Avaliar o método proposto, buscando uma acurácia próxima dos avaliadores humanos.

2.10.1 Principais temas sobre avaliação automática de ensaios

Numa revisão de literatura foram encontrados vários temas importantes para serem abordados, entre eles: nomenclaturas da área, engenharia de atributos, método, expansão de vocabulário e métricas de acurácia.

Nomenclatura. Para ensaios, o termo *Automated Essay Scoring* (AES) gradativamente vem sendo substituído pelo termo *Automated Essay Evaluation* (AEE). O termo AEE surge para ser usado porque o processo automatizado permite que os alunos recebam *feedback* construtivo sobre seu ensaio, conforme relata Zupanc e Bosnic (2016).

Engenharia de atributos. Um dos principais temas da área está associado à engenharia de atributos que compreende em um processo de extração do conhecimento dos atributos associados a dimensões linguísticas a serem avaliadas em um texto. No projeto dos atributos deve-se desenvolver um processo de coleta, seleção e avaliação da contribuição de cada atributo na acurácia de avaliação de uma determinada dimensão. Nos níveis mais altos da hierarquia das dimensões se tem estilo, organização, conteúdo, coerência, semântica, sofisticação léxica, gramática, entre outros. Estas dimensões de atributos vão se detalhando em dezenas até várias centenas de características (atributos): 60 (sessenta) atributos (IEA); 90 (noventa) atributos (Bokette); 150 (cento e cinquenta) atributos (SAGE); até mais de 400 (quatrocentos) atributos (IntelliMetric). Trabalhos recentes como SAGE (ZUPANC; BOSNIC 2017; ZUPANC, 2018) descrevem mais de 130 (cento e trinta) atributos classificados nas seguintes dimensões: Linguístico (sofisticação lexical, gramática, mecânica), conteúdo e semântico (coerência do discurso).

Recentemente surgiram também abordagens automáticas de coleta de atributos baseadas em *Convolutional Neural Networks* (DONG; ZHANG 2016; DASGUPTA *et al*., 2018). Estas abordagens têm alcançado e até superado as acurácias do escore final das abordagens de coleta de atributo manuais, porém elas não permitem dar *feedback* para o estudante com o valor de cada atributo ou com o valor da classe de atributos, por exemplo, listar os erros gramaticais ou dizer que um ensaio possui pontuação baixa em sofisticação léxica, direcionando o esforço de aprendizagem do estudante nos pontos fracos da sua avaliação.

Método. Um dos desafios dos sistemas é combinar dezenas e até centenas de atributos num escore final ou num vetor de valores associados às diferentes dimensões a serem avaliadas. Nos primeiros trabalhos a principal abordagem para combinar atributos foi a

regressão múltipla (*Project Essay Grade* – PEG (PAGE, 1966)). Com os avanços da área de aprendizagem de máquina surgiram novas abordagens tais como redes de *bayes* (*Bayesian Essay Test Scoring sYstem* - BETSY (RUDNER; LIANG, 2002)), *random forest* (*Semantic Automated Grader for Essays* – SAGE (ZUPANC; BOSNIC, 2014)), redes neurais (Bokette (RICH; SCHNEIDER; D'BROT, 2013)), modelos híbridos (*IntilliMetric* (SCHULTZ, 2013)), entre outros. Por outro lado, houve uma evolução na metodologia para avaliar a dimensão do conteúdo, com o uso de variações do modelo espaço vetorial como LSA (*Latent Semantic Analysis*), LDA (*Latente Dirichlet Allocation*) e CVA (*Contente Vector Anlaisys*) (ZUPANC; BOSNIC, 2016).

Expansão do vocabulário. O LSA produz melhores resultados quando compara grandes quantidades de texto, o IEA (LANDAUER *et al*., 1997) em versões mais recentes utiliza em sua etapa de treinamento, além de ensaios pré-avaliados, longos textos com conteúdo do domínio como uma forma de expandir seu vocabulário. Outra forma de expansão de vocabulário é por meio de recursos como *WordNet* (MILLER, 1998), uma espécie de banco de dados lexical para inglês gerando-se paráfrases a partir dos sinônimos.

Métricas para avaliação da acurácia. Existem várias abordagens para validação da acurácia, as três mais utilizadas são: correlação (de *Pearson* ou de *Spearman)* para escore contínuo ou numérico (RUDNER; GARCIA; WELCH, 2006; MOHLER; MIHALCEA, 2009; RABABAH; AL-TAANI, 2017; KE; NG, 2019); erro médio para faixa de valores (SANTOS; FAVERO, 2015); e kappa quadrático (ZUPANC; BOSNIC, 2017; PALMA; ATKINSON, 2018). Esta última vem se consolidando como a mais utilizada.

2.10.2 Revisão dos sistemas automatizados de avaliação de ensaios

No campo da AEE existem muitos desafios em aberto, pois ela considera a avaliação de uma forma abrangente, desde a emissão de um escore até a geração de um *feedback* mais detalhado, considerando aspectos como verificação do estilo, gramaticalidade, coerência, entre outros (HIGGINS *et al*., 2004; PALMA; ATKINSON, 2018; ZUPANC; BOSNIC, 2017). Nas subseções seguintes, será apresentada uma revisão dos sistemas de AEE, apresentando as características conhecidas de sistemas proprietários, bem como de dois sistemas com abordagens disponíveis para a comunidade acadêmica.

A Tabela 4 (quatro) apresenta a lista de sistemas revisados destacando algumas de suas características. Na primeira coluna se tem o nome do sistema e data de sua criação. Na

segunda coluna se tem a classe dos atributos que podem ser divididos em quatro grupos: estilo, conteúdo, discurso e expansão do vocabulário. A terceira coluna foca nos métodos usados na coleta dos atributos, que compreendem estatístico de superfície (contagem de palavras), estatístico (não só de superfície), PLN, LSA, etc.; a PLN permite identificar nomes de entidades e relações associadas às etiquetas das palavras das frases. A quarta coluna mostra o modelo de predição: regressão linear múltipla, aprendizagem de máquina, modelos matemáticos, redes neurais, Medida *Lexile* e *Random Forest*. A quinta coluna mostra o tamanho da base onde o sistema foi testado. Por fim, as três últimas colunas apresentam a acurácia humano contra humano, acurácia sistema contra humano e métricas aplicadas, servindo para verificar se os sistemas alcançam a acurácia humana.

Na tabela pode-se ver que existem sistemas onde o valor de *SxH* já supera a o valor de *HxH*, isto é, a acurácia do sistema já supera a acurácia humana, por exemplo, o PEG em 2001 apresentou uma correlação de 0.87 superando os avaliadores humanos (*HxH*) com 0.73. Apesar disso, a correlação não é uma medida muito precisa, pois, dois valores em escala diferente podem ter a mesma correlação. O pior desempenho é do sistema *Lexile*, apresentou uma Kappa Quadrático (KQ) de 0.63, valor considerado razoável na tabela de referência do KQ (ver Tabela 17). O melhor desempenho é do sistema SAGE, com um valor KQ de 0.93, considerado um ótimo desempenho.

Tabela 4 – Sistemas informatizados para avaliação automática de ensaio na literatura.

Sistema	Classe	Método	Modelo	Base	HxH*	SxH	Métrica (SxH)
PEG (1966-2001)	Estilo	Estatístico Superfície	Regressão Linear Múltipla	100-400	0.73	0.87	*corr*
E-rater (1998-2006)	Estilo e Conteúdo	PLN	Regressão Linear Múltipla	270	0.94	0.77	*agreem*
IEA (1999-2003)	Conteúdo+ Expansão Vocabulário	LSA e PLN	Machine Learning	200-500	0.88	0.73	*kq*
IntelliMetric	Estilo+ Conteúdo	LSA+PLN	Modelos matemáticos + Machine	300	0.95	0.76	*kq*

(1999-2015)	+Discurso		Learning				
Bookette (2013)	Estilo+ Discurso	PLN+ Estatístico	Rede Neural	250-500	-	0.70	*kq*
CRASE (2013)	Estilo+ Discurso	PLN+ Estatístico	Machine Learning	-	-	0.75	*kq*
AutoEscore (2013)	Estilo+ Discurso	PLN+ Estatístico	Modelos Estatísticos	-	-	0.73	*kq*
Lexile (2009-2014)	Estilo+ Discurso	PLN+ Estatístico	Medida Lexile	-	-	0.63	*kq*
LightSIDE (2010-2013)	Estilo+ Discurso	PLN+ Estatístico	Machine Learning	918-1805	0.85	0.75	*corr*
SAGE (2017)	Estilo+ Conteúdo +Discurso	LSA+PLN + Estatístico	Random Forest	12.978	0.75	0.93	*kq*

Fonte: Adaptado de (SHERMIS; BURSTEIN, 2003; SHERMIS; HAMNER, 2012; ZUPANC; BOSNIC, 2017) e próprio autor (2020).

*acurácia dos resultados (*acc*), correlação de regressão múltipla (*corr*) e porcentagem de concordância entre as notas produzidas pelos sistemas e as notas atribuídas por especialistas humanos (*agreem*), kappa quadrático (*kq*).

**medidas com correlação (*corr*)

Complementando os dados apresentados na Tabela 4 (quatro), na sequência apresenta-se uma breve descrição de cada um dos sistemas mencionados, destacando seus principias aspectos tecnológicos.

O *Project Essay Grade* (PEG) é um sistema AES desenvolvido na *Measurement* Inc. (PAGE, 1966). Passou por inúmeras melhores até possuir uma interface gráfica via *web* (SHERMIS *et al*., 2001). O PEG foca na análise de estilo das características linguísticas de um texto (qualidade da escrita), sem levar em conta o seu conteúdo. O sistema analisa e pontua as características linguísticas da superfície do ensaio por meio das medições chamadas "*trins*" e "*proxes*". Um "*trin*" é uma variável de nível essencial no texto, como por exemplo,

pontuação, dicção (variação no comprimento da palavra), gramática (partes do discurso, estrutura da frase) entre outros. Os "*trins*" não podem ser medidos diretamente, precisando ser aproximado das medidas chamadas "*proxes*". A pontuação da "*trins*" é medida por meio do número de "*proxes*" de erros de pontuação e do número de diferentes pontuações usadas. A abordagem utilizada neste sistema foi estatística, um modelo de previsão de análise de regressão. Ele usou um conjunto de treinamento de 100 (cem) a 400 (quatrocentos) ensaios, alcançando uma acurácia de correlação 0.87.

E-rater é um sistema de pontuação automática de ensaios desenvolvido por (BURSTEIN *et al*., 1998) de propriedade da *Educational Testing Service* (ETS). O sistema identifica e extrai várias classes de atributos de um texto usando métodos de PLN e estatísticos fundamentados em regras (ATTALI; BURSTEIN, 2006; BURSTEIN *et al*., 2004). O mecanismo de pontuação é projetado para identificar características no ensaio do aluno que refletem características especificas nos guias de pontuação do leitor. Os leitores devem ler rapidamente para uma impressão total e levar em consideração a variedade sintática, uso da gramática, mecânica (refere-se às regras da linguagem escrita, como maiúsculas, pontuação e ortografia) e estilo, organização e desenvolvimento e uso de vocabulário (BURSTEIN; CHODOROW; LEACOCK, 2003).

Desde a sua concepção, o sistema evoluiu e atualmente está incorporado a outro denominado *Criterion*, uma versão em tempo real baseada na *web*. O sistema possui um módulo de *feedback* para melhorar a habilidade de escrita; foi treinado em uma coleção de 270 (duzentos e setenta) ensaios que foram marcados manualmente por avaliadores humanos. Para atribuir uma pontuação final ao ensaio, o sistema usa um modelo de regressão incluindo a detecção de similaridade de ensaios e avisos que indicam se um ensaio está fora do tópico; ele alcançou uma acurácia de concordância de 0.77 (BURSTEIN *et al*., 2012).

O *Intelligent Essay Assessor* (IEA) também é um sistema de AEE. Ele baseia-se em uma abordagem que combina LSA e PLN para produzir uma pontuação geral (FOLTZ; LAHAM; LANDAUER, 1999). Ele compreende um modelo de aprendizagem mecânica que induz a semelhança semântica de palavras e pontos por análise de grandes *corpora* de texto sendo relevante para o domínio (LANDAUDER *et al*., 2003). O IEA concentra-se mais em conteúdo do que em qualidade de escrita; possui um módulo de *feedback* chamado *Write To Learn*, baseado num modelo de previsão centrado em técnicas de aprendizagem de máquina; ele alcança uma acurácia KQ de 0.73 (SHERMIS; HAMNER, 2012).

O *Intellimetric* foi projetado e lançado em 1999. Ele é um sistema proprietário para pontuar ensaios (ELLIOT, 2003; SCHULTZ, 2013). O sistema analisa elementos do discurso para formar um sentido de significado composto. Esses elementos podem ser divididos em duas categorias: a) conteúdo - discurso/retórica e atributos de conteúdo/conceito; e b) estrutura de atributos sintático-estruturais e mecânicos. Os atributos de conteúdo avaliam o tópico abordado, a amplitude do conteúdo, suporte para conceitos avançados, coesão, consistência em propósito, tema e lógica do discurso. Os atributos da estrutura avaliam a gramática, a ortografia, a completude da sentença, a pontuação, a variedade sintática, a complexidade da sentença, o uso, a legibilidade e concordância do sujeito e verbo (SCHULTZ, 2013; ZUPANC; BOSNIC, 2016). O sistema usa múltiplas abordagens inteligentes com base em vários modelos matemáticos, incluindo análise linear, analise bayesiana e LSA para prever o resultado combinando os modelos em um único desfecho de ensaio final (RUDNER; GARCIA; WELCH, 2006). O sistema alcança uma acurácia KQ de 0.76 (SHERMIS; HAMNER, 2012).

O *Bookette* (RICH *et al*., 2013) foi projetado pelo *California Testing Bureau* (CTB). O Bookette usa técnicas de processamento de linguagem natural para extrair 90 (noventa) atributos que descrevem a qualidade do texto produzido pelo aluno. As combinações desses atributos descrevem características da escrita como organização, desenvolvimento, estrutura de sentenças, escolha de palavras/uso gramatical e mecânica. O sistema usa redes neurais para modelar pontuações avaliadas por especialistas humanos, podendo criar modelos específicos imediatos, bem como modelos genéricos que podem ser muito úteis nas salas de aula para fins formativos. O sistema treina seu algoritmo num conjunto de 250 (duzentos e cinquenta) a 500 (quinhentos) ensaios com pontuação humana. O sistema fornece *feedback* que reproduz o desempenho dos atributos, inclui também *feedback* holístico e comentários sobre as convenções de gramática, ortografia. (RICH *et al*., 2013). O sistema alcança uma acurácia KQ de 0.70 (SHERMIS; HAMNER, 2012).

O sistema CRASE é de propriedade da *Pacific Metrics* (LOTTRIDGE *et al*., 2013), passa por três fases no processo de pontuação: identificação de tentativas inadequadas, extração de atributos e pontuação. A etapa de extração de atributos é organizada em torno de seis características da escrita: ideias, fluência de frases, organização, voz, escolha de palavras, convenções e apresentação escrita. O sistema analisa uma amostra das respostas dos alunos já pontuadas para produzir um modelo do comportamento de pontuação dos alunos. CRASE é

um aplicativo baseado na linguagem *Java* que é executado como um serviço da *web*. O sistema é personalizável em relação às configurações usadas para criar modelos de aprendizado de máquina, bem como a combinação de pontuação humana e de máquina (ou seja, modelos híbridos derivados) (LOTTRIDGE *et al*., 2013). O aplicativo também produz comentários baseados em texto e numéricos que podem ser usados para melhorar os ensaios. O sistema alcança uma acurácia KQ de 0.75 (SHERMIS; HAMNER, 2012).

O *Autoscore* é um sistema AEE proprietário, projetado pelo *American Institute for Research* (AIR). O sistema analisa medidas baseadas em conceitos que discriminam documentos com notas altas e baixas, medidas que indicam a coerência de conceitos dentro e entre parágrafos e uma variedade de medidas de uso de palavras. Detalhes sobre o sistema nunca foram publicados, no entanto, o sistema foi avaliado em (SHERMIS; HAMNER, 2012) alcançando uma acurácia KQ de 0.73.

O *Lexile Writing Analyzer* (SMITH, 2009) foi desenvolvido pela *MetaMetrics*. O sistema é independente de pontuação, gênero, *prompt* e pontuação; ele utiliza a medida do escritor *Lexile*, que é uma estimativa da capacidade do aluno de expressar o idioma por escrito, com base em fatores relacionados à complexidade semântica e sofisticação sintática (como as palavras são escritas em frases). O sistema usa um pequeno número de atributos que representam aproximações para a capacidade de escrita. *Lexile* percebe a capacidade de escrever como uma característica individual subjacente. A fase de treinamento não é necessária, uma vez que é empregada uma escala vertical para medir os ensaios dos alunos (SMITH *et al*., 2014). O sistema alcança uma acurácia KQ de 0.63 (SHERMIS; HAMNER, 2012).

Mayfield e Rosé (2013) lançaram o sistema *LightSIDE* (MAYFIELD; PENSTEIN-ROSÉ, 2010), um mecanismo de avaliação automatizado com código compilado e código-fonte disponível publicamente. O *LightSIDE* foi desenvolvido como uma ferramenta para que não especialistas usem efetivamente a tecnologia de mineração de texto para uma variedade de propósitos, incluindo avaliação de ensaios. Permite escolher o conjunto de atributos e algoritmos para construir o modelo de previsão (por exemplo, regressão linear, *Naïve Bayes*, máquinas de vetores de suporte linear). O conjunto de atributos é focado principalmente em *n*-gramas, etiqueta morfossintática e atributos de "contagem". No entanto, o sistema permite que os usuários insiram manualmente o código para novos atributos. O sistema treina seu

algoritmo num conjunto entre 918 (novecentos e dezoito) a 1805 (mil oitocentos e cinco) ensaios com pontuação humana e o sistema alcança uma acurácia de correlação 0.75.

Em seu trabalho (ZUPANC; BOSNIC, 2017) propõem uma extensão de um sistema de avaliação de ensaio automatizado existente SAGE (*Semantic Automated Grader for Essays*) que incorpora atributos semânticos adicionais. Neste caso, foram projetados novos atributos, transformando janelas sequenciais de um ensaio num espaço semântico e medindo as mudanças entre elas para estimar a coerência do texto. O sistema alcança uma acurácia (KQ) significativamente mais alta (0.93) em comparação com outros 8 (oito) sistemas de avaliação de ensaios automatizados de última geração, treinando seu algoritmo num conjunto de 12.978 (doze mil novecentos e setenta e oito) ensaios e usando 132 (cento e trinta e dois) atributos.

Nesta revisão da literatura foram encontrados três trabalhos voltados para a língua portuguesa, mas com os trabalhos para as outras línguas, principalmente o inglês, se podem estabelecer algumas diretrizes para nosso estudo que é focado em ensaios em português: a tecnologia indica que se pode alcançar uma acurácia próxima dos avaliadores humanos; um dos desafios é conseguir uma base de ensaios com o escore de dois avaliadores humanos; devem-se escolher dimensões linguísticas e definir e/ou reusar os inúmeros atributos listados na literatura, partindo da língua Inglesa para o Português. Pode-se pensar em quatro dimensões: léxica, sintática, semântica e de coerência.

2.11 Avaliação automática respostas curtas

A avaliação automática de respostas curtas sucedeu ao estudo de avaliação de ensaios, pois a avaliação automatizada de respostas curtas usando as mesmas técnicas empregadas na classificação de ensaios não alcançou desempenhos satisfatórios (PRIBADI *et al*., 2017). Assim, replicar as decisões de um avaliador humano para respostas curtas continua sendo um grande desafio.

O avanço na avaliação de respostas curtas dependia de três fatores: *corpora* para fazer testes, onde as questões fossem avaliadas simultaneamente por dois avaliadores independentes; tecnologia já conhecida para avaliar ensaios; e novas tecnologias focadas especificamente para respostas curtas, objetivando principalmente resolver o problema de fazer uma comparação de textos com respostas com poucas palavras.

Com o tempo foram surgindo alguns *corpora* públicos para que a tecnologia proposta fosse avaliada e comparada. Por exemplo, o Texas *corpus* e o ASAP *corpus*.

Texas *Corpus* – *for short answers* (MOHLER; MIHALCEA, 2009). É composto por dez tarefas entre quatro e sete perguntas cada e duas tarefas com dez perguntas cada. Essas tarefas foram atribuídas para um curso introdutório de Ciência da Computação na *University of North Texas* (UNT). As respostas dos alunos foram coletadas por meio de um ambiente de aprendizado *on-line*. O conjunto de dados como um todo contém 2.442 (dois mil quatrocentos e quarenta e duas) respostas de alunos em 80 (oitenta) perguntas. As respostas foram pontuadas por dois juízes humanos, na escala de 0 a 5 (detalhes na Tabela 5). Subconjuntos destas respostas foram utilizados em alguns dos trabalhos relacionados mencionados (MOHLER; BUNESCU; MIHALCEA, 2011; ZIAI; OTT; MEURERS, 2012; PRIBADI; PERMANASARI; ADJI, 2018).

Tabela 5 - Detalhes sobre as pontuações do padrão referência no *corpus* do Texas.

Escore	Quantidade	Escore	Quantidade
0.00	24	3.25	1
0.50	3	3.50	187
1.00	23	3.625	1
1.50	46	3.75	1
1.75	1	4.00	220
2.0	93	4.125	2
2.25	2	4.50	310
2.5	125	4.75	1
3.0	164	5.0	1238

Fonte: Adaptado de Mohler e Mihalcea (2009).

ASAP *corpus* – *for short answers*. As perguntas e respostas dos alunos neste conjunto de dados, Tabela 6 (seis), vieram de contribuições de vários departamentos estaduais de educação nos Estados Unidos. O conjunto de dados contém respostas para 10 (dez) perguntas, diferindo em características como área de assunto e escala de pontuação. As primeiras 9 (nove) perguntas eram do 10º ano, e a 10ª (décima) pergunta era do 8º ano. Algumas respostas foram inseridas diretamente no computador; mas também para outras perguntas as respostas foram

manuscritas (e posteriormente transcritas para o meio digital). Cada resposta foi pontuada por dois avaliadores humanos independentes.

Tabela 6 – Características do conjunto de dados ASAP.

Característica	Tipo/Quantidade
Língua	Inglês
Perguntas	10
Respostas	>2000
Domínio	Ciência, ELA (*English Language Arts*) e Biologia

Fonte: Adaptado de HORBACH; STENNMANNS; ZESCH (2018).

Abaixo foram revisados 7 (sete) trabalhos sobre avaliação automática de respostas curtas. No final do relato será apresentada uma tabela que resume algumas das características destes trabalhos.

Leacock e Chodorow (2003) descrevem o desenvolvimento de um sistema automático de pontuação de respostas curtas a partir de resposta de referência de especialistas chamados *E-rater*. O sistema pontua respostas abertas, com ideias específicas, de áreas como ciências, matemática, compreensão de leitura e gerenciamento de banco de dados. O *E-rater* analisa as respostas utilizando técnicas de PLN que tentam combinar os conceitos linguísticos identificados aos que são representados pelo modelo de resposta de referência, trabalhando com sinônimos e paráfrases e atribuindo uma pontuação, dependendo do número de conceitos correspondentes existentes na resposta. O sistema foi testado em programas de avaliação em larga escala (NAEP *Math Online Project* e na Avaliação final de um curso de inglês no estado de Indiana) com um *corpus* total de 16.625 (dezesseis mil seiscentos e vinte e cinco) respostas alcançou uma acurácia de concordância 84% em relação aos humanos.

Christian Gütl (Gütl, 2007) desenvolveu um protótipo baseado na *Web* para avaliação automática de textos que gera questões e avalia respostas do tipo curtas chamado *e-Examiner*. O principal módulo de avaliação é o *Assessment Test Management* (ATM), que monitora todo o ciclo dos procedimentos de avaliação: da geração de perguntas ao armazenamento de itens de avaliação (pergunta e resposta de referência) ao desempenho da avaliação. O módulo identifica automaticamente conceitos importantes do conteúdo, extraindo uma resposta de

referência e criando uma pergunta simples. O protótipo se baseia na comparação da resposta do estudante com uma resposta de referência utilizando vários *plug-ins* para seu funcionamento: i) GATE e ANNIE: pré-processamento que inclui tokenização, etiqueta morfossintática, análise morfológica, *stopword,* etc.; ii) *COSIN Text Similarity* que estima a similaridade entre a respostas do aluno e uma ou mais respostas de referência com base no modelo de espaço vetorial; iii) *ROUGE Statistics* calcula uma variedade de características estatísticas no nível de palavras para determinar automaticamente a qualidade de uma resposta comparando com a resposta de referência; iv) *Assessment* Escore *Builder* elabora a pontuação final por uma combinação linear de características fornecidas pelos dois *plug-ins* mencionados anteriormente. O algoritmo do protótipo foi treinado em 368 (trezentos e sessenta e oito) respostas alcançando uma medida de correlação *Pearson* de 0.81.

Mohler e Mihalcea (2009) exploram técnicas supervisionadas para a tarefa de classificação automática de respostas curtas. Foram abordados os problemas de classificação de uma perspectiva de similaridade de texto. Foram examinadas várias medidas de similaridade semântica de texto para classificar as respostas curtas dos alunos. Foi utilizado um conjunto de dados (Texas *Corpus*) com 630 (seiscentos e trinta) respostas de estudantes. No experimento, foram combinadas oito medidas baseadas em conhecimento de similaridade semântica (e.g., *WordNet*) e duas medidas baseadas em *corpus* (e.g., LSA). Os resultados alcançaram um coeficiente de correlação *Pearson* de 0.50 (sistemas *versus* humano) contra uma correlação de 0.64 de (humanos *versus* humano).

Rodrigues e Araújo (2012) desenvolveram um sistema para avaliar respostas dissertativas curtas em língua portuguesa, explorando técnicas de PLN com uma etapa de tradução de frases para formas canônicas (listas de palavras versus etiqueta) via a substituição de sinônimos, como o uso de um tesauro (repositório de palavras com significados semelhantes). As respostas dos alunos foram comparadas com as respostas de referência. Na etapa de classificação utilizaram o modelo espaço vetorial alcançando um coeficiente de correlação de 0.78 entre a média do avaliador e o escore dado pelo sistema. A diferença máxima entre as pontuações do sistema e as do professor é de 1.70 pontos em uma pergunta de 6 (seis) pontos, o que dá uma concordância de 0.28. O *corpus* é composto de 16 (dezesseis) perguntas de história dividido em três categorias (enumerar, conhecimento específico e ensaio).

Benonram e Aziz (2013) executam outro experimento usando uma parte do *corpus* Texas, 360 (trezentos e sessenta) respostas, 3 (três) tarefas, 120 (cento e vinte) respostas/tarefa (um subconjunto das 630 respostas do *corpus* mencionado anteriormente). Foi utilizada uma abordagem em duas etapas. Primeiro, o vocabulário é expandido com um dicionário de sinônimos (semelhança baseada no conhecimento). O sistema alcançou um coeficiente de correlação de 0.82 com a média da avaliação dos humanos; a correlação medida entre dois avaliadores humanos foi de 0.72.

Gomaa e Fahmy (2014) descrevem o desenvolvimento de um sistema para pontuação automática de respostas curtas para língua Árabe. A pesquisa foca na aplicação de várias medidas de similaridade em combinação. Para isto foram utilizados três tipos (*String similarity*, *Corpus-based similarity* e *Knowledge-based similarity*) para comparar as respostas dos alunos com a resposta de referência para produzir a pontuação final. Em uma base de 610 (seiscentos e dez) respostas (com 536 execuções: 256 usando *String-Based Similarity*, 64 usando *Corpus-Based Similarity*, e outros 216 usando *Knowledge-Based Similarity measures*) obteve-se um coeficiente correlação *Pearson* de 0.68 (*SxH*) contra 0.86 (*HxH*).

Para respostas curtas, Pribadi et al., (2016) utilizaram uma sobreposição simples de palavras, centrada na similaridade dos coeficientes *Dice*, *Jaccard* e Cosseno. Eles concluem sobre avaliar o conteúdo de respostas curtas que “a medição de similaridade não pode apenas depender a sobreposição de palavras”, porque as respostas curtas têm um número limitado de palavras.

Em resumo, para respostas curtas, os trabalhos mencionados apresentaram três abordagens principais: técnicas de similaridade texto, como *n*-gramas e LSA; similaridade de texto com vocabulário expandido com o conhecimento de dicionários de sinônimos (por exemplo, *WordNet*); e similaridade de texto com vocabulário expandido baseado em um *corpus* relacionado, por exemplo, *Wikipédia*. Para fugir desse tipo de problema utiliza-se a técnica de expansão do vocabulário, seja com *corpus* relacionado, ou seja, com dicionários de sinônimos.

Tabela 7 – Sistemas para avaliação automática de respostas curtos na literatura.

Sistema	Abordagem	Modelo	Base	*HxH*	*SxH*	Métrica
Leacock e Chodorow (2003)	PLN	*Maximum entropy*	16.625	0.87	0.84	*acc*
Gütl (2007)	ROUGE metrics	*Regressão Linerar*	368	NA	0.81	*r*
Mohler e Mihalcea (2009)	Similaridade Semântica e corpus	*Stacking*	630	0.64	0.50	*r*
Rodrigues e Araújo (2012)	PLN	*Vector Space Model (VSM)*	NA	NA	0.78	*r*
Omran e Aziz (2013)	Similaridade de texto	*Stacking*	630	0.64	0.82	*r*
Gomaa e Fahmy (2014)	Similaridade de texto	*Stacking*	2273	0.86	0.68	*r*
Pribadi et al. (2016)	Similaridade de sobreposição	*Stacking*	2273	0.86	0.87	*r*

(r) Coeficiente de correlação; (*acc*) acurácia de erro médio

Fonte: Próprio autor (2020).

A Tabela 7 (sete) fornece uma comparação de características para alguns sistemas de avaliação automática de respostas curtas. Numa análise sobre a performance, a proposta de Pribadi *et al*., (2016) apresentou uma correlação com melhor desempenho dos sistemas levantados. O pior desempenho é da proposta de Mohler e Mihalcea (2009), apresentou um coeficiente correlação *Pearson* de 0.50, valor considerado razoável.

3 *Corpus* de estudo

Os dados da pesquisa foram constituídos pelos seguintes conjuntos: um voltado para resposta de tipo curta e outro para ensaio. Um dos conjuntos constitui-se de três bases de textos escritos para respostas curtas: uma questão discursiva conceitual de Biologia, uma questão discursiva argumentativa de Geografia e uma questão discursiva argumentativa de Filosofia; e um conjunto de textos escritos para respostas compostas por redações.

3.1 Conjunto de dados para respostas do tipo curtas

As questões de Biologia e Geografia constam no boletim EDITAL 016/2007 da terceira fase do Processo Seletivo Seriado (PSS) do ano de 2008 da Universidade Federal do Pará (UFPA). Este boletim era constituído por 75 (setenta e cinco) questões analítico-discursivas de 25 (vinte e cinco) disciplinas, com uma média de três questões por disciplina, além de uma redação, sendo que cada candidato deveria escolher e responder apenas 1 (uma) questão de cada disciplina. As respostas eram escritas em boletins específicos para cada disciplina. De um total de 15.154 (quinze mil cento e cinquenta e quatro) boletins de respostas foram selecionados aleatoriamente 1.000 (mil) boletins de respostas, já com as pontuações atribuídas pelos especialistas humanos. A pontuação máxima de cada resposta era 6 (seis) pontos, sendo que só era permitido para o avaliador humano atribuir as pontuações 0, 1, 2, 3, 4, 5 ou 6. Cada resposta era avaliada por dois avaliadores humanos, sendo que cada avaliador não tinha acesso à avaliação do outro. A pontuação era atribuída da seguinte maneira: no caso de coincidência era atribuída à pontuação comum; se a diferença entre as pontuações fosse de 1 (um) ponto, então era atribuída a maior pontuação; uma diferença igual ou superior a 2 (dois) pontos era considerada uma discrepância, e neste caso, após esse processo, a pontuação era atribuída por um terceiro avaliador. Foi optado por trabalhar com as seguintes respostas: 130 (cento e trinta) respostas para uma questão de Biologia e 229 (duzentos e vinte e nove) respostas para uma questão de Geografia, no total de 359 (trezentos e cinquenta e nove) respostas que passaram por um processo de digitalização manual.

A questão de filosofia foi obtida de um ambiente virtual de aprendizagem e é constituída por respostas do tipo curta de uma atividade da disciplina Filosofia e Ética do curso de Administração na modalidade a distância da UFPA. Nesta questão propõe-se argumentar sobre as diferenças das quatro principais épocas da filosofia, onde foram

selecionadas 192 (cento e noventa e duas) respostas. Essas respostas passaram também por um processo de digitalização manual, onde apenas os erros ortográficos foram corrigidos e, nenhum outro ajuste nos aspectos gramaticais dos textos originais foi feito. As respostas foram avaliadas pelo professor da disciplina e cada resposta recebeu uma pontuação contínua entre 0 (zero) e 5 (cinco).

Nas três próximas subseções apresentam-se os enunciados das questões cujas respostas constituem o *corpus* da pesquisa.

3.1.1 Questão de Biologia

A questão de Biologia é de natureza discursivo-conceitual baseado em três conceitos de uma dada taxonomia da Citologia. Abaixo se apresenta o enunciado da questão:

Os tecidos – grupos de células de mesma origem e semelhantes entre si em estrutura e função – são originados nos seres humanos a partir dos três folhetos embrionários. Cite três tipos de tecidos humanos com suas respectivas funções.

Apenas para efeito ilustrativo na Tabela 8 (oito) se tem uma amostra de cinco respostas que foram digitalizadas para esta questão com a respectiva pontuação atribuída por avaliadores humanos.

Tabela 8 – Respostas escritas por estudantes para a questão de biologia com seus respectivos escores (notas no intervalo de 0 a 6).

