

UNIVERSIDADE FEDERAL DE ALAGOAS  
INSTITUTO DE COMPUTAÇÃO  
PROGRAMA DE PÓS GRADUAÇÃO EM INFORMÁTICA

JÁRIO JOSÉ DOS SANTOS JÚNIOR

**Modelos e Técnicas para melhorar a qualidade da avaliação automática para atividades  
escritas em Língua Portuguesa Brasileira**

Maceió-AL  
Janeiro de 2017

JÁRIO JOSÉ DOS SANTOS JÚNIOR

**Modelos e Técnicas para melhorar a qualidade da avaliação automática para atividades escritas em Língua Portuguesa Brasileira**

Dissertação apresentada como requisito parcial para obtenção do grau de Mestre pelo Programa de Pós-Graduação em Informática do Instituto de Computação da Universidade Federal de Alagoas.

Orientador: Ig Ibert Bittencourt  
Coorientador: Ranilson Oscar Araújo Paiva

Maceió-AL

Janeiro de 2017

**Catálogo na fonte**  
**Universidade Federal de Alagoas**  
**Biblioteca Central**  
**Divisão de Tratamento Técnico**

Bibliotecária Responsável: Helena Cristina Pimentel do Vale

S237m Santos Júnior, Jário José dos.

Modelos e técnicas para melhorar a qualidade da avaliação automática para atividades escritas em língua brasileira portuguesa / Jário José dos Santos Júnior. – 2018.

74 f. : il.

Orientador: Ig Ibert Bittencourt Santana Pinto.

Coorientador: Ranilson Oscar Arújo Paiva.

Dissertação (mestrado em Informática) – Universidade Federal de Alagoas. Instituto de Computação. Programa de Pós-Graduação em Informática. Maceió, 2017.

Bibliografia: f. 69-74.

1. Tecnologias educacional. 2. Processamento de linguagem natural (PLN). 3. Atividade escrita – Avaliação automática. 4. Atividade subjetiva – Avaliação automática. I. Título.

CDU: 004.43



Membros da Comissão Julgadora da Dissertação de Jário José dos Santos Júnior, intitulada: “Modelos e Técnicas para Avaliação Automática de Atividades Escritas em Língua Portuguesa Brasileira”, apresentada ao Programa de Pós-Graduação em Informática da Universidade Federal de Alagoas em 20 de outubro de 2017, às 13h, na Sala de Reuniões do CEPETEC, do Instituto de Computação da UFAL.

#### COMISSÃO JULGADORA

**Prof. Dr. Ig Ibert Bittencourt Santana Pinto**  
UFAL – Instituto de Computação  
Orientador

**Prof. Dr. Ranilson Oscar Araújo Paiva**  
UFAL – Instituto de Computação  
Coorientador

**Prof. Dr. Patrick Henrique da Silva Brito**  
UFAL – Instituto de Computação  
Examinador

**Prof. Dr. Jorge Artur Peçanha de Miranda Coelho**  
UFAL – Faculdade de Medicina  
Examinador

*Dedico esta Dissertação à minha mãe, meus irmãos que me ajudaram bastante e me incentivaram à escrita. Aos meus orientadores, Ranilson Paiva, por estar fazendo parte do desenvolvimento e aperfeiçoamento deste trabalho desde a época da Monografia, e, em especial ao Ig Ibert, por estar presente em todo o processo de desenvolvimento tanto da escrita quanto implementação do modelo aqui proposto e claro, do seu refinamento na transição Monografia-Dissertação. Aos meus amigos do Núcleo de Excelência em Tecnologia Sociais - NEES, em especial, Sérgio e Esther. E a todos os que contribuíram direta ou indiretamente para a realização deste trabalho.*

## **AGRADECIMENTOS**

Agradeço primeiramente a Deus, por sentir ele sempre comigo.

Agradeço a minha mãe e meus irmãos por estarem sempre me incentivando no desenvolvimento desta abordagem.

Agradeço ao meu orientador, por todos os conselhos, pela paciência e ajuda nesse período.

Aos meus amigos do grupo de pesquisa NEES, em especial, Sérgio e Esther, por terem facilitado o acesso à tecnologia na nuvem.

Aos professores de toda a graduação, bem como no mestrado, pois o conhecimento adquirido tanto em áreas da matemática quanto programação e inteligência artificial tornaram o desenvolvimento dos modelos propostos possíveis.

Ao Núcleo de Excelência em Tecnologias Sociais, por prover de um ambiente com equipamentos e conforto para desenvolvimento da abordagem. Ao CNPQ pelo apoio financeiro para realização deste trabalho de pesquisa.

*“Meus filhos terão computadores, sim, mas antes terão livros. Sem livros, sem leitura, os nossos filhos serão incapazes de escrever, inclusive a sua própria história.”*  
*(Bill Gates)*

## RESUMO

Diante da sobrecarga gerada pela avaliação de atividades escritas em ambientes EAD, diferentes sistemas, como Coursera e edX, vêm adaptando suas abordagens para a avaliação das devidas atividades. Contudo, utilizando abordagens distintas, ambos os sistemas fazem uso dos mais populares métodos de avaliação dessas atividades, avaliação por pares e avaliação automática. Entretanto, cada técnica apresenta suas vantagens e desvantagens, principalmente sob o olhar na aprendizagem do aluno, uma vez que além de propiciar um aumento na aprendizagem, avaliação por pares se destaca por possuir qualidade na avaliação semelhante ao de um especialista. Mas, avaliação automática ganha olhares por se tratar de um método que provê rápido *feedback*, e que, quanto maior a base de treino, a técnica apresenta melhor eficácia, além de ser utilizado com o próprio sistema EAD em diferentes aspectos, como por exemplo, acompanhamento das limitações existentes nos alunos, entre outras palavras, quais são as dúvidas que determinado aluno possui, que são extraídas através de características oriundas da avaliação de atividades subjetivas. Com isso, o presente trabalho propõe um modelo de avaliação com qualidade de atividades subjetivas, com o intuito de diminuir a sobrecarga gerada pela avaliação de atividades subjetivas no professor e auxiliar no processo da aprendizagem dos alunos. O sistema foi calibrado com aproximadamente 5407 avaliações e redações, onde passou por um experimento contendo 60 redações que indicou que as avaliações realizada pela abordagem são semelhantes às avaliações realizadas pelo especialista com 95% de nível de confiança. Além de mostrar um precisão mais elevada em relação à avaliação provida pela abordagem.

**Palavras-chaves:** Processamento de Linguagem Natural, PLN, Avaliação de atividades subjetivas, Sobrecarga do professor, Avaliação de atividades escritas.



## ABSTRACT

Against the workload generated by the evaluation of activities written in online learning, different systems, like Coursera and edX, have adapted their approaches to the evaluation of the appropriate activities. However, using different approaches, both systems do use of the most popular methods of evaluating such activities, peer review, and automatic evaluation. However, each technique presents its advantages and disadvantages, especially under the student learning perspective, since in addition to providing an increase in learning, peer evaluation stands out for having quality in the assessment similar to that of a specialist. But, automatic essay evaluation is a method that provides fast feedback, and how larger the training base is, the technique has better efficacy, in addition to being used with the online learning system in different aspects, such as monitoring the existing limitations in students, among other words, what are the doubts that a student has, which are extracted from patterns derived from the evaluation of subjective activities. Thus, the present work proposes a model to evaluate efficiently subjective activities, reducing the workload generated by the evaluation of subjective activities by the teacher. The automatic essay score system was calibrated with approximately 5407 evaluations and essays, where it underwent an experiment containing 60 essays which indicated that the evaluations carried out by the approach are similar to the expert's evaluations with 95% level of trust. In addition to showing a higher accuracy in relation to the evaluation provided by the approach.

**Keywords:** Automatic essay score. Natural Language Processing. Teacher workload.

## LISTA DE ILUSTRAÇÕES

Figura 1 – Aumento no números de matrículas em ambientes EAD . . . . .	13
Figura 2 – Carga horária dos cursos totalmente a distância oferecidos em 2014 . . . . .	14
Figura 3 – Carga horária dos cursos semipresenciais oferecidos em 2014 . . . . .	14
Figura 4 – Quantidade de alunos por turma em cursos totalmente a distância . . . . .	15
Figura 5 – Quantidade de alunos por turma em cursos semipresenciais . . . . .	15
Figura 6 – Sistemas EAD . . . . .	16
Figura 7 – Árvore de Derivação Sintática . . . . .	24
Figura 8 – Relação matricial por vetores entre palavras por documentos . . . . .	26
Figura 9 – Distância Vetorial . . . . .	27
Figura 10 – Fluxo da extração de Conhecimento por RBC . . . . .	27
Figura 11 – Estrutura de uma Rede Neural simples . . . . .	30
Figura 12 – Rede Neural para operação AND . . . . .	31
Figura 13 – Extração dos Trabalhos Relacionados . . . . .	34
Figura 14 – Processo de avaliação da sentença . . . . .	35
Figura 15 – Busca de conceitos através dos parágrafos . . . . .	36
Figura 16 – Inserção de Texto referência . . . . .	37
Figura 17 – Extração de conceitos com base no Texto referência . . . . .	37
Figura 18 – Resultado da mineração com base nas respostas dos alunos . . . . .	38
Figura 19 – Avaliação de discussão . . . . .	38
Figura 20 – Mapeamento de principais n-grams . . . . .	39
Figura 21 – Processo de Recomendação Pedagógica . . . . .	42
Figura 22 – Processo de Detecção de padrões . . . . .	48
Figura 23 – Kernel do Sistema . . . . .	50
Figura 24 – Satélites do Sistema . . . . .	51
Figura 25 – Casos de uso do Sistema . . . . .	51
Figura 26 – Diagrama de Fluxo: Avaliação e Recomendação funcionando de forma dinâmica	52
Figura 27 – Boxplots comparando as avaliações . . . . .	60
Figura 28 – Histograma comparando as avaliações da competência 1 . . . . .	61
Figura 29 – Histograma comparando as avaliações da competência 2 . . . . .	61
Figura 30 – Histograma comparando as avaliações da competência 3 . . . . .	62
Figura 31 – Histograma comparando as avaliações da competência 4 . . . . .	62
Figura 32 – Histograma comparando as avaliações da competência 5 . . . . .	63
Figura 33 – Dispersão das avaliações dadas pelo Especialista . . . . .	64
Figura 34 – Dispersão das avaliações dadas pelo Sistema . . . . .	65
Figura 35 – Dispersão das avaliações dadas pelo Sistema . . . . .	66

## LISTA DE TABELAS

Tabela 1 – Comparação entre Avaliação automática e Avaliação por pares calibrada de atividades escritas . . . . .	18
Tabela 2 – Quantidade de artigos por base de dados . . . . .	34
Tabela 3 – Comparação entre os sistemas de avaliação de atividades escritas . . . . .	41
Tabela 4 – Definição dos níveis dos Fatores . . . . .	56
Tabela 5 – Definição formal das Hipóteses . . . . .	56
Tabela 6 – Teste de normalidade Shapiro-Wilk . . . . .	63
Tabela 7 – Teste de Mann-Whitney Wilcoxon . . . . .	64
Tabela 8 – Teste de Correlação de Pearson . . . . .	66

## SUMÁRIO

<b>1</b>	<b>INTRODUÇÃO . . . . .</b>	<b>13</b>
<b>1.1</b>	<b>Motivação e contextualização . . . . .</b>	<b>13</b>
<b>1.2</b>	<b>Problemática e Justificativa . . . . .</b>	<b>17</b>
<b>1.3</b>	<b>Objetivos . . . . .</b>	<b>19</b>
<b>1.4</b>	<b>Escopo do Trabalho . . . . .</b>	<b>19</b>
<b>1.5</b>	<b>Contribuições do Trabalho . . . . .</b>	<b>19</b>
<b>1.6</b>	<b>Organização do trabalho . . . . .</b>	<b>19</b>
<b>2</b>	<b>FUNDAMENTAÇÃO TEÓRICA . . . . .</b>	<b>21</b>
<b>2.1</b>	<b>Avaliação de Atividades Escritas - Modelo do Exame Nacional do Ensino Médio . . . . .</b>	<b>21</b>
<b>2.2</b>	<b>POS Tagging - Etiquetagem: Analisadores Sintáticos . . . . .</b>	<b>24</b>
<b>2.3</b>	<b>Dicionário: Analisadores Ortográficos . . . . .</b>	<b>25</b>
<b>2.4</b>	<b>Extração de Termos Semânticos . . . . .</b>	<b>25</b>
<b>2.5</b>	<b>Raciocínio Baseado em Casos . . . . .</b>	<b>27</b>
<b>2.5.1</b>	<b>Referencial Histórico . . . . .</b>	<b>27</b>
<b>2.5.2</b>	<b>Tipos de Conhecimento . . . . .</b>	<b>28</b>
<b>2.6</b>	<b>Validação cruzada . . . . .</b>	<b>28</b>
<b>2.7</b>	<b>Aprendizagem Profunda - Deep Learning . . . . .</b>	<b>29</b>
<b>3</b>	<b>TRABALHOS RELACIONADOS . . . . .</b>	<b>32</b>
<b>3.1</b>	<b>Revisão da literatura . . . . .</b>	<b>32</b>
<b>3.1.1</b>	<b>Protocolo da Revisão Sistemática . . . . .</b>	<b>32</b>
<b>3.1.2</b>	<b>CrITÉRIOS de Inclusão e Exclusão . . . . .</b>	<b>33</b>
<b>3.1.3</b>	<b>Extração e Seleção dos Estudos Primários . . . . .</b>	<b>33</b>
<b>3.2</b>	<b>Sistemas de avaliação automática de atividades subjetivas . . . . .</b>	<b>35</b>
<b>3.2.1</b>	<b>Automarking: Automatic Assessment of Open Questions . . . . .</b>	<b>35</b>
<b>3.2.2</b>	<b>SAGE - Semantic Automated Grader for Essays . . . . .</b>	<b>35</b>
<b>3.2.3</b>	<b>Automatic Chinese Essay Scoring Using Connections between Concepts in Paragraphs . . . . .</b>	<b>36</b>
<b>3.2.4</b>	<b>MineraFórum - Automatic analysis of messages in discussion forums . . . . .</b>	<b>37</b>
<b>3.2.5</b>	<b>Towards identifying unresolved discussions in student online forums . . . . .</b>	<b>38</b>
<b>3.3</b>	<b>Comparação entre os trabalhos relacionados . . . . .</b>	<b>39</b>
<b>4</b>	<b>PROPOSTA . . . . .</b>	<b>42</b>
<b>4.1</b>	<b>Metodologia . . . . .</b>	<b>42</b>
<b>4.2</b>	<b>Detectar práticas . . . . .</b>	<b>43</b>

4.2.1	Técnicas para avaliar a Competência 1 . . . . .	43
4.2.2	Técnicas para avaliar a Competência 2 . . . . .	44
4.2.3	Técnicas para avaliar a Competência 3 . . . . .	45
4.2.4	Técnicas para avaliar a Competência 4 . . . . .	46
4.2.5	Técnicas para avaliar a Competência 5 . . . . .	47
4.3	Descobrir padrões . . . . .	48
4.4	Recomendar . . . . .	48
4.5	Monitorar e Avaliar . . . . .	49
4.6	Arquitetura do Sistema . . . . .	49
<b>5</b>	<b>DESIGN DE EXPERIMENTO . . . . .</b>	<b>53</b>
5.1	Situando o Problema . . . . .	53
5.2	Objetivos da Investigação . . . . .	53
5.3	Questões de Pesquisa e Hipóteses . . . . .	54
5.4	Fatores e Variáveis de Resposta . . . . .	55
5.5	Níveis dos Fatores . . . . .	55
5.6	Definição formal das Hipóteses . . . . .	55
5.7	Unidades Experimentais . . . . .	56
5.8	Plano de execução . . . . .	56
5.9	Coleta dos Dados . . . . .	57
5.10	Execução do Experimento . . . . .	57
5.11	Análise dos Resultados . . . . .	57
5.12	Instrumentação . . . . .	57
5.13	Ameaças à validade . . . . .	58
5.13.1	Ameaças à validade de constructo . . . . .	58
<b>6</b>	<b>RESULTADOS E DISCUSSÃO . . . . .</b>	<b>59</b>
6.1	Análise Descritiva . . . . .	59
6.2	Análise Inferencial . . . . .	62
<b>7</b>	<b>CONCLUSÃO E LIMITAÇÕES . . . . .</b>	<b>69</b>
	<b>REFERÊNCIAS . . . . .</b>	<b>71</b>

## 1 INTRODUÇÃO

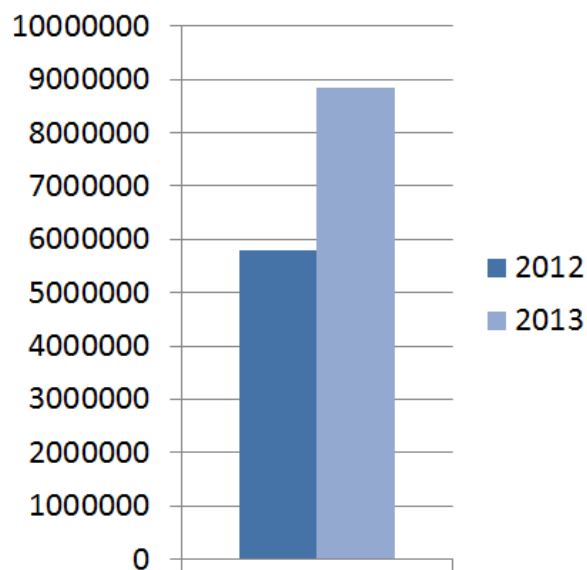
Neste capítulo, introduziremos nosso trabalho. Serão apresentadas a motivação contextualizando o problema a ser abordado, bem como os objetivos a serem alcançados, escopo, contribuições, e organização do presente texto.

### 1.1 Motivação e contextualização

Com o passar dos anos, o ensino tradicional ganha novos olhares: A Educação a Distância (EAD) se torna cada vez mais ativa e presente em ambientes presenciais de ensino-aprendizagem (MORAN, 2008). Como consequência desse fato, vários países adaptam suas abordagens educacionais, promovendo assim, cursos em diversas modalidades: presenciais ou a distância. Na modalidade de ensino a distância, o aluno é um dos principais responsáveis por seu próprio aprendizado (SILVA, 2004). Assim, o uso de ambientes EAD é de fundamental importância para manter o contato aluno-professor fora da instituição física.

A Associação Brasileira de Educação a Distância (ABED<sup>1</sup>) divulga anualmente os resultados obtidos pelo Censo EAD BR<sup>2</sup>. Em 2013, o censo correspondente ao ano de 2012 mostrou que o número de matrículas em ambientes EAD era de 5,8 milhões no Brasil. Em 2014, dados correspondentes ao ano de 2013, a ABED relatou que este número aumentou 52,2%, como mostra a Figura 1.

Figura 1 – Aumento no números de matrículas em ambientes EAD



Fonte – Adaptada do *Censo EAD*

<sup>1</sup> <<http://www.abed.org.br/site/pt/>>

<sup>2</sup> <[http://www.abed.org.br/site/pt/midiateca/censo\\_ead/1193/2013/09/](http://www.abed.org.br/site/pt/midiateca/censo_ead/1193/2013/09/)>

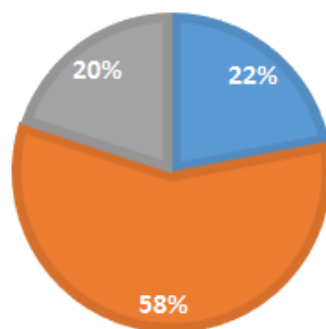
No mesmo censo, a ABED apresenta a quantidade de cursos por nível educacional criada para atender, ao máximo, a demanda de matrículas. O número de cursos regulamentado totalmente a distância oferecidos em 2014 foi de 1840, e o total de cursos semipresenciais registrado foi de 3453.

A carga horária dos cursos ainda é um fator que influencia bastante. Em 2014, a carga horária dos cursos regulamentados totalmente a distância pode ser observada na Figura 2. A figura mostra que aproximadamente 58% dos cursos ofertados possuem mais de 700 horas e que aproximadamente 22% apresentam uma carga horária entre 360 a 659 horas.

Figura 2 – Carga horária dos cursos totalmente a distância oferecidos em 2014

### PORCENTAGEM DA CARGA HORÁRIA

■ Entre 360 a 659 horas ■ Mais de 700 horas ■ Outra



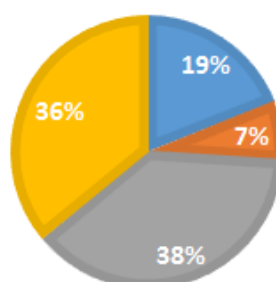
Fonte – Adaptada do *Censo EAD*

Em cursos regulamentados semipresenciais, a carga horária da maioria dos cursos, aproximadamente 38% dos cursos, foi mais de 700 horas, de acordo com a Figura 3.

Figura 3 – Carga horária dos cursos semipresenciais oferecidos em 2014

### PORCENTAGEM DA CARGA HORÁRIA

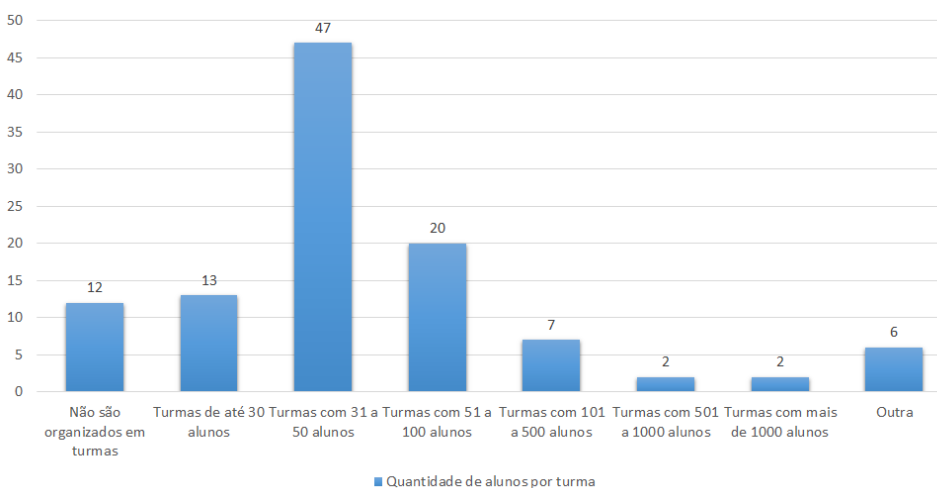
■ Entre 161 e 359 horas ■ Entre 360 e 659 horas  
■ Mais de 700 horas ■ Outra



Fonte – Adaptada do *Censo EAD*

Levando em consideração a quantidade de alunos por turma em cursos totalmente a distância, a ABED, em 2014, registrou que aproximadamente 47 turmas possuem entre 31 e 50 alunos e com 20% das turmas, a quantidade de alunos está entre 51 e 100 alunos por turma. A Figura 4 apresenta a porcentagem das turmas com os respectivos intervalos de alunos.

