UNIVERSIDADE DE ITAÚNA
PRÓ-REITORIA DE ENSINO
COORDENAÇÃO DE CIÊNCIA DA COMPUTACAO
BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

Eugênio Cunha

APRENDIZADO DE MÁQUINA APLICADO À VALORAÇÃO DE REDAÇÕES

ITAÚNA
2017

EUGÊNIO CUNHA

APRENDIZADO DE MÁQUINA APLICADO À VALORAÇÃO DE REDAÇÕES

Projeto submetido à Coordenadoria do Curso de Bacharelado em Ciência da Computação da Universidade de Itaúna - Campus Verde, como requisito parcial para obtenção do grau de Bacharel em Ciência da Computação.

Área de pesquisa: Aprendizagem de Máquina

Orientador: Prof. Dr. Marco Túlio Alves N Rodrigues

Itaúna
2017

UNIVERSIDADE DE ITAÚNA
COORDENAÇÃO DE CIÊNCIA DA COMPUTAÇÃO

EUGÊNIO CUNHA

Este projeto foi julgada adequada para a obtenção do Grau de Bacharel em Ciência da Computação, sendo aprovada pela coordenação de ciência da computação do curso de Bacharelado em Ciência da Computação do Campus verde da Universidade de Itaúna e pela banca examinadora:

_____

Orientador: Prof. Dr. Marco Túlio Alves N Rodrigues
Universidade de Itaúna- UIT

_____

Avaliador: Coord. Prof. Dr. Felipe Domingos da Cunha
Universidade de Itaúna- UIT

Itaúna, 19 de Junho de 2017

Dedico este trabalho a Davi, meu filho, sempre preocupado em proporcionar um minuto de pausa para brincadeiras durante minhas horas de estudo, “meus melhores minutos”!

“Os computadores são incrivelmente rápidos, precisos e burros; os homens são incrivelmente lentos, imprecisos e brilhantes; juntos, seus poderes ultrapassam os limites da imaginação”.

Albert Einstein

Resumo

Este trabalho baseia-se na avaliação das competências exigidas em um texto de redação do tipo dissertativo-argumentativo com temas diversificados de ordem social, científica, cultural ou política. Fundamenta-se no estudo das técnicas de Aprendizado de Máquina supervisionado que provê uma gama diversificada de algoritmos poderosos para classificação de textos.

O objetivo deste trabalho é classificar as competências exigidas em um texto de redação do tipo dissertativo-argumentativo a partir do treinamento de um algoritmo de Aprendizado de Máquina, com base em um *corpus* de redações avaliadas.

A compilação de um *corpus* de redações para treinamento e teste de um algoritmo de Aprendizado de Máquina exigiu a prática de extração de informações que compreende técnicas e algoritmos que realiza duas tarefas importantes: a identificação de informações desejadas a partir de documentos estruturados e não-estruturados, e o armazenamento dessas informações em um formato apropriado para uso futuro.

Afim de se avaliar a eficácia dos classificadores, vários experimentos foram executados usando um *corpus* extraído do banco de redações de um serviço que estimula o estudante treinar a produção de textos, em especial do gênero dissertativo-argumentativo.

O resultado geral de classificação das competências exigidas em um texto de redação obtidas experimentalmente mostraram que o sistema proposto é comparável à avaliação manual de avaliadores capacitados.

Palavras-chaves: Aprendizado de máquina, Banco de Redações, Classificação, Redação

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Abreviações

ADABOOST ***Algoritmo Adaptive Boosting***

AM	Aprendizado de Máquina
BOW	*Bag of Words*
CEBRASPE	Centro Brasileiro de Pesquisa em Avaliação e Seleção e de Promoção de Eventos
CSF	Ciência sem fronteiras
ENEM	Exame Nacional do Ensino Médio
GPL	*General Public License*
HTML	*HyperText Markup Language*
IA	Inteligência artificial
INEP	Instituto Nacional de Estudos e Pesquisas Educacionais
JSON	*JavaScript Object Notation*
PLN	Processamento de linguagem natural
SISU	Sistema de Seleção Unificada
UNB	Universidade de Brasília
TP	True Positive
FN	False Negative
FP	False Positive
TN	True Negative
ROC	Receiver Operating Characteristic Curve

TPR True Positive Rate

FPR False Positive Rate

CAPÍTULO 1

INTRODUÇÃO

O desenvolvimento de uma redação e uma atividade prática presente na cultura civilizada desde a invenção da escrita. (LARA, 1994) cita que na década de 70 iniciou-se processo de redemocratização que consequentemente restitui a palavra ao estudante. O decreto 79.298, de 24 de Fevereiro de 1977 definiu a volta da redação à escola pela "inclusão obrigatória da prova ou questão de redação em língua portuguesa" nos concursos e vestibulares (Art. 1º, alínea d).

Um bom desempenho em redação no Exame Nacional de Ensino Médio - ENEM é um requisito para ser aprovado no processo seletivo de acesso a inúmeras universidades públicas (SISU, 2017) e a importantes programas de governo como Ciência Sem Fronteiras (CSF, 2017).