Texto	Nota
Tecido muscular esquelético responsável pela sustentação dos Tecido epitelial proteção e revestimento Tecido muscular cardíaco bombeamento de fluxo sanguíneo, circulação sanguínea	*6.0*
Tecido ósseo sustentação do corpo tecido cartilaginoso proteção da pele tecido muscular responsável pelas articulações	*5.0*
Tecido Ósseo serve para a estrutura dos ossos tecido Nervoso serve para circulação do sangue Tecido epitelial serve para	*4.0*
Epiderme tecido do revestimento	*0.0*

Fonte: Centro de Processos Seletivos da UFPA.

Para efeito ilustrativo, a Figura 6 (seis) apresenta as 20 (vinte) palavras do *corpus* de biologia mais frequentes. Neste caso, “tecido” é a palavra com mais frequente dentro do conjunto de dados, com mais de trezentas ocorrências, em segundo a palavra “função” com cem ocorrências e em terceiro a palavra “corpo” com um pouco menos de cem.

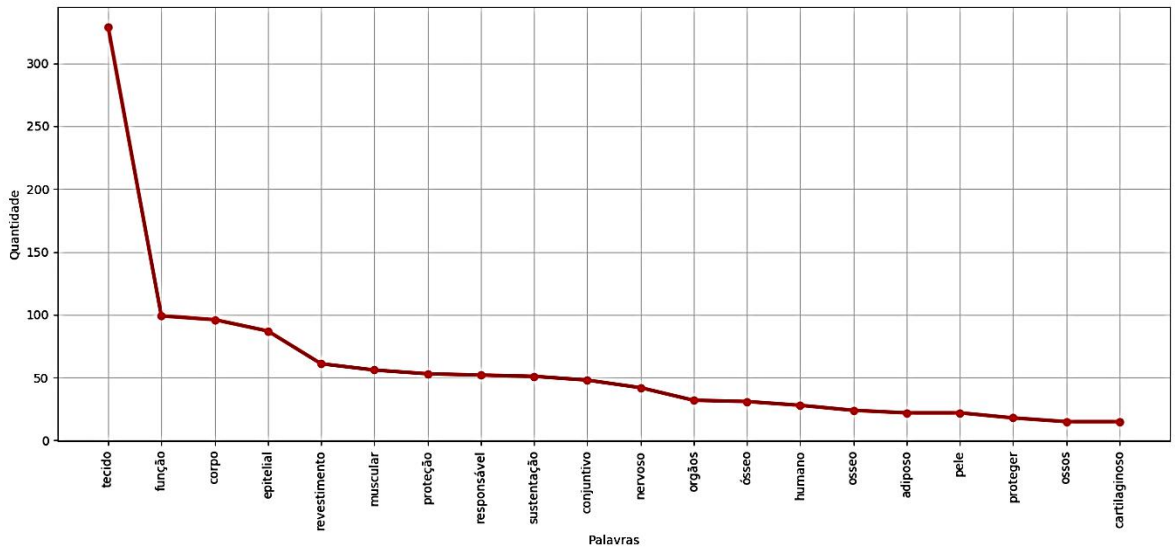


Figura 6 – As 20 palavras mais frequentes no *corpus* de Biologia.

A Figura 7 (sete) apresenta as palavras em formato *Word Cloud* (Nuvens de palavras), também conhecida como nuvens de *Tags*. A nuvem é composta por termos (palavras) mais frequentes de um texto, neste caso na base de Biologia. É plotada levando em consideração a frequência das palavras de um *corpus*, de forma que as palavras mais frequentes possuem um destaque maior em relação às demais palavras.

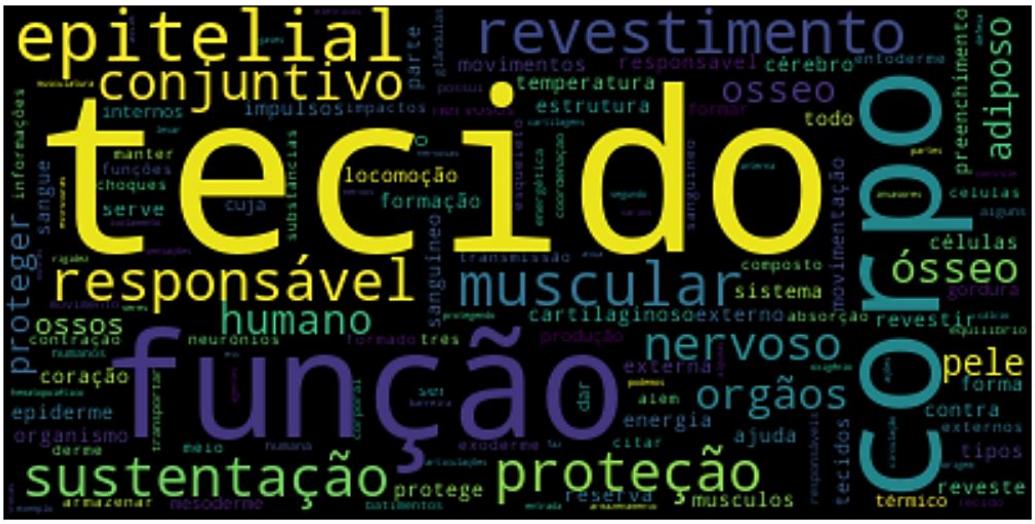


Figura 7 – *Corpus* de Biologia ilustrado no formato nuvens de palavras.

Para efeito ilustrativo, na Figura 8 (oito) apresentam-se por meio de histograma as notas das respostas discursivas do tipo curtas para uma representação gráfica da pontuação atribuída pelos avaliadores humanos.

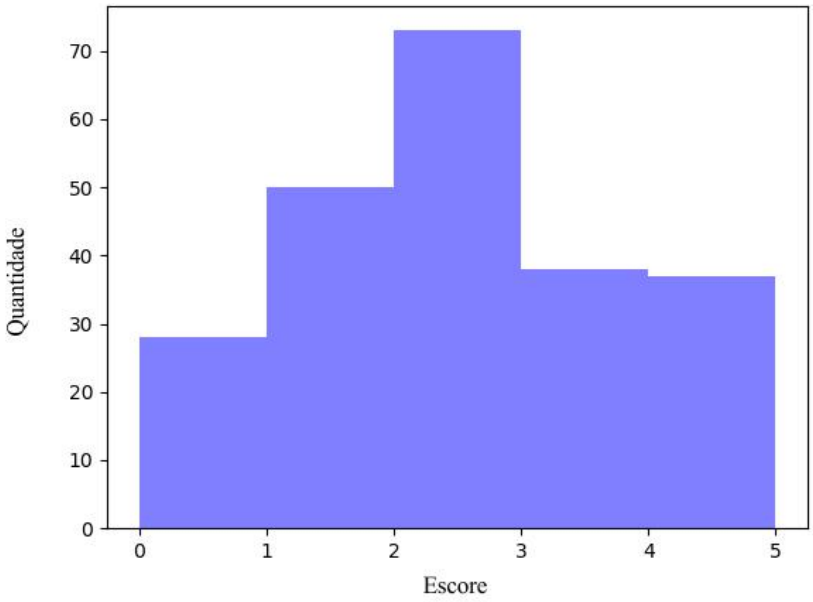
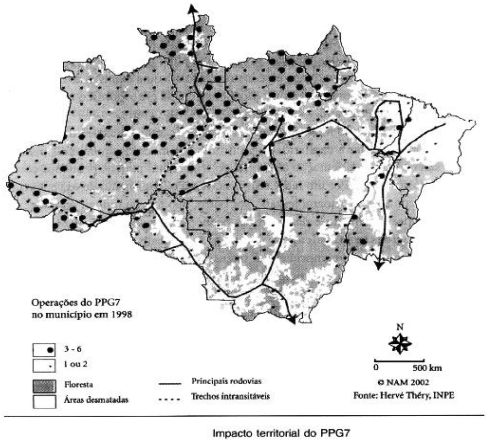


Figura 8 - Representação de um histograma dos escores do *corpus* de Biologia.

3.1.2 Questão de Geografia.

A questão de Geografia, de natureza discursivo-argumentativa, propunha a elaboração de argumentação em defesa de dado ponto de vista formado a respeito da Geografia Humana e Econômica da Região. Abaixo se apresenta o enunciado da questão:

> *Sobre o desmatamento na Amazônia, leia o texto e o mapa abaixo: "De fato, pelas imagens de satélite a possibilidade de imprecisão em torno do desmatamento é grande e os pesquisadores trabalham com aproximações e com cenários projetados. Nestes termos, esse fenômeno, mais as queimadas e a exploração madeireira são das realidades que mais se observa quando se está em trabalho de campo, quer no interior, quer na periferia das cidades. E, para além de uma sociedade em geral insensível quanto à importância dos recursos naturais, dentre os quais os florestais, tem-se um Estado que age, porém, em descompasso com a celeridade dos processos produtivos. O mesmo também se apresenta sempre enfraquecido quanto à questão da garantia dos direitos ambientais definidos constitucionalmente e em leis específicas, o que termina sustentando a impunidade nessa área". (SIMONIAN, L. Tendências recentes quanto à sustentabilidade no uso dos recursos naturais pelas populações tradicionais amazônidas. In: ARAGON, L. E. (ORG.). População e Meio ambiente na Pan-Amazônia. Belém: UFPA/NAEA,2007, p. 30).*



Impacto territorial do PPG7

Considerando as informações acima e seus conhecimentos sobre a realidade amazônica:

***a)** identifique as áreas com maior impacto de desmatamento; **b)** explique o processo de intensificação do desmatamento na Amazônia, valendo-se de dois fatores que estão diretamente relacionados com esse processo?*

Apenas para efeito ilustrativo na Tabela 9 (nove) se tem uma amostra de três respostas que foram digitalizadas para esta questão com a respectiva pontuação atribuída pelos avaliadores humanos:

Tabela 9 – Respostas escritas por estudantes para a questão de Geografia com seus respectivos escores atribuído pelos avaliadores humanos (notas no intervalo de 0 a 5).

Texto	Nota
As áreas leste e sul da região amazônica, têm o maior impacto de desmatamento. O processo de desmatamento da Amazônia, iniciou se a décadas e intensifica se com o fácil acesso às terras, devido aos cartórios fraudulentos, o que leva a grilagem e a partir daí o desmatamento para o desenvolvimento de atividades. Além disso, a infraestrutura (rodovias, incentivos fiscais) oferecida pelo estado, para implantação de projetos também demanda áreas florestais. Na medida em que há o desenvolvimento urbano ao entorno das rodovias e dos projetos, responsável pela atração de atores sociais, culturalmente distintos, e que implantam novas atividades (pecuária, madeireiras, e agricultura de grãos) na região	*4.0*
A região oriental da Amazônia. Com o aproveitamento das rodovias, intensificam se a chegada de madeireiras e fazendeiros que irão desmatar áreas da Amazônia, provocando vários impactos ambientais como: a lixiviação, o aumento do fluxo de CO2 para a atmosfera, contribuindo para o agravamento do efeito estufa.	*2.0*
Os estados do Pará e Matogrosso. Um dos principais motivos do desmatamento é a exploração de minérios que continua intenso com a presença de empresas estrangeiras e um outro motivo é a agricultura que vem desmatando grandes áreas com a expansão da soja.	*1.0*

Fonte: Centro de Processos Seletivos da UFPA.

A Figura 9 (nove) apresenta as 20 (vinte) palavras do *corpus* de Geografia mais frequentes. A palavra mais frequente é “desmatamento” com mais de 350 (trezentos e cinquenta) ocorrências. A palavra “áreas” é a segunda mais frequente com mais de 200 (duzentas) ocorrências e em terceiro a palavra “rodovias” com mais de 100 (cem) ocorrências no *corpus*.

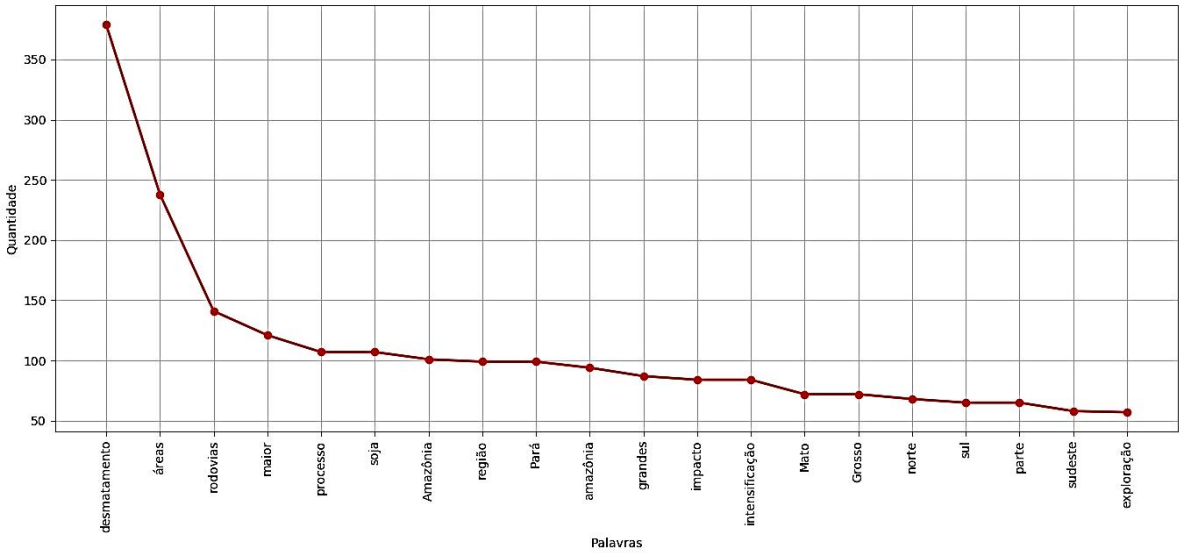


Figura 9 - As 20 (vinte) palavras mais frequentes no *corpus* de Geografia.

Como ilustrado para questão de Biologia, a Figura 10 (dez) representa a questão de Geografia em formato nuvens de palavras.

Figura 10 - *Corpus* de Geografia ilustrado no formato nuvens de palavras.

A Figura 11 (onze) apresenta-se por meio de histograma as notas das respostas discursivas do tipo curtas para uma representação gráfica da pontuação atribuída pelos avaliadores humanos.

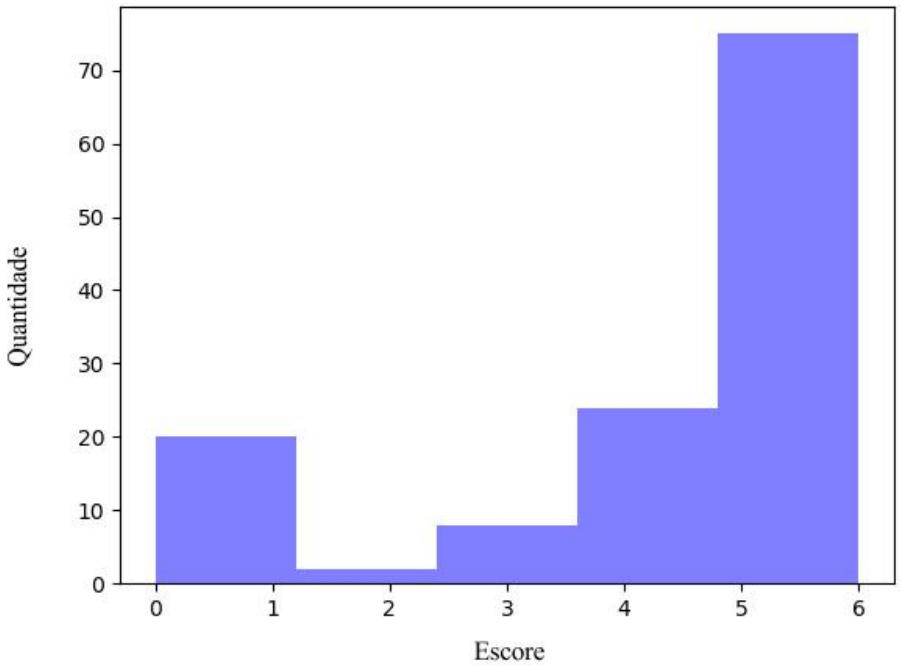


Figura 11 - Representação de um histograma dos escores do *corpus* de Geografia.

3.1.3 Questão de Filosofia

A questão de Filosofia é de natureza discursivo-argumentativa e expõe sobre as diferenças das quatro principais épocas da Filosofia. Abaixo se apresenta o enunciado da questão:

> *Qual a principal diferença entre as visões dos quatro grandes períodos da Filosofia?*

Apenas para efeito ilustrativo na Tabela 10 (dez) se tem uma amostra de três respostas que foram digitalizadas para esta questão com a respectiva pontuação atribuída pelos avaliadores humanos.

Tabela 10 - Respostas escritas por estudantes para a questão de Filosofia com seus respectivos escores atribuídos pelos avaliadores humanos (notas no intervalo de 0 a 5).

Texto	Nota
"Os quatro grandes períodos da Filosofia diferenciam-se essencialmente pelo fato de que na Idade Antiga prevaleceu a visão Fisiocêntrica (o fundamento último é a natureza), na Idade Média a visão Teocêntrica (visão cristã, Deus é a última razão), na Idade Moderna a visão Antropocêntrica (fundamento é o ser humano e sua razão) enquanto que na Idade Moderna não existe a convicção que seja necessário haver um único fundamento."	*5.0*
A principal diferença entre os grandes quatros períodos é que cada período tinha uma visão do mundo e buscava estudar cada um em seu tempo, tendo como foco o que mais lhes instigavam em determinado período exemplos: A Filosofia antiga tinha como sua principal preocupação a origem do mundo e as causas das transformações da natureza, a Filosofia Medieval tinha como principal característica o pensamento, a Filosofia Moderna tinha como objetivo associar as mudanças e ênfase nos seguintes valores: antropocentrismo, racionalismo e o individualismo e por fim, a Filosofia Contemporânea que tem seu conhecimento ampliado e faz surgir daí um novo objeto de estudo o próprio homem	*4.0*
A principal diferença é que em cada um dos períodos o modo de pensar parte de um determinado ponto. Questiona-se ou fortalece algo que reflete bastante o que se vive em tal momento, pelas frases e focos conseguimos distinguir um período de outro, como por exemplo, "Thomas Hobbes (1588-1679), filósofo inglês, além de defender uma visão materialista (tudo é apenas corpo) e mecanicista (toda a realidade funciona como se fosse uma grande máquina)", ao lermos fica bastante claro que estamos falando da Filosofia Moderna	*3.0*

Fonte: Centro de Processos Seletivos da UFPA.

Para efeito ilustrativo, a Figura 12 (doze) apresenta as 20 (vinte) palavras do *corpus* de Filosofia mais frequentes. A palavra mais frequente é "filosofia" com mais de 600 (seiscentas) ocorrências. A palavra "homem" é a segunda mais frequente com 300 (trezentas) ocorrências e em terceiro a palavra "período" com mais de 200 (duzentas) ocorrências.

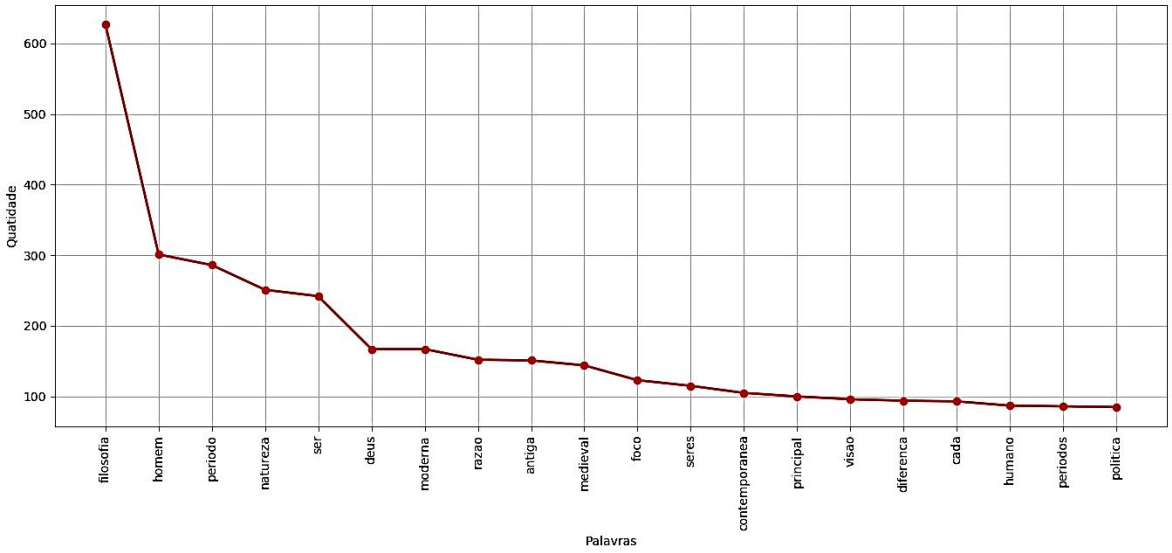


Figura 12 - As 20 (vinte) palavras mais frequentes no *corpus* de Filosofia.

Como efeito ilustrativo para questão de Filosofia, a Figura 13 (treze) representa a questão de geografia em formato *Word Cloud*.

Figura 13 - *Corpus* de Filosofia ilustrado no formato nuvens de palavras.

Para efeito ilustrativo, na Figura 14 (quatorze) apresentam-se por meio de histograma as notas das respostas discursivas do tipo curtas para uma representação gráfica da pontuação atribuída pelos avaliadores humanos.

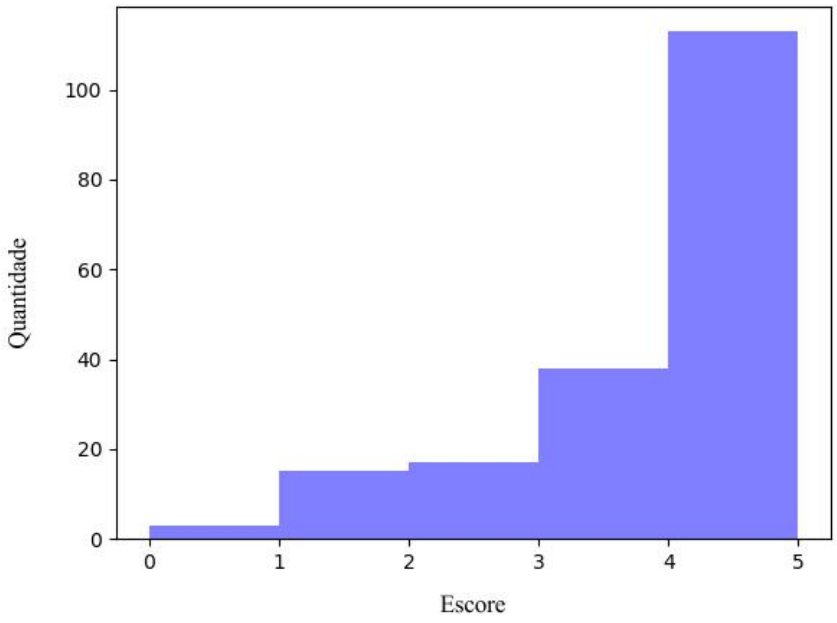


Figura 14 - Representação de um histograma dos escores do *corpus* de Filosofia.

A Tabela 11 (onze) mostra a representação para cada *corpus*, a quantidade de respostas; o número mínimo, máximo e médio de palavras por resposta em cada conjunto; e a precisão humana *versus* humana.

Tabela 11 – Características do conjunto de dados das questões de respostas curtas.

Corpus	**Quantidades**	**Palavras**	***HxH***
Biologia	131	Min = 4, Max = 56, Média = 28.48	93.84
Geografia	192	Min = 11, Max = 269, Média = 149.37	84.93
Filosofia	230	Min = 9, Max = 189, Média = 74.56	-

Fonte: próprio autor (2020).

3.1.4 Respostas de Referência para resposta do tipo curta

O *corpus* da pesquisa foi constituído por respostas escritas por candidatos de um processo seletivo para ingresso na Universidade Federal do Pará. Logo após a realização das provas, equipes de avaliadores são formadas para correção das provas. Estas equipes são responsáveis pela discussão de uma grade de correção para cada disciplina.

Na Tabela 12 (doze) apresenta-se a grade de correção utilizada como referência para a questão de Biologia. O critério para atribuição das pontuações de cada uma das respostas era 1 (um) ponto para cada tecido e 1 (um) ponto para cada função.

Tabela 12 - Grade de correção da questão de Biologia.

Tecidos	Funções
Epitelial ou glandular epitelial	*Revestimento interno (trocas, absorção de substâncias), ou externo (proteção, perda de água,) proteção (mecânica), percepção de estímulos, substâncias*
Conjuntivo	*Preenchimento, suporte, nutrição dos epitélios, proteção contra infecção, transporte de substância, armazenamento e produção de substâncias, cicatrização de tecidos lesados*
Adiposo	*Preenchimento e reserva energética*
Cartilaginoso	*Sustentação, proteção ou revestimento das articulações*
Ósseo	*Proteção, sustentação, armazenamento de Cálcio*
Hematopoiético mielóide	*Produção de glóbulos vermelhos e plaquetas*
Hematopoiético linfoide	*Produção de glóbulos brancos*
Sanguíneo	*Transporte de gases, nutrientes*
Linfático	*Defesa*
Muscular	*Movimento, contração*
Nervoso	*Regulação e integração interna e coordenação corporal, homeostase, raciocínio, memória, irritabilidade, condução de impulsos nervosos*

Fonte: Centro de Processos Seletivos da UFPA.

Na Tabela 13 (treze) apresenta-se a grade de correção utilizada como referência para a questão de Geografia e o critério para atribuição das pontuações de cada uma das respostas.

Tabela 13 – Grade de correção da questão de Geografia

Desempenho	Pontuação
Identifica as áreas considerando os agrupamentos, que podem ser: arco do desmatamento ou arco do povoamento consolidado, ou Amazônia oriental e meridional ou leste e sul da Amazônia. OU identifica as áreas no interior dos territórios dos Estados membros na sua totalidade	*2.0*
Identifica parcialmente as áreas no interior dos territórios dos Estados membros, como, por exemplo, sudeste e sul do Pará, ou norte e noroeste do Tocantins	*1.0*
Explica que as políticas territoriais implementadas pelo governo federal a partir dos anos sessenta priorizaram a instalação de um modelo baseado na exploração em larga escala dos recursos naturais, por meio da abertura de eixos rodoviários, a exemplo da Belém-Brasília e da Transamazônica, o que facilitou o avanço de frentes de expansão, tais como o extrativismo madeireiro, a mineração e a agropecuária, atividades essas responsáveis por grande parte do desmatamento na região; e refere o processo de povoamento decorrente das migrações com a proliferação de vilarejos, povoados e cidades, o que também contribui para a intensificação do desmatamento	*4.0*
Respostas explicativas que apresentarem apenas um fator, como, por exemplo, a atividade madeireira e sua relação com o desmatamento	*2.0*

Fonte: Centro de Processos Seletivos da UFPA.

Para efeito de comparação em cada grupo, uma resposta de referência específica foi usada; para o conjunto de Biologia, a resposta de referência foi um texto produzido por um especialista contendo todos os conceitos "possíveis" e corretos; para os conjuntos de Geografia e Filosofia, foi formado a resposta referência considerando aquelas respostas que obtiveram a maior pontuação por avaliadores humanos.

3.2 Conjunto de dados de respostas do tipo ensaio

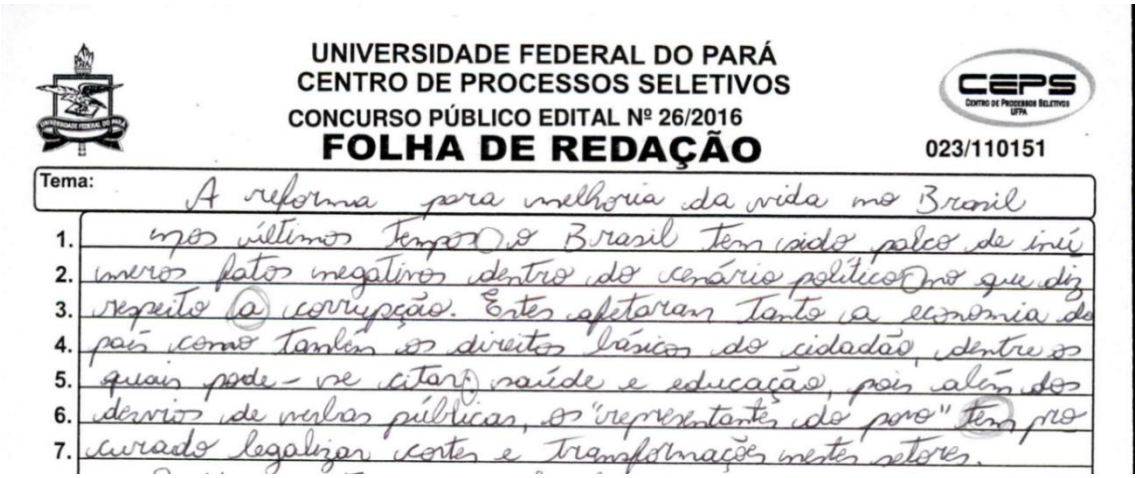
Para o experimento de ensaios, o *corpus* da pesquisa foi composto por uma amostra de redações de um concurso público do edital nº 26/2016- UFOPA para admissão na carreira de

técnico administrativo em educação: foi elaborada uma redação sobre o tema "*A atual crise político-social do Brasil e ações políticas para seu enfrentamento*". Estas redações passaram por um processo de digitalização manual onde não foi feito nenhum tipo de correção ortográfica e nem alterações nos aspectos gramaticais do texto original. Todas as 4.235 (quatro mil duzentas e trinta e cinco) redações foram previamente avaliadas por dois avaliadores humanos, recebendo uma pontuação inteira entre 0 (zero) e 10 (dez), com passo de 0.25, sendo que cada avaliador não conhece a pontuação do outro. Foram feitas verificações de discrepâncias: se as duas pontuações divergem por mais de um ponto, então um terceiro avaliador atribui uma pontuação para ser comparada com as duas pontuações anteriores. Para composição do *corpus* para o experimento, foi seleciona uma amostra de 1.000 (mil) redações com pontuações atribuídas em todas as faixas de pontuações.

A questão de ensaio de natureza em prosa é avaliada sobre a frase "*há um estado febril em nosso país e ele precisa ser curado*", considerando o mais grave sintoma desse estado doentio no Brasil e aponte ações políticas que, em sua opinião, propiciariam melhor qualidade de vida aos brasileiros. Abaixo se apresenta o enunciado da questão:

> *O texto "Patriotismo e Nacionalismo" assinala que "o Nacionalismo, uma forma aguda de sentimento patriótico, que emerge nos momentos de crise nacional, propicia e acompanha as fases de mais intenso desenvolvimento". Também afirma que "Hoje, há um estado febril em nosso país e ele precisa ser curado. Por esse motivo (...) é chegado o momento de resgatar o nosso Nacionalismo". Entende-se que o texto se refere, nesses trechos, a um tipo de Nacionalismo – ao chamado nacionalismo cívico – , que define a nação como uma associação de pessoas que se identificam como pertencentes a ela, que têm direitos políticos iguais e compartilhados e fidelidade a procedimentos políticos semelhantes.*

Apenas para efeito de ilustração, a Figura 15 (quinze) mostra o exemplo de uma das mil redações digitalizadas manualmente para composição do conjunto de dados dos textos dissertativos.



UNIVERSIDADE FEDERAL DO PARÁ
CENTRO DE PROCESSOS SELETIVOS
CONCURSO PÚBLICO EDITAL Nº 26/2016
FOLHA DE REDAÇÃO
023/110151

Tema: A reforma para melhoria da vida no Brasil

1. nos últimos Tempos o Brasil tem sido palco de inú
2. meros fatos negativos dentro do cenário político no que diz
3. respeito à corrupção. Estes afetaram tanto a economia do
4. país como também os direitos básicos do cidadão, dentre os
5. quais pode-se citar a saúde e educação, pois além dos
6. desvios de verbas públicas, os "representantes do povo" tem pro
7. curado legalizar certas transformações nestes setores.

Figura 15 – Exemplo de uma redação do *corpus* das questões dissertativas tipo ensaio.

Para efeito ilustrativo, a Figura 16 (dezesseis) apresenta as 20 (vinte) palavras mais frequentes dentro do *corpus* dos ensaios.

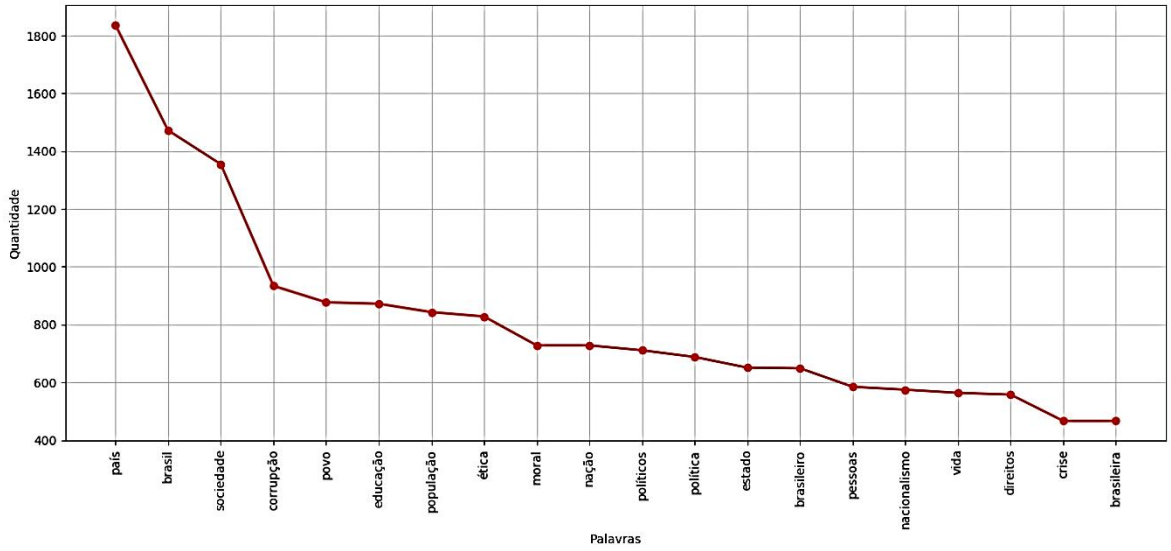


Figura 16 - As 20 (vinte) palavras mais frequentes no *corpus* de ensaio (redação).