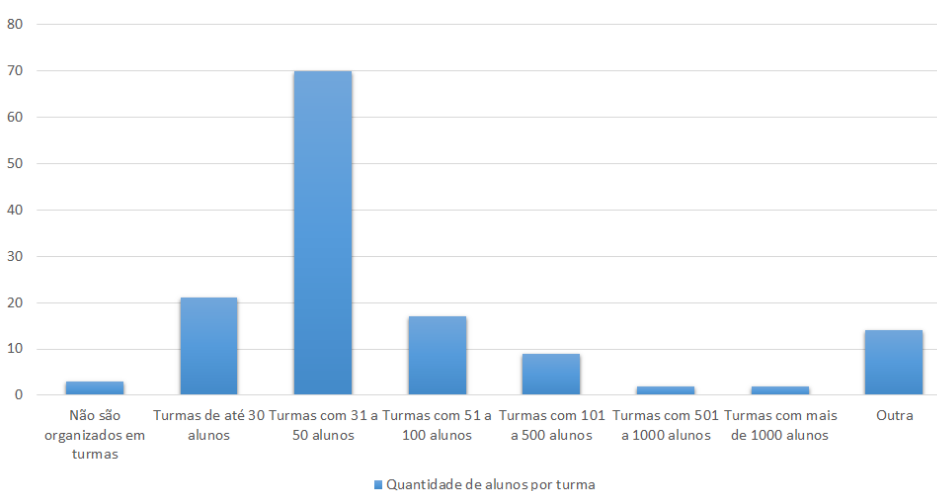
Figura 4 – Quantidade de alunos por turma em cursos totalmente a distância



Fonte – Adaptada do *Censo EAD*

Em cursos semipresenciais de 2014, a quantidade de alunos em aproximadamente em 70 turmas, representando a maioria, estava em um intervalo entre 31 a 50 alunos, e em aproximadamente 21 turmas, a quantidade era até 30 alunos, como pode ser visto na Figura 5.

Figura 5 – Quantidade de alunos por turma em cursos semipresenciais



Fonte – Adaptada do *Censo EAD*

Com o avanço da EAD surgem vários ambientes para o auxílio da aprendizagem (ALMEIDA, 2003). Atualmente, diferentes tipos de tecnologias digitais de informação e comunicação (TDIC) apoiam a modalidade de ensino a distância. Como tecnologias, nós temos os Sistemas



de gerenciamento de aprendizagem (do inglês *Learning Management Systems - LMS*) (SCLATER, 2008), os sistemas tutores inteligentes - STI (SLEEMAN; BROWN, 1982), os sistemas educacionais de hipermídia adaptativa (do inglês *Adaptive Educational Hypermedia Systems*) (BRUSILOVSKY, 1998), os sistemas de aprendizagem colaborativa apoiados por computador (do inglês *Computer-Supported Collaborative Learning - CSCL*) (STAHL; KOSCHMANN; SUTHERS, 2006) e, mais recentemente, os cursos online abertos e massivos (do inglês *Massive Online Open Courses - MOOCs*) (MARTIN, 2012), apresentados na Figura 6, que agregam novos rumos e estratégias de ensino, auxiliando aos alunos em desenvolvimento de provas e atividades. Com a mediação da tecnologia, buscando permitir o aprendizado a partir de qualquer lugar, a qualquer momento e para qualquer pessoa (AAAL<sup>3</sup>), há um processo de mudança no paradigma de aprendizagem, o processo de ensino e aprendizagem antes vinculados às aulas presenciais sofre uma descentralização (BITTENCOURT et al., 2008).

As tecnologias educacionais, entre outras palavras, os sistemas educacionais, possuem limitações em avaliar e dar feedback para trabalhos complexos dos estudantes com avaliações escritas, tais como provas matemáticas, projeto de problemas e redações (PIECH et al., 2013). Atualmente, os ambientes EAD oferecem soluções mais simples diante da dificuldade mencionada, como por exemplo, questões objetivas, variando entre verdadeiro e falso, múltipla escolha ou relacionar colunas, dado que este tipo de questão é facilmente avaliado de forma automática.

Figura 6 – Sistemas EAD



Fonte – Elaborada pelo autor

<sup>3</sup> Do inglês Anywhere, Anytime, Anyone Learning

## 1.2 Problemática e Justificativa

No contexto de atividades subjetivas, como redações, diante de milhões de usuários em ambientes EAD, geraria grande sobrecarga para o professor/tutor (VRASIDAS; MCISAAC, 1999) avaliar tais atividades, pois seria grande a quantidade de dados gerados (FOLEY; KO-BAISSI, 2006). Ainda que os sistemas EAD são limitados para realizar este tipo de avaliação, fica a responsabilidade do professor a correção manual, o que por consequência, quanto mais alunos nestes ambientes, implica em mais atividades escritas para serem corrigidas gerando muita sobrecarga para o professor, sendo uma atividade custosa e que demanda bastante tempo.

Eliminar atividades escritas nesses ambientes não é uma boa solução. Segundo (NAGIN, 2003), a escrita é uma ferramenta essencial para a aprendizagem. Em seu livro, (NAGIN et al., 2012) afirma que diversos fatores socioculturais estão ligados com a escrita, sendo um deles a comunicação e que seus impactos influenciam na aprendizagem das crianças. Em uma edição da editora Abril (ABRIL, 2013)<sup>4</sup>, há ênfase para a importância da escrita, sendo de extrema relevância nas relações sociais, difusão de ideias e informações.

Sabendo que eliminar atividades escritas dos ambientes EAD não é solução, nos deparamos com nosso problema geral:

**(Problema Geral)** Como diminuir a sobrecarga no professor gerada pela avaliação de atividades subjetivas presentes em ambientes EAD?

Diante disto, têm-se o surgimento de diversas soluções para diminuir a sobrecarga gerada pela avaliação das atividades escritas sobre o professor e, segundo (BALFOUR, 2013), dois possíveis caminhos podem ser adotados:

- (i) Avaliação por pares calibrada (ROBINSON, 2001) e (KOLLER; NG, 2012): Avaliação específica onde cada estudante avalia determinadas atividades escritas. Tal solução é usada pelo Coursera<sup>5</sup>.
- (ii) Sistema de Avaliação automática (MARKOFF, 2013): Utiliza algoritmos baseados em técnicas de Processamento de Linguagem Natural, extração de informação e recuperação da informação (JACKSON; MOULINIER, 2007). Atualmente, o EdX<sup>6</sup> e MIT<sup>7</sup> fazem uso desta tecnologia.

Avaliação por pares calibrada e avaliação automática de atividades escritas possuem seus pontos positivos e suas limitações, como pode ser observado na Tabela 1. Visando uma

<sup>4</sup> <<http://educarparacrescer.abril.com.br/comportamento/importancia-escrita-559518.shtml>> : Acessado pela última vez em agosto de 2015

<sup>5</sup> <<https://www.coursera.org/>>

<sup>6</sup> <<https://www.edx.org/>>

<sup>7</sup> <<http://web.mit.edu/>>

comparação explícita entre ambas as técnicas, pode ser destacada a qualidade obtida em ambas as avaliações. Além de outros pontos que divergem ambas os métodos para avaliação, segundo (TENÓRIO et al., 2016), a avaliação por pares calibrada possui qualidade que se assemelha a qualidade de um especialista, em nosso caso, um professor, diferentemente da qualidade da avaliação automática, que ainda é uma incógnita. Contudo, segundo (BALFOUR, 2013), a avaliação automática além de ser mais viável comercialmente e prover feedback para o aluno quase que imediato. Com essa forma de avaliação também é possível extração de características de cada aluno, identificando quais são suas deficiências e limitações.

Tabela 1 – Comparação entre Avaliação automática e Avaliação por pares calibrada de atividades escritas

Fatores	Avaliação automática - PLN	Avaliação por pares calibrada
<i>Avaliação dos textos</i>	Redações niveladas ou tópicos Focado em redações Redação estruturada é melhor avaliada Avalia de forma literal que figurativa	Tópicos simples Redações curtas Não necessita de muita estruturação para avaliação da redação Pode ser usada para alguns textos figurativos
<i>Consistência da avaliação</i>	Elevada Consistência	3 avaliações comentadas divergem Qualidade das avaliações é determinada parcialmente pela calibragem das notas
<i>Comentários providos</i>	Importantes elementos como organização, estilo e criatividade Baseado em uma análise estatística ou por pesquisa semântica Pode haver perda de elementos sutis	Pode ser avaliada seguindo uma sequência dos fatos Comentários providos por humanos Depende da habilidade e da capacidade do revisor
<i>Intervenção de um especialista</i>	Requer uma base de treino com +100	A avaliação se torna inviável com milhares de redações

Fonte – Adaptado de (BALFOUR, 2013)

Com isso, tem-se o problema específico:

**(Problema Específico)** Como melhorar a qualidade da avaliação automática para atividades subjetivas escritas em Língua Portuguesa Brasileira?

Para tanto, é apresentado como questão de pesquisa (**QP**), o seguinte questionamento:

**QP:** De que maneira a avaliação automática apresenta a mesma qualidade, ou melhor, apresenta avaliações semelhantes se comparada às correções realizadas por um especialista?

### 1.3 Objetivos

O objetivo geral deste trabalho é **melhorar a qualidade da avaliação automática de atividades subjetivas em língua portuguesa brasileira**. Desta maneira, a partir das questões de pesquisas apresentadas na seção 1.2, foram propostos os seguintes objetivos específicos:

- (i) Realizar uma revisão sistemática a fim de comparar as técnicas que vêm sendo utilizadas pela comunidade científica que obtiveram os melhores resultados em avaliação de atividades subjetivas;
- (ii) Definir e implementar estratégias para avaliar as atividades subjetivas:
  - Criar um modelo conceitual com base na revisão sistemática realizada;
  - Consolidar o modelo conceitual em um modelo arquitetural;
- (iii) Avaliar empiricamente o impacto do modelo de avaliação de atividades subjetivas;

### 1.4 Escopo do Trabalho

O escopo desta pesquisa visa alcançar os objetivos definidos na seção 1.3 com avaliações de atividades subjetivas através de textos compreendidos por computador, nesse contexto, redações e atividades desenvolvidas em Fóruns educacionais online. Entretanto, afirmamos estar **fora do escopo desta pesquisa** abordar as seguintes questões:

1. Considerar avaliações subjetivas de atividades manuscritas;
2. Considerar avaliações subjetivas em geral (chat, artigos, entre outros);

### 1.5 Contribuições do Trabalho

O presente trabalho pretende contribuir para as áreas de Informática na Educação, com a redução da sobrecarga por parte do professor através de uma avaliação com qualidade de atividades subjetivas, mediante o escopo apresentado na seção 1.4. Contribuir, também, para a área de Inteligência Artificial, mas especificamente, na subárea de Processamento de Linguagem Natural.

### 1.6 Organização do trabalho

Essa proposta de dissertação encontra-se organizada da seguinte maneira: no Capítulo 2 apresentamos conceitos, técnicas e abordagens importantes para uma inicial compreensão do que está sendo abordado nesse trabalho. No Capítulo 3, apresentamos os trabalhos relacionados levantados frutos de uma Revisão Sistemática, listando suas técnicas e limitações, e por fim

uma tabela comparativa para o modelo de avaliação de atividades escritas. A proposta da dissertação é apresentada no Capítulo 4, onde poderá ser encontrada a Metodologia aplicada, todo o processo de avaliação de atividades subjetivas, mencionando métricas e técnicas. No Capítulo 5, é encontrado o Design de Experimento que tem como objetivo avaliar o modelo que está sendo proposto neste trabalho. No Capítulo 6, estão disponíveis os resultados e discussões levantadas acerca do uso da abordagem aqui desenvolvida, e, por fim, no Capítulo 7, está a conclusão do trabalho, listando as principais contribuições do trabalho, bem como suas limitações para futuras melhorias.

## 2 FUNDAMENTAÇÃO TEÓRICA

Neste capítulo serão abordados os principais tópicos contendo os fundamentos teóricos dessa dissertação, necessários para uma apropriada compreensão da mesma. Tais tópicos servirão como uma base teórica para a análise e compreensão dos elementos do modelo apresentado, bem como o entendimento completo de sua implementação em um ambiente e processo de sua validação.

### 2.1 Avaliação de Atividades Escritas - Modelo do Exame Nacional do Ensino Médio

O Exame Nacional do Ensino Médio - ENEM<sup>1</sup> é uma avaliação composta de questões de múltipla escolha e por uma questão dissertativa-argumentativa realizada nacionalmente pelo Ministério da Educação - MEC (CASTRO; TIEZZI, 2005) para ingresso dos estudantes em redes de ensino superior. Entretanto, cada parte do ENEM (questões de múltipla escolha e a questão dissertativa) é avaliada com diferentes critérios. Para avaliação da parte composta por questões de múltipla escolhas, o aluno recebe uma nota proporcional a quantidade de acertos seguidas de um peso de acordo com a dificuldade da questão correspondente. Em contrapartida, a avaliação da questão dissertativa ou redação acontece com o auxílio de um modelo de avaliação que está relacionado as três dimensões do discurso da linguística (FONSECA, 1993): (i) Sintaxe<sup>2</sup>, (ii) Semântica<sup>3</sup> e (iii) Pragmática<sup>4</sup>.

O Modelo de avaliação é composto por 5 competências gerais que são divididas em 6 níveis de desempenho com a sua determinada pontuação. São elas:

- (i) **Demonstrar domínio da modalidade escrita formal da Língua Portuguesa:** O Aluno deverá ter consciência da diferença entre a escrita formal e informal, evitando o emprego repetido de palavras ao relacionar ideias (*e, aí, daí, então*), e obedecer às regras de concordância nominal e verbal, regência nominal e verbal, pontuação, flexão nominal e verbal, colocação pronominal, grafia das palavras e divisão silábica:

**0 ponto** Não demonstra conhecimento da modalidade escrita formal da Língua Portuguesa;

**40 pontos** Demonstra domínio precário da modalidade de escrita formal da Língua Portuguesa, de forma sistemática, com diversos desvios gramaticais, de escolha de registro e de convenções da escrita;

**80 pontos** Demonstra domínio insuficiente da modalidade de escrita formal da Língua Portuguesa, com muitos desvios gramaticais e convenções da escrita;

<sup>1</sup> <<http://portal.inep.gov.br/enem>>

<sup>2</sup> Estudo da combinação entre as palavras de uma determinada linguagem.

<sup>3</sup> Estudo da interpretação entre as palavras de uma determinada linguagem.

<sup>4</sup> Estudo das relações entre as palavras, frases ou períodos de uma linguagem.

- 120 pontos** Demonstra domínio mediano da modalidade escrita formal da Língua Portuguesa, com alguns desvios gramaticais e de convenções da escrita;
- 160 pontos** Demonstra bom domínio da modalidade escrita formal da Língua Portuguesa, com poucos desvios gramaticais e de convenções da escrita;
- 200 pontos** Demonstra excelente domínio da modalidade escrita formal da Língua Portuguesa. Os desvios gramaticais serão aceitos somente como excepcionalidade e quando não caracterizem reincidências.

(ii) **Compreender a proposta de redação e aplicar conceitos das várias áreas de conhecimento para desenvolver o tema, dentro dos limites estruturais do texto dissertativo-argumentativo em prosa:** Etapa responsável em avaliar a compreensão da proposta de redação. Exige que o texto escrito pelo aluno seja dissertativo-argumentativo<sup>5</sup>. É preciso apresentar um texto que exponha um aspecto relacionado ao tema, defendendo uma posição, uma tese<sup>6</sup>, com recursos argumentativos, de modo a convencer o leitor, como exemplos, dados estatísticos, pesquisas, fatos comprováveis, citações ou depoimentos de pessoas especializadas no assunto, alusões históricas e comparações entre fatos, situações e épocas:

- 0 ponto** Fuga ao tema ou não atendimento à estrutura dissertativo-argumentativa;
- 40 pontos** Apresenta o assunto, tangenciando o tema, ou demonstra domínio precário do texto dissertativo-argumentativo, com traços constantes de outros tipos textuais;
- 80 pontos** Desenvolve o tema recorrendo à cópia de trechos dos textos motivadores ou apresenta domínio insuficiente do texto dissertativo-argumentativo, não atendendo à estrutura com proposição, argumentação e conclusão;
- 120 pontos** Desenvolve o tema por meio de uma argumentação previsível e apresenta domínio mediano do texto dissertativo-argumentativo, com proposição, argumentação e conclusão;
- 160 pontos** Desenvolve o tema por meio de uma argumentação consistente e apresenta bom domínio do texto dissertativo-argumentativo, com proposição, argumentação e conclusão;
- 200 pontos** Desenvolve o tema por meio de uma argumentação consistente, a partir de um repertório sociocultural produtivo, e apresenta excelente domínio do texto dissertativo-argumentativo.

(iii) **Selecionar, relacionar, organizar e interpretar informações, fatos, opiniões e argumentos em defesa de um ponto de vista:** Nesta etapa, o aspecto avaliado no texto é a relação, organização e interpretação das informações, fatos, opiniões e argumentos acerca da proposta:

<sup>5</sup> Tipo de texto que demonstra a verdade de uma ideia ou tese, segundo MEC.

<sup>6</sup> Ideia que será defendida no texto. Deve estar relacionada ao tema e apoiada em argumentos ao longo da redação

- 0 ponto** Apresenta informações, fatos e opiniões não relacionados ao tema e sem defesa de um ponto de vista;
- 40 pontos** Apresenta informações, fatos e opiniões pouco relacionados ao tema ou incoerentes e sem defesa de um ponto de vista;
- 80 pontos** Apresenta informações, fatos e opiniões relacionados ao tema, mas desorganizados ou contraditórios e limitados aos argumentos dos textos motivadores, em defesa de um ponto de vista;
- 120 pontos** Apresenta informações, fatos e opiniões relacionados ao tema, limitados aos argumentos dos textos motivadores e pouco organizados, em defesa de um ponto de vista;
- 160 pontos** Apresenta informações, fatos e opiniões relacionados ao tema, de forma organizada, com indícios de autoria, em defesa de um ponto de vista;
- 200 pontos** Apresenta informações, fatos e opiniões relacionados ao tema proposto, de forma consistente e organizada, configurando autoria, em defesa de um ponto de vista.

(iv) **Demonstrar conhecimento dos mecanismos linguísticos necessários para a construção da argumentação:** Esta competência tem como objetivo avaliar aspectos ligados à estruturação lógica e formal entre as partes da redação. Nesta etapa, a organização textual exige que as frases e os parágrafos estabeleçam entre si uma relação que garanta a sequenciação coerente do texto:

- 0 ponto** Ausência de marcas de articulação, resultando em fragmentação das ideias;
- 40 pontos** Articula as partes do texto de forma precária;
- 80 pontos** Articula as partes do texto, de forma insuficiente, com muitas inadequações e apresenta repertório limitado de recursos coesivos;
- 120 pontos** Articula as partes do texto, de forma mediana, com inadequações e apresenta repertório pouco diversificado de recursos coesivos;
- 160 pontos** Articula as partes do texto com poucas inadequações, e apresenta repertório diversificado de recursos coesivos;
- 200 pontos** Articula bem as partes do texto e apresenta repertório diversificado de recursos coesivos.

(v) **Elaborar proposta de intervenção para o problema abordado, respeitando os direitos humanos:** Nesta etapa, o aspecto a ser avaliação na redação é a apresentação de uma proposta de intervenção social ao problema abordado, apoiada em argumentos consistentes:

- 0 ponto** Não apresenta proposta de intervenção ou apresenta proposta de intervenção não relacionada ao tema ou ao assunto;



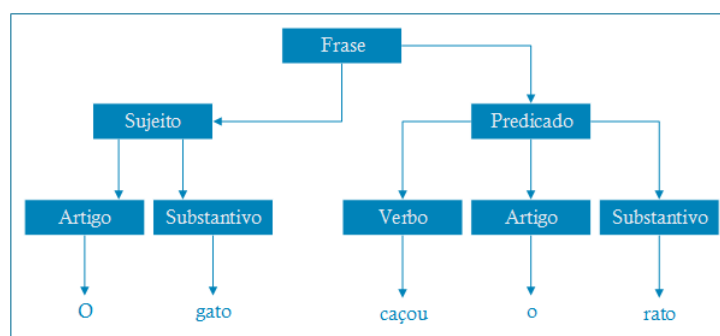
- 40 pontos** Apresenta proposta de intervenção vaga, precária, ou relacionada apenas ao assunto;
- 80 pontos** Elaborar, de forma insuficiente, proposta de intervenção relacionada ao tema ou não articulada com a discussão desenvolvida no texto;
- 120 pontos** Elaborar, de forma mediana, proposta de intervenção relacionada ao tema e articulada à discussão desenvolvida no texto;
- 160 pontos** Elaborar bem a proposta de intervenção relacionada ao tema e articulada à discussão desenvolvida no texto;
- 200 pontos** Elaborar muito bem a proposta de intervenção, detalhada, relacionada ao tema e articulada à discussão desenvolvida no texto.

Como pode ser observado, a nota final varia entre 0-1000. Contudo, há alguns fatores que influenciam exclusivamente na nota *zero*. São eles: (i) a redação que não obedece à estrutura dissertativo-argumentativa e (ii) a redação que não respeite os direitos humanos.

## 2.2 POS Tagging - Etiquetagem: Analisadores Sintáticos

A POS Tagging ou etiquetagem é a marcação de uma classe gramatical (MARTIN; JURAFSKY, 2000). Ela é responsável pela identificação dos itens lexicais e da classe que determinada palavra, em uma sentença, possui, gerando uma árvore de derivação sintática, como mostra a Figura 7. A principal função desempenhada pela etiquetagem é a resolução das ambiguidades.