Em todo processo seletivo é comum o uso de marcações em gabaritos afim de automatizar o processo de correção, uma alternativa rápida e segura, até mesmo aplicações de provas eletrônicas são cada vez mais comum. É notável que todo o processo evoluiu com objetivo de agilidade, confiança e segurança do resultado. Entretanto segundo o edital do ENEM 2016, a avaliação das competências definidas na Tabela 1 de um texto de redação, ainda depende exclusivamente da supervisão de duas ou mais pessoas envolvidas (INEP, 2016).

A redação é aplicada no ENEM desde a primeira edição 1998, hoje o maior exame do Brasil, que na edição de 2016 contou com 8.627.195 escritos confirmados, e a participação direta de 11.360 profissionais externos na correção de 5.825.134 redações, entre eles, 378 supervisores e 10.982 avaliadores de acordo com a (CEBRASPE, 2016).

Segundo o edital do ENEM 2016, cada redação foi avaliada por, pelo menos, dois avaliadores, de forma independente, contabilizando um número mínimo de 11.650.268 avaliações manuais, das competências exigidas em um texto de redação pelo ENEM (INEP, 2016).

A hipótese desta monografia é que a classificação das competências de uma redação por um algoritmo de Aprendizado de Máquina pode ser tão eficiente e seguro quanto o processo de avaliação manual.

1.1 Definição do Problema de Pesquisa

Dado um corpus de redações classificar as competências exigidas em um texto de redação do tipo dissertativo-argumentativo.

1.2 Motivação

Com crescente volume e variedade de dados disponíveis, o processamento computacional que está mais barato e mais poderoso, e o armazenamento de dados de forma acessível, o Aprendizado de Máquina está no centro de muitos avanços tecnológicos, alcançado áreas antes exclusivas de seres humanos. Os carros autônomos do Google são o exemplo de uma atividade antes exclusiva de um humano e hoje exercida e aperfeiçoada por algoritmos de Aprendizado de Máquina (WAYMO, 2017).

Aplicações de Aprendizado de Máquina estão presentes na nossa vida cotidiana como, resultados de pesquisa web, análise de sentimento baseado em texto e na detecção de fraudes em operações com cartões de crédito (BATISTA *et al.*, 1999).

1.3 Objetivos Gerais e Específicos

Este trabalho tem como objetivo geral aplicar Aprendizado de Máquina na classificação das competências exigidas em um texto de redação do tipo dissertativo-argumentativo.

1.3.1 Objetivos Específicos

O método de construção do conhecimento deste trabalho terá como fundamentos processos de pesquisas relacionadas às áreas descritas. O mesmo será dividido em

etapas dentro do escopo geral de forma detalhada e refinada para alcançar o objetivo geral acima, são particularizadas como os seguintes objetivos específicos:

- Percorrer o banco de redações, filtrar e coletar redações avaliadas;
- Normalizar os textos coletados, separar o tema, título, texto e competências avaliadas em uma estrutura de dados;
- Montar um fluxo de trabalho utilizando a ferramenta para mineração de dados *Orange* (DEMŠAR *et al.*, 2013);
- Ajustar e treinar os modelos classificadores com o corpus de redações;
- Realizar testes de acurácia, *overfitting* e *noise* sobre o modelo induzido;
- Representar e comparar graficamente os resultados obtidos;

1.4 Contribuições

O presente estudo contribuirá na área do Aprendizado de Máquina e diretamente no processo de classificação de um texto em prosa do tipo dissertativo-argumentativo.

1.5 Organização do trabalho

Capitulo 2: Trabalhos Relacionados cita alguns dos trabalhos lidos para embasamento teórico que serviram de base para solucionar o problema proposto.

Capitulo 3: Fundamentação Teórica apresenta todo o embasamento teórico utilizado para o desenvolvimento do estudo.

Capitulo 4: Método proposto apresenta as etapas passo a passo para desenvolver e resolver o problema proposto deste trabalho.

Capitulo 5: Resultados Preliminares apresenta os resultados obtidos do trabalho desta pesquisa.

Capitulo 6: Plano de Trabalho organiza um conjunto de objetivos, processos e etapas que visa a conclusão do trabalho.

CAPÍTULO 2

TRABALHOS RELACIONADOS

Este capítulo é destinado a listar uma sequência de artigos científicos nas áreas abordadas neste estudo com o objetivo de adquirir conhecimento para elaboração deste trabalho.

2.1 Matriz de Referência da Redação do ENEM

De acordo com (SILVA; CARVALHO, 2017), à prova de redação do ENEM é avaliada levando em conta uma matriz de referência listada na Tabela 1. Essa matriz, desenvolvida pelo (INEP, 2016), com a colaboração de especialistas, foi elaborada com o objetivo de operacionalizar o exame. A matriz apresenta cinco competências, para cada competência expressa para redação existem níveis de conhecimento associados de 0 a 5.