A palavra mais frequente é "país" com mais de 1800 (mil e oitocentas) ocorrências dentro do *corpus*. A palavra "Brasil" é a segunda mais frequente com mais de 1400 (mil e quatrocentas) ocorrências e em terceiro a palavras "sociedade" com mais de 1200 (mil e

duzentas) ocorrências. A Figura 17 (dezessete) apresenta as palavras mais frequentes no *corpus* de ensaio no formato *Word Cloud*.

Figura 17 - *Corpus* de ensaio ilustrado no formato nuvens de palavras.

Para efeito de esclarecimento, na Figura 18 (dezoito) apresentam-se por meio de histograma as classes de notas do *corpus* da redação para uma representação gráfica do volume de pontuação atribuída pelos avaliadores humanos.

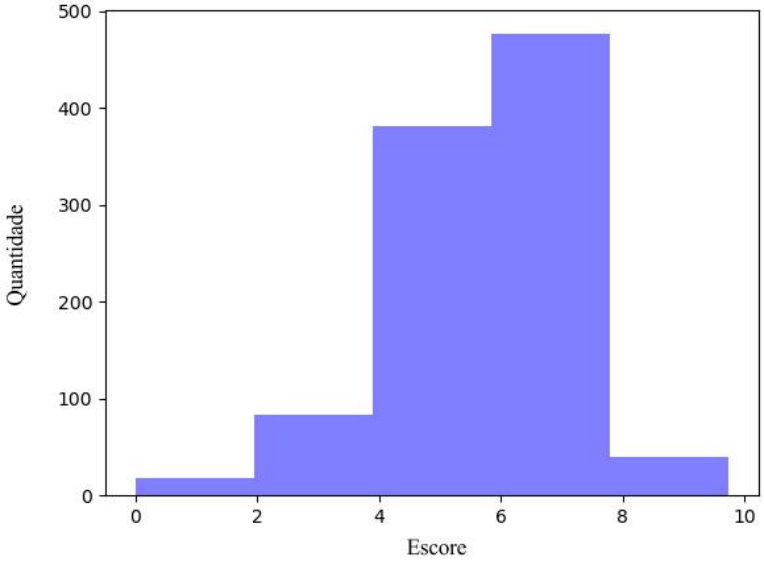


Figura 18 - Representação de um histograma dos escores do *corpus* de ensaio (redação).

Na subseção seguinte, menciona-se a grande dificuldade de encontrar *corpus* de pesquisa principalmente na língua portuguesa e a disponibilidade de um novo repositório para esse campo.

3.3 Disponibilidade de *corpus*

A Tabela 14 (quatorze) mostra o *corpus* que se encontra no repositório e algumas de suas características. Nesta tabela, apresenta-se a acurácia quando as respostas de uma mesma prova são avaliadas por dois especialistas humanos de forma independente, assim se pode medir a acurácia Humano *versus* Humano entre eles.

Tabela 14 – Características dos *corpora* disponibilizados no repositório do Laboratório de pesquisa.

Corpus	**Quantidade**	**Tipo**	***HxH****	***HxH*** **
Biologia	*131*	Resposta curta	*0.94*	*0.89*
Geografia	*230*	Resposta curta	*0.85*	*0.72*
Filosofia	*192*	Resposta curta	-	-
Redações	*1000*	Resposta ensaio	*0.95*	*0.56*

*Erro Médio; **Kappa Quadrático

Fonte: próprio autor (2020)

3.3.1 Corpus disponível para respostas curtas

Uma das grandes dificuldades na pesquisa de AAT é a falta de *corpus*. Um conjunto de dados (*corpus*) para avaliação de respostas discursivas compreende uma coleção dessas respostas avaliadas – preferencialmente – por dois avaliadores humanos disponíveis em formato digital.

Questão de pesquisa sobre *corpus* (QP1): *Como criar um conjunto de corpus público para acelerar a pesquisa na área avaliação automática de texto, especialmente em Português?*

No início deste trabalho, foram compilados os *corpora* listados na Tabela 14 (quatorze). Estas questões estão disponibilizadas em um repositório no *site* do Laboratório de Inovação Interdisciplinar – LabX (http://www.labx.ufpa.br/dataset.html). Com este *corpus* público contribui-se para a pesquisa na área, onde outros pesquisadores poderão reutilizar este material sem o esforço de preparar e digitalizar os conteúdos. Além disso, com as publicações, os pesquisadores também terão parâmetros para as suas pesquisas, como a acurácia alcançada pelos experimentos.

4 Método: a proposta da tese

Neste capítulo é apresentada a proposta da tese centrada numa arquitetura *pipeline* para Avaliação Automática de Texto (AAT). Esta arquitetura depende de uma etapa principal chamada de coleta de atributos linguísticos em quatro dimensões, que por sua vez dependem de muita tecnologia.

Neste capítulo, primeiro mostra-se como coletar atributos dentro das dimensões: **léxica, sintática, semântica e de coerência.** Foi usada uma subseção para cada dimensão de atributos; segundo, foi apresentada uma arquitetura comentando os aspectos conceituais de cada etapa a serem realizados sobre os conjuntos de respostas curta ou ensaio (redação), sempre com o objetivo de aproximar a acurácia da avaliação do sistema com a acurácia dos avaliadores humanos; por fim, foram detalhadas as tecnologias utilizadas na coleta dos atributos de conteúdo: analise semântica latente (LSA), *n*-gramas para medir a similaridade de textos e o método de pesagem de palavras TF-IDF.

4.1 Coleta de atributos

Um modelo preditivo matemático requer atributos numéricos. Assim, o texto das respostas discursivas em si não pode ser usado pelos modelos de aprendizagem de máquina. A etapa de coleta de atributos transforma os textos em vetores numéricos que são à entrada do modelo de predição. Um dos principais desafios desta pesquisa é relacionado à complexidade da tarefa de coleta de atributos.

Segundo Dong e Liu (2018) atributos (*features*) são variáveis usadas para descrever algum aspecto individual de um dado objeto, sendo à base de entrada para modelos preditivos. Para eles, os atributos são divididos em três tipos de domínios:

1. **Categórico** – Variável de um conjunto de domínio sobre valores discretos. Ex. Sim/Não, Chove/Faz-Sol;
2. **Ordinal** – Variável de um conjunto de valores ordenados. Ex. Regular/Bom/Ótimo;
3. **Numérico** – Variável de um conjunto de valores numéricos (quantitativo ou contínuo).

A engenharia de atributos trata do desenvolvimento de métodos e técnicas para a coleta manual ou automática de atributos associados a medidas de características dos textos tais como erros morfológicos e conteúdo; também trata de técnicas para seleção e ou

combinação dos atributos por meio da escolha do método de predição buscando a criação de modelos otimizados. Algumas questões nesta área estão abertas, como, quais são os melhores atributos, quantos são necessários para o modelo preditivo? Qual acurácia é possível alcançar com o modelo? Qual o melhor modelo? Quais são os melhores atributos para medir uma competência de avaliação? (VAJJALLA, 2018)

No contexto da AAT a engenharia de atributos busca proporcionar ferramentas para representar um conjunto de atributos preditores $\{x_1, x_2, x_3 \ldots x_n\}$ que representam a qualidade de um texto, resultando numa boa estrutura de modelagem matemática preditiva. Busca-se determinar automaticamente o escore de um avaliador humano. Para alcançar esse objetivo é necessário coletar os atributos que capturam melhor as informações ocultas dentro do domínio do problema e, em seguida, selecionar os atributos certos para criar um modelo preditivo eficaz.

Na literatura foram encontradas propostas que trabalham com 60 (sessenta) atributos (IEA (LANDAUER, 1999)) e até com mais de 400 (quatrocentos) atributos (*IntelliMetric* (ELLIOT, 2003)). Com base em trabalhos recentes (ZUPANC; BOSNIC, 2017; PALMA; ATKINSON, 2018; VAJJALLA, 2018) foram definidas quatro dimensões de atributos: Léxica, Sintática, Semântica e de Coerência. Dentro destas quatro dimensões foram extraídos mais de 140 (cento e quarenta) atributos que servem de entrada para os modelos preditivos. Nas sessões seguintes, os atributos associados às quatro dimensões serão descritas.

4.2 Dimensão Léxica

A dimensão léxica descreve a coleta de atributos em um aspecto individual das palavras. Esta dimensão tem três principais categorias: (i) estatística de superfície, coleta estatística baseado em contagem de palavras; (ii) diversidade, coleta medidas que representam o quanto é diverso o vocabulário utilizado; (iii) legibilidade, mede o grau de facilidade da leitura do texto.

i. Estatística de superfície

Está relacionada a aspectos que não fazem parte do conteúdo essencial do texto, mas medem alguma característica relacionada com o "estilo", como por exemplo, número de caracteres, número de diferentes palavras, número de palavras, número de palavras curtas,

número de palavras longas, número médio de palavras, número de *stopword*, número de sentenças, comprimento de palavra mais frequente, número de sílaba, entre outros.

ii. Diversidade

Está relacionada com a coleta de medidas que representam o quanto é diverso o vocabulário utilizado no texto. Abaixo se listam cinco medidas de diversidade.

Type-token ratio - TTR. É o relacionamento entre o tipo de *tokens* e o número de *tokens*. O tipo de *token* pode ser a redução da palavra ao radical, obtida pelo processo de *stemmer*. O TTR fornece uma visão básica da quantidade de variação lexical no texto/*corpus*, que pode ser um indicador da complexidade de um texto/*corpus* (RICHARDS, 1987) (Equação 1).

$$TTR = \frac{n° \; de \; tipos \; de \; tokens}{n° \; de \; tokens} \tag{1}$$

Guiraud's index. É uma estrutura sintática paralela entre o número de tipos de *tokens* dividido pela raiz quadrada do número de *tokens* (GUIRAUD, 1954) (Equação 2).

$$Guiraud's \; index = \frac{n° \; de \; tipos \; de \; tokens}{\sqrt{n° \; de \; tokens}} \tag{2}$$

Yule's K. É uma estrutura sintática paralela de repetição lexical que se constitui de variáveis que são utilizadas para determinar a riqueza lexical dos textos. A fórmula de Yule é considerada bem confiável por ser independente do tamanho do texto, onde *K* mede a taxa de repetição lexical (YULE, 1944) (Equação 3).

$$K = 10^4 \times \frac{(\sum_{x=1}^{x} \int x \, X^2) - n° \; de \; tokens}{n° \; de \; tokens^2} \tag{3}$$

The D estimate. É uma estrutura sintática paralela que mede a diversidade lexical usando um modelo não linear (MALVERN *et al*., 2004). Para calcular esta medida é executado o seguinte procedimento: i) pegue uma amostra aleatória de N palavras do texto; ii) calcule o TTR; iii) encontre o valor de D (Equação 4); iv) o valor de D encontrado é uma estimativa da diversidade do texto.

$$TTR = \frac{D}{N}\left[\sqrt{\left(1 + 2\frac{N}{D}\right) - 1}\right] \quad (4)$$

Hapax legomena. É uma estrutura sintática paralela que calcula o número de palavras que ocorrem apenas uma vez no texto. Essa função pega um dicionário (tipo de estrutura de dados do *Python*) de contagens (como o retornado por uma contagem básica) e retorna o número de itens com uma contagem de 1 (um).

iii. Legibilidade

Essas medidas fornecem uma noção da facilidade da leitura do texto. A razão para usar medidas de legibilidade é que um avaliador humano considera a dificuldade de ler um ensaio ao avaliá-lo. Abaixo se apresenta nove medidas de legibilidade.

Gunning Fog Index (GFI). O GFI calcula um índice na faixa de 0 (zero) a 20 (vinte). A fórmula estima os anos de educação formal que seriam necessários para que o leitor entenda o texto em uma primeira leitura (GUNNING, 1968; DUBAY, 2007). Por exemplo, se um pedaço de texto tiver uma pontuação de legibilidade no nível da série 6, isso deve ser facilmente legível por aqueles que foram educados até a 6ª série (Equação 5).

$$GFI = 0.04 \times \left[\left(\frac{total\ de\ palavras}{total\ de\ sentenças}\right) + 100\ \left(\frac{n^{o}\ Palavras\ complexas*}{total\ de\ plavras}\right)\right] \quad (5)$$

(*) palavras complexas são definidas como aquelas que contêm três ou mais sílabas

Flesch Reading Ease (FRE). É uma medida que estima como é fácil ler um texto e também o nível educacional necessário para sua compreensão (FLESCH, 1948; FARR; JENKINS; PATERSON, 1951). Esta medida é complementar ao índice de *Gunning Fog Index* e é calculada da seguinte forma (Equação 6):

$$FRE = 206.835 - 1.015\left(\frac{total\ de\ palavras}{total\ de\ sentenças}\right) - 84.6\left(\frac{total\ de\ silabas}{total\ de\ palavras}\right) \quad (6)$$

Flesch Kincaid grade level (FKGL). Essa medida é um teste de legibilidade projetado para indicar o quão difícil é entender uma passagem em uma língua (KINCAID *et al*., 1975). É uma melhoria do índice de *Flesch Reading Ease*. A razão para o uso de ambas as medidas é que, em algumas situações, a primeira poderia ser um melhor preditor da nota e vice-versa. Esta medida é calculada como (Equação 7):

$$FKGL = 0.39\left(\frac{total\ de\ palavras}{total\ de\ sentenças}\right) + 11.8\left(\frac{total\ de\ silabas}{total\ de\ palavras}\right) - 15.59 \quad (7)$$

Dale-Chall readability (DCR). Mede um texto em relação a um número de palavras consideradas familiares (DUBAY, 2004). De acordo com sua escala, quanto mais palavras desconhecidas forem usadas, maior será o valor do nível de leitura (Equação 8).

$$DCR = 0.1579\left(\frac{n°\ Palavras\ complexas}{100}\right) + 0.0496\left(\frac{total\ de\ palavras}{total\ de\ sentenças}\right) + 3.6365 \quad (8)$$

Autometed readability index (ARI). O índice ARI é semelhante a algumas das fórmulas de legibilidade. É uma medida projetada para estimar quão compreensível é um texto (SENTER; SMITH, 1967). O que o torna um pouco diferente é que, em vez de contar as sílabas em uma palavra, conta os caracteres em uma palavra. Quanto mais caracteres houver em uma palavra, menos fácil será a leitura. Também levam em

consideração as sentenças em seus cálculos. O índice fornece uma aproximação do nível de educação necessário para entender um texto (Equação 9).

$$ARI = 4.71\left(\frac{n° \; de \; caracteres}{n° \; de \; palavras}\right) + 0.5\left(\frac{n° \; de \; palavras}{n° \; de \; senteças}\right) \text{-}21.43 \qquad (9)$$

Läsbarhetsindex (LIX). Mede a legibilidade com base na contagem de letras, em vez de usar o método de contagem de sílabas de muitas outras fórmulas (BJORNSSON, 1968). Como a contagem de sílabas pode se mostrar imprecisa em idiomas diferentes do inglês, o método de contagem de letras de LIX é mais adequado para esse objetivo (Equação 10).

$$\boldsymbol{LIX} = \frac{n° \; palavras}{n° \; sentenças} + \left(\frac{(n° \; palavras{>}6)}{n° \; palavras} \times 100\right) \qquad (10)$$

Word variation index (OVIX). Baseado em variáveis logarítmicas, essa medida extrai a proporção de *tokens* exclusivos em um texto e é usada para indicar a densidade da ideia. OVIX (HULTMAN; WESTMAN, 1977) é calculado pela Equação 11 (onze):

$$OVIX = \frac{\log\left(n° \; de \; palavras\right)}{\log\left(2 - \frac{\log\left(n° \; de \; palavras \; raras\right)}{\log\left(n° \; de \; palavras\right)}\right)} \qquad (11)$$

Nominal Ratio (NR). É uma medida da densidade da informação, neste caso, comparando o número de substantivos, preposições e particípios com o número de pronomes, advérbios e verbos (HULTMAN; WESTMAN, 1977). A NR mede qualidade do texto em relação aos recursos lexicais e gramaticais. Essa medida é dependente de um recurso linguístico de etiquetagem morfossintática e é calculado pela Equação 12 (doze):

$$NR = \frac{n° \; substantivo + n° \; preposições + n° \; participios}{n° \; pronomes + n° \; adverbio + n° \; verbo} \quad (12)$$

Simple Measure of Gobbledygook (SMOG-index). Esta medida estima os anos de educação que uma pessoa precisa para entender a estrutura de legibilidade de um texto que pode ser usada para analisar como é legível (MCLAUGHLIN, 1969) (Equação 13).

$$SMOG = 1.0430\sqrt{n° \; palavras \; polissilabas \times \frac{30}{n° \; sentença}} + 3.1291 \quad (13)$$

4.3 Dimensão Sintática

A maioria dos atributos da dimensão sintática depende de uma análise sintática do texto, gerada por um analisador de marcação gramatical. Esses atributos sintáticos retratam o aspecto individual de cada sentença e baseiam-se em uma visão **morfossintática** da sentença. Aqui, se entende por morfossintática quando se está utilizando somente a classificação das palavras em suas categorias gramaticais. Assim, esta dimensão sintática não tratará da formação e/ou estudo das árvores sintáticas. Essa dimensão compreende as seguintes categorias:

- **Número de cada etiqueta morfossintática**: A proporção do número de etiquetas é calculada com uma simples contagem dessas *tags*.
- **Número de diferentes etiquetas morfossintáticas**: A proporção do número de diferentes de etiquetas é calculada com uma simples contagem da diferenciação de marcação gramatical dessas *tags*.
- **Erros de sintaxe**: Os erros de concordância, pontuação e de sentenças mal formuladas são coletados; para o modelo de predição apenas os números são utilizados; mas os erros em si podem ser um *feedback* para o estudante.
- **Erros ortográficos:** Conta o número de erros ortográficos identificados em cada texto.

- **Número de etiquetas por categoria sintática**: As etiquetas são agrupadas em categorias para a coleta destes números, conforme a Tabela 15 (quinze).

Tabela 15 – Representação da marcação sintática da etiquetagem conforme o corpus Tycho Brahe3.

Classe Morfológica	Etiqueta morfossintática
SR=ser, HV=haver, ET=estar, TR=ter, VB=verbo	-I -P -SP -D -RA -SD -R -SR -G -PP -NA
Gênero	None -F -G
Número	None -P
Substantivo	N
Nome próprio	NPR
Pronomes	PRO
Preposição + pronomes	P+PRO
Possessivo	PRO$
Clítico	CL
Determinante	D
Demonstrativo	DEM
Adjetivo	ADJ
Advérbio	ADV
Quantificador	Q
Conjunção	CONJ
Conjunção subordinada	C
Relativo	WPRO
Interrogativo	WQUE
Determinantes interrogativos	WD
Preposição	P

Fonte: Próprio autor (2020)

[3] http://www.tycho.iel.unicamp.br/~tycho/corpus/manual/pos2016.html

4.4 Dimensão Semântica

Nesta dimensão os atributos estão relacionados com a interpretação semântica do texto. A semântica do texto examina o seu conteúdo dando um significado para cada sentença. Dois textos são semanticamente similares, caso seus conteúdos transmitem a mesma mensagem. Tipicamente, compara-se a resposta do estudante com a resposta de referência, utilizando-se de diferentes estratégias e técnicas, tais como: Análise Semântica Latente, *n*-gramas, pesagem TF-IDF, métricas de similaridade como Cosseno e Distância Euclidiana.

Quanto à composição da resposta de referência existem algumas estratégias, neste trabalho foram utilizadas as seguintes:

- Uma única resposta de referência escrita por um especialista humano;
- Uma única resposta de referência formada a partir de um conjunto das respostas mais bem avaliadas;
- Várias respostas de referência, baseadas em agrupamentos feitos em relação ao escore (ZUPANC; BOSNIC, 2017).

Para cada uma das duas primeiras opções é gerado apenas um atributo. Para a terceira opção são gerados vários atributos. Por exemplo, nas respostas curtas de Geografia com uma faixa de valores de 0...6 foi criado 7 (sete) vetores resposta de referência, um para cada escore. Aqui foram aplicadas as medidas (Cosseno e Distância Euclidiana) contra estes vetores de respostas, incluindo também as variações no tipo de pré-processamento (SSW, CST, CSW). Desta estratégia resultam 2x3x7, 42 atributos.

4.5 Dimensão de Coerência

Nesta dimensão foi avaliada a coerência do texto, que é um **fenômeno semântico** (MANN; THOMPSON, 1988), que representa o fluxo de informação contido num texto. Estes atributos descrevem os aspectos referentes à coerência local dentro de uma resposta assim como a coerência global sobre várias respostas. Para isso, foi utilizada uma abordagem baseada em janelas sobrepostas (ZUPANC; BOSNIC, 2017; PALMA; ATKINSON, 2018). Foi utilizado quatro modelos combinando técnicas de pré-processamento e as duas medidas (Cosseno e Distância Euclidiana), geraram 3(a, b, c) x2(min/máx. med.) x2(cos, dist) x3(pre), que resultou em 36 (trinta e seis) atributos:

a. Distâncias entre duas janelas contíguas, medindo cada janela com a sua adjacente (Figura 19).

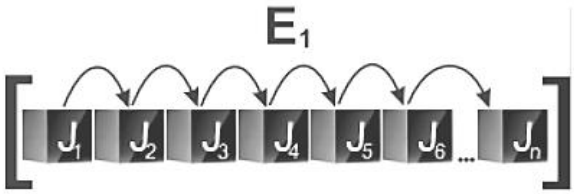


Figura 19 – Extração de Atributos entre janelas vizinhas.

b. Distâncias de todas as janelas contra todas, medimos cada janela com todas as outras (Figura 20).

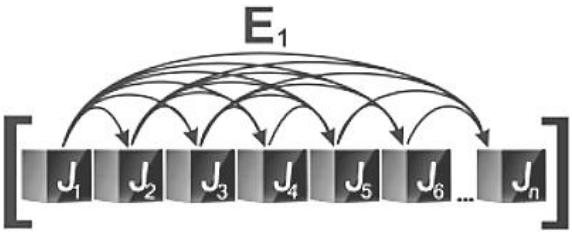


Figura 20 - Abordagem janela sobreposta medindo com todas as janelas.

c. Centro local, todas as janelas contra o centro local, medimos as janelas com os textos mais frequentes (Figura 21).

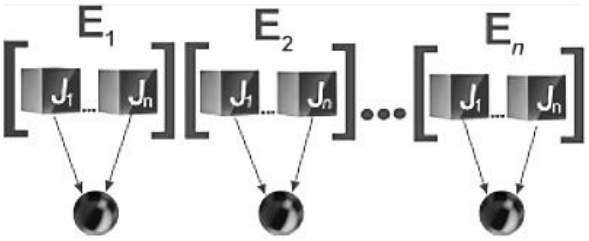


Figura 21 - Abordagem janela sobreposta medindo com os textos mais frequentes.

d. Centro global, todas as janelas contra o centro global, medimos as janelas de um texto com todos os textos mais frequentes de todo o conjunto (Figura 22).

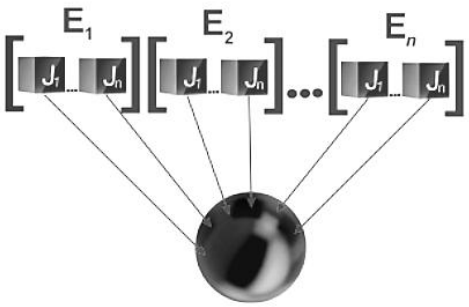


Figura 22 - Abordagem janela sobreposta medindo com o texto global.

4.6 Arquitetura de *pipeline*

Na literatura existe uma tendência para utilizar uma arquitetura de *pipeline* similar à da Figura 23 (vinte e três) para sistema de AAT (BURROWS; GUREVYCH; STEIN, 2015) que contém 5 (cinco) etapas: (1) preparação de *corpus*, (2) pré-processamento, (3) coleta de atributos, (4) modelo de predição e (5) avaliação. Em cada uma dessas etapas os pesquisadores utilizam diferentes bases, métodos e técnicas para no final gerar o escore para cada tipo de resposta.

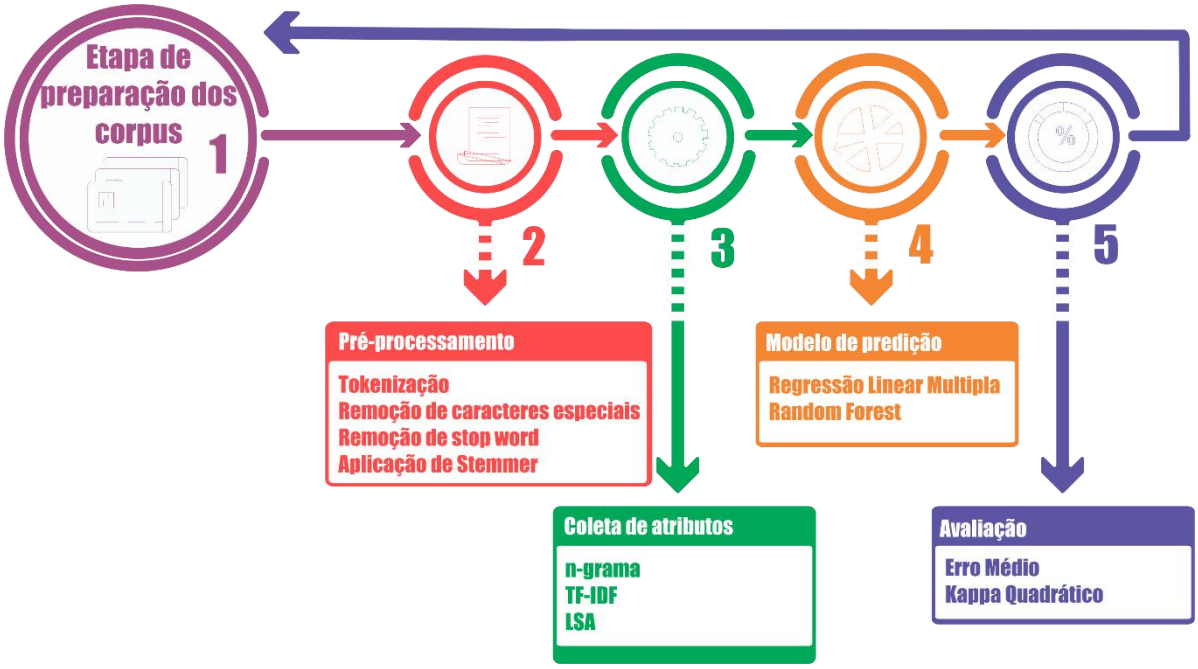


Figura 23 – Arquitetura em *pipeline* composta de cinco etapas para avaliação de textos.

A concatenação destes componentes é um '*pipeline*', no qual a saída de uma etapa é entrada para a próxima. A arquitetura *pipeline* é bem comum em vários campos da pesquisa de PLN, incluindo extração informação, extração de relações e preenchimento de modelos (WACHSMUTH; STEIN; ENGELS, 2011). Na sequência, examina-se cada um dos componentes desta arquitetura focando nos procedimentos adotados nesta pesquisa. O componente da coleta de atributos já foi detalhado nas seções anteriores.

4.6.1 Etapa de preparação do *Corpus*

A primeira etapa da preparação do *corpus* (Figura 24) de respostas discursivas do tipo texto, as respostas das questões abertas são digitadas e organizadas numa coleção. Cada resposta da base deverá ter pelo menos um escore de um avaliador humano, o ideal é dois escores humanos. Se possível para redações é bom ter o escore por tema da dimensão de avaliação, como por exemplo, as 5 (cinco) competências de avaliação da prova dissertativa do ENEM (1 - Uso correto do Português; 2 - Compreender e Desenvolver o Tema no estilo Dissertativo-Argumentativo; 3 - Defender seu Ponto de Vista com argumentos; 4 - Demonstrar capacidade de argumentação; 5 - Elaborar a Proposta de Intervenção). Os escore*s* nestas dimensões permitem trabalharmos com a engenharia de atributos, validando qual atributo pode contribuir na avaliação numa determinada dimensão.

Figura 24 – Primeira etapa da arquitetura

Quando existem os dois escores dos avaliadores humanos calculam-se a acurácia entre eles, que neste caso chama-se de Humano *versus* Humano ($H \times H$). Também com o escore dos humanos é possível calcular a acurácia do sistema contra o avaliador humano, que

neste caso foi chamado de Sistema *versus* Humano (*S*×*H*). A comparação destes dois escores permite verificar a eficácia do método sendo pesquisado.

4.6.2 Etapa de pré-processamento.

Na etapa de pré-processamento (Figura 25) busca-se uma representação "normalizada" do documento para extração do conhecimento, deixando apenas a informação relevante para o processo de avaliação (REZENDE; MARCACINI; MOURA, 2011). O principal objetivo desta etapa é tornar os dados menos esparsos, característica conveniente para o processamento computacional. Segundo Burrows; Gurevych; Stein, (2015) muitas técnicas de pré-processamento linguístico podem ser necessárias dependendo do tipo de texto e da área do problema a ser resolvido.

Figura 25 – Segunda etapa da arquitetura

Na Tabela 16 (dezesseis) apresenta-se uma taxonomia das técnicas de pré-processamento linguístico aplicadas nesta pesquisa. A primeira coluna mostra os níveis de aplicação no processo. A segunda coluna reflete os níveis de processamento da grande área de PLN. A terceira coluna apresenta técnicas para cada subárea: de superfície (ex. remoção de pontuação), léxico (ex. correção ortográfica e remoção de *stopword*), morfológico (ex. *stemming*) e sintático (classificar os *tokens*). A quarta coluna descreve os conceitos relacionados a cada subárea.

Tabela 16 – Representação de uma taxonomia para pré-processamento na área de AAT.

Níveis de Aplicação	Nível de PLN	Tipo	Descrição
1° Aplicação	Superfície	Normalização	Normalizar textos por meio da transformação de letras maiúsculas para minúsculas, remoção de caracteres especiais remover pontuações, numerais, acentuação, entre outros.
2° Aplicação	Léxico	Tokenização e Remoção de *stopword e*	Dividindo as palavras e removendo as que provocam ruídos.
		Stemmer	Reduzir a palavra ao seu radical, isso reduz vocabulário.
3° Aplicação	Morfossintático	Classes sintáticas	Divisão de frases em partes

Fonte: próprio autor (2020)

No início do pré-processamento foi utilizado um corretor gramatical para língua portuguesa brasileira de código aberto denominado CoGrOO (Corretor Gramatical para *Open Office*). Segundo (SILVA, 2013) o *software* é capaz de identificar erros como: colocação pronominal, concordância nominal, concordância sujeito-verbo, uso da crase, concordância nominal e verbal, e outros erros comuns de escrita. O CoGrOO funciona realizando uma análise híbrida: inicialmente o texto é anotado usando técnicas estatísticas de PLN e, em seguida, identifica os possíveis erros gramaticais. Este processamento com o corretor gramatical coleta os primeiros atributos referentes a erros gramaticais, os quais são utilizados na etapa de classificação. Além disso, nesta etapa, decidiu-se também fazer uma correção de palavras, substituindo as palavras inexistentes no vocabulário pela palavra mais indicada pelo corretor. Este ajuste facilita a comparação dos textos das respostas via LSA, eliminando palavras com grafia incorreta.

No pré-processamento, a transformação de letras maiúsculas para minúsculas contribui para o processo de normalização, assim como também a remoção de caracteres especiais, por exemplo, hífen, apóstrofes, entre outros e a remoção de pontuação, por exemplo, vírgulas, espaço em branco, colchetes, parênteses, entre outros.

O método de remoção de *stopwords* é bastante utilizado no campo da PLN. Cada palavra no texto é um termo, alguns desses termos possuem uma função gramatical, mas não

possuem um valor para efeito de comparação de similaridade de conteúdo (SALTON, 1989). Nos modelos de aprendizagem de máquina, ao remover esses termos economiza-se espaço e também, em alguns casos, espera-se bons resultados. Por exemplo, na língua portuguesa brasileira as palavras: 'a', 'as', 'ao', 'aos', 'aquela', 'aquelas', 'aquele', 'aqueles', 'aquilo', entre outras, são consideradas *stopwords*. Fizeram-se experimentos com e sem *stopwords*, para medir se devem ou não serem consideradas as *stopwords*.

O *stemming* é outro método bastante utilizado no campo da PLN na recuperação de informação. Segundo Orengo e Huyck (2001) define-se como um processo que reduz as formas variantes de uma palavra em uma representação ao seu radical. Por exemplo, as palavras: “apresentação”, “apresentando” e “apresentar”, todas são reduzidas para uma representação comum “apresent”. Existem vários algoritmos para *stemming*, nesta pesquisa foi utilizado o algoritmo *RSLPStemmer* para o português do Brasil (ORENGO; HUYCK, 2001).

Na organização da investigação para a etapa de pré-processamento, foi gerado experimentos com variações entre as técnicas: remoção de caracteres especiais e pontuação (+RCE); remoção de *stopwords* (+RSW); e remoção de sufixos (*stemming*) (+RSU). Estas técnicas foram combinadas de quatro modos:

- Com remoção de caracteres especiais (+RCE, –RSW, –RSU);
- Com remoção de caracteres especiais e *stopword* (+RCE, +RSW, –RSU);
- Com remoção de caracteres especiais, *stopword* e aplicação de *stemmer* (+RCE, +RSW, +RSU);

A técnica de tokenização é necessária em todos os casos, no entanto as outras três atividades são opcionais, portanto separando-as se pode medir a contribuição de cada uma delas na acurácia final.

Após a tokenização, os *tokens* foram analisados sintaticamente por um etiquetador morfossintático, o *software* livre Aelius (DRURY; ROSSI; DE ANDRADE LOPES, 2014; ALENCAR, 2013, p. 7; ALENCAR, 2015, p. 233;). O etiquetador Aelius é uma ferramenta para anotação automática de *corpora* que possui uma arquitetura híbrida, ou seja, recorre às regras, formuladas manualmente em expressões regulares, para etiquetar as palavras inexistentes no *language model* (ALENCAR, 2010, p. 3). O etiquetador utiliza vários tipos de *tag-sets* para treinamento de *corpus,* neste trabalho foi utilizado o modelo **AeliusRUBT**

treinado no *corpus Tycho Brahe Parsed Corpus of Historical Portugueses* com arquitetura NLTK. Segundo Alencar (2010, 2012), para o português contemporâneo o nível alcançado de acurácia da etiquetagem supera o de ferramentas análogas de *software* livre, tornando-se uma boa opção para anotação sintática dos nossos *corpora*. Acurácia do Aelius é de 95,4% (OTHERO; AYRES, 2014).