Figura 7 – Árvore de Derivação Sintática



Fonte – Elaborada pelo autor

Os etiquetadores podem ser construídos (i) baseados em regras (VOUTILAINEN, 1995), restrições (CHANOD; TAPANAINEN, 1995), casos (DAELEMANS et al., 1996) e em árvores de decisões não probabilísticas, denominados simbólicos, (ii) baseados em modelos probabilísticos (SCHMID, 2013) podendo utilizar de redes neurais (MA et al., 1999), máxima entropia (REYNAR; RATNAPARKHI, 1997), Modelo de Markov (WILKENS; KUPIEC, 1995) e árvores

de decisão probabilísticas (SCHMID, 2013) para calcular qual a probabilidade de determinada palavra receber a etiqueta com seu respectivo contexto, denominado estocásticos, ou (iii) híbridos, que são baseados na combinação dos dois modelos apresentados anteriormente, ou seja, o processo de etiquetagem deste modelo emprega tanto os modelos baseados em regras quanto os modelos estocásticos. Um modelo bastante conhecido na literatura que utiliza esse tipo de abordagem é o etiquetador TBL (BRILL, 1994), (BRILL, 1995a), (BRILL, 1995b). Tal etiquetador foi construído com intuito de maximizar vantagens e minimizar as desvantagens.

Denomina-se analisadores sintáticos (em inglês, parsers) como os sistemas que realizam a análise estrutural e seus constituintes. Os parsers são responsáveis pelo reconhecimento de estruturas válidas a partir de um termo léxico cuja responsabilidade é definir o vocabulário e o conjunto de regras, compondo assim, a gramática da língua (YOUNGER, 1967). O analisador léxico é usado pelo sintático a fim de reunir os itens lexicais da língua e de sua respectiva gramática (AHO; ULLMAN, 1972).

### 2.3 Dicionário: Analisadores Ortográficos

Utilização de dicionários para verificar a grafia correta de uma determinada palavra é uma estratégia que reduz bastante o processamento, uma vez que se a palavra estiver presente na base de dados, não será possível realizar substituições por novas. Contudo, o tamanho do dicionário pode ser imenso, o que traz a necessidade de haver regras para analisar as palavras por afixos, recuperando o radical da palavra através de algoritmos genéticos (JOSÉ; PAIVA; BITTENCOURT, 2015), por exemplo.

Analisadores ortográficos utilizam regras para extrair sufixos e afixos de palavras, tendo apenas que possuir o radical da palavra em uma base de dado, economizando espaço em disco.

### 2.4 Extração de Termos Semânticos


Modelo estatístico cujo principal objetivo é identificar termos a partir de parâmetros observáveis (RAMOS, 2003). Esses termos podem ser combinados originando termos compostos denominados de *N-Gramas*. N-gramas é uma técnica responsável em criar sequências de palavras a partir de uma sentença (BROWN et al., 1992) utilizada principalmente na análise conceitual entre partes (palavras, frases, parágrafos) do texto.

A análise conceitual das relações entre as partes de um texto pode ser realizada através do uso de frequências dos termos no texto (CARROLL; DAVIES; RICHMAN, 1971). A frequência de termos em um texto vem usada desde a década de 80 com diferentes abordagens, uma delas é detecção de memorização de palavras que não são comuns (GREGG, 1976) e outra seria a detecção de ambiguidades presentes nos idiomas (RAYNER; DUFFY, 1986).

Entretanto, o uso de frequência de palavras presentes em um texto teve seu ápice com o surgimento da Recuperação de Informação (MANNING, 1995) utilizando modelos de vetores

como identificação semântica entre termos. Os documentos passaram a ser tratados como um conjunto de vetores, e estes vetores, compostos por palavras, como pode ser visto na Figura 8.

Figura 8 – Relação matricial por vetores entre palavras por documentos



w/d	d1	d2	d2	d3	d4	d5
w1	1	0	2	1	0	1
w2	1	0	1	4	1	1
w3	1	3	1	0	1	0

Fonte – Elaborada pelo autor

Sabendo que os documentos já poderiam ser expressos como coordenadas em um plano cartesiano, calcular a relação semântica entre esses textos seria o mesmo que calcular a distância entre eles (GUPTA; JAIN, 1997). Diferentes abordagens, como Distância Euclidiana (CANCHO, 2004) ou Cálculo do Cosseno (ZHANG; CALLAN; MINKA, 2002), apresentadas da Equação 2.1 e Equação 2.2 respectivamente, poderiam então ser usadas para o cálculo dessa distância. A Figura 9 mostra graficamente a distância existente entre dois documentos representados por vetores através do cosseno.

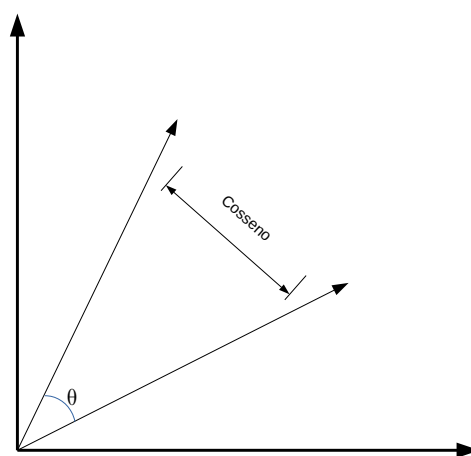
$$\sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \dots + (p_n - q_n)^2} = \sqrt{\sum_{i=1}^n (p_i - q_i)^2} \quad (2.1)$$

$$\cos \theta = \frac{\langle v, u \rangle}{|v| \cdot |u|} \quad (2.2)$$

Em contrapartida, para comparar documentos, houve a necessidade de ponderação de alguns termos, como por exemplo artigos, preposições ou qualquer outra classe gramatical que não seja significativamente semântica. Além disso, documentos longos teriam vantagens absurdas sobre documentos pequenos (CALLAN, 1994). Com efeito, novas abordagens relacionadas à remoção de tais classes gramaticais e ao ponderamento de palavras que possuem altas frequências, nos documentos, foram criadas: (i) eliminação de *Stopwords* (WILBUR; SIROTKIN, 1992) e (ii) ponderação *TF-IDF*<sup>7</sup> (SALTON; BUCKLEY, 1988), respectivamente. Diante das técnicas criadas para equilibrar comparações entre documentos longos e curtos, Distância Euclidiana e Distância através do cosseno, para o cálculo da relação semântica expressa pela distância entre os vetores do documentos, teriam suas vantagens e desvantagens (STREHL; GHOSH; MOONEY, 2000).

<sup>7</sup> do Inglês: term frequency–inverse document frequency

Figura 9 – Distância Vetorial

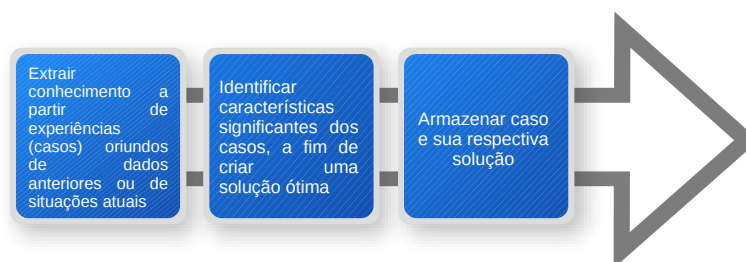


Fonte – Elaborada pelo autor

## 2.5 Raciocínio Baseado em Casos

Raciocínio baseado em casos (**RBC**) é uma técnica para extração de conhecimento com base em experiências obtidas anteriormente (WANGENHEIM; WANGENHEIM, 2003). Através de *RBC* é possível (i) extrair conhecimento, (ii) verificar significância dos casos propondo uma solução e, por fim, (iii) armazenar o caso e sua respectiva solução. A Figura 10 apresenta o fluxo gerado por essas três fases.

Figura 10 – Fluxo da extração de Conhecimento por RBC



Fonte – Elaborada pelo autor

### 2.5.1 Referencial Histórico

Por volta da década de 80, os Sistemas Especialistas Baseados em Regras<sup>8</sup> foram destaque na Inteligência Artificial (HAYES-ROTH, 1985). Tais sistemas eram aplicados em resoluções

<sup>8</sup> do Inglês: Rule-Based Expert Systems (RBES)

de problemas onde havia a necessidade de conhecimento específico, problemas de hardware, exploração geográfica e até mesmo na medicina (STANFILL; WALTZ, 1986). Entretanto, apesar do sucesso com o uso dos Sistemas Especialistas Baseados em Regras, de acordo com Schank (1987), estes sistemas possuem diversos problemas: (i) a construção da Base de conhecimento exige conhecimento bastante técnico oriundo de Especialistas, o que dificulta a construção; (ii) incapacidade de lidar com problemas que não estão explicitamente presentes nas regras; e (iii) necessidade de explicitar, formalizar e estabelecer bem as regras.

Diante dessa dificuldade, uma nova forma de abordar tais problemas foi criada: *Raciocínio Baseado em Casos*. Com esse tipo de paradigma, os sistemas não necessitavam mais de um modelo de domínio explícito, e a base de conhecimento gerada por meio de experiências obtidas anteriormente (casos), pois identificar características significativas é mais fácil do que criar um modelo de domínio explícito (AAMODT; PLAZA, 1994). Por fim, tal abordagem poderia aprender novos conhecimentos que seriam obtidos com casos futuros (SCHANK, 1983).

### 2.5.2 Tipos de Conhecimento

Para construção de um sistema RBC, é necessário conhecimento sob quatro aspectos que estão relacionados aos dados (RICHTER, 1995):

- (i) **Vocabulário:** Inclui o conhecimento necessário para a escolha das características utilizadas para descrever os casos. Entretanto, existe um *trade-off* para a escolha das características. Estas precisam levar em consideração características que sejam úteis para resoluções de problemas semelhantes, ao mesmo tempo, evitar escolher características que sejam utilizados em casos muito diferentes, o que poderia gerar uma falsa solução para o problema;
- (ii) **Medidas de Similaridade:** Inclui o conhecimento adequado para a escolha da medida de similaridade sob o aspecto de prover a organização mais eficiente da base e o método de recuperação do caso mais adequado;
- (iii) **Conhecimento adaptativo:** Inclui o conhecimento sob o processo de implementação de adaptação e estágios de avaliação no sistema RBC. É nessa etapa que deve ser levado em consideração sobre como as diferenças entre os problemas afetam nas soluções;
- (iv) **Casos:** Conhecimento adquirido com base em problemas e soluções desenvolvidas anteriormente.

### 2.6 Validação cruzada

A técnica de validação cruzada consiste em avaliar a capacidade de generalização de um modelo dado um determinado conjunto de dados (GOLUB; HEATH; WAHBA, 1979). Em sistemas onde o objetivo é predição, a validação cruzada estima o quão preciso o modelo é.

Entre outras palavras, expressar o desempenho para qualquer conjunto de dados (KOHAVI et al., 1995). O conceito central da validação cruzada é o particionamento dos dados em subconjuntos, e, após tal particionamento, ter subgrupos que possam ser usados para formação de Base de conhecimento, Base de treino ou Base para testes.

Existem diversas maneiras de dividir os dados, contudo, segundo Kohavi et al. (1995), três dessas formas são as mais utilizadas: (i) *holdout* (KIM, 2009). (ii) *k-fold* (BENGIO; GRAND-VALET, 2004), e (iii) *leave-one-out* (KEARNS; RON, 1999).

A validação cruzada *holdout* consiste em dividir o conjunto total dos dados em dois subconjuntos: (i) um destinado para formação da base de conhecimento<sup>9</sup>, e (ii) outro para validação<sup>10</sup>. Normalmente a divisão dos dados é executada entre  $\frac{2}{3}$  para a base de conhecimento e  $\frac{1}{3}$  para o conjunto de teste.

O método de validação cruzada *k-fold* consiste na divisão em  $k$  subconjuntos do mesmo tamanho, e um determinado grupo  $k$  é usado para teste, enquanto os  $k-1$  grupos são utilizados para o calibragem do sistema e cálculo da acurácia.

A validação cruzada por *leave-one-out* é um caso particular da validação *k-fold*, tendo como valor de  $k$  o tamanho total da amostra  $N$ . Este modelo apresenta uma validação completa sobre os dados, entretanto, o modelo apresenta um alto custo computacional, sendo viável em amostras pequenas.

A Equação 2.3 apresenta a função responsável por calcular o erro final de um modelo, onde, por meio dessa função é possível calcular, de forma quantitativa, a generalização do modelo. Tal função é usada nos tipos de validação apresentados anteriormente.

$$Err_f = \frac{1}{v} \sum_{i=1}^v \varepsilon_{y_i, \hat{y}_i} = \frac{1}{v} \sum_{i=1}^v (y_i - \hat{y}_i) \quad (2.3)$$

onde:

$v$  é a quantidade da amostra utilizada para o conjunto de validação

$y_i$  é o valor real esperado

$\hat{y}_i$  é o valor predito

$\varepsilon_{y_i, \hat{y}_i}$  é o resíduo existente entre o valor real e o valor predito

## 2.7 Aprendizagem Profunda - Deep Learning

Aprendizagem profunda, também conhecida como *Deep Learning*<sup>11</sup>, é uma subárea de Aprendizagem de Máquina que investiga técnicas para modelar abstrações em relação ao com-

<sup>9</sup> Conjunto de treinamento

<sup>10</sup> Conjunto de teste

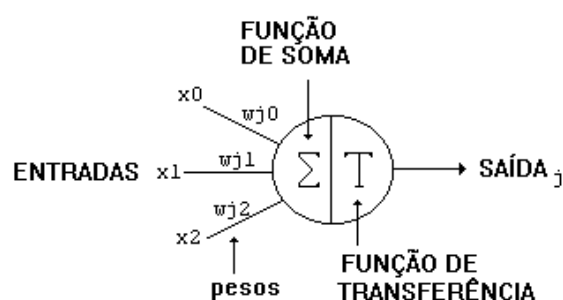
<sup>11</sup> Variação da palavra no Idioma Inglês

portamento cerebral humano em tarefas como reconhecimento e processamento de linguagem natural<sup>12</sup>(SPYNS, 1996), por exemplo.

Segundo Bishop (2007), para modelar abstrações de alto nível de dados, é necessário ter o conhecimento de características mensuráveis relacionadas ao contexto. Algoritmos de aprendizagem de máquina utilizam conjunto de dados para treinamento construído a partir de um processo de engenharia de características, entre outras palavras, extração de *features* significativas ao contexto. Tal processo pode ser feito apenas pela máquina através de técnicas de aprendizagem por representação (BENGIO; COURVILLE; VINCENT, 2013) ou mediados por um especialista de domínio.

Logo, sabendo que a modelar abstrações de alto nível de dados necessita-se conhecimento de características mensuráveis, as Redes Neurais (FUNAHASHI, 1989), técnicas computacionais que apresentam um modelo matemático inspirado na estrutura neural de organismos inteligentes, são capazes de realizar reconhecimento de padrões e auxiliar no processo de aprendizagem por máquina. Uma Rede Neural Artificial é composta por várias unidades de processamento denominadas neurônios. Os neurônios (MCCULLOCH; PITTS, 1943) são conectados por canais de comunicação que estão relacionados as valores reais denominados pesos específicos. A Figura 11 apresenta um modelo de rede neural simples, composto por uma função de ativação, no caso, soma, suas entradas e seus respectivos pesos e a função de transferência. A Rede neural mais simples possui apenas uma camada composto por apenas um neurônio.

Figura 11 – Estrutura de uma Rede Neural simples



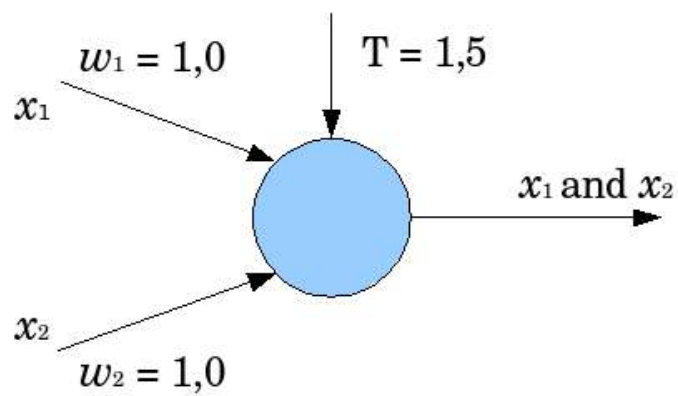
Fonte – Elaborada por McCulloch e Pitts (1943)

As redes neurais podem variar de acordo com a quantidade de camadas que estas possuem. A rede neural composta por apenas um neurônio é bastante limitada, usada basicamente em atividades onde o problema é binário, por exemplo, comutação de circuitos com funções booleanas. A Figura 12 apresenta uma estrutura que visa executar a operação binária AND. O funcionamento desta rede neural é bastante simples: (i) entradas variando entre 0 e 1; (ii) função de ativação sendo a soma entre esses valores; (iii) função de transferência verifica se o resultado da soma é menor que 1.5; (iv) retorna 1 caso verdadeiro ou 0 caso falso. Note também que para

<sup>12</sup> Subárea da inteligência artificial e da linguística que estuda os problemas da geração e compreensão automática de línguas humanas naturais mediadas pela tecnologia

mudar a operação binária para OR, basta apenas alterar o limite<sup>13</sup> da função de transferência para 0.5.

Figura 12 – Rede Neural para operação AND



Fonte – Elaborada pelo autor

---

<sup>13</sup> do Inglês *threshold*



### 3 TRABALHOS RELACIONADOS

Neste capítulo detalharemos alguns trabalhos que consideramos relacionados às ideias apresentadas nessa proposta de dissertação. Pesquisamos por aqueles que abordassem sistemas/ferramentas para avaliação de atividades subjetivas. Fomos capazes de encontrar trabalhos de grande significância a fim de realizar uma análise comparativa com o que está sendo proposto. No final da seção 3.2, é apresentada a Tabela 3 para sumarizar as contribuições e limitações de cada trabalho.

#### 3.1 Revisão da literatura

Nesta seção, detalharemos o processo de extração para os trabalhos relacionados aqui listados. Importante ressaltar que apenas os trabalhos relacionados à avaliação de atividades subjetivas foram considerados. Para tanto para a extração desses trabalhos, foi realizada uma Revisão sistemática (BUCHWALD et al., 2004) para levantamento de estudos primários.

##### 3.1.1 Protocolo da Revisão Sistemática

Com o intuito de levantar o estado da arte de Processamento de Linguagem Natural para avaliação de atividades escritas, temos a seguinte questão: *Como está sendo a avaliação de atividades subjetivas com técnicas de Processamento de Linguagem Natural?*. Portanto, nosso objetivo principal para realização desta Revisão Sistemática é identificar a efetividade do uso de processamento de linguagem natural para a avaliação de atividades subjetivas que estão presentes em *Forúns, redações e questões dissertativas/argumentativas, Wikis* e, por fim, *Blogs*.

Para tanto, temos as seguintes questões que procuramos responder no final desta revisão sistemática:

- Qual a efetividade do uso de técnicas de processamento de linguagem natural para avaliação de atividades subjetivas?
- Quais são os desafios em usar processamento de linguagem natural visando avaliar a ortografia visando sintaxe, semântica e pragmática das atividades subjetivas?
- Quais são os critérios de qualidade que podem ser usados para avaliar atividades subjetivas utilizando processamento de linguagem natural?
- Quais técnicas são usadas para avaliar a sintaxe das atividades subjetivas?
- Quais técnicas são usadas para avaliar a semântica das atividades subjetivas?
- Quais técnicas são usadas para avaliar a pragmática das atividades subjetivas?

Com isso, foi desenvolvida a seguinte **string de busca** (O quê está sendo usado, com que finalidade, para o quê está sendo usado): ("*Natural Language Processing*"OR "*NLP*") AND (*assess\** OR "*evaluation*"OR "*review*"OR "*revision*"OR "*correction*") AND ("*essay*"OR "*wiki*"OR "*forum*"OR "*Blog*"). Como bases de busca, nesse trabalho foram usadas *ACM*<sup>1</sup>, *IEEE*<sup>2</sup>, *Springer Link*<sup>3</sup>, *Science Direct*<sup>4</sup>, *Web of Science*<sup>5</sup> e *Scopus*<sup>6</sup>. Foram extraídos 9318 artigos, classificados conforme a subseção 3.1.2.

### 3.1.2 Critérios de Inclusão e Exclusão

Os critérios de inclusão admitidos para esta pesquisa foram:

- Estudos publicados somente em Língua Inglesa;
- Estudos publicados entre janeiro de 2009 até 3 de março de 2016, data que se justifica pelo início do estudo;
- Estudos que foram avaliados empiricamente;
- Estudos completos publicados;
- Estudos publicados em Journal;

Contudo, para os critérios de exclusão, nós temos:

- Estudos sem resultados empíricos;
- Estudos que não estão escritos em Língua Inglesa;
- Estudos que estão duplicados;
- Estudos que não apresentam nenhuma relação com o intuito da pesquisa;
- Estudos publicados como short-papers ou secundários, como por exemplo, workshops.

### 3.1.3 Extração e Seleção dos Estudos Primários

Com o objetivo da pesquisa definido (ver subseção 3.1.1) e os critérios de exclusão e inclusão dos estudos (ver subseção 3.1.2), o passo seguinte é a extração dos artigos em cada base de dados apretnetadas no final da subseção 3.1.1.