4.6.3 Etapa de Coleta de atributos

Na etapa de coleta de atributos (Figura 26) busca-se uma representação da expansão do vocabulário do documento para extração do conhecimento. Nesta etapa trabalha-se com LSA, n-gramas e TF-IDF.

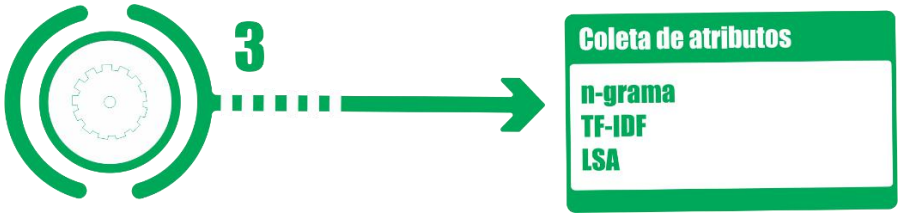


Figura 26 – Terceira etapa da arquitetura

4.6.3.1 Análise Semântica Latente (LSA)

LSA é uma tecnologia para recuperação de informação ou extração de conhecimento. No contexto de AAT permite uma comparação fina entre respostas de estudantes e respostas de referência, superando comparações baseadas apenas na sobreposição de palavras. A técnica captura uma "semântica latente" existente nos espaços entre as palavras do texto. LSA utiliza Álgebra Linear combinando modelo de espaço vetorial com o método científico matemático de Decomposição a Valores Singulares (SVD) para recuperação de informações textuais a partir de outros textos semanticamente associados (DEERWESTER *et al*., 1990).

A informação é tratada em um domínio estatístico presumindo que existe uma estrutura semântica latente no uso das palavras dos textos, que se entende como uma ocorrência linguística, que tem um sentido completo, dotada de certas formalidades que permitem estabelecer uma comunicação entre as partes. Esta abordagem baseia-se no fato de que as estimativas desta estrutura semântica são utilizadas para representar e recuperar informações.

Em ambientes computacionais os conteúdos dos textos normalmente são tratados por meio de indexações. LSA usa a representação vetorial para aquisição de conhecimentos contidos em textos. A representação é chamada vetorial pelo fato de que cada texto está representado por um vetor cujas coordenadas são as frequências das palavras que compõem o próprio texto. A primeira etapa de um método LSA está na criação de uma matriz inicial $A_{m \times n}$, sendo *m* o número de palavras diferentes e *n* o número de textos; assim, o *corpus* da pesquisa está representado pelo espaço das colunas da matriz *A*. A matriz *A* é submetida a uma transformação preliminar onde cada entrada é ponderada por uma função peso que estime a importância da palavra no texto em que ela está contida e seu grau de influência no *corpus*, por exemplo, com o método TF-IDF.

O próximo passo é o cálculo da SVD da matriz *A*. Esta decomposição revela a arquitetura das correlações entre as palavras nos textos. É a partir desta decomposição que o espaço das colunas de *A* é reduzido para formar o espaço semântico onde a etapa seguinte de classificação entre dados textuais é realizada. É consenso na literatura que o cálculo da SVD é a etapa de maior contribuição em todo método LSA.

4.6.3.1.1 Exemplo de uso de LSA num conjunto de dados textuais

Consideram-se como dados textuais nove títulos de memorandos técnicos: cinco sobre interação computador-homem e quatro sobre teoria gráfica de matemática retirada de (DEERWESTER *et al*., 1990).

C1: *Human machine interface for ABC computer applications*
C2: *A survey of user opinion of computer system response time*
C3: *The EPS user interface management system*
C4: *System and human system engineering testing of EPS*
C5: *Relation of user perceived response time to error measurement*
M1: *The generation of random, binary, ordered trees*
M2: *The intersection graph of paths in trees*
M3: *Graph minors IV: Widths of trees and well-quasi-ordering*
M4: *Graph minors: A survey*

A primeira etapa de um método LSA é a construção da matriz que representa os dados textuais. Considerando apenas os termos que aparecem em pelo menos dois títulos, a matriz inicial é a seguinte:

$$A = \begin{bmatrix} & C1 & C2 & C3 & C4 & C5 & M1 & M2 & M3 & M4 \\ human & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ interface & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ computer & 1 & 1 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ user & 0 & 1 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ system & 0 & 1 & 1 & 2 & 0 & 0 & 0 & 0 & 0 \\ response & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ time & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ EPS & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ survey & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ trees & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 \\ graph & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\ minors & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \end{bmatrix}$$

As colunas da matriz *A* são representações vetoriais dos respectivos títulos no espaço gerado pelas suas colunas. Sejam θ_1 e θ_2 os ângulos entre os vetores coluna C1, C2 e C1, M1, respectivamente. Então $\cos\theta_1 = 0{,}3086$ e $\cos\theta_2 = 0$, e assim, os títulos C1 e C2 apresentam alguma similaridade, enquanto que os títulos C1 e M1 não apresentam qualquer similaridade.

Na segunda etapa, a matriz inicial é submetida a uma transformação preliminar, chamada função ponderação (esta etapa é opcional) onde a matriz *A* não é submetida à transformação de ponderação. Na terceira etapa, é feito o cálculo do SVD da matriz *A*. Considerando apenas os vetores singulares associados aos autovalores singulares não-nulos, a SVD das colunas numéricas da matriz *A* é dada por $A = USV^t$, onde:

$$U = \begin{bmatrix} -0.2022 & -0.0547 & 0.0738 & 0.6496 & -0.1989 & -0.1013 & 0.0878 & -0.1598 & 0.1276 \\ -0.2454 & 0.02550 & -0.1885 & 0.1858 & -0.3340 & 0.64803 & -0.1472 & 0.0171 & -0.4051 \\ -0.48745 & -0.0124 & -0.2388 & 0.3739 & -0.0002 & -0.0905 & -0.1367 & 0.0490 & 0.3345 \\ -0.4187 & -0.0056 & -0.1748 & -0.0222 & 0.2421 & -0.4035 & 0.3589 & -0.1141 & -0.6559 \\ -0.4854 & -0.1307 & 0.6463 & -0.3060 & -0.1013 & 0.2441 & -0.1311 & 0.0283 & -0.0814 \\ -0.2852 & 0.0423 & -0.3126 & -0.2757 & 0.1987 & 0.0108 & -0.2244 & 0.0650 & 0.2068 \\ -0.2852 & 0.0423 & -0.3126 & -0.2757 & 0.1987 & 0.0108 & -0.2244 & 0.0650 & 0.2068 \\ -0.2211 & -0.1055 & 0.4542 & 0.0788 & 0.0624 & -0.2525 & 0.0519 & -0.0024 & 0.2256 \\ -0.1942 & 0.2439 & -0.1057 & -0.3218 & -0.4972 & 0.0684 & 0.6428 & -0.1553 & 0.3140 \\ -0.0064 & 0.5168 & 0.0133 & 0.2105 & 0.5953 & 0.3896 & 0.2155 & -0.3247 & 0.1112 \\ -0.0023 & 0.6481 & 0.1279 & 0.0299 & -0.0806 & -0.1442 & -0.0350 & 0.7218 & -0.1116 \\ -0.0213 & 0.4664 & 0.0806 & -0.0487 & -0.3145 & -0.3126 & -0.4913 & -0.5693 & -0.1039 \end{bmatrix}$$

$$S = \begin{bmatrix} 3.6874 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 2.5318 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 2.3494 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1.747 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1.4835 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1.2071 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0.6289 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.5545 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.2399 \end{bmatrix}$$

$$V = \begin{bmatrix} -0.2535 & -0.6512 & -0.4921 & -0.3232 & -0.4004 & -0.0081 & -0.0139 & -0.0648 & \\ -0.0164 & 0.0810 & -0.1221 & -0.1449 & 0.0262 & 0.2041 & 0.4601 & 0.6443 & 0.5365 \\ -0.1504 & -0.2923 & 0.3237 & 0.7435 & -0.4421 & 0.0568 & 0.1113 & 0.1456 & 0.0437 \\ 0.6920 & -0.3674 & 0.4430 & -0.3051 & -0.1143 & 0.1204 & 0.1376 & 0.1097 & -0.1949 \\ -0.3593 & -0.1361 & 0.0642 & 0.0284 & 0.4310 & 0.4013 & 0.3469 & 0.1349 & -0.6015 \\ 0.3778 & 0.4044 & -0.5002 & 0.1952 & -0.3913 & 0.3227 & 0.2033 & -0.0556 & -0.3217 \\ -0.3117 & 0.2192 & 0.3669 & -0.3343 & -0.3604 & 0.3427 & 0.2870 & -0.4941 & 0.1851 \\ 0.0904 & -0.0809 & -0.0993 & 0.0980 & 0.1170 & -0.5855 & 0.7160 & -0.3105 & -0.0051 \\ 0.2376 & -0.3348 & -0.2070 & 0.2611 & 0.3844 & 0.4636 & -0.0018 & -0.4351 & 0.4101 \end{bmatrix}$$

Estas matrizes refletem uma fatoração da matriz original em autovetores linearmente independentes.

4.6.3.1.2 Continuação da aplicação de um modelo LSA

Na quarta etapa de classificação se deve escolher a dimensão *k* do espaço semântico. Tomando *k* = 2, isto é, escolhendo as duas primeiras colunas da matriz *U*, as duas primeiras linhas e colunas da matriz *S* e as duas primeiras colunas da matriz *V*, obtêm-se a matriz:

$$A_k = \begin{bmatrix} & C1 & C2 & C3 & C4 & C5 & M1 & M2 & M3 & M4 \\ human & 0.16 & 0.40 & 0.38 & 0.47 & 0.18 & -0.05 & -0.12 & -0.16 & -0.09 \\ interface & 0.14 & 0.37 & 0.33 & 0.40 & 0.16 & -0.03 & -0.07 & -0.10 & -0.04 \\ computer & 0.15 & 0.51 & 0.36 & 0.41 & 0.24 & 0.02 & 0.06 & 0.09 & 0.12 \\ user & 0.26 & 0.84 & 0.61 & 0.70 & 0.39 & 0.03 & 0.08 & 0.12 & 0.19 \\ system & 0.45 & 1.23 & 1.05 & 1.27 & 0.56 & -0.07 & -0.15 & -0.21 & -0.05 \\ response & 0.16 & 0.58 & 0.38 & 0.42 & 0.28 & 0.06 & 0.13 & 0.19 & 0.22 \\ time & 0.16 & 0.58 & 0.38 & 0.42 & 0.28 & 0.06 & 0.13 & 0.19 & 0.22 \\ EPS & 0.22 & 0.55 & 0.51 & 0.63 & 0.24 & -0.07 & -0.14 & -0.20 & -0.11 \\ survey & 0.10 & 0.53 & 0.23 & 0.21 & 0.27 & 0.14 & 0.31 & 0.44 & 0.42 \\ trees & -0.06 & 0.23 & -0.14 & -0.27 & 0.14 & 0.24 & 0.55 & 0.77 & 0.66 \\ graph & -0.06 & 0.34 & -0.15 & -0.30 & 0.20 & 0.31 & 0.69 & 0.98 & 0.85 \\ minors & -0.04 & 0.25 & -0.10 & -0.21 & 0.15 & 0.22 & 0.50 & 0.71 & 0.62 \end{bmatrix}$$

Como foi visto na seção anterior, não foram utilizadas as colunas da matriz A_k para estimar similaridades entre os títulos, mas as colunas da matriz a seguir:

$$S_k \cdot V_k^t = \begin{bmatrix} C1 & C2 & C3 & C4 & C5 & M1 & M2 & M3 & M4 \\ 0.9348 & -0.0605 & -0.5546 & 2.5517 & 1.3249 & 1.3931 & -1.1494 & 0.3333 & 0.8761 \\ -1.6487 & 0.2051 & -0.7400 & -0.9302 & 0.3446 & 1.0238 & 0.5549 & -0.2048 & -0.8476 \end{bmatrix}$$

As colunas da matriz $S_k \cdot V_k^t$ são as coordenadas de cada título no espaço semântico 2-dimensional. Calculando novamente o cosseno dos ângulos θ_1 e θ_2, entre os vetores coluna $C1$, $C2$ e $C1$, $M1$ no espaço 2-dimensional, se obtêm $\cos\theta_1 = 0.6948$ e $\cos\theta_2 = 0.2583$. Portanto, o modelo LSA revela um aumento na similaridade entre os títulos $C1$ e $C2$ e alguma correlação entre os títulos $C1$ e $M1$.

4.6.3.2 *N*-gramas para comparação de respostas

Um *n*-grama pode ser simplesmente definido como uma sequência contígua de *n* itens de uma determinada sequência de texto. O *n* pode ser qualquer valor, mas geralmente apenas o valor de *n* até 3 (três) é praticamente útil, pois 4-gramas entre duas respostas são bem raros, por exemplo, trigrama serve por exemplo para identificar plágio. Assim, foi trabalhado apenas com 1-2-3-gramas, respectivamente unigramas, bigramas e trigramas, como na Tabela 17 (dezessete).

Tabela 17 – Representação matemática de *n*-gramas.

***n*-gramas**	**Equação**
Unigrama	$P(w)$
Bigrama	$P(W_i \mid W_{i-1})$
Trigrama	$P(W_i \mid\mid W_{i-1} \mid W_{i-2})$

Fonte: Próprio autor (2020)

Para efeito de ilustração a Figura 27 (vinte e sete) apresenta exemplos de 1-2-3-gramas. Note que sempre é necessário quebrar a sentença em *tokens* (unigramas) para formar os

bigramas e trigramas. Esta técnica pode ser aplicada com ou sem as opções de pré-processamento já descritas.

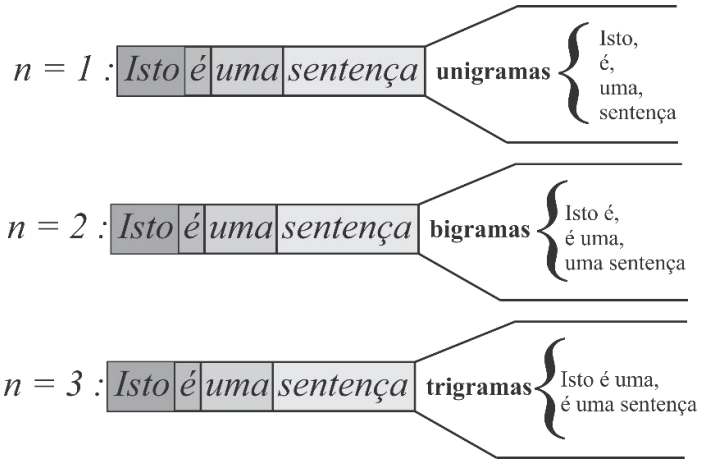


Figura 27 – Um exemplo da funcionalidade o *n*-gramas.

4.6.3.3 Métricas para medir a similaridade entre textos

Na coleta das métricas de similaridade foram trabalhadas medidas de frequência de termos e com medidas da teoria dos conjuntos. Na frequência de termos foi trabalhado com três métricas: cosseno-FT (Frequência de Termos) e distância euclidiana (como similaridade é 1/ (distância)); e também foram combinadas estas duas variáveis via regressão, numa única variável. Na teoria dos conjuntos foi trabalhado com cinco métricas: *overlap* mínimo, *overlap* máximo, *jaccard*, *dice* e cosseno-Conj. Nestas medidas são utilizadas as operações sobre teoria dos conjuntos, card ( ), inter ( ) e union ( ). Para efeito de ilustração a Tabela 18 (dezoito) mostra as fórmulas destas métricas.

Tabela 18 – Fórmulas das métricas de similaridade, baseadas em teoria dos conjuntos e em frequência dos termos.

Duas listas de termos *A* e *B*		**Duas listas numéricas *A* e *B* (vetores)**	
Jaccard	$\frac{card(A \cap B)}{card(A \cup B)}$	Distância Euclidiana	$\sqrt{\sum_i (a_i - b_i)^2}$
$Overlap_{min}$	$\frac{card(A \cap B)}{\min(card(A), card(B))}$	Cosseno-FT	$\frac{< A,B >}{\|A\|\|B\|}$
$Overlap_{max}$	$\frac{card(A \cap B)}{\max(card(A), card(B))}$		
Dice	$\frac{card(A \cap B)}{card(A) + card(B)}$		
Cosseno	$\frac{card(A \cap B)}{\sqrt{card(A) \times card(B)}}$		

Fonte: Próprio autor (2020).

4.6.3.4 *Term frequency–inverse document frequency* (TF-IDF)

O TF-IDF é um dos algoritmos mais reconhecidos na pesquisa de mineração de texto (XIA; CHAI, 2011) para a pesagem dos termos, substituindo os valores brutos de frequência por valores pesados dentro da teoria TF-IDF. Frequência de termo (TF) é o número em que uma palavra pode ser encontrada em um ensaio ou documento e IDF é baseado no *log* da probabilidade inversa de uma palavra ser encontrada em qualquer ensaio (LIU; YANG, 2012).

O TF-IDF é calculado pela Equação 14 (quatorze) para um determinado termo *x* dentro do documento *y*.

$$W_{x,y} = tf_{x,y} \times \log\left(\frac{N}{df_x}\right) \tag{14}$$

Onde,

- $tf_{x,y}$ = frequência de *x* em *y;*
- df_x = número de documentos contendo *x;*
- N = número total de texto.

Detalhando a Equação 14 (quatorze) nas Equações 15 (quinze) e 16 (dezesseis):

- **Term Frequency (TF)**: pontuação da frequência da palavra no documento atual.

$$\boldsymbol{TF(termo)} = \frac{número\ de\ vezes\ que\ o\ termo\ aparece\ em\ um\ documento}{número\ total\ de\ itens\ no\ documento} \quad (15)$$

- **Inverse Term Frequency (ITF):** pontuação de quão rara é a palavra nos documentos.

$$\boldsymbol{IDF(term)} = log\left(\frac{número\ total\ de\ documentos}{número\ de\ documentos\ com\ termo}\right) \quad (16)$$

Usando as equações anteriores se pode calcular o TF-IDF (Equação 17).

$$\boldsymbol{TF - IDF} = TF(term) \times IDF(term) \quad (17)$$

A Tabela 19 apresenta um exemplo do TF-IDF:

Tabela 19 – Exemplo do método TF-IDF

Sentença 1:	*um exemplo simples com gatos e ratos*
Sentença 2:	*outro exemplo simples com caes e gatos*
Sentença 3:	*outro txt com ratos e queijo*
Frequência locais, GLO global:	
{'caes': 0, 'queijo': 0, 'txt': 0, 'outro': 0, 'com': 1, 'e': 1, 'gatos': 1, 'exemplo': 1, 'simples': 1, 'ratos': 1, 'um': 1}	
{'caes': 1, 'queijo': 0, 'txt': 0, 'outro': 1, 'com': 1, 'e': 1, 'gatos': 1, 'exemplo': 1, 'simples': 1, 'ratos': 0, 'um': 0}	
{'caes': 0, 'queijo': 1, 'txt': 1, 'outro': 1, 'com': 1, 'e': 1, 'gatos': 0, 'exemplo': 0, 'simples': 0, 'ratos': 1, 'um': 0}	
Ponderação local ajustada: TF	
{'caes': 0.0, 'queijo': 0.0, 'txt': 0.0, 'outro': 0.0, 'com': 0.14, 'e': 0.14, 'gatos': 0.14, 'exemplo': 0.14, 'simples': 0.14, 'ratos': 0.14, 'um': 0.14}	
{'caes': 0.14, 'queijo': 0.0, 'txt': 0.0, 'outro': 0.14, 'com': 0.14, 'e': 0.14, 'gatos': 0.14, 'exemplo': 0.14, 'simples': 0.14, 'ratos': 0.0, 'um': 0.0}	
{'caes': 0.0, 'queijo': 0.17, 'txt': 0.17, 'outro': 0.17, 'com': 0.17, 'e': 0.17, 'gatos': 0.0, 'exemplo': 0.0, 'simples': 0.0, 'ratos': 0.17, 'um': 0.0}	
Ponderação global de cada palavra: IDF	
{'caes': 1.1, 'queijo': 1.1, 'txt': 1.1, 'outro': 0.41, 'com': 0.0, 'e': 0.0, 'gatos': 0.41, 'exemplo': 0.41, 'simples': 0.41, 'ratos': 0.41, 'um': 1.1}	
Ponderação Local x Global: tf_idf	
{'caes': 0.0, 'queijo': 0.0, 'txt': 0.0, 'outro': 0.0, 'com': 0.0, 'e': 0.0, 'gatos': 0.06, 'exemplo': 0.06, 'simples': 0.06, 'ratos': 0.06, 'um': 0.15}	
{'caes': 0.15, 'queijo': 0.0, 'txt': 0.0, 'outro': 0.06, 'com': 0.0, 'e': 0.0, 'gatos': 0.06, 'exemplo': 0.06, 'simples': 0.06, 'ratos': 0.0, 'um': 0.0}	
{'caes': 0.0, 'queijo': 0.19, 'txt': 0.19, 'outro': 0.07, 'com': 0.0, 'e': 0.0, 'gatos': 0.0, 'exemplo': 0.0, 'simples': 0.0, 'ratos': 0.07, 'um': 0.0}	

Fonte: próprio autor (2020)

4.6.4 Etapa de predição

A Figura 28 (vinte e oito) apresenta à quarta etapa, os primeiros experimentos foram realizados com regressão linear múltipla, que permite combinar diversas variáveis num escore. À medida que o número de atributos fora crescendo, precisou-se evoluir para outros métodos e, o mais promissor foi o algoritmo *Random Forest* (BREIMAN, 2001). Este é um método de aprendizagem supervisionada que permite a combinação de inúmeros atributos num escore. Ele permite regressão e/ou classificação para encontrar padrões nos dados. Embutido nele também tem uma fase de preparação de atributos, assim após o experimento é possível listar os atributos que mais contribuíram com a classificação. Pode-se a partir da listagem fazer experimentos, também, com apenas os principais atributos para ver se a acurácia se mantém. De certo modo, o algoritmo produz bons resultados ajustando e selecionando os melhores atributos.

Figura 28 - Quarta etapa da arquitetura.

O algoritmo de *Random Forest* é baseado na técnica de árvore de decisão que f modelos preditivos de alta precisão, estabilidade e facilidade de interpretação. Ele cria um grande número de árvores de decisão individuais que funcionam como um subproblema, onde cada sub-árvore é treinada por um subconjunto diferente de dados de treinamento. Numa etapa final a montagem das sub-árvores compõe o resultado final (Figura 29).

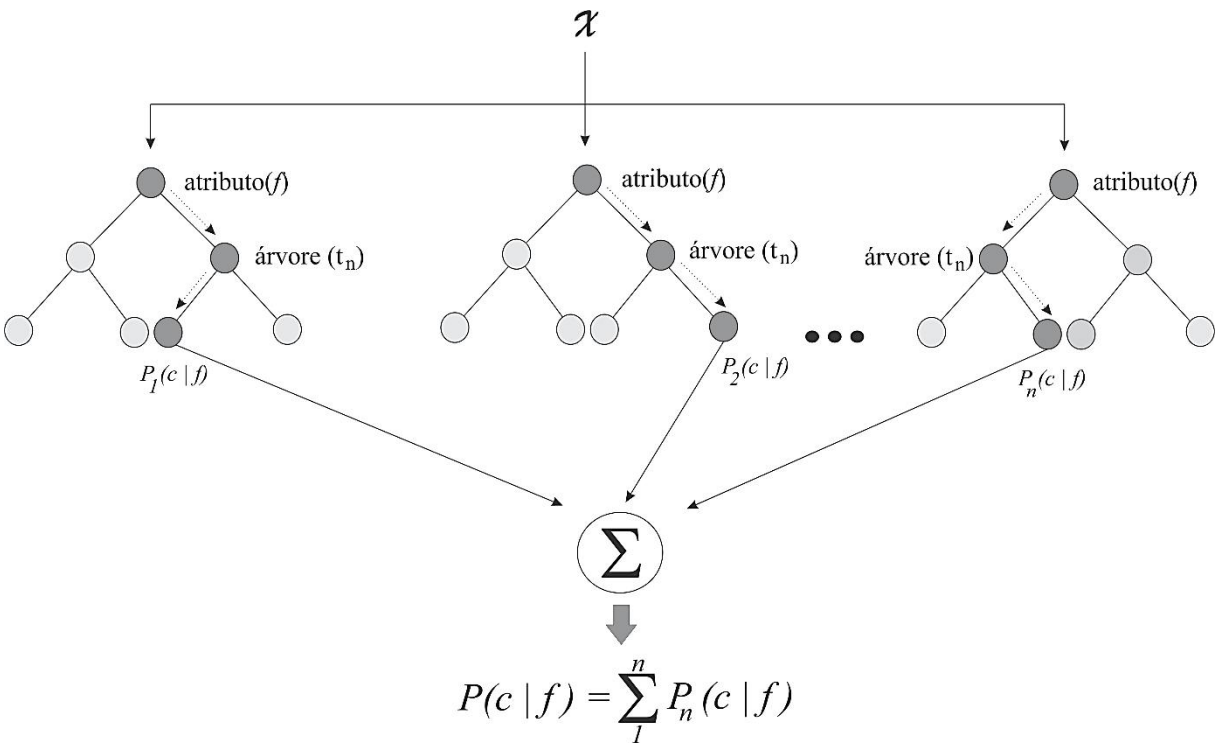


Figura 29 – Estrutura do algoritmo *Random Forest* baseado em árvore de decisão.

Para validação foi utilizado a abordagem *Cross-validation*, particionando o conjunto de dados em 5 (cinco) *folds* (Figura 30); cada *fold* é rodado duas vezes coletando a média entre dez testes, no final avalia-se acurácia do experimento.

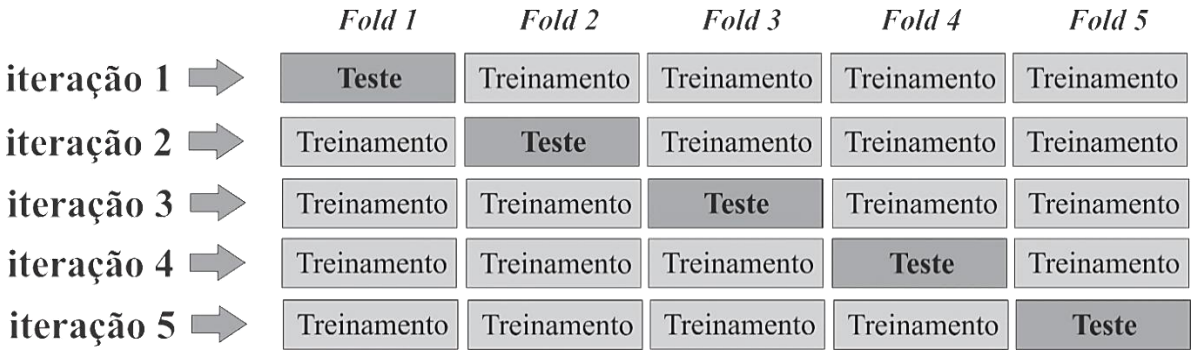


Figura 30 – Exemplo do método *cross-validation* em 5 (cinco) *folds.*

4.6.5 Etapa de avaliação

Na etapa de avaliação (Figura 31), procura-se selecionar as melhores combinações das etapas anteriores buscando maximizar a precisão.

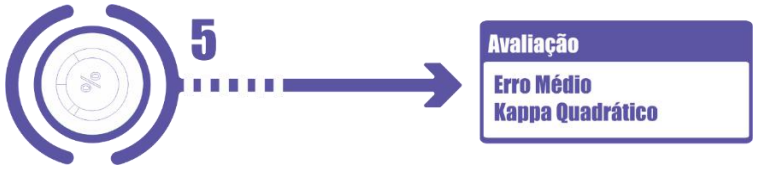


Figura 31 – Quinta etapa da arquitetura

Num primeiro experimento sobre respostas curtas foi utilizada a medida de erro médio (Equação 18).

$$acuracia = \frac{6 - erro_{medio}}{6} \times 100 \quad (18)$$

Nos outros experimentos, considerando que nos trabalhos mais recentes os pesquisadores estavam utilizando correlação e Kappa quadrático, no qual passou a utilizar o kappa quadrático. O Kappa Quadrático - KQ (FLEISS; COHEN, 1973) mede o grau de concordância entre duas classes com certa flexibilidade em relação à concordância exata. O KQ mede também a concordância parcial: se devia predizer 6 (seis), mas se resultou em 5 (cinco), não é totalmente errado. Essa métrica geralmente varia de 0 (pouca concordância entre avaliadores) a 1 (concordância completa entre avaliadores). Caso a concordância entre os avaliadores seja abaixo do mínimo esperado, essa métrica também pode resultar em valores negativos.

O KQ é calculado criando-se uma matriz de acordo com a Equação 19 (dezenove) e 20 (vinte). Neste caso, a matriz O contém as pontuações, de tal modo que a classificação i é dada pelo avaliador humano e j dada pelo modelo. $W_{i,\ j}$ contém os pesos como derivado na Equação 15 (quinze) e a matriz E contém as pontuações esperadas dos avaliadores humanos, obtidas pela multiplicação dos vetores de histograma das duas pontuações. Os subscritos em

matriz $O_{i,\,j}$ correspondem ao número de respostas que pontuam i do avaliador humano e j do sistema.

$$W_{i,j} = \frac{(i-j)^2}{(N-1)^2} \tag{19}$$

No final do processo KQ calculado como:

$$K = 1 - \frac{\sum_{i,j} W_{i,j} O_{i,j}}{\sum_{i,j} W_{i,j} E_{i,j}} \tag{20}$$

A interpretação dos resultados de acurácia, entre 0 (zero) e 1 (um), entre pouca e muita concordância, pode ser um tanto subjetiva. Portanto, foi citada abaixo uma interpretação recomendada por Landis e Koch (LANDIS; KOCH, 1977) que considera seis faixas de valores (Tabela 20):

Tabela 20 – Representação da força de concordância do Kappa Quadrático.

KQ	Força do acordo
< 0.00	Pobre
0.00 - 0.20	Fraco
0.21 - 0.40	Razoável
0.41 - 0.60	Moderado
0.61 - 0.80	Substancial
0.81 - 1.00	Quase perfeito

Fonte: Adaptado em Landis e Koch (1977).

5 Avaliação de respostas tipo curtas

Neste capítulo descrevem-se dois experimentos com avaliação de respostas curtas. O foco é avaliação de questões discursivas na língua portuguesa brasileira e o primeiro desafio foi conseguir um *corpus* de questões curtas. O *corpus* a ser utilizado foi detalhado no Capítulo 3 (três), onde foram trabalhadas as questões de Biologia, Geografia e Filosofia. As duas primeiras são oriundas de um processo seletivo da UFPA e a terceira é oriunda de um curso *on-line* da mesma instituição.

Um dos problemas das respostas curtas é que elas possuem um número limitado de palavras, por exemplo, para a questão de Biologia em média foram 28 (vinte e oito) palavras, para a de Geografia em média 149 (cento e quarenta e nove) palavras e para a de Filosofia em média 74 (setenta e quatro) palavras. Assim, muitos dos atributos descritos no Capítulo 4 (quatro), no item coleta de atributos, não se aplicam para as respostas curtas. Autores sugerem apenas o trabalho com a dimensão do conteúdo (GOMAA; FAHMY, 2014). Portanto, foram feitos dois experimentos, sendo um somente com atributos de conteúdo e outro com mais dimensões linguísticas:

a) Somente com atributos da dimensão de conteúdo, com variações em: i) formas de pré-processamento, ii) medidas de similaridade, da teoria dos conjuntos e de frequência de termos; iii) cosseno e distância euclidiana; iv) unigramas e bigramas;

b) Com atributos da dimensão de conteúdo mais outros das dimensões léxica e sintática.

Este capítulo está organizado em duas principais seções, um para cada experimento.

5.1 Experimento 1: somente com atributos de conteúdo

Na avaliação automática de questões discursivas foram encontradas algumas abordagens para avaliar respostas do tipo curta (Capítulo 2). Uma dessas abordagens é centrada nas medidas de similaridade entre textos, comparando-se cada resposta de um estudante contra uma resposta de referência, que é a abordagem a ser utilizada neste experimento. Neste

comparativo foram passados por diversas etapas como pré-processamento, coleta de atributos, modelo de predição e avaliação.

Na etapa de pré-processamento, as respostas foram vetorizadas em sentenças e em seguidas tokenizadas, utilizando-se comandos da biblioteca NLTK (`nltk.word_sent` e `nltk.word_tokenize`) (BIRD; KLEIN; LOPER, 2009). As três técnicas de pré-processamento foram utilizadas: (1) Remoção de Caracteres Especiais, pontuação, acentuação e conversão de letras maiúsculas em letras minúsculas (RCE); (2) Remoção de *stopword* (RSW) e; (3) Remoção de sufixos (*stemmer*) (RSU). As técnicas foram combinadas da seguinte forma: a) sem pré-processamento (-RCE, -RSW, -RSU); b) com remoção de caracteres especiais (+RCE, -RSW, -RSU); c) com remoção de caracteres especiais e *stopword* (+RCE, +RSW, -RSU) e; d) com remoção de caracteres especiais, *stopword* e aplicação de *stemmer* (+RCE, +RSW, +RSU).

Associada a esta etapa se tem a questão de pesquisa: **Pré-processamento (QP2)** quais as melhores técnicas de pré-processamento para abordagem de similaridade de texto, para questões curtas? O pré-processamento influencia na acurácia final para abordagem de similaridade de texto?

Quanto à resposta de referência para as questões curtas existe uma abordagem que é criar a resposta de referência com as melhores respostas da base de treinamento ou criar uma resposta de referência com o auxílio de um especialista humano. Em (BURROWS; GUREVYCH; STEIN, 2015; PÉREZ *et al*., 2005) relata-se que a pontuação automática depende diretamente da resposta de referência, como por exemplo, uma resposta com pouco vocabulário leva a resultados medíocres. Para as respostas de Biologia já existia uma resposta de referência dada por um especialista humano, neste caso não foram concatenadas as respostas por se tratar de uma questão conceitual. Já na questão de Geografia, a composição da resposta de referência foi pela concatenação das melhores quatro respostas bem avaliada.

Como já tinha a resposta de referência de Biologia propõe-se uma questão de pesquisa**:** **Resposta de Referência (QP3)** O que é melhor, ter uma única resposta de referência dada por um especialista humano, ou compor uma resposta de referência a partir da concatenação das melhores respostas?

Para medir a similaridade entre textos se podem utilizar medidas da teoria dos conjuntos como *Overlap* e *Dice* ou se pode utilizar medidas que levam em consideração a frequência dos termos. Em (PRIBADI *et al*., 2017) argumenta-se que para respostas curtas as medidas de

similaridade com frequência de termos não têm relevância, basta trabalhar só com as métricas de similaridade de conjuntos. Para testar esta hipótese se tem a questão de pesquisa: **medidas da teoria dos conjuntos *versus* frequência de termos (QP4):** É melhor trabalhar com interseção de conjuntos ou frequência dos termos? Qual é a melhor medida de similaridade de conjuntos?

Ainda sobre métricas, alguns autores (OLMOS *et al*., 2011) argumentam que o cosseno mede uma dimensão de proximidade angular enquanto que a distância euclidiana mede a quantidade de conteúdo, e que deverias se utilizar as duas medidas combinadas. Para testar esta hipótese se tem a questão de pesquisa: **Unigrama *versus* Bigrama (QP5**): O uso de bigramas minimiza o problema? Combinando bigramas com unigramas se tem uma boa acurácia?

Por fim, se tem a questão de pesquisa: **Acurácia (QP6)**: O método de avaliação centrado em atributos de conteúdo alcança a acurácia dos avaliadores humanos?

5.2 Resultados e discussão do Experimento 1: atributos de conteúdo

Esta seção foi organizada a partir das questões de pesquisa. Para cada questão são mostrados os resultados e é feita uma breve discussão.

5.2.1 Questão de pesquisa 2: Pré-processamento (QP2) - *Quais as melhores técnicas de pré-processamento para abordagem de similaridade de texto? O pré-processamento influencia na acurácia final para abordagem de similaridade de texto?*

Foram utilizadas três técnicas de pré-processamento: (1) Remoção de Caracteres Especiais, pontuação, acentuação e conversão de letras maiúsculas em letras minúsculas (RCE); (2) Remoção de *stopword* (RSW) e; (3) Remoção de sufixos (*stemmer*) (RSU). Estas técnicas foram agrupadas da seguinte forma:

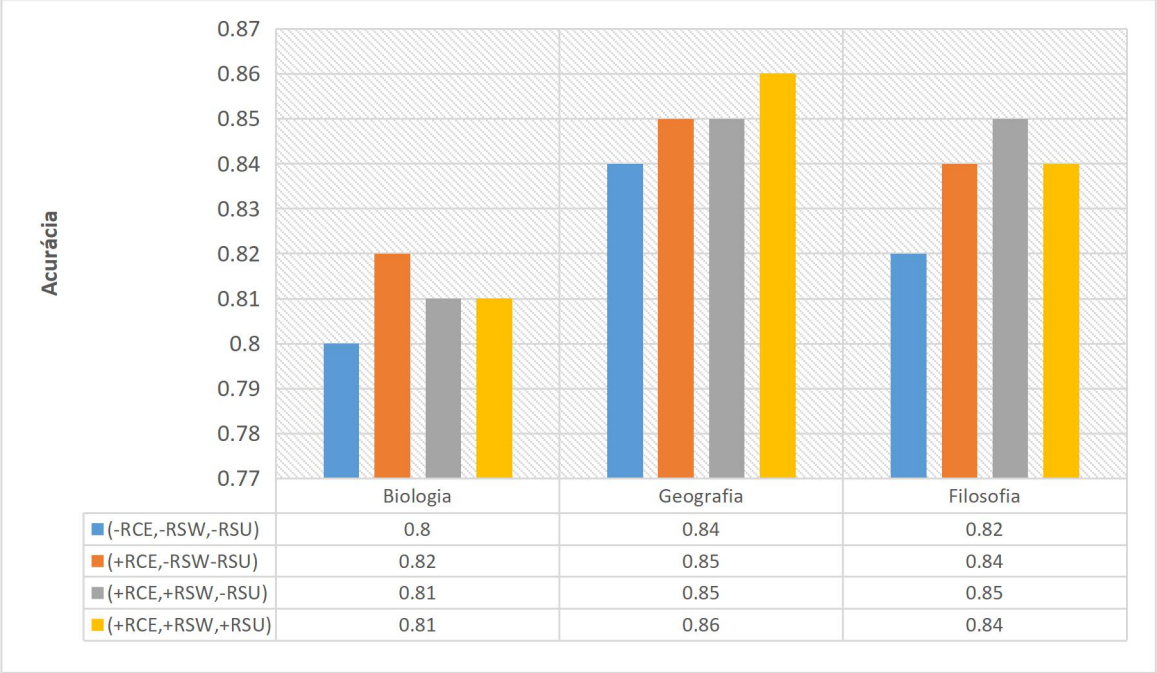
a) sem pré-processamento (-RCE, -RSW, -RSU);

b) com remoção de caracteres especiais (+RCE, -RSW, -RSU);

c) com remoção de caracteres especiais e *stopword* (+RCE, +RSW, -RSU) e;

d) com remoção de caracteres especiais, *stopword* e aplicação de *stemmer* (+RCE, +RSW, +RSU).

A Figura 32 (trinta e dois) apresenta os resultados do experimento realizado para as questões de Biologia, Geografia e Filosofia considerando as variações nas técnicas de pré-processamento e o resultado de acurácia. Nestes experimentos foi utilizada a métrica do erro médio.



	Biologia	Geografia	Filosofia
(-RCE,-RSW,-RSU)	0.8	0.84	0.82
(+RCE,-RSW-RSU)	0.82	0.85	0.84
(+RCE,+RSW,-RSU)	0.81	0.85	0.85
(+RCE,+RSW,+RSU)	0.81	0.86	0.84

Figura 32 – Comparativo das técnicas de pré-processamento para cada uma das questões de respostas curtas (**Métrica:** erro médio).

Neste experimento, o pré-processamento com remoção de caracteres especiais, *stopword* e aplicação de *stemmer* (+RCE, +RSW, +RSU), obteve bons resultados em média em relação às demais combinações das técnicas de pré-processamento. A única observação em geral com relação a este experimento foi o desempenho próximo da combinação do pré-processamento com remoção de caracteres especiais e *stopword* (+RCE, +RSW, -RSU).

O pré-processamento mostra-se bem relevante para cada *corpus*, com uma diferença entre 0.2 e 0.01 pontos com relação à acurácia de melhor e de pior resultado. Isso demonstra que independente de tipo de questão (conceitual e argumentativo) as técnicas de pré-processamento influenciam um pouco nos resultados da acurácia.

As diferentes técnicas de pré-processamento apresentaram valores de acurácia divergentes. No entanto, as diferenças não são tão significativas dentro das bases, sendo a diferença do menor para o maior valor 0.02 para Biologia, 0.02 para Geografia e de 0.03 para Filosofia. Em relação a cada base, em média a diferença entre cada combinação é de 0.05. A diferença entre o menor valor de Biologia para o maior valor de Geografia é de 0.07 e o inverso é de 0.02.

Para Biologia a melhor técnica foi com remoção de caracteres especiais (+RCE, -RSW, -RSU) obtendo uma acurácia de 0.82; Para Geografia a melhor técnica foi com remoção de caracteres especiais, *stopword* e aplicação de *stemmer* (+RCE, +RSW, +RSU) obtendo uma acurácia de 0.86 e para a prova de Filosofia a melhor técnica foi com remoção de caracteres especiais e *stopword* (+RCE, +RSW, -RSU) obtendo uma acurácia de 0.85.

5.2.2 Questão de pesquisa 3: Resposta de Referência (QP3) - *O que é melhor, ter uma única resposta ouro dada por um especialista humano, ou compor uma resposta ouro a partir da concatenação das melhores respostas?*

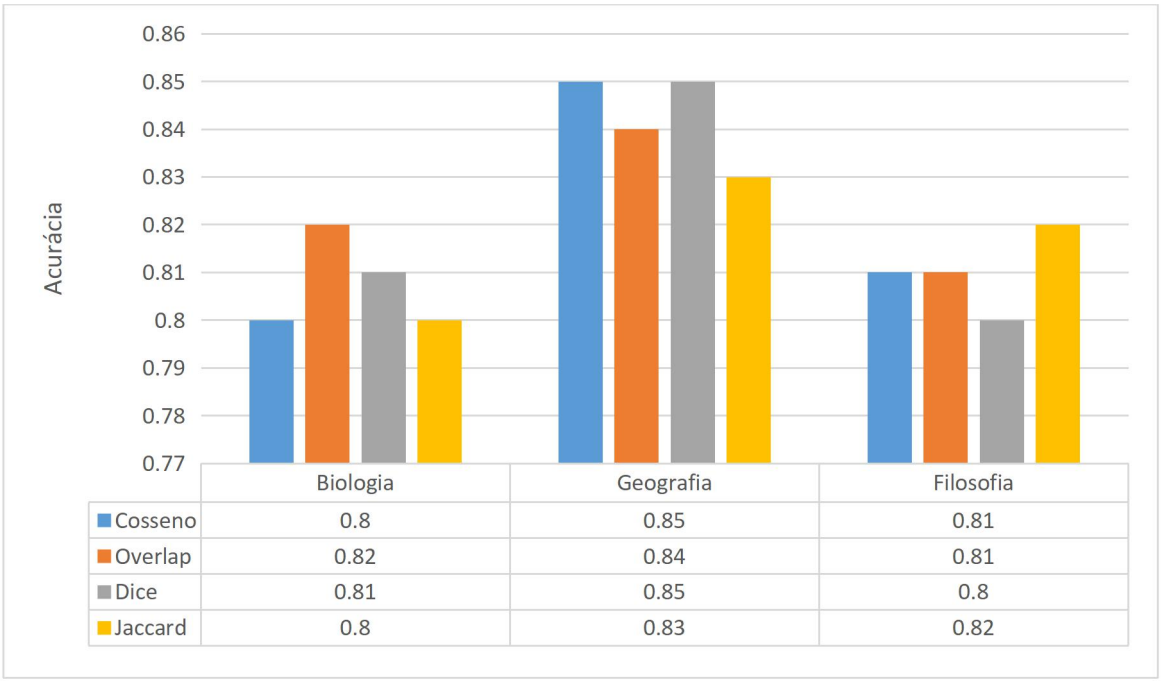
Para responder essa questão foram feitos experimentos juntando as quatro respostas melhores avaliadas que resultaram numa acurácia de 0.82 contra 0.84 da resposta do especialista, portanto pelo resultado é melhor compor uma resposta de referência com mais texto juntando as melhores respostas do *corpus*, pois neste caso se tem um acréscimo de vocabulário, isso contribui para uma maior concordância no processo de medir a similaridade.

5.2.3 Questão de pesquisa 4: frequência de termos ou teoria dos conjuntos (QP4) - *É melhor trabalhar com interseção de conjuntos ou frequência dos termos? Qual é a melhor medida de similaridade de conjuntos?*

Na coleta das métricas de similaridade foram trabalhados com medidas de frequência de termos e com medidas da teoria dos conjuntos. Na frequência de termos foram trabalhados com três métricas: cosseno-FT e distância euclidiana; e também foram combinadas estas duas variáveis via regressão, numa única variável. Na teoria dos conjuntos fora trabalhado com cinco métricas: *overlap* mínimo, *overlap* máximo, *jaccard*, *dice* e cosseno-Conj.

A Figura 33 (trinta e três) mostra as acurácias para diferentes medidas de similaridade baseadas em teoria dos conjuntos. Para Biologia o melhor resultado foi *Overlap*, para

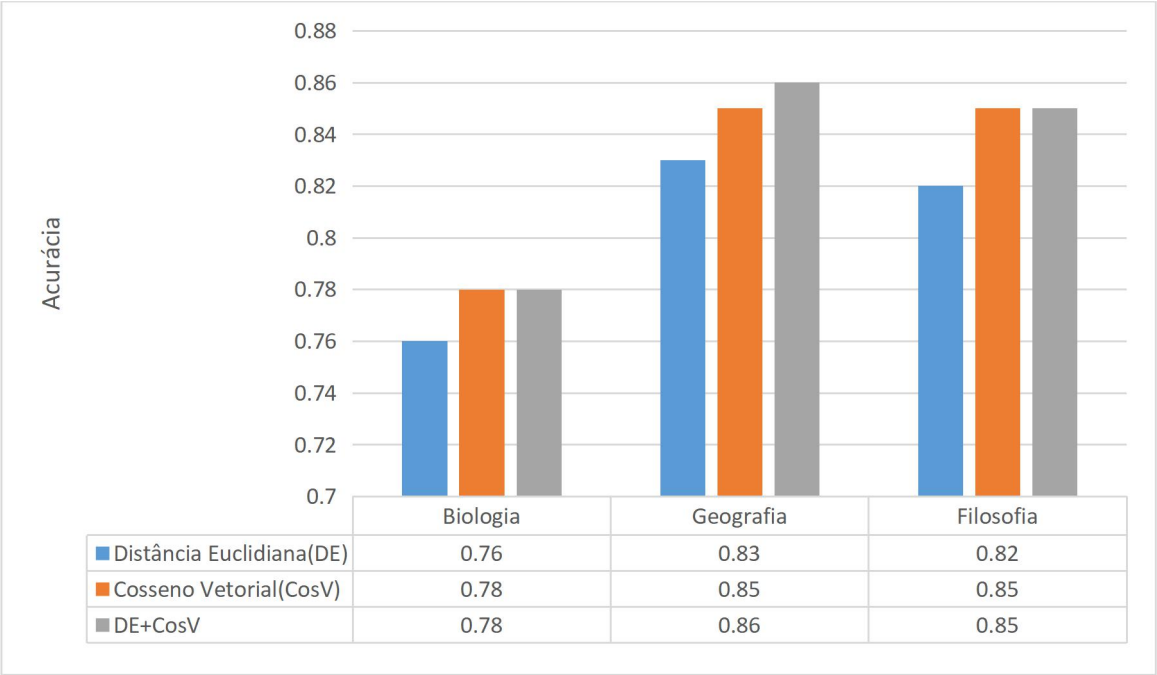
Geografia venceram Cosseno e *Dice*, e para Filosofia venceu *Jacard*. Portanto, pode-se observar que entre a pior e a melhor medida a diferença é de apenas 0.02 pontos para os três conjuntos de dados.



	Biologia	Geografia	Filosofia
Cosseno	0.8	0.85	0.81
Overlap	0.82	0.84	0.81
Dice	0.81	0.85	0.8
Jaccard	0.8	0.83	0.82

Figura 33 – Comparativo das medidas de teoria de conjuntos para as provas de Biologia, Geografia e Filosofia (**Métrica:** erro médio).

A Figura 34 (trinta e quatro) mostra as acurácias para diferentes medidas de similaridade baseadas em frequência de termos. A combinação entre cosseno-FT e distância euclidiana foi o melhor resultado nas três provas. Nas provas de Geografia e Filosofia esta medida superou à de teoria dos conjuntos: para Geografia 0.83 *versus* 0.86 e para Filosofia 0.83 *versus* 0.85. No entanto, para a prova de Biologia que tinha pouco texto, a medida de Teoria dos Conjuntos venceu 0.82 *versus* 0.78.



	Biologia	Geografia	Filosofia
Distância Euclidiana(DE)	0.76	0.83	0.82
Cosseno Vetorial(CosV)	0.78	0.85	0.85
DE+CosV	0.78	0.86	0.85

Figura 34 – Comparativo das medidas de teoria de conjuntos para as provas de Biologia, Geografia e Filosofia (**Métrica:** erro médio).

Sobre as medidas de frequência de termos, o resultado justifica a recomendação de (OLMOS *et al*., 2011) onde argumenta que o cosseno mede uma dimensão de proximidade angular enquanto que a distância euclidiana mede a quantidade de conteúdo, e que deveria utilizar as duas medidas combinadas.

5.2.4 Questão de Pesquisa 5: unigramas x bigramas (QP5) - *O uso de bigramas minimiza o problema? Combinando bigramas com unigramas se tem uma boa acurácia?*

Neste procedimento, além de coletar as medidas de unigramas e bigramas individualmente, coleta-se também à medida que combina Unigramas com Bigramas. Ainda, se coleta por meio de regressão a combinação de Cosseno-FT com Distância Euclidiana. No final desta etapa, com o uso dessas métricas, as listas de unigramas e bigramas são transformados em valores numéricos. A próxima etapa, que é a classificação recebe como entrada esses valores em forma de vetores numéricos.

Para sair do modelo de saco de palavras (dos unigramas) e construir abordagens mais robustas se pode combinar unigramas com bigramas. A Figura 35 (trinta e cinco) apresenta

acurácias para só unigrama, só bigramas e unigramas combinado com bigrama. Nos três conjuntos de dados a combinação unigrama mais bigrama alcançou maiores acurácias.

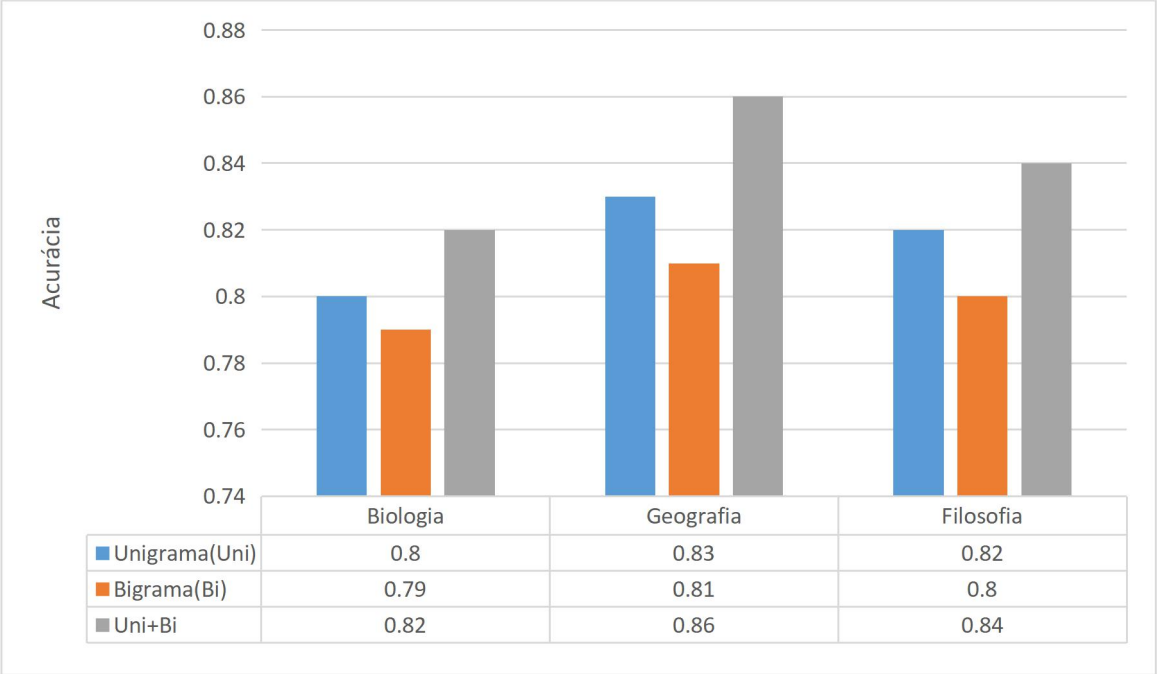


Figura 35 – Resultados das acurácias para somente unigrama, somente bigramas e unigrama combinado com bigrama (**Métrica:** erro médio).

Esta combinação permite que a solução seja mais robusta, pois considera a ordem de escrita do texto.

5.2.5 Questão de pesquisa 6: Acurácia (QP6) - *O método de avaliação centrado em atributos de conteúdo alcança a acurácia dos avaliadores humanos?*

Neste experimento de similaridade a meta é maximizar a acurácia *SxH* buscando uma aproximação com a acurácia *HxH*. No comparativo (Tabela 21) da acurácia do sistema *SxH* contra a acurácia dos humanos *HxH* para a prova de Biologia a acurácia *SxH* ficou um pouco longe da *HxH*, 0.82 *versus* 0.94; porém para a prova de Geografia a acurácia *SxH* superou a acurácia *HxH*: 0.86 *versus* 0.85. Medindo a acurácia *SxH*, o *H* é a média dos dois humanos e medindo a acurácia *HxH*, cada *H* é a nota do especialista humano.

Tabela 21 – Resultado da acurácia do experimento das questões do tipo curtas (**Métrica:** erro médio).

Corpus	*SxH*	*HxH*
Biologia	0.82	**0.94**
Geografia	**0.86**	0.85
Filosofia	0.81	-

Fonte: Próprio autor (2020).

5.3 Experimento 2: Atributos linguísticos (Léxica, Sintática e Semântica).

No experimento 2 (dois), diferente do experimento 1 (um), foi utilizado uma abordagem centrada na coleta de atributos em três dimensões: léxica, sintática e semântica. Neste comparativo, passa-se pelas diversas etapas da arquitetura *pipeline* como: (1) preparação de *corpus*, (2) pré-processamento, (3) coleta de atributos, (4) modelo de predição e (5) avaliação. A proposta deste experimento é maximizar o valor $S \times H$ buscando uma aproximação com $H \times H$.

A etapa de pré-processamento no experimento 2 (dois) foi semelhante ao do experimento 1 (um). Relacionada a esta etapa se tem a questão de pesquisa**: Pré-processamento com atributos (QP7)** O pré-processamento influencia na acurácia final para abordagem em dimensões linguísticas sobre respostas do tipo curta?

Na etapa de coleta de atributos foi trabalhada a dimensão léxica, que descreve o aspecto individual das palavras. Nesta dimensão se tem 4 (quatro) principais categorias: (1) estatística de superfície, coleta estatística baseado em contagem de palavras. (2) diversidade, coleta medidas que representam o quanto é diverso o vocabulário utilizado. (3) legibilidade, mede o grau de facilidade da leitura do texto (Tabela 22).

Tabela 22– Atributos extraídos da dimensão lexical para respostas curtas.

Lexical	**Estatística de Superfície**	N° de caracteres, n° de diferentes palavras, n° de palavras, n° de palavras curtas, n° de palavras longas, n° média de palavras, n° de *stopword*, n° de sentença, n° comprimento de palavra mais frequente.
	Diversidade	*Type-token-ratio – TTR, Guiraud's index, Yule's K, the D estimate, hapax legomena.*
	Legibilidade	*Gunning Fox Index, Flesch Kincaid grade level, Dale-Chall readability formula, autometed readability index, LIX, word variation index, nominal ratio, SMOG-index*

Fonte: Próprio autor (2020).

Na dimensão sintática foram coletados atributos que retratam o aspecto individual de cada sentença, compreende duas categorias: (1) número de cada POS *tag* (*part-of-speech tagging*), como por exemplo, número de nomes (noun) e verbos (verb) (2) Erro Léxico e Sintático, conta o número de erros de sentenças mal formuladas, por exemplo, erros de concordância e pontuação (3) Erro ortográfico, conta o número de erros ortográficos (Tabela 23).

Tabela 23– Atributos extraídos da dimensão Sintática para respostas curtas conforme descrição na seção sintática.

Sintático	**Número de cada etiqueta morfossintática**	Número de diferentes etiquetas morfossintáticas, Número de etiquetas morfossintáticas por categoria sintática: SR=ser, HV=haver, ET=estar, TR=ter, VB=verbo (I, -P, -SP, -D, -RA, -SD, -R, -SR, -G, -PP, -NA), *Agreement Particle* (*genre* (none = masc, -F = fem, -G = *doublegender*), *number* (none = sing, -P = plural)), N, NPR, PRO, P + PRO, PRO$, CL, D, DEM, ADJ, ADV, Q, CONJ, C, WPRO, WQUE, WD, P
	Erro de sintaxe	Número erros de pontuação e concordância
	Erro ortográficos	Número de erros de ortográficos

Fonte: Próprio autor (2020).

Na dimensão de conteúdo coletam-se atributos que descrevem os aspectos que estão relacionados ao conteúdo do texto, por exemplo, medidas de similaridade entre a resposta do aluno e a resposta de referência. Coletam-se também atributos que descrevem os aspectos que se relacionam a coerência textual, tanto local dentro de uma resposta como global em relação às várias respostas (Tabela 24).

Tabela 24– Atributos extraídos da dimensão Semântica para respostas curtas

Semântico	**Similaridade Cosseno e distância euclidiana com resposta de referência**	Similaridade com Texto Fonte (Pré: SSW, CST, CSW) (Med: Cosseno e Distância Euclidiana)
	Similaridade e distância contra as faixas dos escore**s**	Similaridade (nível: 0, 1, 2, 3, 4, 5, 6) (Pré: SSW, CST, CSW) (Med:Cosseno e Distância Euclidiana)
	Soma ponderada de todos os valores de correlação baseada nos valores de cosseno e distância euclidiana	Correlação Ponderada (Pré: SSW, CST, CSW) (Med: Cosseno e Distância Euclidiana)

Fonte: Próprio autor (2020).

Assim foi relacionado a questão de pesquisa: **atributos preditores (QP8)** quais os melhores atributos preditores para a língua portuguesa brasileira em questões de respostas discursivas curtas?

Vajjalla (2018) faz um estudo sobre quais atributos linguísticos são úteis e consistentes para um modelo de predição. Para este fim utiliza-se um conjunto de dados em língua inglesa que foram escritos por não-nativos em um cenário de testes. As propriedades linguísticas foram codificadas em tipos lexicais e erros de linguagem. Assim foi relacionado a questão de pesquisa: **Importância dos atributos (QP9)** A importância de contribuição dos atributos se repete em diferentes conjuntos de dados nas respostas curtas? Por fim, se tem a questão de pesquisa: **Acurácia (QP10)** O método de avaliação centrado em atributos entre três dimensões alcança a acurácia dos avaliadores humanos?

5.4 Resultados e discussão do Experimento 2: atributos em três dimensões (Léxica, Sintática e Semântica)

Esta seção foi organizada a partir das questões de pesquisa. Para cada questão são mostrados os resultados e, em seguida é feita uma breve discussão.

5.4.1 Questão de pesquisa 7: Pré-processamento com atributos (QP7) - *O pré-processamento influencia na acurácia final para abordagem em dimensões linguísticas sobre respostas do tipo curta?*

Foram repetidas as mesmas três técnicas de pré-processamento do experimento 1 (um). Neste caso, além do pré-processamento descrito no experimento 1 (um), os *tokens* foram etiquetados morfologicamente para classificação conforme suas categorias gramaticais, para isto foi utilizado o *software* Aelius (ALENCAR, 2010; ALENCAR, 2013, p. 7; DRURY; ROSSI; DE ANDRADE LOPES, 2014; ALENCAR, 2015, p. 233).

A Tabela 25 (vinte e cinco) apresenta os resultados do experimento realizado para as questões de Biologia e Geografia considerando as variações nas técnicas de pré-processamento e o resultado de acurácia. Neste experimento foi utilizada a métrica KQ (ver capítulo 4).

Tabela 25 – Pré-processamento das respostas curtas de Biologia e de Geografia.

humano *versus* humano	**0.89**		**0.58**	
***Corpus* e Dimensões**	**Biologia**		**Geografia**	
	Sem	**Lex+Sint+Sem**	**Sem**	**Lex+Sint+Sem**
-RCE, -RSW, -RSU	0.65	0.65	0.70	0.70
+RCE, -RSW, -RSU	0.70	0.70	**0.74**	0.74
+RCE, +RSW, -RSU	0.64	0.64	0.66	**0.76**
+RCE, +RSW, +RSU	**0.71**	**0.72**	0.73	0.73

Fonte: Próprio autor (2020).

As diferentes técnicas de pré-processamento apresentam diferentes valores de acurácia. No entanto, as diferenças são bem significativas dentro de cada conjunto de *corpus*, sendo a diferença do menor para o maior valor 0.08 em Biologia e 0.10 para Geografia, o que responde à questão de pesquisa. Considerando estes valores é importante ter a etapa de pré-processamento nas abordagens de avaliação automática de respostas curtas.

5.4.2 Questão de pesquisa 8: atributos preditores QP8 - *Quais os melhores atributos preditores para a língua portuguesa brasileira em questões de respostas discursivas curtas?*

Na Tabela 26 (vinte e seis) apresentam-se os principais atributos por ordem de importância gerada pelo modelo de predição. Partimos com um conjunto de mais de 140 (cento e quarenta) atributos, onde foi utilizado o método *random forest* para predição da importância dos atributos. Na coleta de atributos de conteúdo uma resposta de aluno é contrastada com a resposta de referência com uso de *n*-gramas (unigrama e bigrama) e com as medidas de Distância Euclidiana e Cosseno. Foi utilizado, também, métodos de ponderação local e global dos textos como TF-IDF. Normalmente, a resposta de referência é formada a partir de um conjunto das respostas mais bem avaliadas. Por outro lado, alguns autores sugerem que se podem utilizar também respostas de referência baseadas em agrupamentos feitos em relação ao escore (ZUPANC; BOSNIC, 2017). Assim, foi criado 7 (sete) vetores resposta de referência, um para cada escore (0...6). Aqui foram aplicadas as medidas – Distância Euclidiana e Cosseno – contra estes vetores de respostas, incluindo também as variações no tipo de pré-processamento; disso resultaram 42 (quarenta e dois) atributos de conteúdo, onde muitos deles estão posicionados com maior relevância neste experimento.

Tabela 26- Importância dos atributos nas respostas curtas de Biologia e Geografia.

	Atributos	Importância
1	Similaridade Cosseno Escore 4	0.23
2	Similaridade Cosseno Escore 6 sem *Stopword*	0.13
3	Similaridade Cosseno Escore 5	0.12
4	Número de caracteres	0.09
5	Similaridade Cosseno Escore 5 sem *Stopword*	0.09
6	Similaridade Cosseno com texto fonte e com *Stopword*	0.07
7	Número de *Stopword*	0.06
8	Número de palavras longas	0.06
9	Similaridade Distância Euclidiana Escore 0	0.06
10	Similaridade Cosseno Escore 6	0.05
11	Similaridade Cosseno Escore 3	0.05
12	Similaridade de Cosseno Escore 3 com *Stopword*	0.05
13	Número de palavras	0.05
14	Similaridade de Cosseno Escore 4 com *Stopword*	0.04
15	Número de pronomes	0.04
16	Número diferente de palavras	0.03

Fonte: Próprio autor (2020).

Foram selecionados os melhores atributos das bases curtas de português (Biologia e Geografia) e foram verificados que atributos de dimensão de conteúdo, como as medidas de cosseno por faixa de escore são os principais, assim alcançando o topo da lista de importância dentro do modelo de predição. Por outro lado, atributos da dimensão léxica (de estatísticas de superfície) ocupam seis posições em uma lista de importância de dezesseis, tais como número de palavras e número de palavras diferentes e número de palavras longas.

5.4.3 Questão de pesquisa 9: Importância dos atributos (QP9) *A importância de contribuição dos atributos se repete em diferentes conjuntos de dados nas respostas curtas?*

Na Tabela 27 (vinte e sete) apresentam-se os principais atributos por ordem de importância.

Tabela 27 - Resultado da importância dos atributos em cada base de dados.

	Biologia	
	Atributos	**Importância**
1	cosseno escore 6 sem *stopword*	0.13
2	cosseno escore 5	0.11
3	cosseno escore 5 sem *stopword*	0.09
4	número de caracteres	0.09
5	cosseno com texto fonte	0.07
6	cosseno escore 6	0.05
7	número de palavras longas	0.04
8	cosseno texto fonte sem *stopword*	0.03
9	número de pronomes	0.03
	Geografia	
1	cosseno escore 4	0.23
2	euclidiana escore 0	0.06
3	cosseno escore 3	0.05
4	número de *stopword*	0.05
5	cosseno escore 3 com *stopword*	0.04
6	cosseno escore 4 com *stopword*	0.04
7	número de palavras	0.03
8	número de palavras diferentes	0.03
9	cosseno escore 2	0.02

Fonte: Próprio autor (2020).

Foram destacados os melhores atributos de cada *corpus* e, foram verificados que neste caso as medidas de cosseno e distância euclidiana por faixa de escore são os principais atributos das duas bases. Por outro lado, os outros atributos mais relevantes são os léxicos de estatísticas de superfície, tais como número de palavras e número de palavras diferentes e número de palavras longas.

5.4.4 Questão de pesquisa 10: Acurácia (QP10) *O método de avaliação centrado em atributos entre três dimensões alcança a acurácia dos avaliadores humanos?*

Neste experimento a meta é maximizar a acurácia KQ *SxH* buscando uma aproximação com a acurácia KQ *HxH*. A Tabela 28 (vinte e oito) compara *SxH* contra *HxH*, no comparativo para a prova de Biologia o sistema ficou um pouco longe: 0.72 *SxH*, contra 0.89 *HxH*; porém para a prova de Geografia obteve-se o resultado de 0.76 *SxH* contra 0.58 *HxH*. Este segundo resultado, o sistema superou a acurácia entre dois humanos.

Tabela 28 - Resultados das acurácias (*HxH versus SxH*) das respostas curtas de Biologia e de Geografia (**Métrica:** KQ).

Corpus	***SxH***	***HxH***
Biologia	0.72	**0.89**
Geografia	**0.76**	0.58

Fonte: Próprio autor (2020).

Uma possível justificativa para a acurácia bem elevada para a resposta de Biologia é a existência de uma resposta **conceitual** modelo de correção (capítulo 3, Tabela 8), onde praticamente todas as opções de respostas são listadas, tornando a questão quase objetiva. A partir desta resposta modelo os dois avaliadores se alinharam na correção destas respostas, elevando a acurácia entre eles. Por outro lado, a resposta de geografia é puramente **argumentativa**, dificultando um alinhamento entre os dois avaliadores humanos, assim a acurácia entre eles é baixa, talvez por isso o sistema tenha superado a acurácia dos humanos.