<sup>1</sup> <<http://dl.acm.org/>>

<sup>2</sup> <<http://ieeexplore.ieee.org/Xplore/guesthome.jsp>>

<sup>3</sup> <<http://link.springer.com/>>

<sup>4</sup> <<http://www.sciencedirect.com/>>

<sup>5</sup> <<https://webofknowledge.com>>

<sup>6</sup> <<https://www.scopus.com/>>

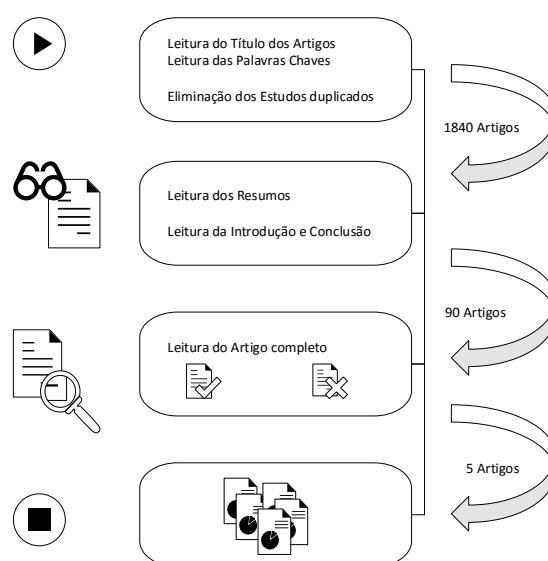
Tabela 2 – Quantidade de artigos por base de dados

Base de dados	Quantidade de artigos
ACM Digital Library	163
IEEE	814
Springer Link	921
Science Direct	1006
Web of Science	38
Scopus	6376
<b>Total</b>	<b>9318</b>

Fonte – Elaborada pelo autor

Para tanto, conforme pode ser visto na Tabela 2, tivemos um total de 9318 artigos extraídos das bases de dados e cada base com a sua respectiva quantidade. Para auxílio desta revisão sistemática, fez-se uso da ferramenta *StArt*<sup>7</sup>.

Figura 13 – Extração dos Trabalhos Relacionados



Fonte – Elaborada pelo próprio autor

A meta-análise oriunda desta revisão sistemática ainda não está finalizada. Contudo, para a extração dos trabalhos relacionados, foram avaliados, até o momento, aproximadamente 1.840 artigos, dentre os quais, foram extraídos, como processo final parcial, 90 trabalhos. A leitura completa destes trabalhos resultou em 5 trabalhos fortemente relacionados com o tema desta proposta, que serão apresentados na seção 3.2. A Figura 13 mostra o processo de forma detalhada.

<sup>7</sup> Mais informações: <[http://lapes.dc.ufscar.br/tools/start\\_tool](http://lapes.dc.ufscar.br/tools/start_tool)>

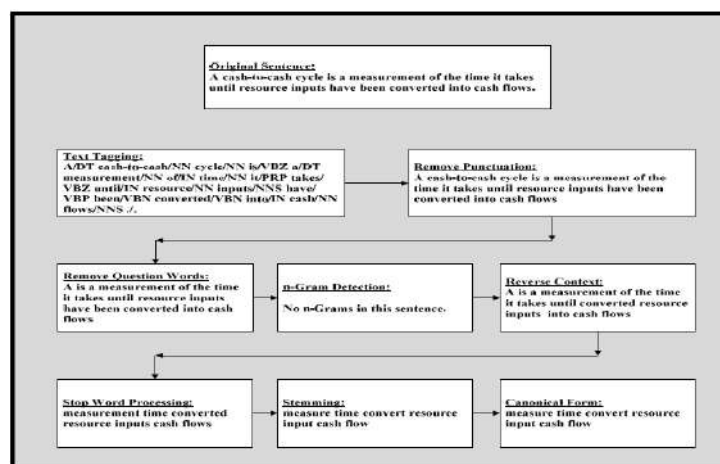
## 3.2 Sistemas de avaliação automática de atividades subjetivas

Nesta seção serão encontrados trabalhos relacionados a avaliação de atividades subjetivas oriundos da seção 3.1. Considerando o escopo do trabalho em questão, foram considerados também trabalhos que avaliassem questões discussivas em Fóruns.

### 3.2.1 Automarking: Automatic Assessment of Open Questions

O sistema **Automarking** (CUTRONE; CHANG, 2010) foi desenvolvido com o objetivo de avaliar a semântica existente entre questões discursivas encontradas em Sistemas EAD com base em respostas enviadas por alunos. Faz uso do dicionário semântico WordNet<sup>8</sup>, sendo utilizado apenas para o Idioma pertencente a Língua Inglesa. A Figura 14 mostra o processo de avaliação de uma sentença pelo Automarking.

Figura 14 – Processo de avaliação da sentença



Fonte – Extraída de (CUTRONE; CHANG, 2010)

Contudo, o Automarking apresenta algumas limitações. O sistema só é capaz de avaliar respostas contendo uma simples sentença, e esta sentença deve ser livre de erros sintáticos e gramaticais.

### 3.2.2 SAGE - Semantic Automated Grader for Essays

O **SAGE** (ZUPANC; BOSNIC, 2014) é um sistema de avaliação baseado em coerência textual baseado em relações semânticas. Baseado em atributos linguísticos léxico-gramatical e atributos de conteúdo, que segundo o próprio autor, são atributos extraídos das redações que

<sup>8</sup> Segundo (SOSA; LOZANO-TELLO; PRIETO, 2008), o WordNet é uma base léxico-semântica com mais de 150.000 palavras. As palavras estão relacionadas de acordo com a semântica entre cada uma delas. Nesta base, pode ser encontrado sinônimos, até as categorias gramaticais existentes, como verbos e substantivos, por exemplo.

serão avaliadas pelo sistema, sendo assim, usados para avaliar cada redação. Para a avaliação semântica, faz uso de aprendizagem de máquina baseada em modelos de regressão<sup>9</sup>.

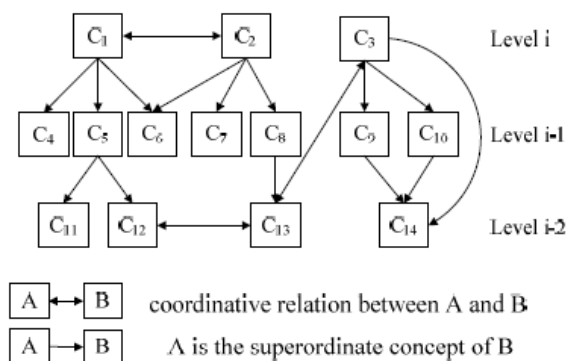
Como limitação, o sistema faz uso de características extraídas das próprias redações, o que carece em detectar a exatidão e credibilidade dos atributos extraídos. Sendo assim, a técnica semântica poderia ser melhorada visando outras características semelhantes aos textos, mas não contidas nas redações.

### 3.2.3 Automatic Chinese Essay Scoring Using Connections between Concepts in Paragraphs

Segundo (CHANG; LEE, 2009), o objetivo principal do trabalho é verificar a relação semântica entre redações baseado em análise conceitual dos parágrafos existentes nelas.

A Figura 15 mostra todo o processo executado pelo sistema dividido em três etapas. Na primeira etapa, as redações são transformadas em um conjunto de treino, onde, na segunda etapa, são extraídos os parágrafos e computados cada similaridade existente entre os demais. Na terceira e última etapa, é computado o score de cada redação com a razão de similaridade encontrada entre os parágrafos.

Figura 15 – Busca de conceitos através dos parágrafos



Fonte – Extraída de (CHANG; LEE, 2009)

O sistema utiliza semântica local baseada em parágrafos, deixando uma lacuna entre as relações que existem entre os demais parágrafos das redações. As relações existentes entre os parágrafos da própria redação não são avaliadas. Contudo, avaliando e comparando os parágrafos com os de outras redações existentes, o sistema necessita de uma grande base de dados para conseguir uma boa acurácia, uma vez que a comparação entre redações sem levar em consideração uma base de treinamento externa.

<sup>9</sup> Definição segundo Microsoft Azure: Algoritmos de regressão são algoritmos que aprendem a prever o valor de uma função real para uma única instância de dados. Algoritmos de regressão podem incorporar a entrada de vários recursos, determinando a contribuição de cada recurso de dados para a função de regressão.

### 3.2.4 MineraFórum - Automatic analysis of messages in discussion forums

MineraFórum<sup>10</sup> (AZEVEDO; BEHAR; REATEGUI, 2011) é uma ferramenta de mineração textual usada em fóruns de discussão que realiza um mapeamento de conceitos relacionados ao tema em debate. A Figura 16 apresenta a interface principal do sistema e o texto que deve ser passado como referência de onde deverá ser feita a extração dos conceitos.

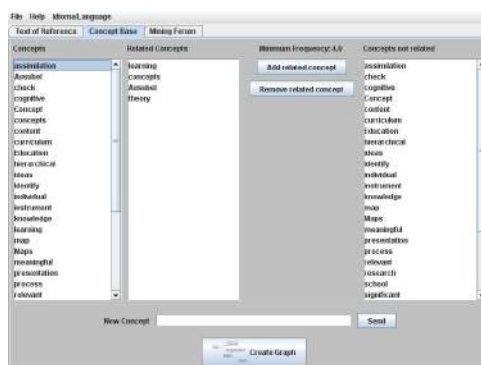
Figura 16 – Inserção de Texto referência



Fonte – Extraída de (AZEVEDO; BEHAR; REATEGUI, 2011)

A Figura 17 apresenta toda a extração de conceitos, com a possibilidade de inserção de novos conceitos pelo usuário, o que permite que palavras informais sejam inseridas preservando a semântica. A extração é realizada com base em um dicionário de sinônimos pertencente à ferramenta. Nesta etapa é calculada a relevância de cada *post* em relação a temática.

Figura 17 – Extração de conceitos com base no Texto referência



Fonte – Extraída de (AZEVEDO; BEHAR; REATEGUI, 2011)

Por fim, como mostra a Figura 18, é informado um texto descritivo para o professor/tutor do sistema. Nele está presente a relevância de cada proposta dos estudantes em relação ao tema, o total de *posts* de cada usuário, a quantidade de *posts* relevantes e não relevantes com a proposição e, por fim, quais são os importantes conceitos usados considerados importantes na discussão.

<sup>10</sup> Disponível em: <[http://www.nuted.ufrgs.br/?page\\_id=1386](http://www.nuted.ufrgs.br/?page_id=1386)>

Figura 18 – Resultado da mineração com base nas respostas dos alunos



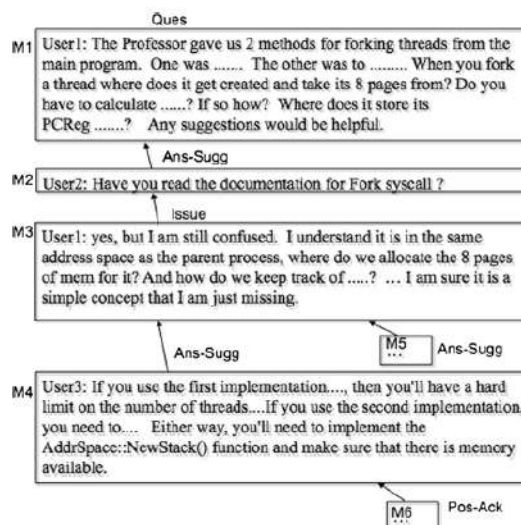
Fonte – Adaptado de (AZEVEDO; BEHAR; REATEGUI, 2011)

MineraFórum possui apenas um dicionário com sinônimos, o que limita a ferramenta. De acordo com o próprio autor, esse dicionário pode ser informado de acordo com o tema proposto, ou usar o próprio dicionário do sistema. Com isso, todas as vezes que o tópico tratado no fórum for diferente, cabe ao professor do ambiente, fornecer uma base de sinônimos para que a avaliação seja efetiva.

### 3.2.5 Towards identifying unresolved discussions in student online forums

(KIM; KANG, 2014) apresenta um sistema com objetivo em avaliar discussões em fóruns educacionais.

Figura 19 – Avaliação de discussão



Fonte – Extraída de (KIM; KANG, 2014)

O foco é identificar quais comentários dos alunos estão relacionados com o tema em debate na discussão, identificando quais resolveram, quais ainda estão sem resposta e quais discussões precisam ser melhoradas. Um alerta é designado para os tutores/professores do ambiente, gerando assim, a intervenção e auxílio para os estudantes que se encontram com problemas na resolução do debate.

A ferramenta de avaliação faz uso de um banco mapeado com palavras de sinalização para cada aspecto de resposta. A Figura 19 apresenta um fórum mapeado. A Figura 20 apresenta a classificação existente com os principais n-grams que podem ser encontrados no texto de cada discussão e seu índice associado com base em algoritmos de SVM<sup>11</sup> e WEKA<sup>12</sup>.

Figura 20 – Mapeamento de principais n-grams

F-SA category	Description	Example cue words from annotations	%	kappa
QUES	A question about a problem, including question about a previous message	“how” “what” “can we” “are”/ “is” “why” “just/were/was wondering” “I/we have a question” “my question”	40.9	0.94
ANS-SUGG	A simple or complex answer to a previous question. Suggestion or advice	“perhaps” “how about” “you might” “you probably” “maybe” “try” “i think” “I am/was thinking” “I’m guessing” “my guess” “it should” “it seems” “look at” “check”	54.6	0.72
ISSUE	Report misunderstanding, unclear concepts or issues in solving problems	“I am still confused” “I was confused” “doesn’t make sense” “I’m not sure” “I’ve no idea” “Not sure” “We ran into” “I am getting fault” “I discovered an error” “problem I am facing” “I am not able to find anything” “I couldn’t find any specific” “I still have no clue”	13.3	0.88
Pos-Ack	An acknowledgement, compliment or support in response to a previous message	“good job” “you got it” “good plan” “good/nice/correct answer” “correct” “thank you/thanks” “i got it:;)” “ok/okay” “I agree” “its fine with me” “i’m okay with...” “good job” “good/nice/correct answer” “correct”	7.7	0.87
Neg-Ack	A correction or objection (or complaint) to/on a previous message	“WRONG” “but not correct” “the above is incorrect.” “Not going to work” “not true” “is actually wrong” “That is true, but” “sure but” “yes, but” “right but” “certainly, but” “I understand but”	2.3	0.85

Fonte – Extraída de (KIM; KANG, 2014)

Como limitação, por possuir uma base anotada de conceitos chaves usados para a classificação, o sistema encontra dificuldade em avaliar mensagens com longas sentenças, palavras informais e respostas em forma de questões.

### 3.3 Comparação entre os trabalhos relacionados

A seguir será apresentada uma tabela comparativa entre os trabalhos relacionados. O intuito desta comparação é destacar a técnica utilizada e qual a limitação encontrada, além de le-

<sup>11</sup> As Máquinas de Vetores Suporte (Support Vector Machines - SVMs) constituem uma técnica embasada na Teoria de Aprendizado Estatístico (HEARST et al., 1998)

<sup>12</sup> WEKA é reconhecido como um sistema de referência em mineração de dados e aprendizagem de máquina (HALL et al., 2009)



vantar os diferentes aspectos de uma avaliação de atividades subjetivas, levando em consideração a sintaxe, semântica e pragmática. Para tanto, alguns critérios foram considerados:

- (i) Qual a técnica utilizada pelo sistema, uma vez que tal critério difere no processo de avaliação da sintaxe, semântica e pragmática;
- (ii) Qual a cobertura do sistema mediante a avaliação sintática, semântica ou pragmática;
- (iii) Limitações encontradas em cada sistema, uma vez que identificadas as limitações, o aperfeiçoamento e melhoria das técnicas utilizadas será de grande importância.

Tabela 3 – Comparação entre os sistemas de avaliação de atividades escritas

Sistema	Técnicas utilizadas	Sintaxe	Semântica	Pragmática	Limitações
<b>Automarking</b>	WordNet. Avaliação de sentenças locais	-	x*	-	Sistema capaz de avaliar somente simples sentenças livres de erros sintáticos e gramaticais
<b>SAGE</b>	Avaliação de atributos linguísticos léxico-gramaticais. Aprendizagem de máquina baseada em modelos de regressão	-	x	-	Extração de atributos das próprias redações, com isso, não avalia características semelhantes aos textos
<b>Automatic Chinese Essay Scoring</b>	Análise conceitual de parágrafos	-	x	-	Usa semântica local, com isso, as relações existentes entre parágrafos distantes são perdidas. Para alcançar uma boa acurácia, o sistema necessita de uma grande base de dados para extrair relações semânticas entre os parágrafos de forma confiável
<b>MineraFórum</b>	Utiliza texto referência para extração de conceitos. Utiliza dicionário de sinônimos para comparação entre termos	-	x	-	Dicionário de sinônimos pode ser limitado, logo, para cada tema pedido, para ter uma boa acurácia na avaliação, deve ser informado um dicionário de acordo com o contexto
<b>(KIM; KANG, 2014)</b>	Banco de anotações com mapeamentos de palavras e classificação. Utiliza algoritmos de mineração do WEKA e classificação por SVM	-	x	-	Base anotada precisa ser grande para uma boa acurácia na avaliação. Dificuldade em avaliar mensagens com longas sentenças, palavras informais e respostas em forma de perguntas

Fonte – Elaborada pelo autor

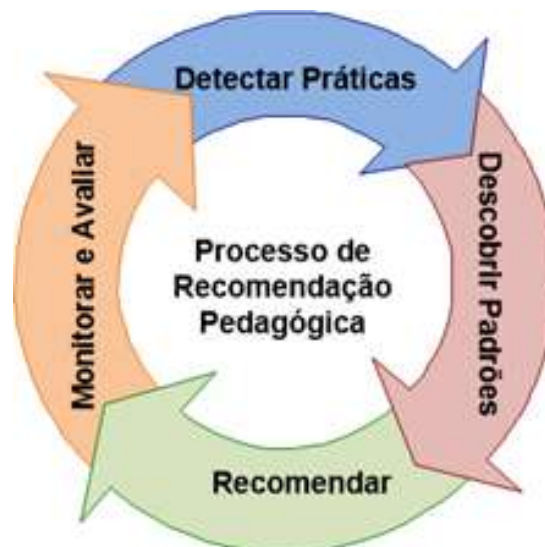
## 4 PROPOSTA

Neste capítulo, serão apresentadas as etapas para realização da proposta aqui apresentada, bem como o modelo utilizado para a fundamentação da pesquisa.

### 4.1 Metodologia

A metodologia deste trabalho é baseada em um processo denominado Processo de Tomada de Decisão Pedagógica - PTDP<sup>1</sup> proposto por (PAIVA et al., 2016), conforme ilustrado na Figura 21. O objetivo deste modelo é combinar e coordenar os esforços da inteligência humana com a inteligência artificial.

Figura 21 – Processo de Recomendação Pedagógica



Fonte – Extraída de Paiva et al. (2016)

O Modelo é de natureza **iterativa e incremental** e constituído por quatro etapas apresentadas logo a seguir.

- (i) **Detectar práticas** (positivas e negativas) no processo de aprendizagem (*O que?*). É a primeira etapa do processo e tem como objetivo avaliar e identificar erros que os alunos comentem na realização de atividades subjetivas.
- (ii) **Descobrir padrões** responsáveis pelas práticas detectadas (*Por quê?*). O foco principal dessa etapa é identificar padrões relacionados aos erros que foram detectados na etapa anterior.

<sup>1</sup> Decisões pedagógicas: Ações de natureza preventiva, ou reativa, associadas a situações pedagógicas definidas. Tais ações podem ser realizadas utilizando os recursos nativos do ambiente de aprendizado, ou através de recursos externos, e têm o objetivo de solucionar os problemas (pedagógicos) identificados, melhorando a experiência de aprendizado dos alunos, segundo (PAIVA et al., 2016).

- (iii) **Recomendar** conteúdo pedagógico com base nos padrões descobertos, com o propósito de solucionar problemas detectados (*Como?*). Essa etapa é responsável por disponibilizar recomendações de recursos educacionais que auxiliem aos alunos com base nos padrões descobertos na etapa anterior.
- (iv) **Monitorar e Avaliar** a efetividade/eficiência das recomendações (*Quais os Resultados?*). Essa etapa é responsável em avaliar se os alunos estão progredindo positivamente em decorrência das avaliações recebidas.

## 4.2 Detectar práticas

Esta etapa é responsável em avaliar as atividades subjetivas. Para avaliação de atividades subjetivas, o processo segue o modelo de avaliação da redação<sup>2</sup> proposto pelo Ministério da Educação - MEC que vem sendo utilizado atualmente no ENEM - Exame Nacional do Ensino Médio<sup>3</sup>. A apresentação deste modelo pode ser encontrada no Capítulo 2 na seção 2.1.

Para cada competência apresentada, o sistema constará com técnicas específicas para a avaliação de cada uma delas, e serão detalhadas logo a seguir.

### 4.2.1 Técnicas para avaliar a Competência 1

A competência 1 refere a habilidade do aluno em demonstrar domínio da modalidade escrita formal da Língua Portuguesa. É nesta etapa que se tem preocupação com a ortografia e gramática do estudante. Para isso, as técnicas utilizadas para realizar essa avaliação são POS Tagging (seção 2.2), Avaliação ortográfica com base em dicionário (seção 2.3), Raciocínio baseado em casos (seção 2.5) e Redes Neurais (seção 2.7).

O modelo terá a seguinte execução:

1. Classificação das palavras presentes no texto;
2. Verificação de palavras sem classe gramatical;
3. Verificação da gramática através de regras mapeadas em XML;
4. Cálculo da nota adequada para a respectiva redação.

Na etapa de **Classificação das palavras presentes no texto**, todo o processo de categorização das palavras com suas respectivas classes gramaticais e relação com aspectos da língua como tempo, gênero e número, será realizado. Apesar de ser uma tarefa apenas de classificação, esta etapa será responsável por facilitar e otimizar o trabalho realizado tanto pela verificação ortográfica, quanto pela avaliação gramatical.

<sup>2</sup> <[http://download.inep.gov.br/educacao\\_basica/enem/guia\\_participante/2013/guia\\_de\\_redacao\\_enem\\_2013.pdf](http://download.inep.gov.br/educacao_basica/enem/guia_participante/2013/guia_de_redacao_enem_2013.pdf)>

<sup>3</sup> <<http://portal.inep.gov.br/enem>>

Uma vez que todas as palavras passaram pela classificação, na etapa de **verificação de palavras sem classe**, todas as palavras que não possuam uma classe associada passarão pelo dicionário com o intuito de validar se tal palavra é um erro ortográfico ou apenas ainda não possui tal classificação. Se a palavra estiver no dicionário indica que ela ainda não está mapeada com sua respectiva classe gramatical. Essa palavra é então adicionada em uma base temporária para posteriormente ser adicionada na base principal. Esta etapa de adição na base principal depende da intervenção humana.

A **Verificação da gramática através de regras mapeadas em XML** será responsável por toda validação gramatical do texto. Identificar relações de concordância entre os termos classificados. Neste etapa, faz-se o uso de um conjunto de regras mapeadas que são responsáveis pela identificação de erros gramaticais. Uma vez que todas as palavras possuem uma classe associada, através dos aspectos identificados, como o gênero da palavra (masculino, feminino ou neutro), ou até mesmo o número da palavra (singular ou plural), regras mapeadas em XML serão responsáveis por verificar a gramática do texto.

Por fim, na etapa do **Cálculo da nota adequada para a respectiva atividade escrita**, o sistema utilizará todas as características extraídas, como erros relacionados à ortografia, gramática e pontuação. Para tanto, se faz uso de uma abordagem com a técnica de Raciocínio baseado em casos, que tem como principal função identificar padrões na avaliação humana e caracterizá-las em uma avaliação por máquina que dizem respeito ao domínio da norma padrão de escrita. Entretanto, para auxiliar na nota final do texto, utiliza-se uma rede neural de regressão para atribuição da nota com base nas características extraídas.

#### 4.2.2 Técnicas para avaliar a Competência 2

Tratando-se da competência 2, que diz respeito a compreensão da proposta de redação e aplicar conceitos das várias áreas de conhecimento para desenvolver o tema, dentro dos limites estruturais do texto dissertativo-argumentativo em prosa, existe a necessidade de técnicas para avaliar a estrutura do texto e Analisar conceitualmente os parágrafos, sendo que o texto dever estar relacionado com a proposta solicitada. Para isso, se faz uso de Extração de Termos Semânticos e N-Gramas, apresentados na seção 2.4.

O modelo terá a seguinte execução:

1. Eliminação de Stopwords;
2. Extração de termos semânticos;
3. Construção de N-Gramas;
4. Cálculo da Similaridade Vetorial;
5. Cálculo da nota adequada para a respectiva redação.

A fase de **Eliminação de Stopwords** simplesmente elimina todos os termos desnecessários que estejam presentes na redação que possam atrapalhar na fase de extração de termos. Os

termos descartados nessa etapa são artigos, preposições e conjunções. Essa é a etapa que garante uma melhora na próxima etapa.

Na etapa de **Extração de termos semânticos**, é realizada uma busca em relação a frequência de verbos, adjetivos e substantivos. A partir disso, são criados conjuntos de palavras, denominado de vetores semânticos. Por meio dos vetores semânticos formados, são construídas novas estruturas com relações semânticas definidas, por exemplo, dois substantivos ou um substantivo e um verbo. Tais formações são denominadas Bi-gramas. Essa fase de construção é denominada **Construção de N-Gramas**. Essa etapa, é responsável por criar tri-gramas ou até mesmo conjuntos maiores que três palavras. Essa junção de termos é fundamental para a fase do Cálculo vetorial, etapa a ser detalhada logo a seguir.

Com o **Cálculo da Similaridade Vetorial** é possível comparar textos com o intuito de identificar o nível de similaridade entre eles através do Cosseno do ângulo formado entre eles. Nesta etapa, existe uma execução paralela com a verificação da proposta e do texto que está sendo avaliado. Entretanto, é também nessa fase que se compara a redação com textos que estão próximos da proposta de redação solicitada. Tais textos são recuperados através de um *WebCrawler*, técnica utilizada para procurar arquivos que estejam disponíveis na WEB com um filtro para determinados conteúdos. Tal processo de busca utiliza um N-Grama formado da etapa anterior que consiga resumir a proposta solicitada em apenas poucas palavras. A comparação resultante entre os textos, e inclusive a proposta, com a redação é armazenada em um vetor, que, posteriormente é usado para calcular a nota final do aluno nesta competência.

Tendo um vetor resultante com o nível de similaridade da etapa anterior, p **Cálculo da nota adequada para a respectiva redação** é realizado com o auxílio de um Rede Neural de regressão específica para prover o *score* respectivo do texto. Essa rede neural é calibrada com base em redações avaliadas anteriormente onde casos gerados através de um componente que faz uso da técnica de RBC. Nesta proposta, tal modelo ajudará na identificação estrutural de um texto dissertativo-argumentativo montando todas as probabilidades possíveis do texto que está sendo avaliado estar na estrutura requerida. Para esta técnica, é fundamental ter redações ou textos dissertativos-argumentativos para que o modelo possa ser treinado, aumentando sua acurácia.

### 4.2.3 Técnicas para avaliar a Competência 3

Como já foi visto (ver seção 2.1), na competência 3, o aluno deverá ser capaz de selecionar, relacionar, organizar e interpretar informações, fatos, opiniões e argumentos em defesa de um ponto de vista. A avaliação para esta competência é semelhante a abordagem apresentada na subseção 4.2.2.

Portanto, a abordagem responsável por avaliar a competência 3 terá a seguinte execução:

1. Eliminação de Stopwords;

2. Extração de termos semânticos locais e globais;
3. Construção de N-Gramas locais e globais;
4. Cálculo da Similaridade Vetorial;
5. Cálculo da nota adequada para a respectiva redação.

Preocupados com todas as palavras que possam não ter uma significância elevada para a extração de termos semânticos, a **Eliminação de Stopwords** visa amenizar todo o efeito por tais palavras. Como já foi dito anteriormente, essa técnica visa refinar todo o texto sob o aspecto da eliminação de artigos, preposições, conjunções ou qualquer outra palavra que possa influenciar na precisão de extração de termos significativos.

Sabendo que nessa competência, o aluno deverá expor fatos e argumentos em defesa de um ponto de vista, há a necessidade de **Extração de termos semânticos locais e globais**. Os termos locais são os termos extraídos de cada parágrafo que compõe o texto como um todo. Esses termos são responsáveis para verificar se os argumentos que são utilizados pelo aluno estão relacionados com o tema solicitado. Já em relação aos termos globais, esses são termos componentes mais utilizados pelo aluno no desenvolvimento de sua redação.

Uma vez que os termos semânticos estão extraídos, na fase de **Construção de N-Gramas locais e globais** são associados vetores com os respectivos termos e os parágrafos onde tais termos foram extraídos. Portanto, os parágrafos agora serão expressos por vetores componentes, assim como, também, todo o texto terá um vetor formado pelos termos mais utilizados, compondo um N-Grama resumo do texto.

Com o **Cálculo da Similaridade Vetorial**, é possível, através dos vetores locais e globais formados pela etapa anterior, calcular a similaridade dos argumentos que estão presentes no texto em relação à proposta. De forma similar à abordagem executada pelo modelo da competência 2, o cálculo da similaridade acontece através do Cosseno para comparação dos textos: proposta, redação e textos recuperados e baseados nos argumentos descritos na redação. Tais textos, irão compor a unidade de avaliação responsável por avaliar fatos, opiniões e os argumentos que compõem os parágrafos componentes da redação.

Para tanto, tendo todos os valores de similaridade entre os argumentos, proposta e textos oriundos dos argumentos, é através da **Cálculo da nota adequada para a respectiva redação**. Este modelo faz uso, assim como os modelos anteriores responsáveis em dar a nota adequada para competência 1 e 2, de uma Rede Neural de Regressão. Esta rede tem como principal funcionalidade receber os scores gerados pela similaridade e balancear através de pesos pela função de regressão linear, e, por fim, retornar a avaliação para a competência 3.

#### 4.2.4 Técnicas para avaliar a Competência 4

Sabendo que nessa competência, o aluno precisa demonstrar conhecimento dos mecanismos linguísticos necessários para a construção da argumentação, será utilizada um modelo de

regressão através de uma Rede Neural. Entretanto, levando em consideração que tal competência visa avaliar a construção da argumentação, e que tal abordagem está ligada à área da Pragmática, precisamos prover alguma estratégia para contornar a limitação existente na máquina para avaliar tal área do Discurso.

A primeira delas é começar a se preocupar com o uso de termos coesivos. Um texto bem estruturado evita repetições de termos substituindo-os por pronomes ou qualquer artifício linguístico. Para tanto, tal modelo terá a seguinte execução:

1. Verificação de termos repetitivos através de Árvores de derivação sintática;
2. Cálculo da nota adequada para a respectiva redação.

**Verificação de termos repetitivos através de Árvores de derivação sintática** é um modelo para encontrar termos que poderiam ser substituídos por outros termos, tendo como principal preocupação não perder a coesão e coerência do texto. Nesta etapa, é utilizado o *Parser*, técnica utilizada na competência 1 para classificar as palavras. O Parser, ou analisador sintático, ajudará na verificação da palavra repetida, identificando possíveis ambiguidades ou palavras utilizadas em contexto diferente, por exemplo, casa (substantivo) ou casa (verbo).

Por fim, para o **Cálculo da nota adequada para a respectiva redação**, um modelo de regressão construído sob uma rede neural recebe os scores do texto sob as três competências anteriores, com o intuito de encontrar um relação entre, a forma de escrita formal, a compreensão da proposta e o uso de argumentos no desenvolvimento do texto. Para tanto, é criada uma relação com base na quantidade de termos repetidos, e, logo em seguida, a nota é criada. Esta abordagem utiliza critérios de que se o texto é bem escrito, possui argumentos em relação ao tema solicitado, faz uso de termos coesivos adequados, então, a nota para avaliação será alta.

#### 4.2.5 Técnicas para avaliar a Competência 5

Uma vez que a competência 5 diz respeito a elaborar uma proposta de intervenção para o problema abordado, respeitando os direitos humanos, uma Rede Neural baseada em modelo de regressão foi utilizada como técnica para resolução do problema.

Analisar a escrita nesta competência exige eficiência nas técnicas. Identificar relação entre avaliações obtidas nas competências anteriores e prever uma avaliação para o texto se torna a principal funcionalidade do modelo de Rede Neural utilizado para estimar uma avaliação para a redação.

Para tanto, a execução do modelo dedicado em resolver esta competência simplesmente limita-se em uma abordagem que faça uso de um modelo de regressão capaz de estimar uma nota. Entretanto, tal modelo deverá reconhecer padrões das avaliações anteriores. Entre outras palavras, o modelo receberá as avaliações obtidas das competências 1, 2, 3 e 4, e, através destes



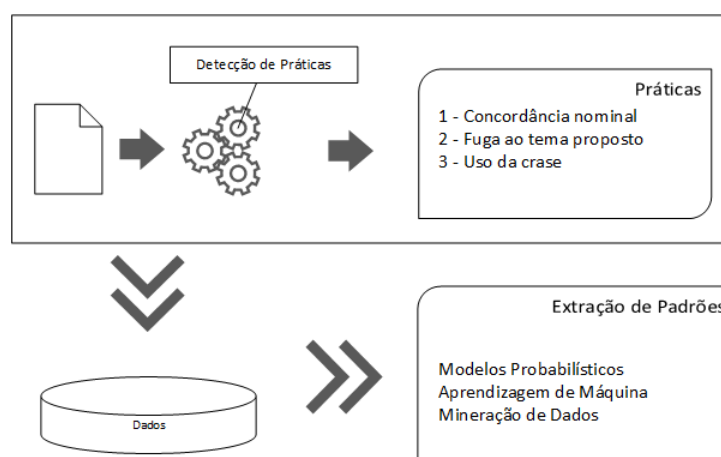
scores, calcular o valor da competência 5, relacionando e balanceando os diferentes mecanismos linguísticos oriundos de tais avaliações das competências anteriores.

### 4.3 Descobrir padrões

Etapa responsável em identificar padrões que foram detectados na etapa anterior (ver seção 4.2), ou seja, descobrir o motivo pelo qual cada prática está acontecendo, entre outras palavras, o aluno está errando? O que ele está errando? Qual a frequência do erro?.

A Figura 22 apresenta de forma detalhada o processo de detecção de padrões em uma atividade subjetiva avaliada mediante a seção 4.2.

Figura 22 – Processo de Detecção de padrões



Fonte – Elaborada pelo autor

A descoberta de padrões é o componente responsável por calibrar e alimentar todas as Redes Neurais criadas para a avaliação das competências definidas na etapa de Detecção de Práticas. É por meio deste componente, que cada Rede Neural aumenta sua acurácia, provendo uma avaliação cada vez melhor.

Entretanto, é por meio deste componente que a máquina procura compreender a relação existente entre o que o avaliador humano levou em consideração para dar determinada nota para um texto. Este modelo atua identificando e ponderando os pesos específicos de cada ponto que encontrar de acordo com a nota do especialista.

### 4.4 Recomendar

Etapa responsável por disponibilizar sugestões de melhoria na escrita do texto. Com base nas avaliações realizadas nas competências mencionadas na seção 4.2, a abordagem será capaz de sugerir possíveis alterações de forma que o aluno consiga melhorar seu texto. Esse componente é responsável por prover todo o *feedback* entre a avaliação provida e os erros detectados.

Entretanto, esse componente atua também como um provedor de textos que possam aperfeiçoar ainda mais o domínio do aluno na escrita de determinados temas onde a avaliação provida pelo sistema seja baixa, identificando assim uma carência de escrita. Essas recomendações são criadas com base na extração de padrões mencionada anteriormente (ver seção 4.3).

#### 4.5 Monitorar e Avaliar

Nesta etapa, todas as interações entre o aluno e as práticas extraídas na etapa anterior são armazenadas de modo que seja capaz avaliar como o aluno está se comportando diante dos padrões detectados. O sistema irá verificar se o padrão extraído voltará a se repetir mediante uma nova interação do aluno com o sistema, entre outras palavras, verificar se o aluno cometerá o mesmo erro e qual a frequência de ocorrência.

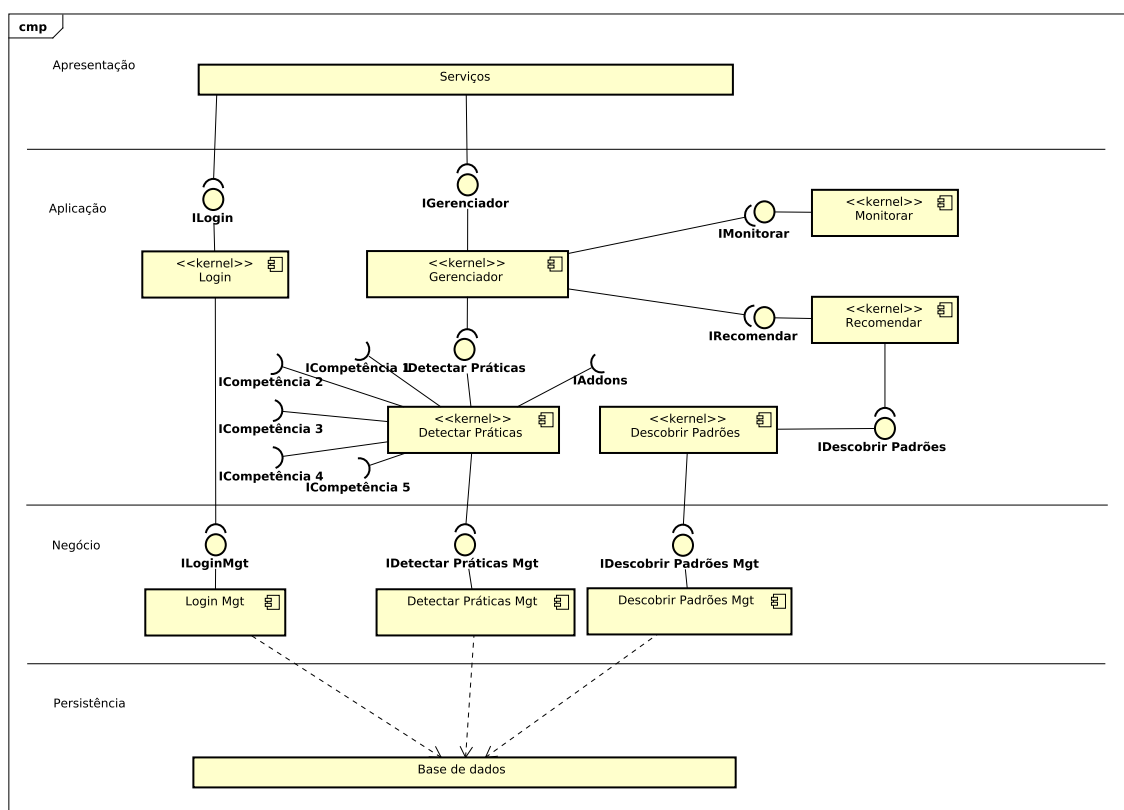
#### 4.6 Arquitetura do Sistema

Nesta seção será encontrada tanto a arquitetura do sistema proposto, quanto as técnicas utilizadas para sua construção. Portanto, para facilitar a compreensão e visualização do sistema como um todo, a arquitetura utilizada é microkernel baseada em COSMOS, que favorece ainda mais o processo de adaptação e implantação de novas técnicas para avaliar cada competência apresentada na seção 4.1, levando em consideração a necessidade de manutenção, controle e diminuição de sobrecarga, e prover a variabilidade do sistema.

Na Figura 23 pode ser visto o núcleo do sistema. Esses são os componentes que são estáticos para a execução do sistema. Funcionalidades básicas necessárias para o funcionamento do sistema. Nela, os componentes responsáveis por manter o sistema funcional são apresentadas. A arquitetura foi desenvolvida sob 4 (quatro) camadas. Cada qual com sua responsabilidade: (i) Apresentação: toda a parte de serviços providos do sistema que podem ser usados tanto quanto API quanto um sistema independente. Para tanto, todos os serviços será REST/JSON; (ii) Aplicação: Camada intermediária entre os serviços e a camada de negócios. Nessa camada, toda a lógica do sistema com relação as funcionalidades providas são encontradas; (iii) Negócio: Toda lógica de comunicação entre a camada de persistência e a camada de aplicação. Esta camada é responsável por gerenciar a persistência que é demandada pela camada de aplicação, fazendo com o que não ocorra sobrecarga no banco; e, por fim, (iv) Persistência: Camada responsável por armazenar todas as entidades que são criadas por operações executadas em componentes de outras camadas.

Contudo, na Figura 24 são exibidos todos os componentes com suas devidas técnicas utilizadas, que foram apresentadas na seção 4.2, seção 4.3 e na seção 4.4. Nela é possível observar que outras técnicas poderão ser adicionadas para enriquecer a análise a ser realizada. Note também, que cada competência possui sua rede neural específica. Tais redes foram treinadas especificamente para corrigir sua respectiva competência. Os **Addons** são componentes auxiliares

Figura 23 – Kernel do Sistema



Fonte – Elaborada pelo autor

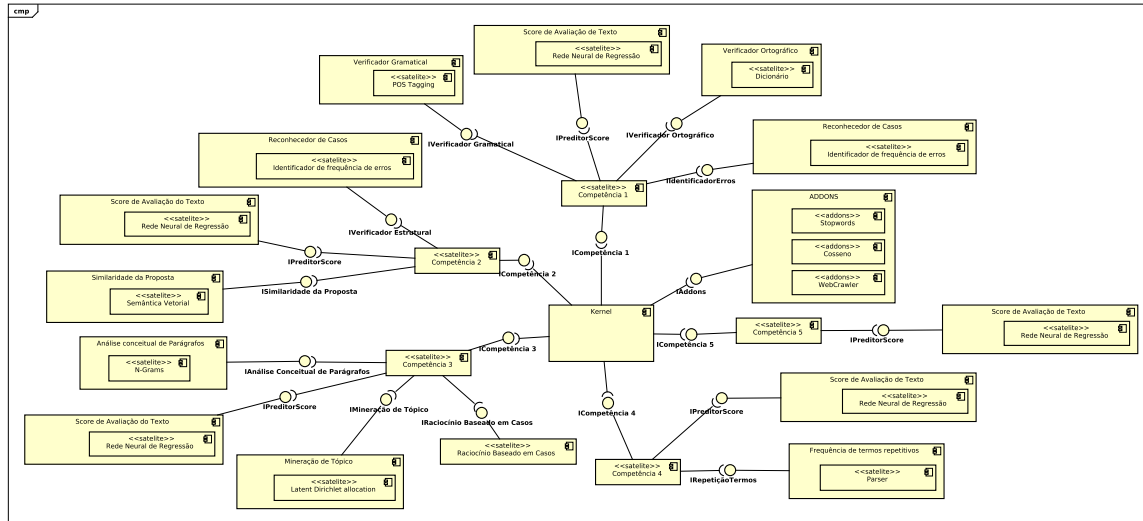
utilizados como ferramentas que podem ser usados em qualquer funcionalidade que demande sua execução.

Com o objetivo de exibir as principais funcionalidades do sistema, na Figura 25 encontram-se as interações do sistema e do usuário. São as atividades que cada ator desempenhará no ambiente.

Para facilitar a compreensão do funcionamento dinâmico do sistema como um todo, a Figura 26 apresenta um diagrama de fluxo de dados exibindo todo o processo para a avaliação de uma redação, de recomendação de um recurso educacional até o monitoramento do mesmo.

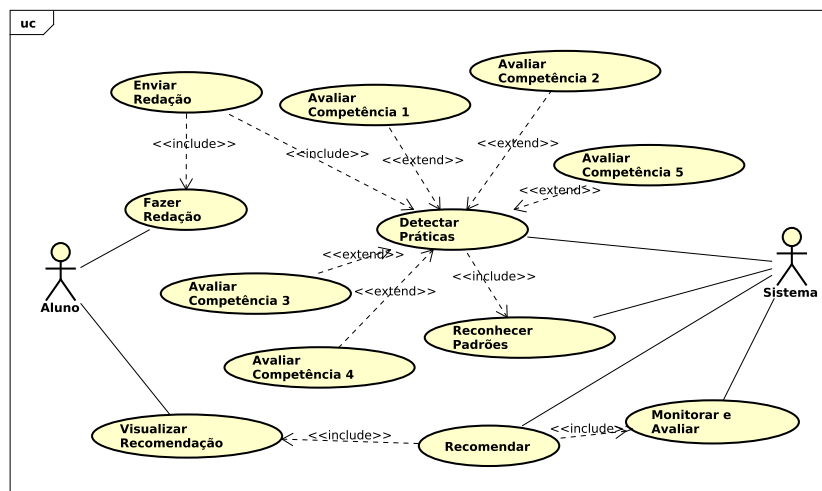
Como pode ser visto, a primeira interação acontece quando o aluno envia sua redação para o sistema. O sistema solicita a avaliação da atividade de acordo com as competências que serão avaliadas e retorna os resultados para o usuário. Com o objetivo de auxiliar no processo de aprendizagem do aluno, o sistema solicita recursos educacionais. Para tanto, os recursos educacionais são criados com base na detecção de padrões que são extraídos com base na avaliação realizada da redação. Por fim, os recursos educacionais são retornados para o usuário, que por sua vez, será monitorado para saber seu progresso e a eficácia da recomendação oferecida para o aluno mediante o padrão identificado.