6 Avaliação de resposta tipo ensaio (redações)

Neste capítulo realizam-se experimentos num *corpus* com 1.000 (mil) redações todas sobre o mesmo tema: "*A atual crise político-social do Brasil e ações políticas para seu enfrentamento*" (CEPS - UFPA). Diferente dos experimentos com questões de respostas curtas onde algumas dimensões linguísticas não foram utilizadas aqui foram utilizadas as quatro dimensões linguísticas: léxica, sintática, semântica e coerência. Foram trabalhados com mais de 140 (cento e quarenta) atributos, a grande maioria adaptados da língua inglesa. Em parte, o que se pretende é verificar se os atributos para língua inglesa são adequados para o português.

Trabalha-se com 7 (sete) questões de pesquisas. Por exemplo, deve-se responder à questão de pesquisa principal - **Acurácia (QP16):** para redações em língua portuguesa qual acurácia final se pode alcançar, na abordagem baseada nas quatro dimensões? Antes de responder esta questão se têm várias outras questões secundárias:

- **(QP11-12-13-14, Léxica, Sintática, Semântica e Coerência)** qual a contribuição de cada uma das quatro dimensões? Quais os melhores atributos para cada uma das dimensões?

- Explorar também o cruzamento entre as dimensões: **(QP15)** qual a acurácia da combinação das dimensões duas a duas: léxica x sintática, léxica x semântica, léxica x coerência, sintática x semântica, sintática x coerência e semântica x coerência? Além disso, como semântica tem maior influência explora-se semântica + léxica + coerência, semântica + sintática + coerência e semântica + léxica + sintática.

- Nas 1.000 (mil) redações o escore humano é gerado a partir de três variáveis associadas às três competências que foram avaliadas nas redações (tema (conteúdo), coerência e regras). **(QP16)** medir por meio da correlação de *Pearson* e também KQ a influência de cada uma das dimensões em relação às três variáveis das competências.

- **(QP17)** por fim, se tem a questão de pesquisa sobre acurácia. Com base nas quatro dimensões, a acurácia final do método alcança a acurácia dos avaliadores humanos?

6.1 Experimento com as redações

A proposta de avaliação segue a arquitetura *pipeline* de cinco passos: (1) Preparação de *corpus*, (2) pré-processamento, (3) coleta de atributos, (4) modelo de predição e (5) avaliação. A **preparação de *corpus*** faz a leitura de um arquivo no formato *comma-separated-values* (csv), cujas colunas representam os escores atribuídos por dois avaliadores humanos e os textos dos respectivos ensaios (redações). Para o **pré-processamento** foi implementada uma classe de tarefas tais como retirada de acentuação, conversão de maiúscula em minúscula, remoção de pontuação e caracteres especiais, entre outros. A opção sem qualquer pré-processamento linguístico foi necessária na extração de alguns atributos, das dimensões léxica e sintática. Em seguida, os ensaios foram tokenizados em sentenças e palavras usando os módulos `word_tokenize()` e `sent_tokenize()` da biblioteca NLTK (BIRD; KLEIN; LOPER, 2009). Estas técnicas foram combinadas com processamento de elementos morfológicos, a saber, remoção de *stopwords* e remoção de sufixos (*stemming*). Foram consideradas as seguintes variações nos textos das redações:

1. Sem qualquer tipo de pré-processamento;
2. Removendo caracteres especiais;
3. Removendo *stopwords*;
4. Removendo *stopwords,* mais um processo de *stemming*.

Na dimensão Semântica, é consenso que o pré-processamento com remoção de *stopwords* junto com um algoritmo de *stemming* fornece bons resultados (PÉREZ *et al*., 2005). Entretanto, como não se tem resultados em língua portuguesa que comprovem tal afirmação, foi optado por testar e considerar todas as opções disponíveis para o pré-processamento.

Na etapa de **coleta de atributos** foram trabalhados nas quatro dimensões linguísticas. Foram usados: a biblioteca NLTK (BIRD; KLEIN; LOPER, 2009), Aelius (DRURY; ROSSI; DE ANDRADE LOPES, 2014; ALENCAR, 2013, p. 7; ALENCAR, 2015, p. 233). Cogroo (KINOSHITA *et al*., 2007), PyEnchant (KELLY, 2014) e inúmeros programas em *Python* especializados na coleta de grupos de atributos, tais como os de coerência.

Este estudo tem uma direção principal que é criar um modelo baseado em atributos linguísticos e, a partir do modelo explorar como alcançar uma acurácia máxima em relação à

acurácia medida entre os dois avaliadores humanos. Como **modelo de predição** foi utilizado o algoritmo de classificação *Random Forest*, que é um algoritmo de aprendizagem de máquina de fácil implementação e utilização, e que produz bons resultados sem a necessidade de ajuste de parâmetros (FERNÁNDEZ-DELGADO *et al*., 2014). Este algoritmo também mede a importância de cada atributo utilizado. O algoritmo divide os dados em duas partições: uma de treinamento e outra de teste. Em seguida, o algoritmo constrói várias árvores de decisão a partir da partição de treinamento para gerar o classificador, que rodando sobre o conjunto de teste gera o vetor de escore*s* do sistema. Na etapa de **avaliação**, esse vetor é contrastado com o vetor de escore dos humanos. Neste experimento foi utilizado o coeficiente *Weighted Cohen's Kappa* para avaliar a acurácia.

6.2 As quatro dimensões

6.2.1 Dimensão Léxica

Nesta dimensão foram coletados atributos relacionados com aspectos lexicais, os quais indicam o uso eficiente de recursos linguísticos, a facilidade de leitura e a diversidade lexical de um texto (JOHANSSON, 2009). Estes aspectos foram divididos em três subclasses: nível de palavras, medidas de legibilidade e diversidade lexical (VAJJALLA, 2018; ZUPANC; BOSNIC 2017):

- No nível de **palavras** foram coletados os seguintes atributos: número de sílabas (nsyllable); número de palavras (snwords); número de palavras curtas (nshortswords); número de palavras longas (nlongwords); comprimento mais frequente de palavras (mostfreqwordslen); comprimento médio de palavras (averagewordslen); número de sentenças (nsentences); número de palavras diferentes (ndifwords); número de *stopwords* (nstopwords) e número de caracteres (ncharacteres). Ao todo, são 10 (dez) atributos.

- Os índices de **legibilidade** estimam a dificuldade de leitura de um texto (READABLE, 2019). O trabalho de DuBay (2004) apresenta mais de 200 (duzentas) fórmulas para medidas de legibilidade. Entretanto, como sugerido por (ZUPANC; BOSNIC 2017), para esta pesquisa foram coletados como atributos as seguintes medidas de legibilidade: *Coleman–Liau index* (CLindex); *Gunning Fog*

index (GFindex); *Flesch reading ease* (FREindex); *Flesch Kincaid grade level* (FKGindex); *Dale-Chall readability formula* (DCindex); *automated readability index* (ARindex); *simple measure of Gobbledygook* (SMOGindex) e *LIX* (LIXindex). Ao todo, são 8 (oito) atributos.

- A **diversidade lexical** é uma medida de quantas palavras diferentes são utilizadas em um texto. Para esta pesquisa foram selecionados como atributos a medida *Type-token Ratio* (TTR) e variações *Corrected Type-token Ratio* (CTTR); *Measure of Textual Lexical Diversity* (MTLD); *Moving Average Type-token Ratio* (MATTR) e *Mean Segmental Type-token Ratio* (MSTTR); uma variante *Guiraud's index* (Gindex); além dos índices *Yule's K* (YK); *hapax legomena* (hápax); *Hypergeometric Distribution Diversity* (HDD); *Lexical Diversity* (LD). Ao todo, são 10 (dez) atributos.

Ao final, foram coletados 28 (vinte e oito) atributos com a utilização das bibliotecas *numpy* e NLTK do *Python*, do pacote *textstat* para cálculo de estatísticas do texto (BALIKAS, 2018), do pacote *Readability calculator*, que fez o papel de uma calculadora de legibilidade e do pacote *LexicalRichness*, que fornece medidas de riqueza lexical.

A Tabela 29 (vinte e nove) mostra os atributos que tiveram maior influência dentro da dimensão lexical, por ordem de importância.

Tabela 29 – Lista por ordem de importância dos atributos da dimensão lexical.

	Atributo	Importância
1	Nsyllable	0.17
2	Snwords	0.09
3	Nstopwords	0.07
4	Nlongwords	0.06
5	YK	0.05
6	LIXindex	0.05
7	Ncharacteres	0.04
8	Nshortswords	0.04
9	CTTR	0.04
10	Hapax	0.04
11	FREindex	0.04
12	Nsentences	0.03
13	Ndifwords	0.03
14	MATTR	0.03
15	MTLD	0.03
16	CLindex	0.03
17	SMOGindex	0.02
18	LD	0.02
19	HDD	0.02
20	GFindex	0.02
21	FKGindex	0.02
22	ARindex	0.02
23	Mostfreqwordslen	0.01
24	Averagewordslen	0.01
25	Gindex	0.01

Fonte: Próprio autor (2020).

Os três atributos mais importantes foram: Número de Sílaba (nsyllable), Número de palavras (snwords) e Número de *stopwords* (nstopwords) que são da subclasse de nível de palavras. Os atributos *Type-token Ratio* (TTR) e sua variação *Mean Segmental Type-token Ratio* (MSTTR) não apresentaram qualquer importância nos resultados obtidos na distribuição do desempenho desses atributos, como pode ser mais bem observada na Figura 36 (trinta e seis) abaixo.

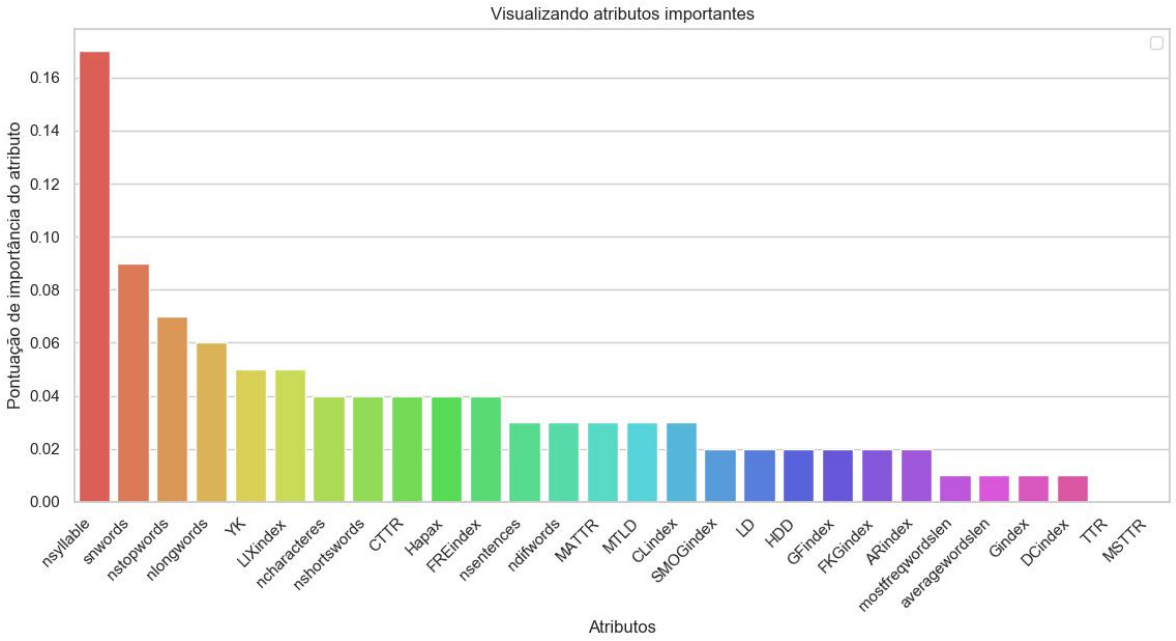


Figura 36 – Contribuição dos atributos da dimensão lexical organizados por ordem de importância.

6.2.2 Dimensão sintática

Em termos de Gramática, a sintaxe é caracterizada por postulados gramaticais que fazem o estudo de palavras de acordo com a função que desempenham em determinado contexto linguístico (ORLANDI, 2017). Para esta pesquisa foram coletados atributos sintáticos de três classes: erros de concordância gramatical, erros de verificação ortográfica e etiquetas morfossintáticas (anotar uma palavra em um texto conforme sua definição e seu contexto).

Erros de concordância gramatical e de ortografia. Para coleta desse atributo foi utilizada uma interface de modo texto de acesso ao corretor gramatical do CoGroo na linguagem *Python*; com esta interface foi possível identificar erros como: colocação

pronominal, concordância nominal, concordância sujeito-verbo, uso da crase, concordância nominal e verbal e outros erros comuns da escrita em Português do Brasil (SILVA, 2013). A Tabela 30 (trinta) mostra um exemplo deste tipo de erro que corresponde a um dos atributos.

Tabela 30 – Exemplo de erro gramatical com base em um trecho da redação com a saída do corretor gramatical.

Entrada	Saída
"Uma exclusão de governantes seria a solução, com auditorias e punições mais severas. Seguindo a ideia de eficiência, mantendo assim, os que produzissem mais (ou algo). E filtrando candidatos a eleições"	[xml:114] Os determinantes concordam com o substantivo a que se referem. Categoria: Concordância Determinante-Substantivo Grupo: determinante singular + substantivo plural

Fonte: Próprio autor (2020).

Erros de verificação ortográfica. Para coleta deste atributo foi utilizado o módulo PyEnchant na linguagem *python*, o qual fornece um conjunto de ligações da linguagem com a biblioteca de verificação ortográfica Enchant. Esta biblioteca é bastante flexível e lida como vários dicionários e idiomas, entre eles o português do Brasil. A Tabela 31 (trinta e um) mostra um exemplo de erros de verificação ortográfica de uma redação que corresponde a um atributo.

Tabela 31 - Exemplo de erros de ortografia de uma redação com a saída da verificação ortográfica.

Entrada	***Saída***
'o que dimiminuiu a moral do eleitor e gerou um receso economico e grave exercício'	*ERROR: dimiminuiu* *ERROR: receso* *ERROR: econômico*

Fonte: Próprio autor (2020).

Atributos de etiquetagem morfossintática (POS *tagging)*. Nesta classe, os atributos estão relacionados com aspectos de análise superficial dos textos de cada ensaio por meio da marcação POS *tagging* das palavras e posterior contagem dessas marcações. Para a coleta

destes atributos foi utilizado o *software* Aelius[4], que é um projeto de *software* livre para análise sintática superficial do português do Brasil. Cada redação foi tokenizada em palavras e submetida ao analisador que retornou uma lista de palavras anotadas com sua classe gramatical. Com isso, desses atributos foi contado o número de diferentes classes de POS-*tag* e o número total de cada POS-*tag*. Devido à grande quantidade de marcações morfológicas (POS e flexões), foi optado por agrupar as *tags* considerando apenas a parte infinitiva (em contextos verbais e nominais): por exemplo, as *tag* SR-G (verbo ser-gerúndio) e são SR-P (verbo ser-presente) foram agrupados pela *tag* SR. Deste modo, as *tags* consideradas foram: SR (verbo ser); HV (verbo haver); ET (verbo ter); TR (verbo estar); VB (verbos em geral); N (nome); PRO (pronomes); CL (clíticos); D (determinantes); ADJ (adjetivos); ADV (advérbios); Q (quantificadores); CONJ (conjunções coordenativas); C (conjunções subordinativas); WPRO (pronomes relativos); P (preposições); OUTRO (outro, outros, outra, outras); FP (partículas de foco); NUM (número cardinal); NEG (negação); e INTJ (interjeição). Para extração de atributos, por exemplo, as *tags* SR-G (verbo ser-gerúndio) e são SR-P (verbo ser-presente) foram agrupados pela *tag* SR. Por fim, são 21 (vinte e um) atributos de etiquetas, mais dois relacionados a erros (pontuação e ortografia), totalizando 23 (vinte e três).

A Tabela 32 (trinta e dois) mostra os atributos que tiveram maior influência na dimensão sintática:

[4] O sistema Aelius é livremente disponível no endereço http://sourceforge.net/projects/aelius/files/.

Tabela 32 - Atributos da dimensão sintática classificados por importância.

	Atributo	Importância
1	ADJ	0.18
2	N	0.15
3	P	0.08
4	misspellings	0.08
5	Ndifpostag	0.04
6	SR	0.04
7	VB	0.04
8	PRO	0.04
9	D	0.04
10	ADV	0.04
11	CONJ	0.04
12	CL	0.03
13	WPRO	0.03
14	NEG	0.03
15	grammar_erros	0.03
16	ET	0.02
17	TR	0.02
18	Q	0.02
19	HV	0.01
20	OUTRO	0.01
21	FP	0.01
22	NUM	0.01
23	INTJ	0.01

Fonte: Próprio autor (2020).

Os três principais atributos foram Adjetivos (ADJ), Nome (N) e Preposição (P). O número de erros léxicos foi relevante ficando no quarto lugar. Já os erros de sintaxe (grammar_erros) ficaram na posição quinze. Na Figura 37 (trinta e sete) os dados da tabela estão representados num gráfico.

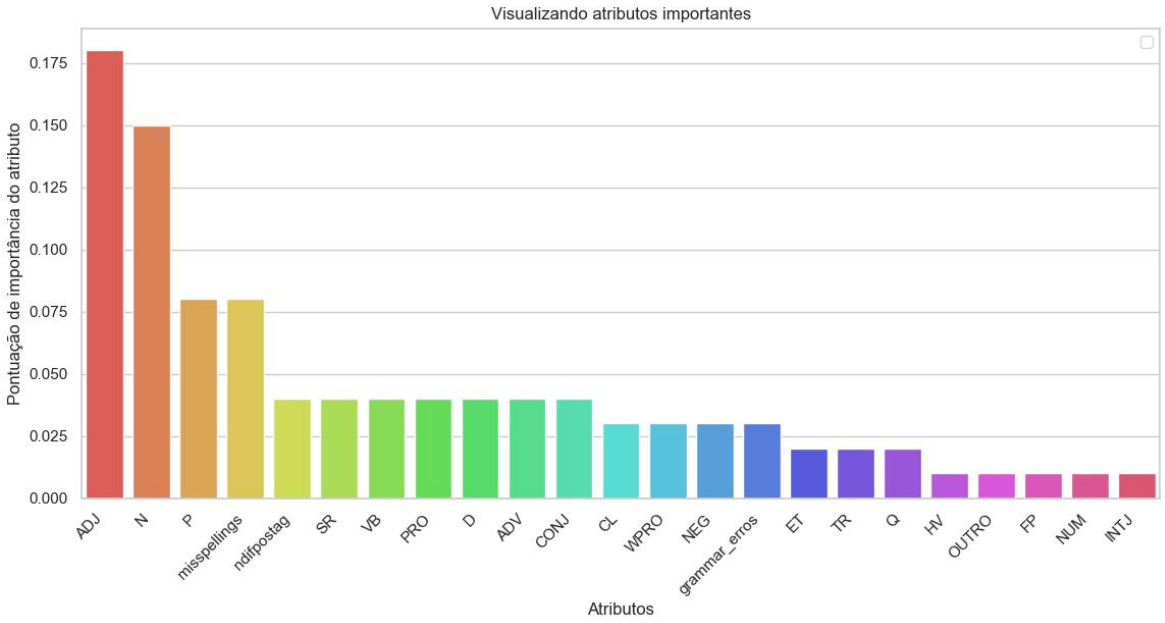


Figura 37 – Visualização dos atributos da dimensão sintática classificados por importância.

6.2.3 Dimensão Semântica

Nesta dimensão foram coletados atributos relacionados ao sentido das palavras e interpretação de sentenças ou textos (FRAWLEY, 2013). Uma das abordagens para avaliação automática de textos é baseada em similaridade semântica. O trabalho de Olmos *et al.*, (2011) afirma que a LSA pode comparar o conteúdo de um ensaio escrito por um aluno com um texto de referência por meio do cosseno do ângulo ou da distância euclidiana. Com LSA os atributos extraídos estão relacionados com a similaridade latente dos textos de cada ensaio. Na literatura é bem frequente o uso da seguinte **abordagem para um modelo LSA** (JORGE-BOTANA *et al.*, 2015; LEON *et al.*, 2013; REFAAT *et al.*, 2012; SANTOS; FAVERO, 2015):

i) Construção da matriz termo documento;
ii) Pesagem TF-IDF;
iii) Cálculo da SVD;

iv) Redução do espaço semântico; e

v) Classificação.

Esta abordagem de cinco passos foi adotada nesta pesquisa considerando duas métricas: cosseno do ângulo e a distância euclidiana entre dois vetores (ZUPANC; BOSNIC, 2017).

Para **teste do modelo de LSA**, foi utilizada a técnica computacional k-*fold cross validation*, que utiliza todo *corpus* disponível como amostras de treinamento e de teste (DUCHESNEEET, 2005). Como a base de dados consta de 1.000 (mil) redações, definimos k=5 e a base foi dividida em cinco partições com 200 (duzentas) redações. Após a divisão, foi utilizada uma partição para teste do modelo e as demais partições utilizadas como treinamento. O processo de validação foi repetido 5 (cinco) vezes, de modo que cada uma das partições foi utilizada exatamente uma vez como teste.

Nesta dimensão foram considerados diversos atributos. Os primeiros 10 (dez) atributos vieram de subpartições relacionadas com as faixas: escore [0..1], (1..2], (2..3]...(9..10]; intervalo aberto em "(" e fechado em "]". O valor real dos escores é de zero a dez, com passo de 0.25. Por exemplo, entre 6.0 e 7.0, se pode ter, 6.0, 6.25, 6.50, 6.75 e 7.0. Outros quatro atributos foram agrupados por faixas (f1..f4): [0..3], (3..6], (6..8], (8..10]. Até aqui são 14 (quatorze) atributos. Mas foram considerados mais 2 (dois): um texto de referência dado para o estudante se embasar para escrever a redação; e juntando todos os textos de todas as redações da partição de treinamento. São 16 (dezesseis) atributos que foram avaliados com: cosseno, distância euclidiana e cosseno mais distância euclidiana. O atributo cosseno mais distância euclidiana foi obtido por meio de um processo de regressão linear múltipla. Os 16 (dezesseis) atributos combinados com os três métodos de avaliação e com três variações de pré-processamento resultaram em 16*9, totalizando 144 (cento e quarenta e quatro). Destas combinações, foram consideradas apenas as melhores.

A combinação do cosseno mais distância euclidiana foi a que deu melhor resultado, a Tabela 33 (trinta e três) mostra a influência dos atributos no resultado da classificação considerando tais medidas.

Tabela 33 - Atributos de conteúdo classificados por ordem de importância avaliado com a medida de cosseno mais distância euclidiana (cd)

	Atributo	Importância
1	f4	0.24
2	cd1	0.1
3	f1	0.08
4	cd10	0.08
5	cd2	0.06
6	cd3	0.06
7	f3	0.05
8	cd8	0.05
9	cd9	0.05
10	f2	0.04
11	cd4	0.04
12	cd5	0.04
13	cd7	0.04
14	cd11	0.04
15	cd6	0.02

Fonte: Próprio autor (2020).

Os quatros principais atributos foram f4, cd1, f1 e cd10. Considerando que *corpus* de provas tinha poucos escores com valores baixos e ou com valores muito altos, os atributos mais relevantes foram os dos extremos das faixas de pontuação. A informação da Tabela 33 (trinta e três) também é apresentada na Figura 38 (trinta e oito).

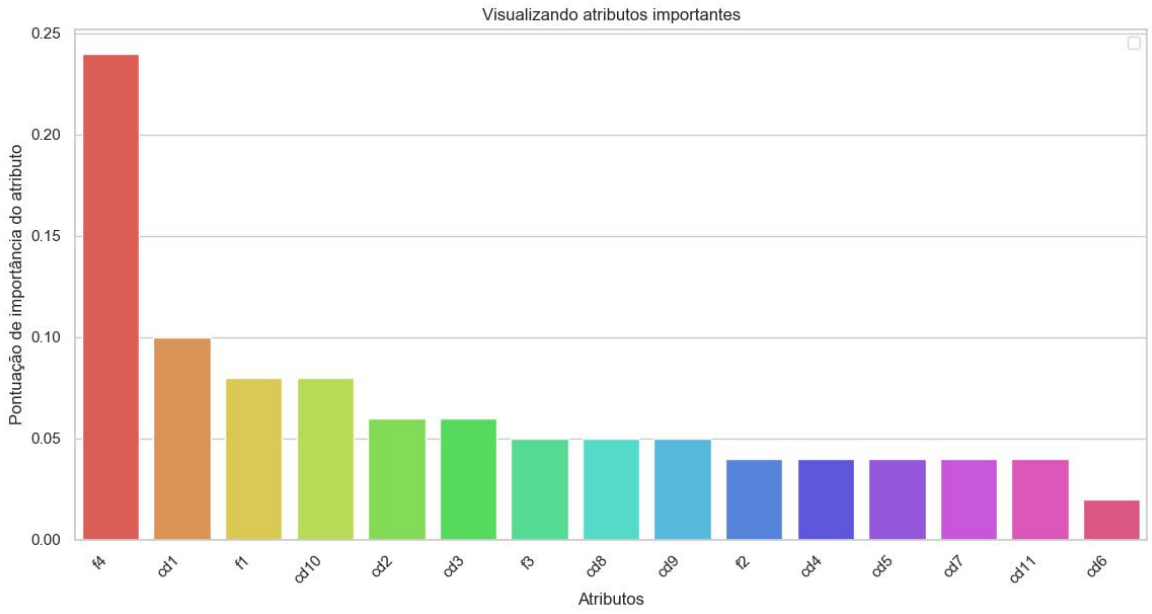


Figura 38 – Atributos de conteúdo classificados por ordem de importância.

6.2.4 Dimensão Coerência

Nesta dimensão foram extraídos atributos relacionados com a coerência entre partes do texto. Na literatura foram encontradas duas classes de modelos que capturam coerência textual. A primeira engloba aspectos sintáticos considerando as diferentes posições que uma determinada POS-*tag* ocorre em sentenças adjacentes (ZUPANC; BOSNIC, 2017; PALMA; ATKINSON, 2018). Neste trabalho, foi optado pela segunda que engloba aspectos semânticos e quantifica a coerência local por meio do grau de conectividade entre partes do texto (ZUPANC; BOSNIC, 2017; PALMA; ATKINSON, 2018). Os primeiros estudos nesta direção foram feitos por Halliday e Hasan (1976) e mostram a importância da coesão lexical na coerência do texto. A principal proposição por trás de cadeias lexicais é que textos coerentes apresentam uma grande quantidade de palavras semanticamente relacionadas. A relação de coerência intra-texto foi feita calculando a similaridade semântica entre partes do texto usando LSA.

Para geração dos atributos de coerência de conteúdo, cada redação foi dividida em partes (que foi chamado de janelas) sequenciais sobrepostas. A primeira janela é formada por 1/4 do comprimento da redação e as demais se movendo cada janela em um passo de 10 (dez) palavras. A última janela é descartada se tem menos de 10 (dez) palavras.

Para definir os atributos, foi considerado cada janela como um ponto no espaço semântico e então foi calculada uma similaridade entre esses pontos. Consideram-se três casos:

- Dentro da redação, todas as janelas contra todas: distâncias mínima, máxima e média (atributo 1);
- Com centro local: todas as janelas contra o centro local (atributo 2); distâncias mínima, máxima e média;
- Com centro global: todas as janelas contra o centro global (atributo 3); distâncias mínima, máxima e média;

Foram extraídos 9 (nove) atributos que foram combinados com Cosseno, Distância Euclidiana e Cosseno mais Distância Euclidiana. São 9*3=27 atributos, sem as combinações de pré-processamento. Os melhores resultados foram com cosseno e sem *stopwords*. A Tabela 34 (trinta e quatro) mostra os atributos que tiveram maior influência no resultado.

Tabela 34 - Atributos da dimensão coerência do conteúdo listados por ordem de importância.

	Atributos	**Importância**
1	Local Centro Cosseno Máximo (c3.1)	0.19
2	Contíguos Distância Máxima (c1.1)	0.08
3	Todos Cosseno Máximo (c2.1)	0.08
4	Todos Distância Mínima (c1.5)	0.08
5	Contíguos Distância Mínima (c2.5)	0.08
6	Contíguos Cosseno Mínimo (c2)	0.07
7	Todos Cosseno Médio (c2.3)	0.07
8	Todos Distância Máxima (c2.4)	0.07
9	Contíguos Cosseno Máximo (c1)	0.06
10	Local Centro Distância Mínima (c3.5)	0.06
11	Contíguos Cosseno Médio (c3)	0.05
12	Local Centro Cosseno Mínimo (c3.3)	0.05
13	Local Centro Distância Máximo (c3.4)	0.04

Fonte: Próprio autor (2020)

Os cinco principais atributos foram c3.1, c1.1, c2.1 e c1.5. Os atributos c1.3 e c1.4 não apresentaram qualquer importância nos resultados obtidos. A distribuição desses atributos pode ser vista no gráfico abaixo (Figura 39).

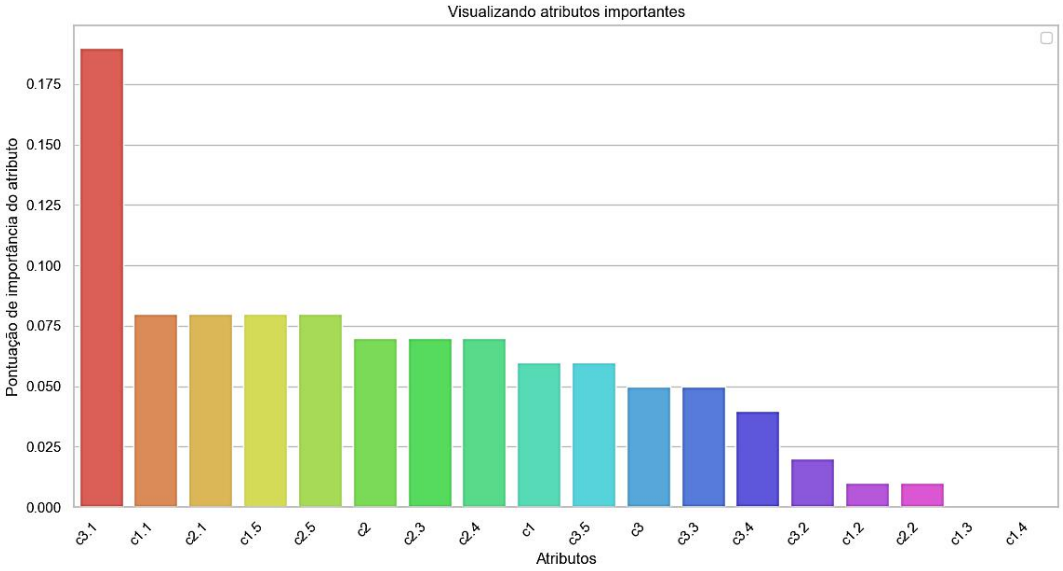


Figura 39 – Atributos da dimensão coerência do conteúdo listados por ordem de importância

6.3 Resultado parcial de acurácia de cada dimensão

Resposta para a questão de pesquisa QP11-12-13-14: Qual a contribuição de cada uma das quatro dimensões? A Figura 40 (quarenta) compara a contribuição das dimensões. A melhor contribuição é de Semântica, depois da dimensão lexical, depois sintática e por último a coerência. Esta comparação foi feita considerando-se somente os atributos de uma dimensão para predizer o escore dos humanos. A dimensão léxica, sintática, semântica e de coerência apresentaram os valores de KQ = 0.42, 0.46, 0.40 e 0.59, respectivamente, que são valores apenas moderados na interpretação do Kappa Quadrático.

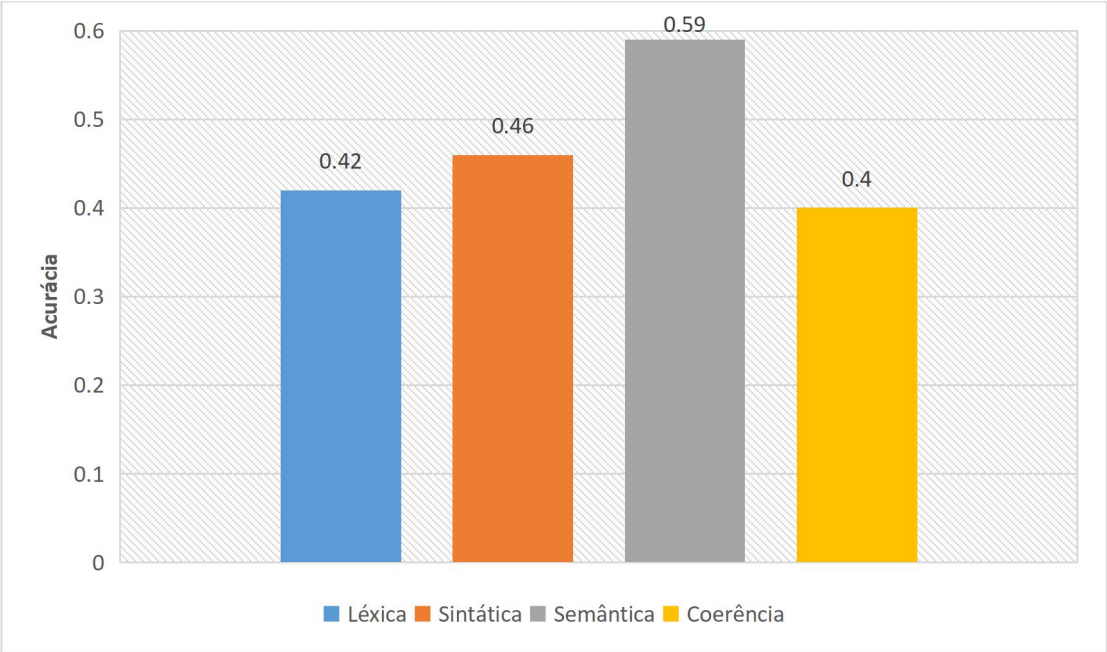


Figura 40 – A contribuição de cada uma das dimensões (Léxica, Sintática, Semântica e Coerência) na acurácia final (**Métrica:** KQ).

6.4 Resultado da combinação das dimensões

Buscando explorar o cruzamento entre as dimensões responda-se à questão de pesquisa **(QP14)** qual a acurácia da combinação das dimensões duas a duas: Léxica x Sintática, Léxica x Semântica, Léxica x Coerência, Sintática x Semântica, Sintática x Coerência e Semântica x Coerência? Além disso, como a dimensão Semântica tem a maior influência explora-se Conteúdo + Léxica + Coerência, Semântica + Sintática + Coerência e Semântica + Léxica + Sintática (Figura 41).

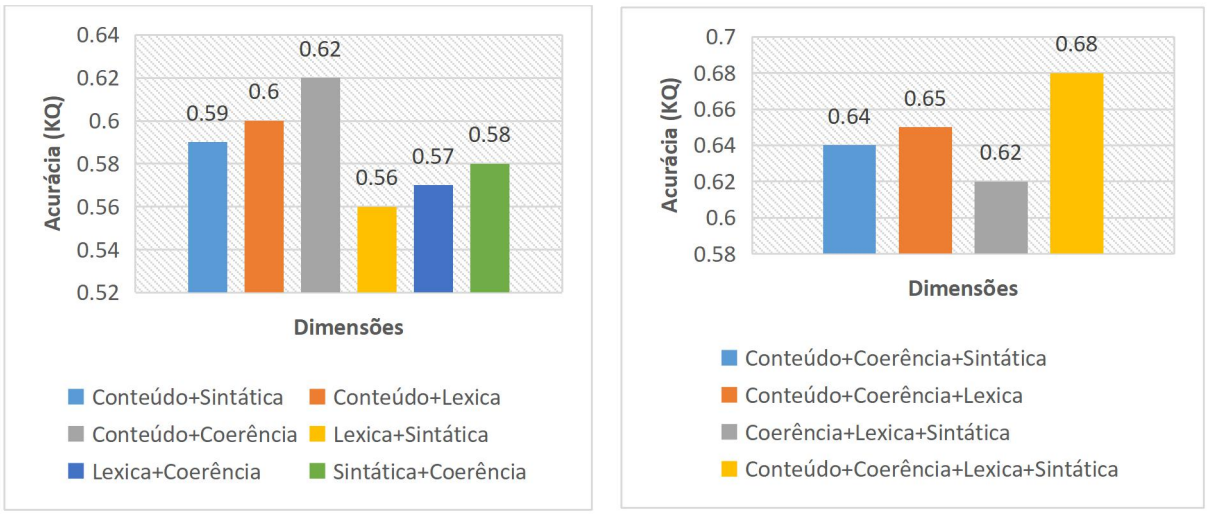


Figura 41 – Explorando a combinação das dimensões 2 a 2 e 3 a 3 na contribuição da acurácia.

Com duas dimensões, a menor acurácia vem da combinação da dimensão léxica e sintática com 0.56 sendo que o melhor desempenho é com a combinação das dimensões de semântica + coerência, com 0.62. Com três dimensões a pior combinação vem de semântica + léxica + sintática com 0.62 e a melhor vem de semântica + coerência + léxica, com 0.65.

Agora então, combinam-se as quatro dimensões para responder à questão de pesquisa **(QP16):** O método de avaliação com base nas quatro dimensões alcança a acurácia dos avaliadores humanos? A combinação das quatro dimensões linguísticas resultou uma acurácia final de KQ de 0.68 contra a acurácia *HxH* com valor de 0.56. Portanto, o sistema supera a acurácia humana.

A Figura 42 (quarenta e dois) apresenta o desempenho das pontuações em comparação entre *SxH*, para uma amostra aleatória de 100 (cem) ensaios na combinação das quatro dimensões. Observa-se que o desempenho das pontuações *HxH* e *SxH* em alguns pontos são similares, significando uma aproximação entre as notas. Por outro lado, existem alguns pontos de discrepâncias no desempenho do *HxH* em relação *SxH*. Para Zupanc e Bosnić, (2018) avaliadores humanos introduzem involuntariamente pontuações subjetivas dificultando a aprendizagem dessas pontuações ruidosas pelo modelo de aprendizagem de máquina.

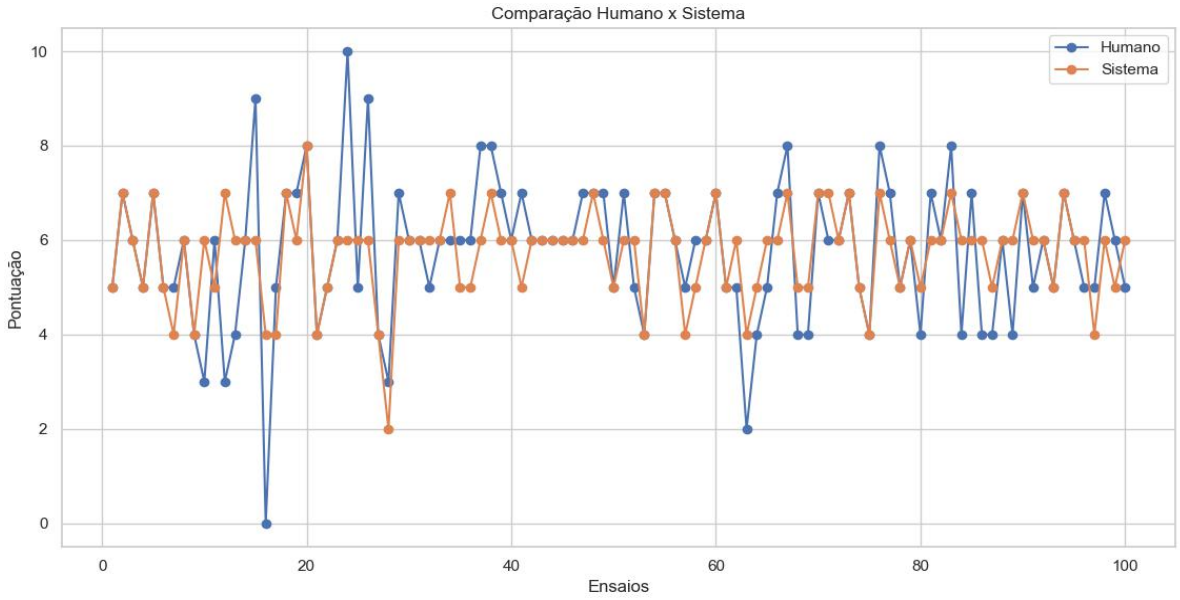


Figura 42 – Desempenho nas notas em comparação *SxH* contra *HxH*, com os atributos das quatro dimensões

6.5 Resultado da influência de cada dimensão nas três competências de avaliação (tema (conteúdo), coerência e regras)

No conjunto de dados das redações, cada escore final é vinculado com três notas parciais referentes às competências: tema, coerência e regras. Neste caso, para responder à questão de pesquisa **(QP15)** medisse por meio da correlação de *Pearson* a influência de cada uma das dimensões em relação ao escore parcial da competência.

Inicialmente, mostra-se novamente a contribuição de cada dimensão com o escore final (Resultado da Figura 43) contrastando KQ com correlação. O que se pode ver são os valores relativamente próximos. Em alguns casos a correção é maior, por exemplo, 0.66 *x* 0.59 e 0.49 *x* 0.42, mas em outro o KQ é superior 0.46 *x* 0.44 e 0.4 *x* 0.35.

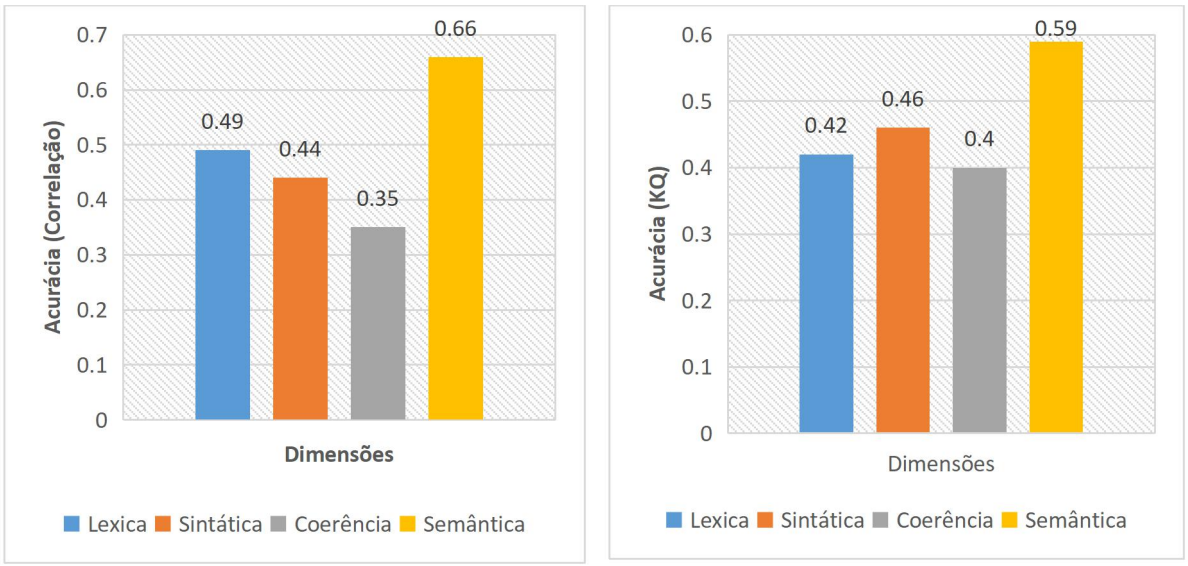


Figura 43 – Comparativo da participação das quatro dimensões na composição do escore final, comparando a Correlação e o Kappa; o gráfico da direita é cópia da Figura 30.

Na Figura 44 (quarenta e quatro) apresenta-se o comparativo com escore parcial do tema, "Tema - competência". O melhor resultado da correlação foi 0.68 e a menor 0.41. De certo modo, a correlação e o kappa possuem valores alinhados, a menor correlação corresponde ao menor kappa e vice-versa.

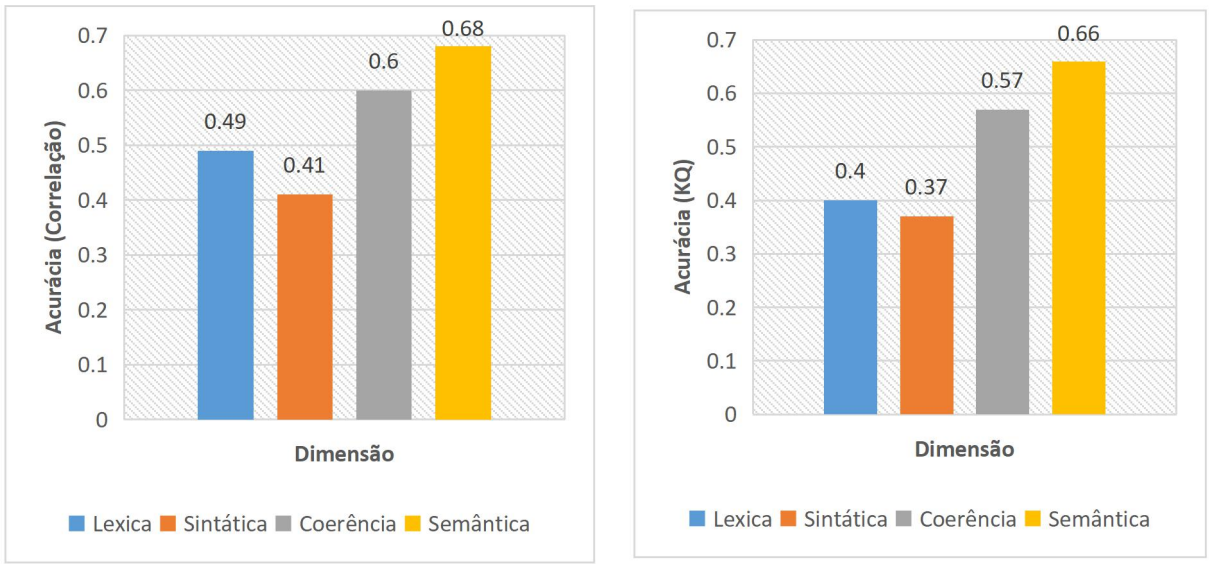


Figura 44 – Comparativo da participação das quatro dimensões na composição do tema-competência.

Na Figura 45 (quarenta e quatro) o experimento está relacionado à competência de "Coerência-competência" onde se avalia a utilização de recursos coesivos da modalidade de escrita. Na correlação o melhor desempenho foi 0.68 para a dimensão semântica e o pior 0.41 para a dimensão sintática. Novamente, os gráficos têm comportamento similar, sendo que o mais bem avaliado em correlação é o melhor em kappa.

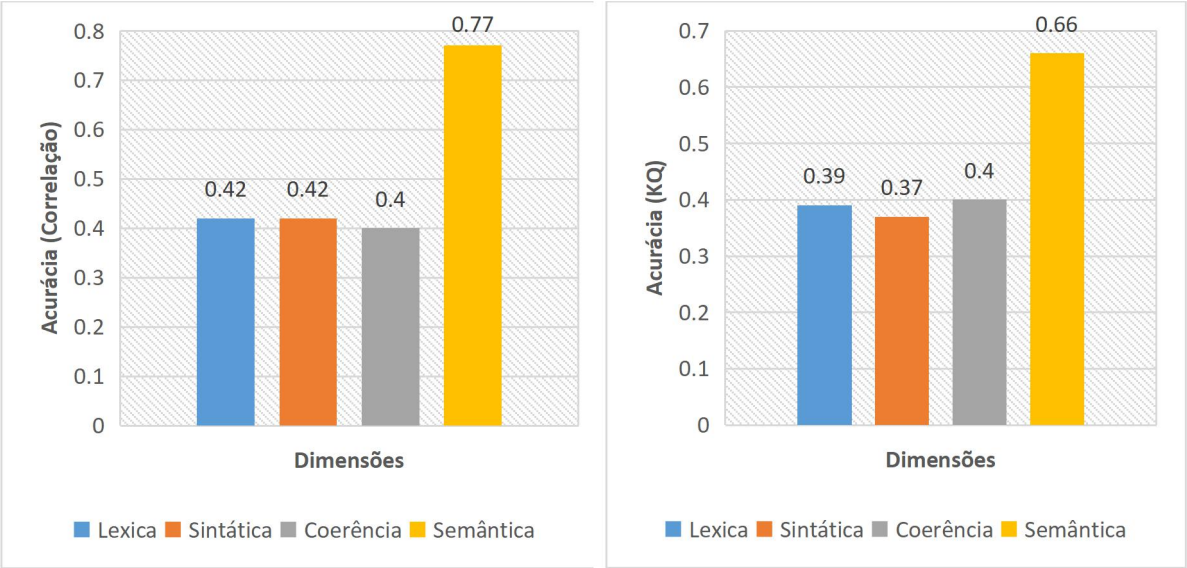


Figura 45 – Comparativo da participação das quatro dimensões na composição do coerência-competência.

Na Figura 46 (quarenta e seis) o experimento foi relacionado à competência de "Regra-competência" onde se avalia a utilização de recursos coesivos da modalidade escrita. Na correlação o pior resultado foi da dimensão léxica com 0.18 e o melhor foi da Semântica com 0.45. Houve pequenas variações no comportamento da correlação *vs* kappa. Para kappa o pior resultado foi a dimensão sintática com 0.12; a melhor contribuição ficou com a Semântica para ambos as métricas.

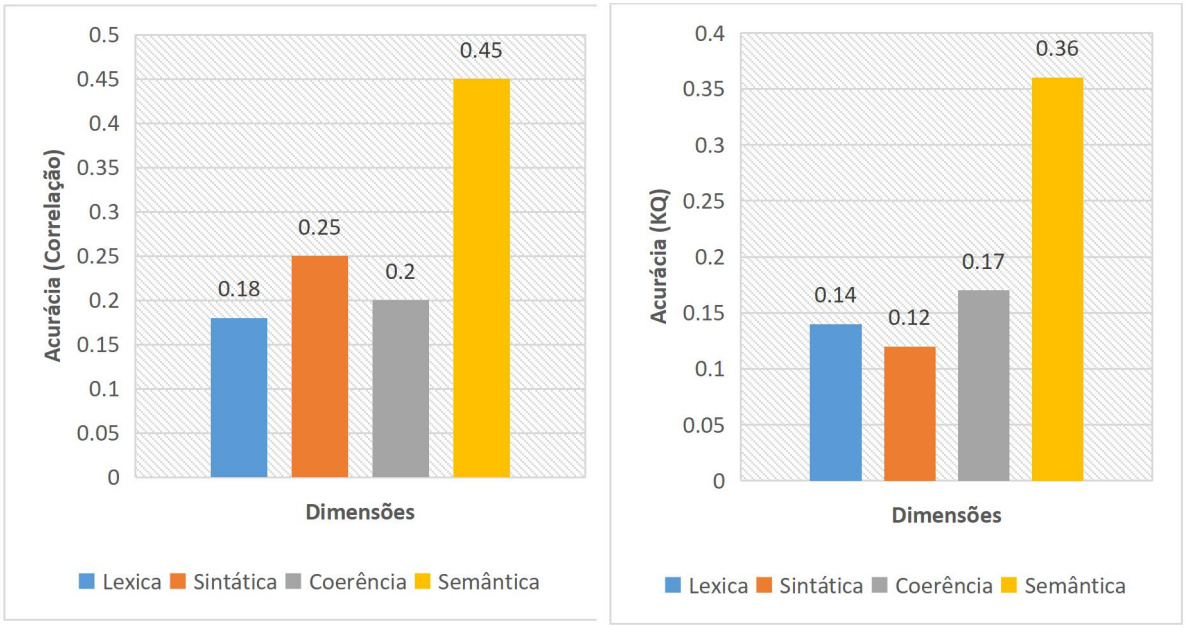


Figura 46 – Comparativo da participação das dimensões na composição do regra-competência.

No geral, para todas as três competências a melhor contribuição foi da dimensão Semântica e a segunda melhor foi com a dimensão coerência, no tema-competência a correlação ficou acima de 0.5 para estas duas dimensões. Por outro lado, as dimensões léxica e sintática em todas as competências ficaram abaixo de 0.5 com métrica de correlação. Na avaliação do KQ os valores dessas dimensões são considerados insuficientes.

7 Conclusão e trabalhos futuros

Contexto. O objetivo deste trabalho foi desenvolver um método de avaliação automática de respostas discursivas, curtas e ensaios (redações), baseado na coleta de atributos em quatro dimensões: léxica, sintática, semântica e coerência. O trabalho inicia com uma revisão da literatura onde a grande maioria dos trabalhos encontrados, com boa acurácia, são de avaliadores automáticos para a língua inglesa. A partir da revisão da literatura foram agrupados numa taxonomia de quatro dimensões mais de 140 (cento e quarenta) atributos. Alguns atributos foram ajustados para a língua portuguesa, por exemplo, os atributos da dimensão sintática dependem diretamente de ferramentas para o Português, como um etiquetador e outras ferramentas de correção léxica e sintática.

Para realização dos experimentos, com respostas curtas e com ensaios (redações), propomos o uso de uma arquitetura *pipeline* linear de 5 (cinco) etapas: preparação de *corpus*, pré-processamento, coleta de atributos, modelo de predição e avaliação.

Foram trabalhados três conjuntos de dados para respostas curtas e um conjunto para redações (*corpus* que está disponibilizado *on-line*):

- 131 (cento e trinta e uma) respostas de uma prova de Biologia, 229 (duzentas e vinte e nove) respostas de uma prova de Geografia e 192 (cento e noventa e duas) respostas de uma prova de Filosofia.
- 1.000 (mil) redações de um concurso público do edital nº 26/2016- UFOPA para admissão na carreira de técnico administrativo em educação.

Engenharia de atributos. Os dois temas desta pesquisa é o estudo da coleta e da importância dos atributos. O algoritmo *Random Forest* permitiu gerar um classificador a partir de um grande número de atributos, além de retornar a relevância de cada atributo na etapa de classificação. Respondem-se as seguintes questões de pesquisa relacionadas com os atributos:

Resultado I. Para respostas **curtas** se fez dois experimentos, um com a medida de acurácia Erro Médio e outro com a medida de acurácia Kappa Quadrático (KQ). Ambos apresentaram resultados similares que superaram a acurácia dos avaliadores humanos, mas em relação ao segundo experimento foram coletados mais atributos. A partir dos atributos coletados o objetivo foi predizer o escore dos avaliadores humanos com uma acurácia

próxima ou até superior a acurácia medida entre os dois avaliadores humanos. Como resultado para respostas curtas se obtive um KQ 0.72 *SxH* contra 0.89 *HxH* para a prova de Biologia e, um valor *SxH* 0.76 contra *HxH* 0.58 para a prova de Geografia. Por assim, na prova de Geografia o sistema superou a acurácia dos humanos. Por outro lado, na prova de Biologia o sistema atingiu um KQ de 0.72 que é considerado "substancial", mesmo sendo inferior ao coletado entre dois avaliadores humanos. Estes resultados mostram que a tecnologia é promissora e já pode ser utilizado na prática em ambientes virtuais de aprendizagem, no qual esta tecnologia já está alcançando um estado de maturidade.

Resultado II. Para as 1.000 (mil) **redações** o resultado principal da acurácia do sistema foi de 0.86 *SxH* (KQ) contra 0.58 *HxH*. Aqui também o sistema superou a acurácia dos avaliadores humanos. A contribuição de cada dimensão no escore final foi de (KQ) 0.42 léxica, 0.46 sintática, 0.40 coerência e 0.59 semântica. Por outro lado, medimos a contribuição de cada dimensão para os três escore*s* parciais das competências de avaliação: tema, coerência e regras. A dimensão Semântica sempre teve a melhor contribuição (0.66 corr tema; 0.77 coerência; 0.45 regras). A dimensão coerência para o tema também teve uma correlação boa 0.60.

Discussão: curta *versus* ensaio. A grande diferença entre respostas curtas e ensaios é a quantidade de texto, em números de palavras. Numa resposta curta com pouco texto, algumas das tecnologias que são promissoras para ensaios, como o método do LSA para o conteúdo e/ou para a coerência de conteúdo não podem ser aplicadas. Numa resposta curta o conteúdo foi avaliado por similaridade de texto simples. Por outro lado, nas respostas curtas foi utilizado também bigramas de palavras que mantém a ordem de escrita do texto. Para os experimentos das redações não foi utilizado a tecnologia dos bigramas, pois o LSA ficaria muito pesado.

A relevância de se trabalhar com quatro dimensões de atributos é poder dar um *feedback* para o estudante que não é apenas formado pelo escore. O estudante pode receber vários indicadores da qualidade da sua resposta, das várias dimensões, por exemplo: da dimensão sintática: os erros, léxicos e gramaticais; da dimensão léxica: algumas medidas de legibilidade e diversidade lexical; da dimensão de coerência de conteúdo: um índice; da dimensão de conteúdo: outro índice.

Outro fato relevante, quanto às métricas combinadas com LSA, foi que a métrica mais utilizada que é o Cosseno, nos nossos experimentos, não alcançou os melhores resultados. Os melhores resultados foram da combinação de Cosseno mais Distância Euclidiana.

7.1 Contribuições

Uma das contribuições do estudo foi a coleta e classificação dos atributos utilizados na pesquisa. Este estudo deverá ser aprofundado em trabalhos futuros, buscando-se conhecer melhor a contribuição de cada atributo, por exemplo, nas redações, relacionando as competências usadas na avaliação. Reconhece-se que muito ainda pode ser feito, como uma continuidade deste trabalho, por exemplo, encontrando atributos que são relevantes para a acurácia e também para dar um *feedback* para o estudante.

Outra contribuição foi clarificar a área de avaliação automática de respostas discursivas para o Português com a discussão de 17 (dezessete) questões de pesquisa, sendo algumas relacionadas com aspectos das tecnologias e, outras relacionadas com a coleta e contribuição de cada dimensão de atributos.

Outra contribuição é a disponibilidade dos *corpora* utilizados nesta pesquisa para outros pesquisadores (http://www.labx.ufpa.br/dataset.html).

Por fim, a melhor e mais importante contribuição são os números encontrados na acurácia que mostram que esta área está amadurecendo e já se podem ter avaliadores para auxiliar na tarefa do professor na correção de textos escritos.

Como resultantes desta pesquisa foram publicados 01 (um) artigo em periódico e 04 (quatro) artigos em eventos como foi mostrado no item 1.4.4.

7.2 Limitações

Uma das limitações foi o número pequeno de questões de respostas curtas que foram encontrados para fazer a pesquisa. Outra limitação similar deve-se as redações. Em nossa Universidade se tem milhões de provas dos processos seletivos, mas estão todas em páginas manuscritas. Desse modo, torna-se um desafio criar uma base grande, pois depende da digitalização manual.

Nas respostas curtas, melhores resultados poderiam ser alcançados com algum método de expansão de vocabulário, por exemplo, com o uso da *wordnet* (BENOMRAN; AB AZIZ, 2013) para gerar os sinônimos.

7.3 Trabalhos futuros

Pode-se aprofundar o estudo deste trabalho em diversas frentes, listam-se algumas abaixo:

- Aumento do *corpus*, como aumentar as bases utilizadas, por exemplo, com mais alguns milhares de redações.
- Aplicar o *pipeline* proposto em outras bases, sobre outros temas, em outros domínios.
- Aprofundar o estudo da importância de cada atributo associado ao escore e/ou as competências.
- Aplicar algumas técnicas de *machine learning* como *deep learning* e rede neural.
- Criar uma versão operacional de um avaliador automático numa plataforma virtual de ensino onde podem ser explorados assuntos relacionados com o *feedback* com os estudantes; múltiplas submissões da resposta a partir do *feedback*, entre outros.
- Fazer aplicabilidade do algoritmo num ambiente virtual em atividade.
- Fazer uma Análise estatística, para cada um dos atributos estudados, comparando as acurácias encontradas nos experimentos para validar se o aumento de acurácias relativo aquele atributo é realmente significativo ou não.

Referências

ALENCAR, L. F. de. (2010) **Aelius: uma ferramenta para anotação automática de corpora usando NLTK.** IX Encontro de Linguística de Corpus. Porto Alegre, PUCRS.

ALENCAR, L. F. de. (2012) **Superando o estado da arte na etiquetagem morfossintática por meio de regras de pós-etiquetagem.** In: Anais do X Encontro de Linguística de Corpus – Aspectos metodológicos dos estudos de corpora. Belo Horizonte: UFMG.

ALENCAR, L. F.; **Aelius: uma ferramenta para anotação automática de corpora usando o NLTK**. In: IBAÑOS, A. M. T.; MOTTIN, L. P.; SARMENTO, S.; SARDINHA, T. B.. (Org.). Pesquisas e perspectivas em linguística de corpus. 1ed.Campinas: Mercado de Letras, 2015, v. 1, p. 233-282.

ALENCAR, L. F.; **Novos recursos do Aelius para o processamento computacional raso do português.** In: Éric Laporte; Aucione Smarsaro; Oto Araújo Vale. (Org.). Dialogar é preciso: linguística para o processamento de línguas. 1ed.Vitória: PPGEL/UFES, 2013, v. , p. 7-20.

AMALIA, A. et al. (2019) **Automated Bahasa Indonesia essay evaluation with latent semantic analysis**. In: Journal of Physics: Conference Series. IOP Publishing. p. 012100.

ATTALI, Y. (2013) **Handbook of Automated Essay Evaluation: Current Applications and New Directions.** [S.l.]: M. D. Shermis and J. C. Burstein.

ATTALI, Y. e BURSTEIN, J. (2004). **Automated essay scoring with e-rater® v. 2.0**. ETS Research Report Series, 2004(2), i-21.

ATTALI, Y., BURSTEIN, J. e ANDREYEV, S. (2003) **E-rater version 2.0: Combining writing analysis feedback with automated essay scoring.** Unpublished manuscript.

AZMI, A. M., AL-JOUIE, M. F., e HUSSAIN, M. (2019). **AAEE–Automated evaluation of students' essays in Arabic language.** Information Processing & Management, 56(5), 1736-1752.

BALIKAS, G. (2018**). Lexical Bias In Essay Level Prediction.** arXiv preprint arXiv:1809.08935.

BENOMRAN, A. e AB AZIZ, M. (2013). **Automatic essay grading system for short answers in english language.** Journal of Computer Science, 9:1369–1382. DOI: 10.3844/jcssp.2013.

BIRD, S., KLEIN, E. e LOPER, E. (2009). **Natural language processing with Python: analyzing text with the natural language toolkit**. " O'Reilly Media, Inc.".

BJORNSSON, C. H. (1968). **Lasbarhet [readability].** Bokförlaget Liber, Stockholm.

BREIMAN, L. (2001) **"Random forests"** Machine learning. Vol. 45, nº 1, pp. 5-32.

BULL, J. e MCKENNA, C. (2001) **A Blueprint for Computer-Assisted Assessment.** Taylor & Francis Editora.

BURROWS, S, GUREVYCH, I. e STEIN, B. (2015) **The eras and trends of automatic short answer grading.** International Journal of Artificial Intelligence in Education.

BURSTEIN, J. e CHODOROW, M. (1999, June). **Automated essay scoring for nonnative English speakers.** In Proceedings of a Symposium on Computer Mediated Language Assessment and Evaluation in Natural Language Processing (pp. 68-75). Association for Computational Linguistics.

BURSTEIN, J., CHODOROW, M. e LEACOCK, C. (2003, August). **CriterionSM Online Essay Evaluation: An Application for Automated Evaluation of Student Essays**. In IAAI (pp. 3-10).

CARLOTTO, M. S. (2002) **A síndrome de burnout e o trabalho docente.** Psicologia em estudo, v. 7, n. 1, p. 21–29.

DASGUPTA, I., GUO, D., STUHLMÜLLER, A., GERSHMAN, S. J., e GOODMAN, N. D. (2018). **Evaluating compositionality in sentence embeddings**. arXiv preprint arXiv:1802.04302.

DEERWESTER, S., DUMAIS, S. T., FURNAS, G. W., LANDAUER, T. K. e HARSHMAN, R. (1990). **Indexing by latent semantic analysis.** Journal of the American society for information science, 41(6), 391-407.

DONG, F., ZHANG, Y. e YANG, J. (2017). **Attention-based recurrent convolutional neural network for automatic essay scoring.** In Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017) (pp. 153-162).

DONG, G., e LIU, H. (2018**). Feature engineering for machine learning and data analytics**. CRC Press.

DRURY, B., ROSSI, R. G., e de ANDRADE LOPES, A. (2014, October). **Identification of Brazilian Portuguese causative verbs through a weighted graph classification strategy**. In International Conference on Computational Processing of the Portuguese Language (pp. 274-279). Springer, Cham.

DUBAY W. H. **Smart Language: Readers, Readability, and the Grading of Text.** ERIC, 2007.

DUBAY, W. H. (2004). **The Principles of Readability**. Online Submission.

ELLIOT, S. (2003). **IntelliMetric: From here to validity. Automated essay scoring: A cross-disciplinary perspective**, 71-86.

FARR, J. N., JENKINS, J. J. e PATERSON, D. G. (1951). **Simplification of Flesch Reading Ease Formula.** Journal of applied psychology, 35(5), 333.

FERNÁNDEZ-DELGADO, M., CERNADAS, E., BARRO, S. e AMORIM, D. (2014). **Do we need hundreds of classifiers to solve real world classification problems?.** The Journal of Machine Learning Research, 15(1), 3133-3181.

FLEISS, J. L., e COHEN, J. (1973). **The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability**. Educational and psychological measurement, 33(3), 613-619.

FLESCH, R. (1948). **A new readability yardstick**. Journal of applied psychology, 32(3), 221.

FOLTZ, P. W., LAHAM, D. e LANDAUER, T. K. (1999). **The intelligent essay assessor: Applications to educational technology.** Interactive Multimedia Electronic Journal of Computer-Enhanced Learning, 1(2), 939-944.

FONSECA, E., MEDEIROS, I., KAMIKAWACHI, D., e BOKAN, A. (2018, September). **Automatically grading brazilian student essays.** In International Conference on Computational Processing of the Portuguese Language (pp. 170-179). Springer, Cham.

FRAWLEY, W. (2013). **Linguistic semantics.** Routledge.

GOMAA, W. H., e FAHMY, A. A. (2014). **Automatic scoring for answers to Arabic test questions.** Computer Speech & Language, 28(4), 833-857.

GUIRAUD, P. (1954**). Les caractères statistiques du vocabulaire**. Presses universitaires de France.

GUNNING, R. (1968). **The technique of clear writing**. New York: McGraw-Hill.

GÜTL, C. (2007, April**). e-Examiner: towards a fully-automatic knowledge assessment tool applicable in adaptive e-learning systems**. In Proceedings of the 2nd international conference on interactive mobile and computer aided learning (pp. 1-10).

HALÁCSY, P., KORNAI, A., e ORAVECZ, C. (2007). **HunPos: an open source trigram tagger.** In Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions (pp. 209-212). Association for Computational Linguistics.

HALEY, D. T. et al. (2007) **Seeing the whole picture: evaluating automated assessment systems**. ITALICS.

HALLIDAY, M. A. K., e Hasan, R. (2014). **Cohesion in english.** Routledge.

HATZLVASSILOGLOU, V., KLAVANS, J. L. e ESKIN, E. (1999). **Detecting text similarity over short passages: Exploring linguistic feature combinations via machine learning.** In: Joint SIGDAT conference on empirical methods in natural language processing and very large *corpora*.

HE, Y., HUI, S. C. e QUAN, T. T. (2009) **Automatic summary assessment for intelligent tutoring systems.** Computers & Education, 2009.

HEARST, M. A. (2000) **The debate on automated essay grading.** IEE Intelligeng Systems archive.

HIGGINS, D., BURSTEIN, J. e ATTALI, Y. (2006). **Identifying off-topic student essays without topic-specific training data.** Natural Language Engineering, 12(2), 145-159.

HORBACH, A., STENNMANNS, S. e ZESCH, T. (2018, June). **Cross-lingual content scoring.** In Proceedings of the thirteenth workshop on innovative use of NLP for building educational applications (pp. 410-419).

HULTMAN, T. G. e WESTMAN, M. (1977). **Gymnasistsvenska**. Lund: Liber Läromedel.

INEP (2018). **Instituto Nacional de Estudos e Pesquisas Educacionais Anisio Texeira**. Disponível em: http://www.inep.gov.br (Acesso em Novembro 13, 2018).

ISLAN, M. M. e HOQUE, M. L. (2012) **Automated essay scoring using generalized latent semantic analysis**. Journal of Computers.