Figura 24 – Satélites do Sistema



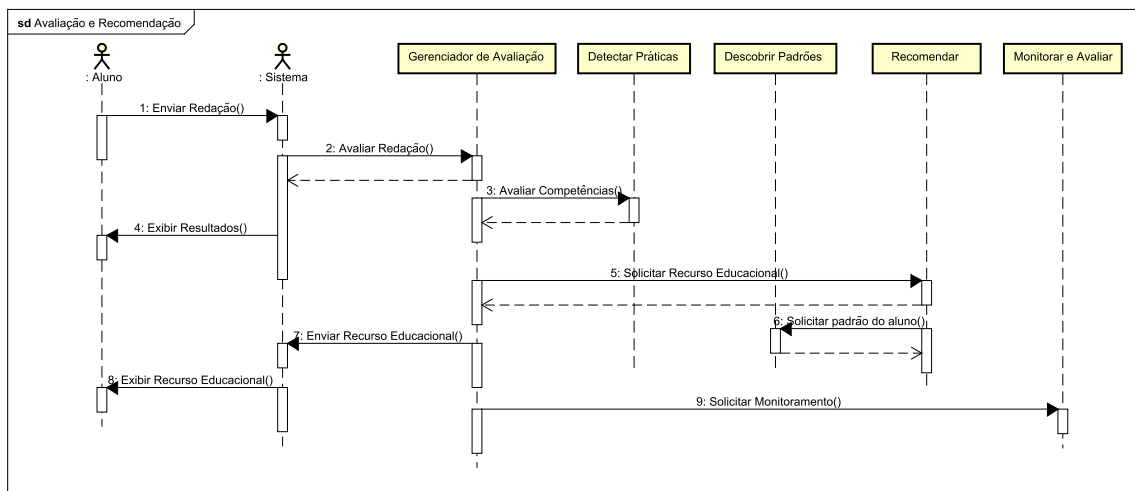
Fonte – Elaborada pelo autor

Figura 25 – Casos de uso do Sistema



Fonte – Elaborada pelo autor

Figura 26 – Diagrama de Fluxo: Avaliação e Recomendação funcionando de forma dinâmica



Fonte – Elaborada pelo autor

## 5 DESIGN DE EXPERIMENTO

Sabendo que o presente trabalho visa desenvolver um modelo conceitual com consolidação em um sistema que seja capaz de efetuar avaliações de atividades subjetivas, avaliando **sintaxe**, **semântica** e **pragmática**, afim de diminuir a sobrecarga por parte do professor, e auxiliar os alunos no processo da aprendizagem mediado pela escrita, neste capítulo, será apresentado um design de experimento que tem como finalidade avaliar o sistema desenvolvido.

### 5.1 Situando o Problema

O desenvolvimento da escrita em ambientes EAD se desenvolve através de duas formas (PAVEZI et al., 2011):

- (i) Atividades Síncronas (Interação por Chat)
- (ii) Atividades Assíncronas (Interação por Fórum, Wiki, Questões dissertativas/discursivas)

Quando atividades subjetivas são avaliativas, acaba gerando sobrecarga por parte dos professores (BALFOUR, 2013). Com isso, retomamos nosso problema geral:

**(Problema Geral)** Como diminuir a sobrecarga no professor gerada pela avaliação de atividades subjetivas presentes em ambientes EAD?

Atualmente, avaliação por pares (KOLLER; NG, 2012) e sistemas de avaliação automática (MARKOFF, 2013) são usados como solução para o problema. Como pode ser visto na Tabela 1, apresentada no Capítulo 1, ambas as soluções apresentam seus pontos positivos e seus pontos negativos. Entretanto, segundo (BALFOUR, 2013), é mais viável comercialmente utilizar avaliação automática, uma vez que, mesmo provendo o rápido *feedback* para os alunos, a técnica utilizada pode ser usada combinando aspectos com o próprio sistema EAD, como por exemplos, recomendações personalizadas de recursos educacionais.

Diante disso, nos deparamos com nosso problema específico:

**(Problema Específico: )** Como melhorar a qualidade da avaliação automática para atividades subjetivas escritas em Língua Portuguesa Brasileira?

### 5.2 Objetivos da Investigação

A pesquisa a ser realizada é de caráter experimental e tem como objetivo geral **avaliar a qualidade da avaliação automática em atividades subjetivas**. O Intuito deste experimento é **aceitar** as hipóteses nulas definidas na seção 5.3 indicando que a qualidade da avaliação automática se assemelha a qualidade de um especialista.

Formalmente, o objetivo da nossa investigação pode ser definido como **analisar** algoritmos de processamento de linguagem natural **com a intenção de** compará-los **a respeito** de sua eficácia **do ponto de vista** de avaliação de atividades subjetivas **no contexto** de atividades subjetivas compreendidas por computador em sistemas EAD **com o fim** de utilizar os melhores algoritmos de forma combinada **provendo** uma melhoria na qualidade de avaliações subjetivas.

Como objetivos específicos, nós temos:

- (i) Comparar as abordagens sob cada dimensão do discurso (Sintaxe, Semântica e Pragmática).
- (ii) Avaliar empiricamente a qualidade dos modelos criados.

### 5.3 Questões de Pesquisa e Hipóteses

Após ter os nossos objetivos apresentados na seção 5.2, nos deparamos com as seguintes questões de pesquisa:

**QP 1:** De que maneira a avaliação automática apresenta a mesma qualidade, ou melhor, apresenta avaliações semelhantes se comparada às correções realizadas por um especialista?

**QP 2:** Como podemos avaliar a qualidade das correções realizadas pelo sistema corretor?

**QP 3:** De que forma as correções podem ser comparadas às correções realizadas por um especialista (professor)?

**QP 4:** Como o modelo de correção automática contribui auxiliando na redução da carga do professor?

Portanto, para ser capaz de responder a tais questionamentos, temos nossas hipóteses sob a métrica de *score* para cada competência:

$H_{1.0}$  : O *score* das avaliações realizadas pelo sistema é equivalente ao *score* das avaliações realizadas pelo especialista em relação à competência 1.

$H_{1.1}$  : O *score* das avaliações realizadas pelo sistema não é equivalente ao *score* das avaliações realizadas pelo especialista em relação à competência 1.

$H_{2.0}$  : O *score* das avaliações realizadas pelo sistema é equivalente ao *score* das avaliações realizadas pelo especialista em relação à competência 2.

$H_{2.1}$  : O *score* das avaliações realizadas pelo sistema não é equivalente ao *score* das avaliações realizadas pelo especialista em relação à competência 2.

$H_{3.0}$  : O *score* das avaliações realizadas pelo sistema é equivalente ao *score* das avaliações realizadas pelo especialista em relação à competência 3.

$H_{3.1}$  : O *score* das avaliações realizadas pelo sistema não é equivalente ao *score* das avaliações realizadas pelo especialista em relação à competência 3.

$H_{4.0}$  : O *score* das avaliações realizadas pelo sistema é equivalente ao *score* das avaliações realizadas pelo especialista em relação à competência 4.

$H_{4.1}$  : O *score* das avaliações realizadas pelo sistema não é equivalente ao *score* das avaliações realizadas pelo especialista em relação à competência 4.

$H_{5.0}$  : O *score* das avaliações realizadas pelo sistema é equivalente ao *score* das avaliações realizadas pelo especialista em relação à competência 5.

$H_{5.1}$  : O *score* das avaliações realizadas pelo sistema não é equivalente ao *score* das avaliações realizadas pelo especialista em relação à competência 5.

#### 5.4 Fatores e Variáveis de Resposta

Com a definição de nossas *hipóteses* apresentadas na seção 5.3, temos os nossos fatores, também conhecidos como variáveis independentes, como sendo:

**Abordagem** : Consiste na forma em que a avaliação será realizada;

**Dimensão** : Qual dimensão do discurso está sendo avaliada. Neste caso, representada pelas Competências.

Como variável de resposta, também conhecidas como variáveis dependentes, nós temos:

**Score** : Para avaliação de atividades subjetivas, indica a nota final dada pelas abordagens.

Nossos fatores são variáveis qualitativas nominais, contudo, nossas variáveis de resposta são quantitativas.

#### 5.5 Níveis dos Fatores

Nossos níveis dos Fatores são apresentados na Tabela 4.

#### 5.6 Definição formal das Hipóteses

Formalmente, todas as hipóteses definidas na seção 5.3 podem ser definidas conforme a Tabela 5.



Tabela 4 – Definição dos níveis dos Fatores

Fatores	Níveis
<i>Abordagem</i>	Avaliação por Especialista
	Avaliação automática
<i>Dimensão</i>	Competência 1
	Competência 2
	Competência 3
	Competência 4
	Competência 5

Fonte – Elaborada pelo autor

Tabela 5 – Definição formal das Hipóteses

Hipótese	Hipótese Nula	Hipótese Alternativa
$H_1$	$H_0:S(C_1(A_1)) = H_0:S(C_1(A_2))$	$H_1:S(C_1(A_1)) \neq H_1:S(C_1(A_2))$
$H_2$	$H_0:S(C_2(A_1)) = H_0:S(C_2(A_2))$	$H_1:S(C_2(A_1)) \neq H_1:S(C_2(A_2))$
$H_3$	$H_0:S(C_3(A_1)) = H_0:S(C_3(A_2))$	$H_1:S(C_3(A_1)) \neq H_1:S(C_3(A_2))$
$H_4$	$H_0:S(C_4(A_1)) = H_0:S(C_4(A_2))$	$H_1:S(C_4(A_1)) \neq H_1:S(C_4(A_2))$
$H_5$	$H_0:S(C_5(A_1)) = H_0:S(C_5(A_2))$	$H_1:S(C_5(A_1)) \neq H_1:S(C_5(A_2))$

Fonte – Elaborada pelo autor

Legenda – S: Score,  $A_1$ : Avaliação automática,  $A_2$ : Avaliação do especialista,  $A_3$ : Avaliação por pares,  $C_1$ : Competência 1,  $C_2$ : Competência 2,  $C_3$ : Competência 3,  $C_4$ : Competência 4,  $C_5$ : Competência 5

## 5.7 Unidades Experimentais

O experimento é do tipo comparativo para ambas as avaliações dos Modelos de avaliação de atividades subjetivas. Contudo, o design para o Modelo de avaliação de atividades subjetiva pertence ao *two-group design*, sendo o nosso grupo de controle as avaliações realizadas pelo especialista, e nosso grupo experimental, as avaliações realizadas pela abordagem automática.

## 5.8 Plano de execução

A execução deste experimento envolve os seguintes passos:

1. Coleta dos dados
2. Calibrar o sistema com conteúdos disponíveis na WEB
3. Execução do experimento
4. Extração das métricas
5. Análise estatística descritiva e inferencial dos resultados

## 5.9 Coleta dos Dados

Na amostra para o experimento do Modelo de avaliação automática, 467 redações, totalizando 23 temas distintos, serão extraídas aleatoriamente da base de dados da UOL<sup>1</sup>. Vale ressaltar que estas redações são avaliadas já por profissionais da área. O experimento ainda constará com 5.000 (cinco mil avaliações) do Exame Nacional do Ensino Médio.

## 5.10 Execução do Experimento

O Experimento envolverá validação cruzada. Nele, os 23 temas de redações dispostos em 467 redações serão escolhidos aleatoriamente para compor a base de conhecimento e a base de validação. Ao total, 3 temas serão escolhidos para compor a base de validação. Já, os 20 temas restantes serão utilizados para calibrar a abordagem. Cada rede neural será calibrada com tais temas, sendo para a avaliação das competências 4 e 5, a calibragem também será realizada com uso das 5.000 avaliações do ENEM.

Todo o processo de avaliação das atividades escritas terá como base o pior dos casos: Toda redação corrigida deverá ser levada em consideração como se fosse o primeiro texto a ser avaliado. Entre outras palavras, para a execução deste experimento, a máquina irá apagar qualquer texto ou avaliação já realizada anteriormente. Isso será fundamental para medir o modelo sob o aspecto de primeiro texto corrigido.

## 5.11 Análise dos Resultados

A análise dos dados será com estatística quantitativa com uso de boxplots e gráficos de dispersão. Serão utilizados testes para a normalidade dos resultados afim de escolher o teste apropriado. À princípio, os testes serão para comparação de dois grupos não pareados (*Unpaired T-test: para amostras normais* ou *Mann-whitney wilcoxon: para amostras não normais*) com 95% de nível de confiança (*p-valor 0.05*), verificando se existe a diferença estatística significativa entre as notas do sistema em relação às do especialista. Logo em seguida, passarão por novos testes que vão medir o nível de confiança e a correlação entre as amostras.

## 5.12 Instrumentação

Para a realização do experimento para avaliação das atividades subjetivas, serão necessários os seguintes instrumentos:

- (i) IDE Netbeans 8.1 para auxílio no desenvolvimento do código responsável pela coleta e execução do experimento;
- (ii) R 3.3.x sendo a ferramenta para análise estatística;

<sup>1</sup> <<http://educacao.uol.com.br/bancoderedacoes>>

- (iii) Hadoop 2.7.2 com Mahout 0.12.2 em ambiente clusterizado com 5 desktops;
- (iv) 5 Desktop com 8GB de RAM, 320GB de HD, Core i5, com SO Ubuntu Server;
- (v) Servidor Glassfish para plataforma online;
- (vi) IBM 4GB de RAM, 40GB de HD, Dual-Core com Suse Enterprise.

### **5.13 Ameaças à validade**

Nos tópicos a seguir, serão apresentadas e detalhadas as ameaças deste experimento.

#### **5.13.1 Ameaças à validade de constructo**

É possível que os níveis de fatores escolhidos para as Dimensões (*sintaxe, semântica e pragmática*) não sejam suficientes para observar diferenças significativas entre a qualidade das avaliações em relação ao sistema e ao especialista.

## 6 RESULTADOS E DISCUSSÃO

Neste capítulo serão abordados os resultados obtidos através da comparação entre a abordagem automática e a avaliação do especialista. Para tanto, nesse capítulo o leitor deverá encontrar os resultados de uma análise descritiva até os resultados com base nos testes estatísticos realizados para se obter as conclusões com base no estudo realizado. Para replicação do experimento, os dados podem ser encontrados neste link<sup>1</sup>.

### 6.1 Análise Descritiva

Visando identificar possíveis indícios e compreender a variância dos dados, nessa seção, os dados serão sumarizados de forma a facilitar a compreensão dos mesmos.

A Análise descritiva realizada nesse trabalho começa com a apresentação dos Boxplots resultantes de cada avaliação, tanto pelo especialista, quanto pelo sistema entre as competências. A Figura 27 apresenta a variância entre as notas. É possível observar que as notas dadas pelo especialista tendem a ser um pouco atenuadas do que as notas do sistema. Podemos destacar leves indícios de padrões entre as notas, por exemplo, a mediana tende a ser aparentemente igual na maioria dos boxplots.

Na competência 1, podemos observar que o sistema aparenta ter avaliado os textos de forma semelhante ao professor. Entretanto, podemos notar que o máximo obtido da avaliação automática foi aproximadamente 150, enquanto que para o especialista, a nota 200 foi a maior obtida.

Na competência 2, uma das competências que cobre a área do discurso da Semântica, notamos que o sistema foi um pouco mais rígido em relação a avaliação, diferente da avaliação do especialista, que foi muito mais dispersa, variando entre 0 e 200, escala de avaliação usada pelo ENEM, e tendo uma mediana um pouco mais baixa.

A competência 3, podemos observar um cenário semelhante ao que aconteceu na competência 2. Entretanto, assim como a avaliação do especialista é reduzida, nota-se também que o sistema reduz ainda mais a variância das notas. Isso pode dar indícios de que a rede neural utilizada conseguiu identificar características e padrões da avaliação humana.

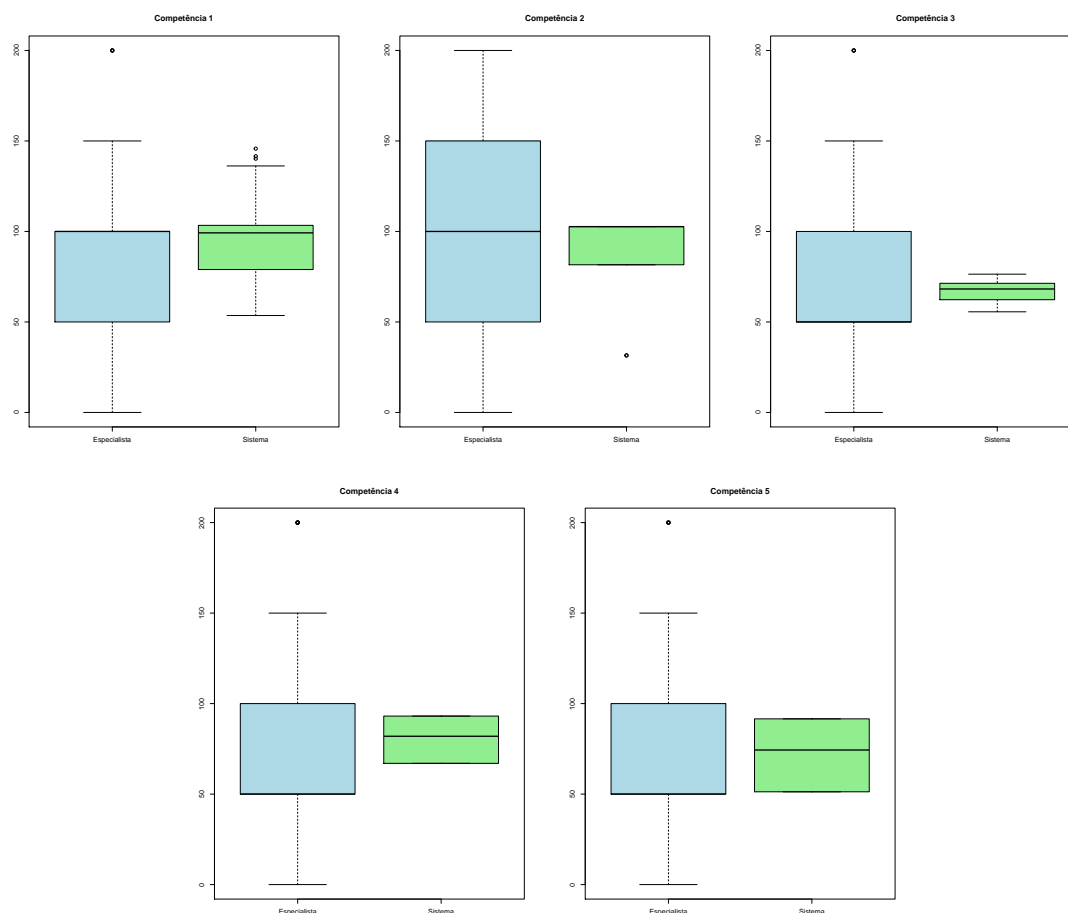
Em relação a competência 4 e 5, podemos observar que existe um padrão entre as notas do Especialista. Tais notas aparentemente tendem a seguir uma escala padronizada. Entretanto, podemos destacar a semelhança na competência 5 entre as avaliações do especialista e do sistema.

Para analisar a distribuição dos dados na competência 1, a Figura 28 apresenta o histograma resultante das avaliações realizadas pelo sistema e especialista. Pode ser observado uma distribuição semelhante com média e mediana próximos de 100. É possível destacar a curva

---

<sup>1</sup> <<https://goo.gl/SYRbfA>>

Figura 27 – Boxplots comparando as avaliações



Fonte – Elaborada pelo autor

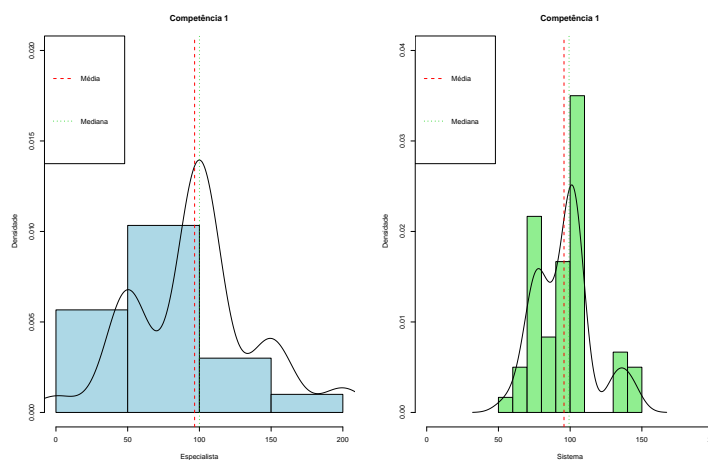
expressa em ambas as distribuições. Nela, podemos notar que o modelo de regressão utilizado como função na rede neural, apresenta uma leve aproximação da avaliação do especialista. Entre outras palavras, há indícios que o sistema consegue avaliar de forma semelhante.

A distribuição dos dados em relação a competência 2 é apresentada na Figura 29. Como pode ser visto, a distribuição apresenta média e mediana próximas. Em contrapartida, é fácil observar que a densidade das avaliações do sistema para notas acima de 100 foi maior do que as notas dadas pelo especialista.

Entretanto, a Figura 30 apresenta a distribuição das avaliações em relação à competência 3. As avaliações feitas pelo especialista mostra um média e mediana entre 50 e 100. Contudo, a mediana encontra-se um pouco afastada da média, e por meio do gráfico também é fácil notar que há concentração maior da nota 50. Em relação a avaliação realizada pelo sistema apresenta média e mediana próximas com uma concentração de notas entre 50 e 100, entretanto, um pouco maior que 50.