JOHANSSON, V. (2009). **Lexical diversity and lexical density in speech and writing: A developmental perspective.** Lund Working Papers in Linguistics, 53, 61-79.

JORGE-BOTANA, G., LUZÓN, J. M., GÓMEZ-VEIGA, I. e MARTÍN-CORDERO, J. I. (2015). **Automated LSA assessment of summaries in distance education: some variables to be considered.** Journal of Educational Computing Research, 52(3), 341-364.

KE, Z. e NG, V. (2019, August). **Automated essay scoring: a survey of the state of the art.** In Proceedings of the 28th International Joint Conference on Artificial Intelligence (pp. 6300-6308). AAAI Press.

KELLY, R. **pyenchant: a spellchecking library for python**, 2014.

KINCAID, J. P., FISHBURNE Jr, R. P., ROGERS, R. L. e CHISSOM, B. S. (1975). **Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel.**

KINOSHITA, J., SALVADOR, L. N., MENEZES, C. E. e SILVA, W. D. C. (2007, October). **Cogroo-an openoffice grammar checker.** In Seventh International Conference on Intelligent Systems Design and Applications (ISDA 2007) (pp. 525-530). IEEE.

LANDAUER, T. K, FOLTZ, P. W. e LAHAM, D. (1998) **An introduction to latent semantic analysis.** Discourse processes, Taylor & Francis, v. 25, n. 2-3, p. 259–284.

LANDAUER, T. K., LAHAM, D., REHDER, B.,e SCHREINER, M. E. (1997). **How well can passage meaning be derived without using word order? A comparison of Latent Semantic Analysis and humans.** In Proceedings of the 19th annual meeting of the Cognitive Science Society (pp. 412-417).

LANDAUER, Thomas K. (2003) **Automatic essay assessment. Assessment in education: Principles**, policy & practice, v. 10, n. 3, p. 295-308.

LANDIS, J. R., e KOCH, G. G. (1977). **The measurement of observer agreement for categorical data.** biometrics, 159-174.

LARKEY, L. S. (1998, August). **Automatic essay grading using text categorization techniques.** In SIGIR (Vol. 98, pp. 90-95).

LEACOCK, C., e CHODOROW, M. (2003). **C-rater: Automated scoring of short-answer questions.** Computers and the Humanities, 37(4), 389-405.

LEE, I. (2014*).* **Teachers' reflection on implementation of innovative feedback approaches** in EFL writing. English Teaching, 69(1), 23-40.

LEÓN, J. A., OLMOS, R., ESCUDERO, I., JORGE-BOTANA, G. e PERRY, D. (2013). **Exploring the assessment of summaries: Using latent semantic analysis to grade summaries written by spanish students.** Procedia-Social and Behavioral Sciences, 83, 151-155.

LIU, M., e YANG, J. (2012). **An improvement of TFIDF weighting in text categorization**. International proceedings of computer science and information technology, 44-47.

MAGNINI, B., RODRIGUEZ, P., PEREZ, D., GLIOZZO, A., ALFONSECA, E. e STRAPPARAVA, C. (2005). **About the effects of combining latent semantic analysis with natural language processing techniques for free-text assessment**. Revista signos: estudios de linguıstica, ISSN 0035-0451, No. 59, 2005, pags. 325-343.

MALVERN D. D., RICHARDS B. J, CHIPERE N. e DURÁN P.(2004) **Lexical diversity and language development**. Houndmills, Hampshire, UK: Palgrave Macmillan.

MANN, W. C., e THOMPSON, S. A. (1988). **Rhetorical structure theory: Toward a functional theory of text organization**. Text-interdisciplinary Journal for the Study of Discourse, 8(3), 243-281.

MAYFIELD, E. e ROSÉ, C. P. (2013) *LightSIDE: Open source machine learning for text*. **In: Handbook of Automated Essay Evaluation**. Routledge. p. 146-157.

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

MILLER, G. A. (1998). **WordNet: An electronic lexical database.** MIT press.

MOHLER M., BUNESCU R., e MIHALCEA R. (2011). **Learning to grade short answer questions using semantic similarity measures and dependency graph alignments**. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, pages 752–762, Portland, Oregon, USA, June. Association for Computational Linguistics.

MOHLER, M. e MIHALCEA, R. **Text-to-text semantic similarity for automatic short answer grading.** EACL'09 - Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics, 2009.

MOHRI, M., ROSTAMIZADEH, A., e TALWALKAR, A. (2018). **Foundations of machine learning**. MIT press.

NOORBEHBAHANI, F. e KARDAN, A. A. (2011) **The automatic assessment of free text answers using a modified bleu algorithm.** Computer & Education.

OLMOS, R., LEÓN, J. A., ESCUDERO, I. e JORGE-BOTANA, G. (2011). **Using latent semantic analysis to grade brief summaries: some proposals.** International Journal of Continuing Engineering Education and Life Long Learning, 21(2-3), 192-209.

ORENGO, V. M., e HUYCK, C. (2001, November). **A stemming algorithm for the portuguese language.** In Proceedings Eighth Symposium on String Processing and Information Retrieval (pp. 186-193). IEEE.

ORLANDI, E. P. (2017). **O que é linguística.** Brasiliense.

OTHERO, G. D. Á., e AYRES, M. R. (2014). **Anotação morfológica automática de corpus de língua falada: desafios ao Aelius.** Texto livre. Belo Horizonte, MG. Vol. 7, n. 2 (primavera 2014), f. 44-60.

PAGE, E. B. (1966) **The imminence of grading essay by computer.** The Phi Delta Kappan.

PALMA, D. e ATKINSON, J. (2018) **Coherence-Based Automatic Essay Assessment. IEEE** Intelligent Systems, v. 33, n. 5, p. 26-36.

PASSERO, G., HAENDCHEN Filho, A., e DAZZI, R. (2016, November). **Avaliaçao do uso de métodos baseados em lsa e wordnet para correçao de questoes discursivas**. In Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na

Educação-SBIE) (Vol. 27, No. 1, p. 1136).

PEREZ, D. et al. (2005a) **About the effects of combining latent semantic analysis with natural language processing techniques for free-text assessment.** Revista signos, Pontificia Universidad Católica de Valparaíso, v. 38, n. 59, p. 325–343.

PEREZ, D. et al. (2005b) **Automatic assessment of students free text answers underpinned by the combination of a bleu inspired algorithm and latent semantic analysis.** Machine Translation, 2005.

PRIBADI, F. S. et al. (2017) **Automatic short answer scoring using words overlapping methods.** API Conference Proceedings.

PRIBADI, F. S., ADJI, T. B., e PERMANASARI, A. E. (2016). **Automated short answer scoring using weighted cosine coefficient.** 2016 IEEE Conference on eLearning, eManagement and e-Services (IC3e), pages 70–74. DOI: 10.1109/IC3e.2016.8009042

PRIBADI, F. S., PERMANASARI, A. E. e ADJI, T. B. (2018). **Short answer scoring system using automatic reference answer generation and geometric average normalized-longest common subsequence (GAN-LCS)**. Education and Information Technologies, 23(6), 2855-2866.

RABABAH, H., e AL-TAANI, A. T. (2017, May). **An automated scoring approach for Arabic short answers essay questions.** In 2017 8th International Conference on Information Technology (ICIT) (pp. 697-702). IEEE.

READABLE. **What is Readability?** Disponível em: https://readable.com/readability/. Acesso em: 8 de jun. 2019

REAFAT, M. M., EWEES, A. A., EISA, M. M. e AB SALLAM, A. (2012). **Automated assessment of students arabic free-text answers.** Int J Cooperative Inform Syst, 12, 213-222.

REZENDE, S. O., MARCACINI, R. M., e MOURA, M. F. (2011). **O uso da mineração de textos para extração e organização não supervisionada de conhecimento**. Embrapa Informática Agropecuária-Artigo em periódico indexado (ALICE).

RICH, Changhua S.; SCHNEIDER, M. Christina e D'BROT, JUAN M**. Applications of automated essay evaluation in West Virginia**. In: Handbook of Automated Essay Evaluation. Routledge, 2013. p. 121-145.

RICHARDS, B. (1987). **Type/token ratios: What do they really tell us?.** Journal of child language, 14(2), 201-209.

RODRIGUES, F. e ARAÚJO, L. (2012) **Automatic Assessment of Short Free Text Answers.** In: CSEDU (2). p. 50-57.

RUDNER, L. M. e LIANG, T. (2002) **Automated essay scoring using Bayes' theorem. The Journal of Technology**, Learning and Assessment, v. 1, n. 2.

RUDNER, L. M., GARCIA, V. e WELCH, C. (2006). **An evaluation of IntelliMetric™ essay scoring system.** The Journal of Technology, Learning and Assessment, 4(4)

SANTOS, H. G. dos. **Comparação da performance de algoritmos de machine learning para a análise preditiva em saúde pública e medicina**. 2018. Tese de Doutorado. Universidade de São Paulo.

SANTOS, J. C. A. e FAVERO, E. L. (2015). **Practical use of a latent semantic analysis (LSA) model for automatic evaluation of written answers**. Journal of the Brazilian Computer Society, 21(1), 21.

SCHULTZ, M. T. (2013). **The IntelliMetric automated essay scoring engine-a review and an application to chinese essay scoring.** Handbook of automated essay scoring: Current applications and future directions, 89-98.

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

SHEHAB, A., FAROUN, M., e RASHAD, M. (2018). **An automatic Arabic essay grading system based on text similarity Algorithms**. Int. J. Adv. Comput. Sci. Appl.(IJACSA), 9(3), 263-268.

SHERMIS, M. D., e BURSTEIN, J. (Eds.). (2013). **Handbook of automated essay evaluation: Current applications and new directions**. Routledge.

SHERMIS, M. D., KOCH, C. M., PAGE, E. B., KEITH, T. Z. e HARRINGTON, S. (2002). **Trait ratings for automated essay grading. Educational and Psychological Measurement**, 62 (1): 5 – 18. DOI: 10.1177/001316440206200101.

SHERMIS, M. e BURSTEIN, J. **Automated essay scoring: A cross disciplinary perspective**. Lawrence Erlbaum Associates, 2003.

SIDDIQI, R, HARRISON, C. J. e SIDDIQI, R. (2010) **Improving teaching and learning through automated short-answer marking.** IEEE Transactions on Learning Technologies.

SILBER, H. G. e MCCOY, K. F. (2002). **Efficiently computed lexical chains as an intermediate representation for automatic text summarization**. Computational Linguistics, 28(4), 487-496.

SILVA, W. D. C. de M. (2013) **Aprimorando o corretor gramatical CoGrOO.** Tese de Doutorado. Universidade de São Paulo.

SUKKARIEH, J. Z. e STOYANCHEV, S. (2009) **Automating model building in c-rater.** Proceedings of the 2009 Workshop on Applied Textual Inference. Association for Computational Linguistics.

TOUTANOVA, K., KLEIN, D., MANNING, C. D. e SINGER, Y. (2003, May). **Feature-rich part-of-speech tagging with a cyclic dependency network.** In Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1 (pp. 173-180). Association for computational Linguistics.

VAJJALLA, S. (2018) **"Automated assessment of non-native learner essays: Investigating the role of linguistic features"**. International Journal of Artificial Intelligence in Education, v. 28, n. 1, p. 79-105.

VALENTI, S., NERI, F. e CUCCHIARELLI, A. (2003) **An overview of current research on automated essay grading.** Journal of Information Technology Education: Research.

WACHSMUTH, H., Stein, B., e ENGELS, G. (2011, October). **Constructing efficient information extraction pipelines.** In Proceedings of the 20th ACM international conference on Information and knowledge management (pp. 2237-2240). ACM.

XIA, T., e CHAI, Y. (2011). **An improvement to TF-IDF: Term Distribution based Term Weight Algorithm.** JSW, 6(3), 413-420.

YANG, W. (2012). **A study of students' perceptions and attitudes towards genre-based ESP writing instruction.** The Asian ESP Journal, 8(3), 50-73.

YULE G.U. (1944). **The Statistical Study of Literary Vocabulary**. Cambridge University Press, Cambridge.

ZEN, K., ISKANDAR, A. e LINANG, O. (2011) **Using latent semantic analysis for automated grading programming assignments.** International Conference on Semantic Technology and Information Retrieval.

ZIAI, R., OTT, N. e MEURERS, D. (2012). **Short answer assessment: Establishing links between research strands.** In Proceedings of the Seventh Workshop on Building Educational Applications Using NLP (pp. 190-200). Association for Computational Linguistics.

ZUPANC, K. (2018) **Semantics-based automated essay evaluation.** 120 f. PhD thesis - Faculty of Computer and Information Science, University of Ljubljana.

ZUPANC, K. e BOSNIC, Z. (2015) *Automated essay evaluation augmented with semantic coherence measures*. **IEEE International Conference on Data Mining (ICDM).**

ZUPANC, K. e BOSNIC, Z. (2017). *Automated essay evaluation with semantic analysis*. **Know.-Based Syst.,** 120(C):118 – 132. DOI: 10.1016/j.knosys.2017.01.006

ZUPANC, K., e Bosnic, Z. (2016). **Advances in the field of automated essay evaluation.** Informatica, 39(4).

ZUPANC, K., e BOSNIĆ, Z. (2018, June). **Increasing accuracy of automated essay grading by grouping similar graders.** In Proceedings of the 8th International Conference on Web Intelligence, Mining and Semantics (pp. 1-6).

Apêndice

Neste apêndice, mostra-se trechos do código utilizado nos experimentos desta pesquisa, todo o material encontra-se no repositório *Github* no endereço https://github.com/labx-ufpa/AAT.

```
#####################  BIBLIOTECAS  ##########################
import pandas as pd
from nltk import *
import textstat
from readcalc import readcalc
from lexicalrichness import *
import collections as col
from Aelius import AnotaCorpus as a
from Aelius.Toqueniza import TOK_PORT as tok
import csv
from numpy.linalg import inv
inf = float(1e-20)  # evitar divisão por zero
import re
from unicodedata import normalize
import nltk
from nltk import *
from nltk.corpus import stopwords
from nltk.util import ngrams
import numpy as np
from sklearn.feature_extraction.text import TfidfTransformer
from sklearn.feature_extraction.text import CountVectorizer
from scipy import linalg
import random
import numpy as np
from sklearn.model_selection import train_test_split, RepeatedKFold
from sklearn.preprocessing import StandardScaler
from sklearn.ensemble import RandomForestRegressor as regressor
from sklearn.metrics import cohen_kappa_escore
import pycm
import warnings
warnings.filterwarnings('ignore')   # To ignore all warnings that arise
here to enhance clarity
from sklearn.utils import shuffle
```

```
#####################  IMPORTAR DADOS  ######################
dados = pd.read_csv("base.csv", usecols = ['Textos','Arquivos'])

notas = pd.read_csv("notas.csv", usecols =
['Pacote','Sequencia','NotaF','TemaF','CoerenciaF','RegrasF'])

notaszip = list(zip(notas['Pacote'],
notas['Sequencia'],notas['NotaF'],notas['TemaF'],notas['CoerenciaF'],notas
['RegrasF']))
```

```
#####################  PRÉ-PROCESSAMENTO  ##################
class ProcLing():
```

```
    def __init__(self, lista):
        self.lista = lista
    def retiraAcentuacao(self):
        lista_sem_acentos = []
        for l in self.lista:
            try:
                l = normalize('NFKD',
l).encode('ASCII','ignore').decode('ASCII')
            except:
                l = l
            lista_sem_acentos.append(l)
        return lista_sem_acentos
    def transformaMinusculas(self):
        lista_minuscula=[]
        for i in range(0,len(self.lista)):
            lista_minuscula.append(self.lista[i].lower())
        return lista_minuscula
    def retiraPontuacao(self):
        lista_sem_pontuacao=[]
        for i in range(0,len(self.lista)):
            lista_sem_pontuacao.append(re.sub(u'["...:,;()!?%&"]',' 
',self.lista[i]))
        return lista_sem_pontuacao
    def filtrar(self):
        self.lista = self.retiraAcentuacao()
        self.lista = self.transformaMinusculas()
        self.lista = self.retiraPontuacao()
        return self.lista

Textos_ssw=[]
for i in range(0,len(Textos)):
    Textos_ssw.append(cleanData(Textos[i],remove_stops = True))

Textos_cst=[]
for i in range(0,len(Textos)):
    Textos_cst.append(cleanData(Textos_ssw[i],stemming = True))
```

```
#####################  DIMENSÃO LEXICAL  ######################
def lenwords(T):
    L=[len(x) for x in word_tokenize(T)]
    return L

def get_yules(s):
    tokens = tokenize(s)
    token_counter = col.Counter(tok.upper() for tok in tokens)
    m1 = sum(token_counter.values())
    m2 = sum([freq ** 2 for freq in token_counter.values()])
    if m1==m2:
        i=1
    else:i = (m1*m1) / (m2-m1)
    k = 1/i * 10000
    return k

def hapaxLegonema(x):
    return(sum([1 for (x,y) in Counter(x).items() if y==1]))

def guiraudIndex(x):return round(len(set(x))/(len(x)**0.5)*10)
def lexical_diversity(text):return len(set(text)) / len(text)
```

```
SMOGindex=[]
for i in range(0,len(Textos)):
    s = readcalc.ReadCalc(Textos[i])
    SMOGindex.append(int(s.get_smog_index()))

TTR=[]
for i in range(0,len(Textos)):
    lex = LexicalRichness(Textos[i])
    TTR.append(lex.ttr)

CTTR=[]
for i in range(0,len(Textos)):
    lex = LexicalRichness(Textos[i])
    CTTR.append(lex.cttr)

MSTTR=[]
for i in range(0,len(Textos)):
    lex = LexicalRichness(Textos[i])
    MSTTR.append(lex.msttr(segment_window=25))

(...)

L=[nsentences,ncharacteres,snwords,nlongwords,nshortswords,
   mostfreqwordslen,averagewordslen,ndifwords,nstopwords,
   nsyllable,SMOGindex,LD,TTR,CTTR,MSTTR,MATTR,MTLD,
   HDD,Gindex,YK,Hapax,GFindex,FREindex,FKGindex,DCindex,
   ARindex,LIXindex,CLindex,NotasF,TemaF,CoerenciaF,RegrasF]

L=list(zip(*L))
H=[]
for i in range(len(L)):H.append(list(L[i]))

df = pd.DataFrame(H,columns=['nsentences','ncharacteres',
                             'snwords','nlongwords','nshortswords',
                             'mostfreqwordslen','averagewordslen',
                             'ndifwords','nstopwords','nsyllable',
                             'SMOGindex','LD','TTR','CTTR','MSTTR',
                             'MATTR','MTLD','HDD','Gindex','YK',
                             'Hapax','GFindex','FREindex','FKGindex',
                             'DCindex','ARindex','LIXindex','CLindex',
                             'NotasF','TemaF','CoerenciaF','RegrasF'])

df.to_csv('lexical.csv', sep=';', encoding='utf-8')
```

```
####################  DIMENSÃO SINTÁTICO  ###################
TAGS=[]
for i in range(0,len(T)):
    t=T[i]
    t=t.decode("utf-8")
    sents=tok.tokenize(t)
    m = a.TAGGER2
    h = a.anota_sentencas([sents], m)
    TAGS.append(extrai_tags(h[0]))

ndifpostag=[]
for i in range(0,len(TAGS)):ndifpostag.append(len(set(TAGS[i])))
```

```
SR=[]
for i in range(0,len(TAGS)):SR.append(TAGS[i].count('SR')+
                                      TAGS[i].count('SR-F')+
                                      TAGS[i].count('SR-I')+
                                      TAGS[i].count('SR-P')+
                                      TAGS[i].count('SR-SP')+
                                      TAGS[i].count('SR-D')+
                                      TAGS[i].count('SR-RA')+
                                      TAGS[i].count('SR-SD')+
                                      TAGS[i].count('SR-R')+
                                      TAGS[i].count('SR-SR')+
                                      TAGS[i].count('SR-G')+
                                      TAGS[i].count('SR-PP'))

HV=[]
for i in range(0,len(TAGS)):HV.append(TAGS[i].count('HV')+
                                      TAGS[i].count('HV-F')+
                                      TAGS[i].count('HV-I')+
                                      TAGS[i].count('HV-P')+
                                      TAGS[i].count('HV-SP')+
                                      TAGS[i].count('HV-D')+
                                      TAGS[i].count('HV-RA')+
                                      TAGS[i].count('HV-SD')+
                                      TAGS[i].count('HV-R')+
                                      TAGS[i].count('HV-SR')+
                                      TAGS[i].count('HV-G')+
                                      TAGS[i].count('HV-PP')+
                                      TAGS[i].count('HV-NA'))

ET=[]
for i in range(0,len(TAGS)):ET.append(TAGS[i].count('ET')+
                                      TAGS[i].count('ET-F')+
                                      TAGS[i].count('ET-I')+
                                      TAGS[i].count('ET-P')+
                                      TAGS[i].count('ET-SP')+
                                      TAGS[i].count('ET-D')+
                                      TAGS[i].count('ET-RA')+
                                      TAGS[i].count('ET-SD')+
                                      TAGS[i].count('ET-R')+
                                      TAGS[i].count('ET-SR')+
                                      TAGS[i].count('ET-G')+
                                      TAGS[i].count('ET-PP'))

TR=[]
for i in range(0,len(TAGS)):TR.append(TAGS[i].count('TR')+
                                      TAGS[i].count('TR-F')+
                                      TAGS[i].count('TR-I')+
                                      TAGS[i].count('TR-P')+
                                      TAGS[i].count('TR-SP')+
                                      TAGS[i].count('TR-D')+
                                      TAGS[i].count('TR-RA')+
                                      TAGS[i].count('TR-SD')+
                                      TAGS[i].count('TR-R')+
                                      TAGS[i].count('TR-SR')+
                                      TAGS[i].count('TR-G')+
                                      TAGS[i].count('TR-PP')+
                                      TAGS[i].count('TR-NA'))

VB=[]
```

```
for i in range(0,len(TAGS)):VB.append(TAGS[i].count('VB')+
                                      TAGS[i].count('VB-F')+
                                      TAGS[i].count('VB-I')+
                                      TAGS[i].count('VB-P')+
                                      TAGS[i].count('VB-SP')+
                                      TAGS[i].count('VB-D')+
                                      TAGS[i].count('VB-RA')+
                                      TAGS[i].count('VB-SD')+
                                      TAGS[i].count('VB-R')+
                                      TAGS[i].count('VB-SR')+
                                      TAGS[i].count('VB-G')+
                                      TAGS[i].count('VB-PP')+
                                      TAGS[i].count('VB-AN'))

AG=[]
for i in range(0,len(TAGS)):AG.append(TAGS[i].count('-F')+
                                      TAGS[i].count('-G')+
                                      TAGS[i].count('-P'))

N=[]
for i in range(0,len(TAGS)):N.append(TAGS[i].count('N')+
                                    TAGS[i].count('N-P')+
                                    TAGS[i].count('NPR')+
                                    TAGS[i].count('NPR-P'))

PRO=[]
for i in range(0,len(TAGS)):PRO.append(TAGS[i].count('PRO')+
                                       TAGS[i].count('P+PRO')+
                                       TAGS[i].count('PRO$')+
                                       TAGS[i].count('PRO$-F')+
                                       TAGS[i].count('PRO$-P')+
                                       TAGS[i].count('PRO$-F-P'))

CL=[]
for i in range(0,len(TAGS)):CL.append(TAGS[i].count('CL')+
                                      TAGS[i].count('CL+CL')+
                                      TAGS[i].count('...+CL')+
                                      TAGS[i].count('...+CL+CL')+
                                      TAGS[i].count('SR-R!CL')+
                                      TAGS[i].count('ET-R!CL')+
                                      TAGS[i].count('HV-R!CL')+
                                      TAGS[i].count('TR-R!CL')+
                                      TAGS[i].count('VB-R!CL'))

D=[]
for i in range(0,len(TAGS)):D.append(TAGS[i].count('D')+
                                     TAGS[i].count('D-F')+
                                     TAGS[i].count('D-P')+
                                     TAGS[i].count('D-F-P')+
                                     TAGS[i].count('D-G')+
                                     TAGS[i].count('D-G-P')+
                                     TAGS[i].count('D-UM')+
                                     TAGS[i].count('D-UM-F')+
                                     TAGS[i].count('D-UM-P')+
                                     TAGS[i].count('D-UM-F-P')+
                                     TAGS[i].count('DEM'))

ADJ=[]
for i in range(0,len(TAGS)):ADJ.append(TAGS[i].count('ADJ')+
```

```
                                        TAGS[i].count('ADJ-F')+
                                        TAGS[i].count('ADJ-G')+
                                        TAGS[i].count('ADJ-P')+
                                        TAGS[i].count('ADJ-F-P')+
                                        TAGS[i].count('ADJ-G-P')+
                                        TAGS[i].count('ADJ-R')+
                                        TAGS[i].count('ADJ-R-F')+
                                        TAGS[i].count('ADJ-R-P')+
                                        TAGS[i].count('ADJ-R-F-P')+
                                        TAGS[i].count('ADJ-R-G')+
                                        TAGS[i].count('ADJ-R-G-P')+
                                        TAGS[i].count('ADJ-S')+
                                        TAGS[i].count('ADJ-S-F')+
                                        TAGS[i].count('ADJ-S-P')+
                                        TAGS[i].count('ADJ-S-F-P'))

(...)

L=[ndifpostag,SR,HV,ET,TR,VB,N,PRO,CL,D,ADJ,ADV,Q,CONJ,WPRO,
   P,OUTRO,FP,NUM,NEG,INTJ,NotasF,TemaF,CoerenciaF,RegrasF]

L=list(zip(*L))
H=[]
for i in range(len(L)):H.append(list(L[i]))

df = pd.DataFrame(H,columns=['ndifpostag','SR','HV','ET',
                             'TR','VB','N','PRO','CL','D',
                             'ADJ','ADV','Q','CONJ','WPRO',
                             'P','OUTRO','FP','NUM','NEG', 'INTJ',
                             'NotasF','TemaF','CoerenciaF','RegrasF'])

df.to_csv('sintatica.csv', sep=';', encoding='utf-8')
```

```
#####################  DIMENSÃO CONTEÚDO  ####################
def TfIdf(matriz):
    transformer = TfidfTransformer()
    tfidf = transformer.fit_transform(matriz)
    return tfidf.toarray()

def LSA(d, y, v):
    L = []
    for i in range(0, len(d)):
        vocabulary = [v, d[i]]
        vectorizer = CountVectorizer(min_df=1, ngram_range=(1, 1))
        dtm = vectorizer.fit_transform(vocabulary)
        df=pd.DataFrame(dtm.toarray(),index=vocabulary,
columns=vectorizer.get_feature_names()).head(len(d))
        m = df.values.T.tolist()
        matriz = np.matrix(m)
        matriz=TfIdf(matriz)
        U, Sigma, Vt = linalg.svd(matriz)
        k = 2
        S = np.array(Sigma[0:k])
        Sk = np.diag(S)
        Vtk = Vt[:, 0:k]
        Vtkt = Vtk.transpose()
        Ak = np.dot(Sk, Vtkt)
```

```
        L1 = []
        if y == str('cosseno'):
            L1.append(abs(CosV(Ak[:, 0], Ak[:, 1])))
        elif y == str('distancia'):
            L1.append(euclD(Ak[:, 0], Ak[:, 1]))
        L.append(L1[0])
    return L

(...)

textos_teste = []
textos_treina = []
notas_teste = []
notas_treina = []
passo = 50
for faixa in range(0, 1000, passo):
    print(faixa, faixa + passo)
    textos_teste = TextosS[faixa:faixa + passo]
    notas_teste = NotasS[faixa:faixa + passo]
    textos_treina = TextosS[0:faixa] + TextosS[faixa + passo:]
    notas_treina = NotasS[0:faixa] + NotasS[faixa + passo:]
    modelo()
    grava()
```

```
####################  DIMENSÃO DE COERÊNCIA  ##################
maxi = 5
def getWind(txt):
    b=0;W=[];
    lenght=round(len(txt)/4)
    while b+lenght<=len(txt):
        w1=wind(b,lenght,txt);W.append(w1);b+=maxi
    if len(W)==1: W+=[txt,['0','0']]
    #w1,w2=wind(b,t1);W.append(w1)
    return W

def windows(txt):
    X1 = word_tokenize(txt)
    L = getWind(X1)
    M = []
    for j in range(0, len(L)):
        M.append(" ".join(L[j]))
    return M

def modelo():
    print('modelo...')
    TextosWind = []
    for             i             in             range(len(textos_teste)):
TextosWind.append(windows(textos_teste[i]))
    CC1c = []
    for          i          in          range(0,          len(TextosWind)):
CC1c.append(LSACONTIGUOS(TextosWind[i], 'cosseno'))
    modes=[max,min,media]
    global simc;simc = []
    for i in range(0,len(CC1c)):
        L=[]
        for m in modes:
           L.append(m(CC1c[i]))
```

```
        simc.append(L)
    res = list(zip(*simc))
    global sim_contiguosc;sim_contiguosc=[]
    for i in range(len(res)):
        sim_contiguosc.append([int(v*100) for v in res[i]])
    CC1d = []
    for          i          in          range(0,          len(TextosWind)):
CC1d.append(LSACONTIGUOS(TextosWind[i], 'distancia'))
    modes = [max, min, media]

    global simd;simd = []
    for i in range(0, len(CC1d)):
        L = []
        for m in modes:
            L.append(m(CC1d[i]))
        simd.append(L)
    res = list(zip(*simd))
    global sim_contiguosd;sim_contiguosd = []
    for i in range(len(res)):
        sim_contiguosd.append([int(v) for v in res[i]])

(...)

def grava():
    sim=list(zip(*simc))
    sim.append(Hum)
    L=sim
    H=[]
    for i in range(len(L)):H.append(list(L[i]))
    H=list(zip(*H))
    cols='localcenterc1,localcenterc2,localcenterc3,Hum'.split(',')
    df = pd.DataFrame(H,columns=[*cols])
    filex='local'+str(faixa)+'c.csv'
    df.to_csv(filex)

    sim = list(zip(*simd))
    sim.append(Hum)
    L = sim
    H = []
    for i in range(len(L)): H.append(list(L[i]))
    H = list(zip(*H))
    cols = 'localcenterd1,localcenterd2,localcenterd3,Hum'.split(',')
    df = pd.DataFrame(H, columns=[*cols])
    filex = 'local' + str(faixa) + 'd.csv'
    df.to_csv(filex)

    sim = list(zip(*simcd))
    sim.append(Hum)
    L = sim
    H = []
    for i in range(len(L)): H.append(list(L[i]))
    H = list(zip(*H))
    cols = 'localcentercd1,localcentercd2,localcentercd3,Hum'.split(',')
    df = pd.DataFrame(H, columns=[*cols])
    filex = 'local' + str(faixa) + 'cd.csv'
    df.to_csv(filex)
```

```
###############  MODELO DE PREDIÇÃO E AVALIAÇÃO  ############
```

```
file_pd = pd.read_csv('conteudocdt.csv')
features = pd.get_dummies(file_pd)
labels = features['NotasF']
features = features.drop('NotasF', axis=1)
#features = features.drop('TemaF', axis=1)
#features = features.drop('CoerenciaF', axis=1)
#features = features.drop('RegrasF', axis=1)


feature_list = list(features.columns)
features = np.array(features)
labels = np.array(labels)

# Training and Testing Sets

def nm_norm(x,min_,max_):nm=max_-min_;return [(v-min_)/nm for v in x]

def erro(v1, v2):
    MIN=min(v1+v2)
    MAX=max(v1+v2)
    v1=nm_norm(v1,MIN,MAX)
    v2=nm_norm(v2,MIN,MAX)
    s=sum(map(lambda x: abs(x[0]-x[1]), zip(v1, v2)))/len(v1)
    return 1-s

def medidas(R1, R2):
    R1a=R1[:];R2a=R2[:]
    cm = pycm.ConfusionMatrix(R1a,R2a, digit=5)
    return (round(erro(R1,R2),2),round(cm.Kappa,2),
            round(cohen_kappa_score(R1, R2, weights='linear'), 2),
            round(cohen_kappa_score(R1,R2,weights='quadratic'),
2),round(cm.AC1,2))

kf = RepeatedKFold(n_splits=5, n_repeats=5, random_state=42)

kappa2=[]
for train_index, test_index in kf.split(features):
    #print("Train:", train_index, "Validation:",test_index)
    X_train, X_test = features[train_index], features[test_index]
    y_train, y_test = labels[train_index], labels[test_index]
    #print(y_test); print(y_train); exit(1)
    sc_x = StandardScaler()
    X_train = sc_x.fit_transform(X_train)
    X_test = sc_x.transform(X_test)
    rf_exp = regressor(n_estimators=200, random_state=13)
    rf_exp.fit(X_train, y_train)
    predictions = rf_exp.predict(X_train)
    errors = abs(predictions - y_train)
    predictions = list(map(round, predictions))
    # Make predictions on test data
    predictions = rf_exp.predict(X_test)
    # Performance metrics
    errors = abs(predictions - y_test)
    #print('Average  absolute  error:',  round(fr.np.mean(errors),  2),
'degrees.');exit(1)
    predictions = list(map(round, predictions))
    predictions = list(map(int, predictions))
    print(predictions)
    print(list(y_test));exit(1)
```

```
    mm = medidas(predictions, y_test)
    #print('teste erro & kappa2',medidas(predictions , y_test));exit(1)
    kappa2.append(mm)
    mape = np.mean(100 * (errors / y_test))
    accuracy = 100 - mape

k1, k2, k3, k4, k5 = zip(* kappa2)
print('MEDIA GERAL DOS FOLDS=', np.mean(k1),'\n',np.mean(k2),'\n',
np.mean(k3),'\n', np.mean(k4),'\n',np.mean(k5))

# Get numerical feature importances
importances = list(rf_exp.feature_importances_)
# List of tuples with variable and importance
feature_importances = [(feature, round(importance, 2)) for feature,
                       importance       in       zip(feature_list[0:50],
importances)]
# Sort the feature importances by most important first
feature_importances = sorted(feature_importances, key=lambda x: x[1],
reverse = True)
#Print out the feature and importances
[print('Variable: {:20} Importance: {}'.format(*pair)) for pair in
feature_importances]
```