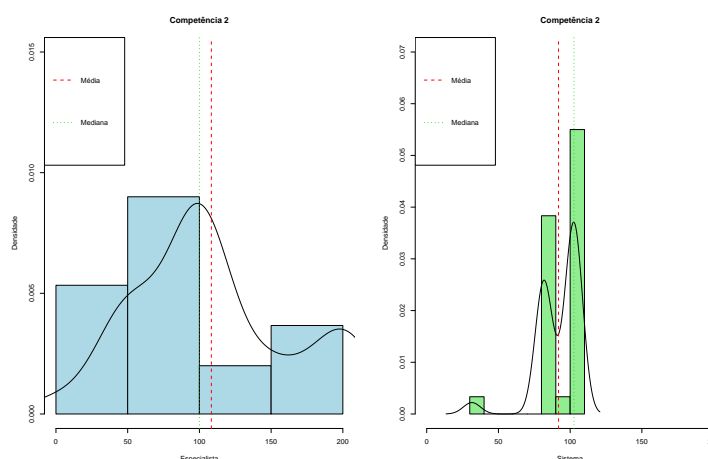
A Figura 31 apresenta os histogramas referentes às avaliações do especialista e do sistema. Com características semelhantes ao histograma da competência 3, a avaliação do especialista

Figura 28 – Histograma comparando as avaliações da competência 1



Fonte – Elaborada pelo autor

Figura 29 – Histograma comparando as avaliações da competência 2



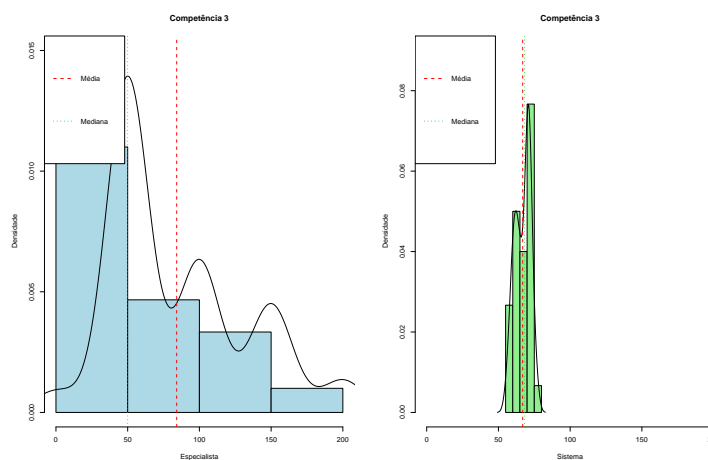
Fonte – Elaborada pelo autor

apresenta densidades maiores entre 50 e 100, com média próximo a 100 e mediana em 50. Já, na avaliação dada pelo sistema, é possível observar que as notas tendem a valores entre 50 e 100, entretanto, com densidades maiores em pontos distantes.

A distribuição das avaliações realizada pelo especialista e pelo sistema pode ser encontrada na Figura 32. O padrão de similaridade com a distribuição apresentada na competência 4 é extremamente alto. É possível observar fortes indícios que ambas as Redes Neurais foram capazes de identificar padrões na avaliação do especialista que fizeram com que a máquina repetisse o padrão. Novamente, as médias e medianas de ambas as distribuições concentram-se entre 50 e 100.

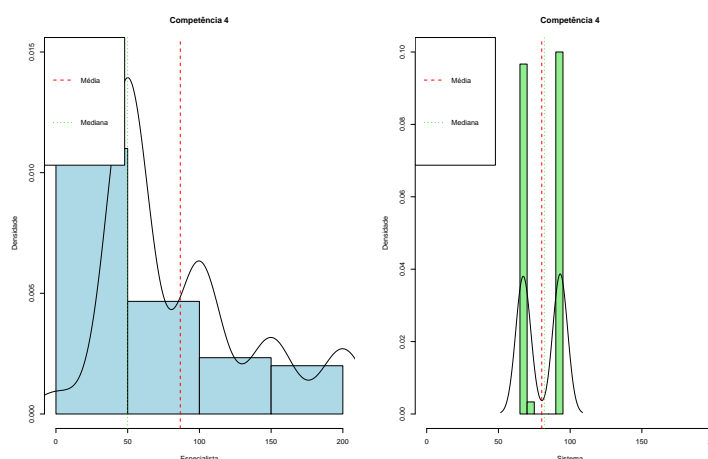
A Figura 33 apresenta a dispersão encontrada na avaliação do especialista. Nela, verifica-se indícios de que a normalidade dos dados não pertença a uma distribuição normal. É possível verificar que o especialista avalia de forma escalar (0 - 100).

Figura 30 – Histograma comparando as avaliações da competência 3



Fonte – Elaborada pelo autor

Figura 31 – Histograma comparando as avaliações da competência 4



Fonte – Elaborada pelo autor

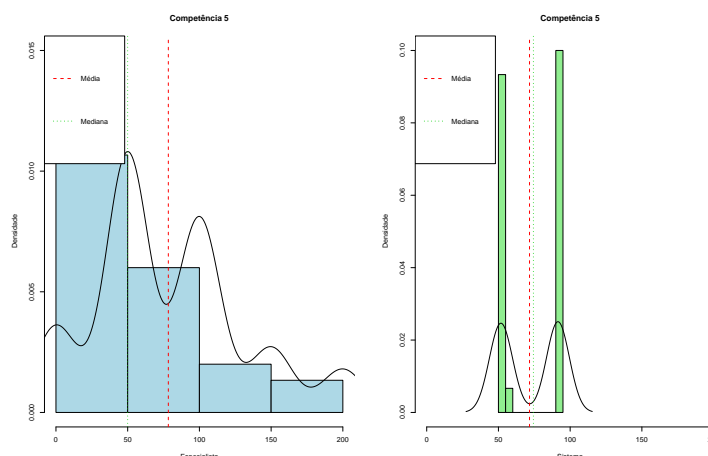
Em relação a dispersão das notas dadas pelo sistema, a Figura 34 apresenta o comportamento da avaliação em todas as 5 competências. Destaca-se o padrão repetitivo na dispersão das competências 4 e 5. Assim como na dispersão das notas do especialista nas respectivas competências.

Os intervalos de confiança relacionados a cada competência são exibidos na Figura 35. Neles é possível observar a taxa de erro em relação a média. Há indícios que a avaliação do sistema seja mais precisa que a avaliação realizada pelo especialista.

## 6.2 Análise Inferencial

Nesta seção, serão apresentados os resultados oriundos da análise estatística inferencial. Nele, os resultados de testes de normalidade, intervalo de confiança e o testes utilizados para a

Figura 32 – Histograma comparando as avaliações da competência 5



Fonte – Elaborada pelo autor

comparação de dois grupos também serão descritos.

A primeira etapa para realizar a análise estatística comparativa entre as avaliações do sistema e do especialista é verificar a normalidade dos dados. Como a própria análise descritiva apresentada anteriormente deu indícios dos dados pertencerem a uma distribuição não-normal, foi utilizado o teste de Shapiro-Wilk com o intuito de validar a normalidade dos dados. A Tabela 6 apresenta os p-valores resultantes do teste de normalidade. Podemos concluir, com os fortes indícios da análise descritiva e os p-valores, que os dados partem de distribuições não-normais. Para tanto, o teste de hipótese escolhido é o Mann-Whitney Wilcoxon.

Tabela 6 – Teste de normalidade Shapiro-Wilk

	Competências	P-valor
<i>Especialista</i>	Competência 1	0.000001514
	Competência 2	0.0000009661
	Competência 3	0.00000005207
	Competência 4	0.00000002026
	Competência 5	0.000004236
<i>Sistema</i>	Competência 1	0.001863
	Competência 2	0.0000000007701
	Competência 3	0.01418
	Competência 4	0.000000001235
	Competência 5	0.000000001235

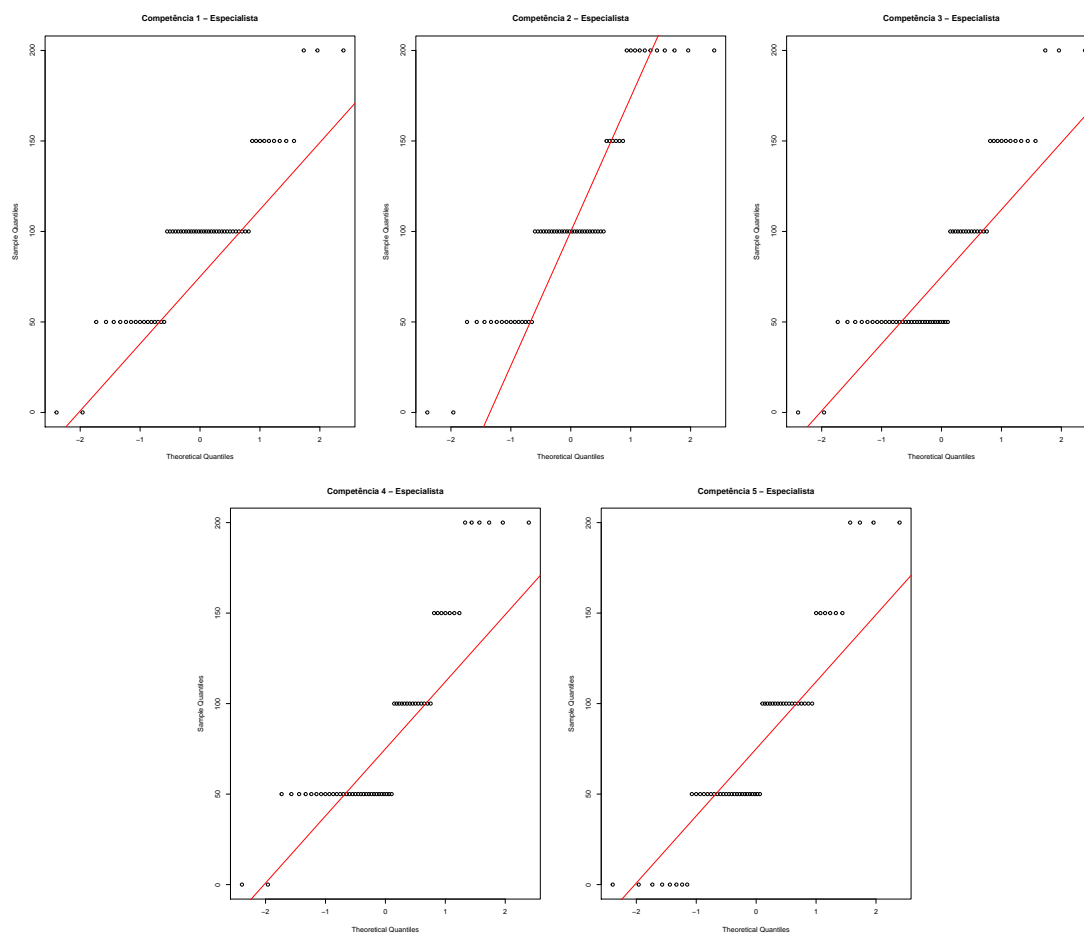
Fonte – Elaborada pelo autor

A Tabela 7 apresenta os p-valores oriundos das comparações entre competências em relação às avaliações do especialista e do sistema. Como pode ser observado, todas as avaliações não apresentaram diferenças estatísticas significativas. Entre outras palavras, tanto a nota do especialista quanto a nota do sistema são praticamente iguais.

Logo, nossas hipóteses nulas são aceitas:



Figura 33 – Dispersão das avaliações dadas pelo Especialista



Fonte – Elaborada pelo autor

Tabela 7 – Teste de Mann-Whitney  
Wilcoxon

Competências	P-valor
Competência 1	0.3214
Competência 2	0.4521
Competência 3	0.1707
Competência 4	0.1697
Competência 5	0.2633

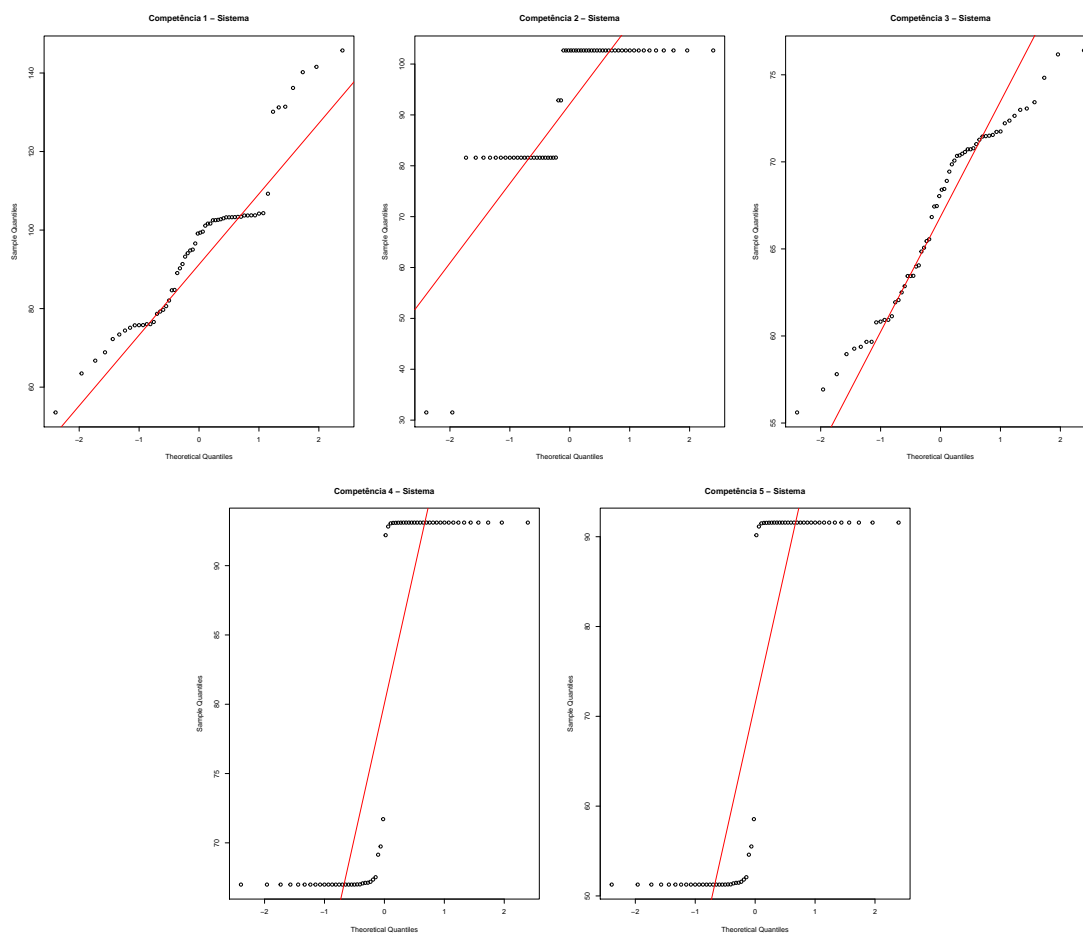
Fonte – Elaborada pelo autor

$H_{1.0}$  : O *score* das avaliações realizadas pelo sistema é equivalente ao *score* das avaliações realizadas pelo especialista em relação à competência 1.

$H_{1.1}$  : O *score* das avaliações realizadas pelo sistema não é equivalente ao *score* das avaliações realizadas pelo especialista em relação à competência 1.

$H_{2.0}$  : O *score* das avaliações realizadas pelo sistema é equivalente ao *score* das avaliações realizadas pelo especialista em relação à competência 2.

Figura 34 – Dispersão das avaliações dadas pelo Sistema



Fonte – Elaborada pelo autor

$H_{2.1}$  : O *score* das avaliações realizadas pelo sistema não é equivalente ao *score* das avaliações realizadas pelo especialista em relação à competência 2.

$H_{3.0}$  : **O *score* das avaliações realizadas pelo sistema é equivalente ao *score* das avaliações realizadas pelo especialista em relação à competência 3.**

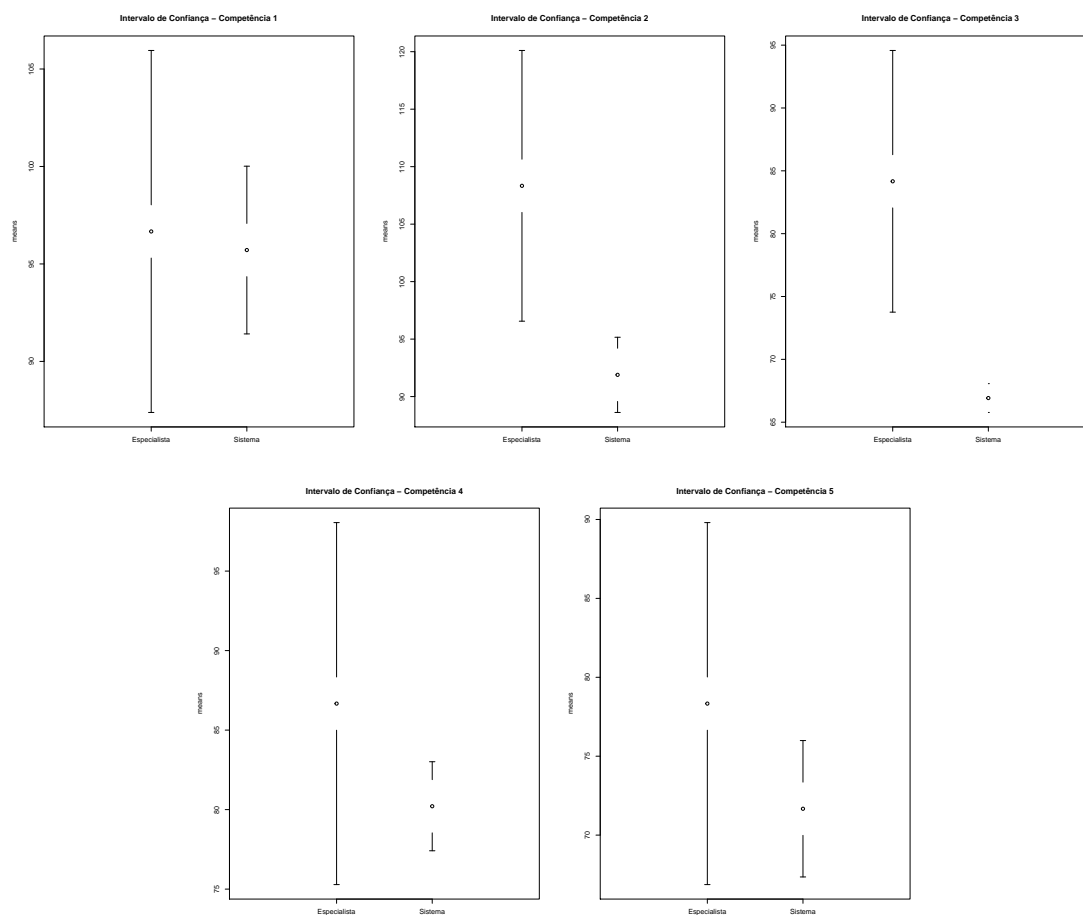
$H_{3.1}$  : O *score* das avaliações realizadas pelo sistema não é equivalente ao *score* das avaliações realizadas pelo especialista em relação à competência 3.

$H_{4.0}$  : **O *score* das avaliações realizadas pelo sistema é equivalente ao *score* das avaliações realizadas pelo especialista em relação à competência 4.**

$H_{4.1}$  : O *score* das avaliações realizadas pelo sistema não é equivalente ao *score* das avaliações realizadas pelo especialista em relação à competência 4.

$H_{5.0}$  : **O *score* das avaliações realizadas pelo sistema é equivalente ao *score* das avaliações realizadas pelo especialista em relação à competência 5.**

Figura 35 – Dispersão das avaliações dadas pelo Sistema



Fonte – Elaborada pelo autor

$H_{5,1}$  : O *score* das avaliações realizadas pelo sistema não é equivalente ao *score* das avaliações realizadas pelo especialista em relação à competência 5.

Entretanto, há necessidade de verificar se existe alguma correlação existente entre as notas do especialista e a nota do sistema. Diante disso, foram analisados as avaliações através do teste de correlação de Pearson. A Tabela 8 apresenta os resultados.

Tabela 8 – Teste de Correlação de Pearson

Competências	P-valor
Competência 1	0.1119424
Competência 2	0.2526567
Competência 3	-0.2676738
Competência 4	0.001223677
Competência 5	0.06915516

Fonte – Elaborada pelo autor

De acordo com a tabela, é possível observar uma correlação entre as notas muito fraca. Com exceção da competência 3, todas as demais competências tiveram uma correlação positiva. Uma vez que o especialista possui uma escala de valores com 0 - 200, faz com o que ele fique preso apenas nesses valores. O Sistema, por sua vez, é capaz de calibrar os valores e avaliar os textos sob aspectos onde o especialista limita-se ao uso de uma escala. Diferenças em relação a escrita, quantidade de erros, argumentação, são fatores que influenciam diretamente o sistema na avaliação. A correlação ser fraca pode ter sido gerada pelo fato de textos com quantidades de erros próximas pertencerem a uma nota igual, perante a escala do especialista, entretanto, o sistema é capaz de avaliar textos com quantidades de erros semelhantes de forma distinta.

É possível perceber, com os gráficos apresentados na seção 6.1, que as avaliações do sistema tendem a seguir um padrão de acordo com a avaliação do especialista. A variância dos dados mostrados nos boxplots apresenta praticamente avaliações com mediana semelhantes e ainda, que mesmo com valores máximos e mínimos maiores na avaliação do especialista, o sistema conseguiu manter sua avaliação próxima. Em contrapartida, nos histogramas, é possível observar que o especialista tende a seguir toda a escala utilizada para avaliar competência por competência. O sistema, por sua vez, criou sua própria escala com nas avaliações realizadas. Nota-se ainda que a escala de avaliação do sistema é contínua, o que faz ter um *score* fidedigno com a avaliação realizada pelo MEC no ENEM.

Sabendo que um terceiro avaliador é utilizado para efetuar a correção de uma redação onde haja uma divergência de notas entre as competências maior que 100 pontos, calculando a diferença entre as notas e efetuando a proporção entre as notas, com a abordagem a taxa de correção para um terceiro avaliador foi de 8.4% enquanto para o ENEM, essa taxa fica, em média maior que 30% segundo Portal MEC nas divulgações anuais da correção do ENEM<sup>2</sup>. Já pra uma diferença entre 80 pontos, a taxa para um terceiro avaliador ser chamado com o sistema corretor é de 21.7%.

Com efeito, podemos então responder nossas questões de pesquisas apresentadas na seção 5.3.

**[QP 1: ]** De que maneira a avaliação automática apresenta a mesma qualidade, ou melhor, apresenta avaliações semelhantes se comparada às correções realizadas por um especialista?

Utilizando a abordagem proposta, foi possível obter resultados na avaliação onde com o nível de confiança de 95%, tais avaliações foram estatisticamente semelhantes. Utilizar modelos para identificar padrões no comportamento humano no que diz respeito à avaliações escritas pode apresentar resultados interessantes. Os modelos devem ser capazes de identificar o padrão de avaliação, mas também conseguir melhorar com a quantidade de textos corrigida.

**[QP 2: ]** Como podemos avaliar a qualidade das correções realizadas pelo sistema corretor?

---

<sup>2</sup> <<http://portal.mec.gov.br/>>

Comparando as correções com a do especialista, foi possível observar uma semelhança entre as notas. Essa semelhança na avaliação identificou não somente que o sistema foi capaz de avaliar seguindo padrões do especialista, mas também foi capaz de ser mais preciso, adotando uma escala contínua para prover a avaliação. Além do mais, o sistema também não somente foi capaz de dar a nota, mas de prover um *feedback* para o aluno, contendo seus erros sobre as áreas do discurso e também sugestões de melhorias.

**[QP 3: ]** De que forma as correções podem ser comparadas às correções realizadas por um especialista (professor)?

A abordagem criada conseguiu comparar as notas do especialista com o sistema através de uma abstração de correção baseado no modelo de avaliação proposto pelo MEC. Com esse modelo, foi possível comparar as correções e indicar o nível de qualidade entre elas.

**[QP 4: ]** Como o modelo de correção automática contribui auxiliando na redução da carga do professor?

Uma vez que a qualidade de avaliação da abordagem proposta foi semelhante a de um especialista, o sistema poderá auxiliar o professor provendo uma análise inicial de correção para que o professor a tenha como base. Assim, o professor terá uma análise onde o que ele precisa fazer é apenas melhorar a análise, e não ter que realizar a análise do princípio.

Portanto, pode ser observado que a nota dada pelo sistema conseguiu ser igual ao do professor e ainda um pouco mais precisa, abrangendo e diferenciando textos com quantidades parecidas de erros, sejam eles relacionados a sintaxe, semântica ou pragmática, mesmo com uma correlação desprezível, o que justifica-se pela diferenciação entre *scores* para cada texto corrigido, levando em consideração diferentes aspectos.

## 7 CONCLUSÃO E LIMITAÇÕES

Neste capítulo será apresentada a conclusão do trabalho desenvolvido bem como as principais limitações da abordagem.

A abordagem proposta se comportou de maneira estável na avaliação dos textos. A avaliação provida pela abordagem teve um nível de similaridade com a avaliação provida pelo especialista em 95% de confiança. Também é possível observar que a precisão das notas, levando em conta o erro estipulado pelo Intervalo de confiança que a máquina tende ser mais precisa que a avaliação do professor.

O Objetivo de melhorar a qualidade da avaliação automática foi atingido, onde o sistema conseguiu prover uma avaliação com *feedbacks* para os textos nas 5 competências utilizadas como abstração para o modelo de correção. A abordagem mostrou-se precisa, sendo uma ferramenta eficaz para auxiliar ao professor no contexto de avaliação de textos. Atuando assim, na diminuição da sobrecarga gerada por atividades escritas em ambientes de Educação Online. Logo, percebemos contribuição na Área de Informática na Educação, auxiliando o professor no processo de avaliação de atividades escritas.

Os modelos e técnicas propostos para cada competência conseguiram identificar padrões na avaliação do especialista que foram capazes de fazer com o que a máquina conseguisse aprender a calibrar as notas para cada texto de forma que a semelhança na avaliação persistisse. Com isso, percebemos contribuição para a área de Processamento de Linguagem natural sob o aspecto da avaliação da pragmática na língua Portuguesa Brasileira.

Além do mais, os modelos criados baseado em Regressão e Redes Neurais são capazes de se aperfeiçoar com o passar do tempo corrigindo textos. Vale ressaltar ainda que as Redes Neurais, juntamente com o modelo de Raciocínio Baseado em Casos consegue aprender e identificar padrões de correções de atividades escritas em diferentes contextos.

As técnicas utilizadas na construção do modelo são capazes de ser generalizadas para diferentes idiomas, tendo a necessidade apenas do dicionário. Entretanto, a abordagem aqui desenvolvida já provê suporte para mais de 20 idiomas para a competência 1, e para as demais competências, a abordagem provê suporte para as línguas que sejam compreensíveis por máquina.

Pelo experimento ter sido executado no pior dos casos, apagando qualquer correção já realizada e considerando a avaliação como sempre sendo a primeira pelo algoritmo, uma vez que a rede neural tem a capacidade de aprender com as correções realizadas, o modelo mostrou-se eficiente em relação à avaliação bem como as sugestões. Foi possível observar que os textos corrigidos relacionados ao mesmo tema se mantiveram com avaliações balanceadas. Tal aspecto também deixou o sistema livre para avaliar e comparar textos disponíveis na WEB. Entretanto, vale ressaltar que os textos passavam por um filtro para verificar a procedência e origem deles.

A abordagem proposta foi capaz também de avaliar os textos de acordo com peculiaridades pertencentes a cada um deles. Entre outras palavras, levando em consideração a quantidade de erros sintáticos, o sistema consegue diferenciar textos com quantidades parecidas de erros. Com isso, a avaliação se torna mais fiel diferenciando textos com mais ou menos erros.

Determinadas técnicas desenvolvidas para a abordagem poderão ser utilizadas para avaliação de outros tipos de textos onde se desejam cobrir determinados aspectos da língua. Entre outras palavras, o sistema poderá ser utilizado, por exemplo, em correção sintática de chat, correção semântica de frases de fórum, entre outros.

A arquitetura do sistema desenvolvida em Kernel-Satélites juntamente com o desenvolvimento de componentes baseado no padrão COSMOS permite a implementação de novas técnicas e modelos para melhorar ainda mais a avaliação. Fazendo com o que o sistema continue tendo suas funcionalidades principais, mas também esteja arquiteturalmente pronto para receber novas funcionalidades, e que tal tarefa não exija muito da implementação e junção dos componentes.

Como limitações, o sistema ainda apresenta uma deficiência na questão de zerar uma redação ou até mesmo de atribuir uma nota 1000, nota máxima para a avaliação. Isto pode ser devido ao fato de que o sistema possar ter que ser treinado também exclusivamente com notas zero e notas 1000. Assim, a abordagem deverá ser capaz de identificar padrões na correção destes tipos de textos. A avaliação da competência 5 ainda precisa ter uma função que seja capaz de identificar se o texto fere algum princípio dos Direitos Humanos. Um trabalho futuro seria implantar uma Rede Neural que fosse calibrada para identificar padrões e correlações entre Direitos humanos e aspectos que vão contra a estes. Também é possível utilizar técnicas de mineração de argumento que sejam capazes de identificar a opinião central do texto e classificá-lo em determinadas classes. Implantar um modelo de correção via OCR capaz de corrigir redações manuscritas, não ficando limitado a textos compreensíveis por máquina.

## REFERÊNCIAS

- AAMODT, A.; PLAZA, E. Case-based reasoning: Foundational issues, methodological variations, and system approaches. *AI communications*, IOS press, v. 7, n. 1, p. 39–59, 1994. Citado na página 28.
- ABRIL, G. Educar para crescer - a importância da escrita. *Acesso em*, v. 5, 2013. Citado na página 17.
- AHO, A. V.; ULLMAN, J. D. *The theory of parsing, translation, and compiling*. [S.l.]: Prentice-Hall, Inc., 1972. Citado na página 25.
- ALMEIDA, M. E. B. de. Educação a distância e tecnologia: contribuições dos ambientes virtuais de aprendizado. In: *Anais do Workshop de Informática na Escola*. [S.l.: s.n.], 2003. v. 1, n. 1, p. 96–107. Citado na página 15.
- AZEVEDO, B. F. T.; BEHAR, P. A.; REATEGUI, E. B. Automatic analysis of messages in discussion forums. *methods*, v. 5, p. 6, 2011. Citado 2 vezes nas páginas 37 e 38.
- BALFOUR, S. P. Assessing writing in moocs: Automated essay scoring and calibrated peer review (tm). *Research & Practice in Assessment*, Research & Practice in Assessment, v. 8, 2013. Citado 3 vezes nas páginas 17, 18 e 53.
- BENGIO, Y.; COURVILLE, A.; VINCENT, P. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, IEEE, v. 35, n. 8, p. 1798–1828, 2013. Citado na página 30.
- BENGIO, Y.; GRANDVALET, Y. No unbiased estimator of the variance of k-fold cross-validation. *Journal of machine learning research*, v. 5, n. Sep, p. 1089–1105, 2004. Citado na página 29.
- BISHOP, C. Pattern recognition and machine learning (information science and statistics), 1st edn. 2006. corr. 2nd printing edn. *Springer, New York*, 2007. Citado na página 30.
- BITTENCOURT, I. I. et al. Research directions on semantic web and education. *Interdisciplinary Studies in Computer Science*, Citeseer, v. 19, n. 1, p. 60–67, 2008. Citado na página 16.
- BRILL, E. Some advances in transformation-based part of speech tagging. *arXiv preprint cmp-lg/9406010*, 1994. Citado na página 25.
- BRILL, E. Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging. *Computational linguistics*, MIT Press, v. 21, n. 4, p. 543–565, 1995. Citado na página 25.
- BRILL, E. Unsupervised learning of disambiguation rules for part of speech tagging. In: SOMERSET, NEW JERSEY: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. *Proceedings of the third workshop on very large corpora*. [S.l.], 1995. v. 30, p. 1–13. Citado na página 25.
- BROWN, P. F. et al. Class-based n-gram models of natural language. *Computational linguistics*, MIT Press, v. 18, n. 4, p. 467–479, 1992. Citado na página 25.



- BRUSILOVSKY, P. Methods and techniques of adaptive hypermedia. In: *Adaptive hypertext and hypermedia*. [S.l.]: Springer, 1998. p. 1–43. Citado na página 16.
- BUCHWALD, H. et al. Bariatric surgery: a systematic review and meta-analysis. *Jama*, American Medical Association, v. 292, n. 14, p. 1724–1737, 2004. Citado na página 32.
- CALLAN, J. P. Passage-level evidence in document retrieval. In: SPRINGER-VERLAG NEW YORK, INC. *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*. [S.l.], 1994. p. 302–310. Citado na página 26.
- CANCHO, R. F. i. Euclidean distance between syntactically linked words. *Physical Review E*, APS, v. 70, n. 5, p. 056135, 2004. Citado na página 26.
- CARROLL, J. B.; DAVIES, P.; RICHMAN, B. *The American Heritage word frequency book*. [S.l.]: Houghton Mifflin Boston, 1971. Citado na página 25.
- CASTRO, M. H. G.; TIEZZI, S. A reforma do ensino médio e a implantação do enem no brasil. *Os desafios da educação no Brasil. Rio de Janeiro: Nova Fronteira*, p. 119–154, 2005. Citado na página 21.
- CHANG, T.-H.; LEE, C.-H. Automatic chinese essay scoring using connections between concepts in paragraphs. In: IEEE. *Asian Language Processing, 2009. IALP'09. International Conference on*. [S.l.], 2009. p. 265–268. Citado na página 36.
- CHANOD, J.-P.; TAPANAINEN, P. Creating a tagset, lexicon and guesser for a french tagger. *arXiv preprint cmp-lg/9503004*, 1995. Citado na página 24.
- CUTRONE, L. A.; CHANG, M. Automarking: automatic assessment of open questions. In: IEEE. *2010 10th IEEE International Conference on Advanced Learning Technologies*. [S.l.], 2010. p. 143–147. Citado na página 35.
- DAELEMANS, W. et al. Mbt: A memory-based part of speech tagger-generator. *arXiv preprint cmp-lg/9607012*, 1996. Citado na página 24.
- FOLEY, B. J.; KOBALISSI, A. Using virtual chat to study in informal learning in online environments. In: *American Educational Research Association: Annual Meeting, San Francisco, CA*. [S.l.: s.n.], 2006. Citado na página 17.
- FONSECA, J. *Estudos de sintaxe-semântica e pragmática do português*. [S.l.]: Porto Editora, 1993. v. 1. Citado na página 21.
- FUNAHASHI, K.-I. On the approximate realization of continuous mappings by neural networks. *Neural networks*, Elsevier, v. 2, n. 3, p. 183–192, 1989. Citado na página 30.
- GOLUB, G. H.; HEATH, M.; WAHBA, G. Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, Taylor & Francis, v. 21, n. 2, p. 215–223, 1979. Citado na página 28.
- GREGG, V. *Word frequency, recognition and recall*. John Wiley & Sons, 1976. Citado na página 25.
- GUPTA, A.; JAIN, R. Visual information retrieval. *Communications of the ACM*, ACM, v. 40, n. 5, p. 70–79, 1997. Citado na página 26.

- HALL, M. et al. The weka data mining software: an update. *ACM SIGKDD explorations newsletter*, ACM, v. 11, n. 1, p. 10–18, 2009. Citado na página 39.
- HAYES-ROTH, F. Rule-based systems. *Communications of the ACM*, ACM, v. 28, n. 9, p. 921–932, 1985. Citado na página 27.
- HEARST, M. A. et al. Support vector machines. *IEEE Intelligent Systems and their Applications*, IEEE, v. 13, n. 4, p. 18–28, 1998. Citado na página 39.
- JACKSON, P.; MOULINIER, I. *Natural language processing for online applications: Text retrieval, extraction and categorization*. [S.l.]: John Benjamins Publishing, 2007. v. 5. Citado na página 17.
- JOSÉ, J.; PAIVA, R.; BITTENCOURT, I. I. Avaliação automática de atividades escritas baseada em algoritmo genético e processamento de linguagem natural: Avaliador ortográfico-gramatical. In: *Anais dos Workshops do Congresso Brasileiro de Informática na Educação*. [S.l.: s.n.], 2015. v. 4, n. 1, p. 95. Citado na página 25.
- KEARNS, M.; RON, D. Algorithmic stability and sanity-check bounds for leave-one-out cross-validation. *Neural computation*, MIT Press, v. 11, n. 6, p. 1427–1453, 1999. Citado na página 29.
- KIM, J.; KANG, J.-H. Towards identifying unresolved discussions in student online forums. *Applied intelligence*, Springer, v. 40, n. 4, p. 601–612, 2014. Citado 3 vezes nas páginas 38, 39 e 41.
- KIM, J.-H. Estimating classification error rate: Repeated cross-validation, repeated hold-out and bootstrap. *Computational statistics & data analysis*, Elsevier, v. 53, n. 11, p. 3735–3745, 2009. Citado na página 29.
- KOHAVI, R. et al. A study of cross-validation and bootstrap for accuracy estimation and model selection. In: STANFORD, CA. *Ijcai*. [S.l.], 1995. v. 14, n. 2, p. 1137–1145. Citado na página 29.
- KOLLER, D.; NG, A. *The Online Revolution: Education at Scale*. [S.l.], 2012. Citado 2 vezes nas páginas 17 e 53.
- MA, Q. et al. Elastic neural networks for part of speech tagging. In: IEEE. *Neural Networks, 1999. IJCNN'99. International Joint Conference on*. [S.l.], 1999. v. 5, p. 2991–2996. Citado na página 24.
- MANNING, D. Introduction. In: *Introduction to Industrial Minerals*. [S.l.]: Springer, 1995. p. 1–16. Citado na página 25.
- MARKOFF, J. *Essay-grading software offers professors a break*. 2013. Citado 2 vezes nas páginas 17 e 53.
- MARTIN, F. G. Will massive open online courses change how we teach? *Communications of the ACM*, ACM, v. 55, n. 8, p. 26–28, 2012. Citado na página 16.
- MARTIN, J. H.; JURAFSKY, D. Speech and language processing. *International Edition*, v. 710, 2000. Citado na página 24.

MCCULLOCH, W. S.; PITTS, W. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, Springer, v. 5, n. 4, p. 115–133, 1943. Citado na página 30.

MORAN, J. M. A educação a distância e os modelos educacionais na formação dos professores. *BONIN, I. et al. Trajetórias e processos de ensinar e aprender: políticas e tecnologias. Porto Alegre: EDIPUCRS*, p. 245–259, 2008. Citado na página 13.

NAGIN, C. National writing project. *Because Writing Matters: Improving Student Writing in Our Schools*, 2003. Citado na página 17.

NAGIN, C. et al. *Because writing matters: Improving student writing in our schools*. [S.l.]: John Wiley & Sons, 2012. Citado na página 17.

PAIVA, R. et al. What do students do on-line? modeling students' interactions to improve their learning experience. *Computers in Human Behavior*, Elsevier, v. 64, p. 769–781, 2016. Citado na página 42.

PAVEZI, A. M. et al. O uso das ferramentas do ambiente virtual de aprendizagem pelos acadêmicos dos cursos de administração e processos gerenciais do nead-cesumar. In: *17º Congresso Internacional de Educação a Distância*. [S.l.: s.n.], 2011. Citado na página 53.

PIECH, C. et al. Tuned models of peer assessment in moocs. *arXiv preprint arXiv:1307.2579*, 2013. Citado na página 16.

RAMOS, J. Using tf-idf to determine word relevance in document queries. In: *Proceedings of the first instructional conference on machine learning*. [S.l.: s.n.], 2003. Citado na página 25.

RAYNER, K.; DUFFY, S. A. Lexical complexity and fixation times in reading: Effects of word frequency, verb complexity, and lexical ambiguity. *Memory & Cognition*, Springer, v. 14, n. 3, p. 191–201, 1986. Citado na página 25.

REYNAR, J. C.; RATNAPARKHI, A. A maximum entropy approach to identifying sentence boundaries. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. *Proceedings of the fifth conference on Applied natural language processing*. [S.l.], 1997. p. 16–19. Citado na página 24.

RICHTER, M. M. On the notion of similarity in case-based reasoning. In: SPRINGER. *Proceedings of the ISSEK94 Workshop on Mathematical and Statistical Methods in Artificial Intelligence*. [S.l.], 1995. p. 171–183. Citado na página 28.

ROBINSON, R. Calibrated peer review: an application to increase student reading & writing skills. *The American Biology Teacher*, BioOne, v. 63, n. 7, p. 474–480, 2001. Citado na página 17.

SALTON, G.; BUCKLEY, C. Term-weighting approaches in automatic text retrieval. *Information processing & management*, Elsevier, v. 24, n. 5, p. 513–523, 1988. Citado na página 26.

SCHANK, R. *Memory-based expert systems*. [S.l.], 1987. Citado na página 28.

SCHANK, R. C. *Dynamic memory: A theory of reminding and learning in computers and people*. [S.l.]: cambridge university press, 1983. Citado na página 28.

- SCHMID, H. Probabilistic part-of-speech tagging using decision trees. In: ROUTLEDGE. *New methods in language processing*. [S.l.], 2013. p. 154. Citado 2 vezes nas páginas 24 e 25.
- SCLATER, N. Web 2.0, personal learning environments, and the future of learning management systems. *Research Bulletin*, v. 13, n. 13, p. 1–13, 2008. Citado na página 16.
- SILVA, A. C. R. D. Educação a distância e o seu grande desafio: o aluno como sujeito de sua própria aprendizagem. 2004. Citado na página 13.
- SLEEMAN, D.; BROWN, J. S. Intelligent tutoring systems. London: Academic Press, 1982. Citado na página 16.
- SOSA, E.; LOZANO-TELLO, A.; PRIETO, Á. E. Semantic comparison of ontologies based on wordnet. In: IEEE. *Complex, Intelligent and Software Intensive Systems, 2008. CISIS 2008. International Conference on*. [S.l.], 2008. p. 899–904. Citado na página 35.
- SPYNS, P. Natural language processing. *Methods of information in medicine*, v. 35, n. 4, p. 285–301, 1996. Citado na página 30.
- STAHL, G.; KOSCHMANN, T.; SUTHERS, D. Computer-supported collaborative learning: An historical perspective. *Cambridge handbook of the learning sciences*, Cambridge, United Kingdom, v. 2006, p. 409–426, 2006. Citado na página 16.
- STANFILL, C.; WALTZ, D. Toward memory-based reasoning. *Communications of the ACM*, ACM, v. 29, n. 12, p. 1213–1228, 1986. Citado na página 28.
- STREHL, A.; GHOSH, J.; MOONEY, R. Impact of similarity measures on web-page clustering. In: *Workshop on artificial intelligence for web search (AAAI 2000)*. [S.l.: s.n.], 2000. v. 58, p. 64. Citado na página 26.
- TENÓRIO, T. et al. A gamified peer assessment model for on-line learning environments in a competitive context. *Computers in Human Behavior*, Elsevier, v. 64, p. 247–263, 2016. Citado na página 18.
- VOUTILAINEN, A. A syntax-based part-of-speech analyser. In: MORGAN KAUFMANN PUBLISHERS INC. *Proceedings of the seventh conference on European chapter of the Association for Computational Linguistics*. [S.l.], 1995. p. 157–164. Citado na página 24.
- VRASIDAS, C.; MCISAAC, M. S. Factors influencing interaction in an online course. *American Journal of Distance Education*, Taylor & Francis, v. 13, n. 3, p. 22–36, 1999. Citado na página 17.
- WANGENHEIM, C. G. V.; WANGENHEIM, A. V. *Raciocínio baseado em casos*. [S.l.]: Editora Manole Ltda, 2003. Citado na página 27.
- WILBUR, W. J.; SIROTKIN, K. The automatic identification of stop words. *Journal of information science*, Sage Publications Sage CA: Thousand Oaks, CA, v. 18, n. 1, p. 45–55, 1992. Citado na página 26.
- WILKENS, M.; KUPIEC, J. Training hidden markov models for part-of-speech tagging'. *Internal document, Xerox Corporation*, 1995. Citado na página 24.
- YOUNGER, D. H. Recognition and parsing of context-free languages in time n<sup>3</sup>. *Information and control*, Elsevier, v. 10, n. 2, p. 189–208, 1967. Citado na página 25.

ZHANG, Y.; CALLAN, J.; MINKA, T. Novelty and redundancy detection in adaptive filtering. In: ACM. *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*. [S.l.], 2002. p. 81–88. Citado na página 26.

ZUPANC, K.; BOSNIC, Z. Automated essay evaluation augmented with semantic coherence measures. In: IEEE. *2014 IEEE International Conference on Data Mining*. [S.l.], 2014. p. 1133–1138. Citado na página 35.