De acordo com (BRAGA, 2015), no texto de redação, o candidato defenderá uma opinião a respeito do tema proposto, de forma coerente e coesa, embasado em argumentos consistentes. O texto será redigido respeitando a escrita formal da Língua Portuguesa. Ao fim, o candidato elabora uma proposta de intervenção social para o problema apresentado no desenvolvimento do texto que respeite os direitos humanos.

Tabela 1: Matriz de referência elaborada pelo INEP.

I	**Demonstrar domínio da norma padrão da língua escrita.**	
	0	Demonstra desconhecimento da modalidade escrita formal da língua portuguesa.
	1	Demonstra domínio precário da modalidade escrita formal da língua portuguesa, de forma sistemática, com diversificados e frequentes desvios gramaticais, de escolha de registro e de convenções da escrita.

	2	Demonstra domínio insuficiente da modalidade escrita formal da língua portuguesa, com muitos desvios gramaticais, de escolha de registro e de convenções da escrita.
	3	Demonstra domínio mediano da modalidade escrita formal da língua portuguesa e de escolha de registro, com alguns desvios gramaticais e de convenções da escrita.
	4	Demonstra bom domínio da modalidade escrita formal da língua portuguesa e de escolha de registro,com poucos desvios gramaticais e de convenções da escrita.
	5	Demonstra excelente domínio da modalidade escrita formal da língua portuguesa e de escolha de registro. Desvios gramaticais ou de convenções da escrita serão aceitos somente como excepcionalidade e quando não caracterizem reincidência.
II	**Compreender a proposta de redação e aplicar conceitos das varias áreas de conhecimento para desenvolver o tema, dentro dos limites estruturais do texto dissertativo-argumentativo em prosa.**	
	0	“Fuga ao tema/não atendimento à estrutura dissertativo-argumentativa”.
	1	Apresenta o assunto, tangenciando o tema ou demonstra domínio precário do texto dissertativo-argumentativo, com traços constantes de outros tipos textuais
	2	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.
	3	Desenvolve o tema por meio de argumentação previsível e apresenta domínio mediano do texto dissertativo-argumentativo, com proposição, argumentação e conclusão.
	4	Desenvolve o tema por meio de argumentação consistente e apresenta bom domínio do texto dissertativo-argumentativo, com proposição, argumentação e conclusão.
	5	Desenvolve o tema por meio de argumentação consistente, a partir de um repertório sócio cultural 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.**	
	0	Apresenta informações, fatos e opiniões não relacionados ao tema e sem defesa de um ponto de vista.
	1	Apresenta informações, fatos e opiniões pouco relacionados ao tema ou incoerentes e sem defesa de um ponto de vista.
	2	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.
	3	Apresenta informações, fatos e opiniões relacionados ao tema, limitados aos argumentos dos textos motivadores e pouco organizados, em defesa de um ponto de vista.
	4	Apresenta informações, fatos e opiniões relacionados ao tema, de forma organizada, com indícios de autoria, em defesa de um ponto de vista.
	5	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.**	
	0	Não articula as informações.
	1	Articula as partes do texto de forma precária.
	2	Articula as partes do texto, de forma insuficiente, com muitas inadequações e apresenta repertório limitado de recursos coesivos.
	3	Articula as partes do texto, de forma mediana, com inadequações, e apresenta repertório pouco diversificado de recursos coesivos.
	4	Articula as partes do texto com poucas inadequações e apresenta repertório diversificado de recursos coesivos.
	5	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.**	
	0	Não apresenta proposta de intervenção ou apresenta proposta não relacionada ao tema ou ao assunto.

1	Apresenta proposta de intervenção vaga, precária ou relacionada apenas ao assunto.
2	Elabora, de forma insuficiente, proposta de intervenção relacionada ao tema, ou não articulada com adiscussão desenvolvida no texto.
3	Elabora, de forma mediana, proposta de intervenção relacionada ao tema e articulada à discussão desenvolvida no texto.
4	Elabora bem proposta de intervenção relacionada ao tema e articulada à discussão desenvolvida no texto.
5	Elabora muito bem proposta de intervenção, detalhada, relacionada ao tema e articulada à discussão desenvolvida no texto.

2.2 Aprendizado de Máquina

Segundo (MONARD; BARANAUSKAS, 2003) “A indução é a forma de inferência lógica que permite obter conclusões genéricas sobre um conjunto particular de exemplos.” Na indução, um conceito é aprendido efetuando-se inferência indutiva sobre os exemplos apresentados. O aprendizado indutivo pode ser dividido em supervisionado e não-supervisionado como ilustrada a Figura 1.

No aprendizado não-supervisionado, o algoritmo de aprendizado analisa os exemplos fornecidos e tenta determinar se alguns deles podem ser agrupados de alguma maneira, formando *clusters* ou agrupamentos. Já no aprendizado supervisionado é fornecido ao algoritmo de aprendizado um conjunto de exemplos de treinamento para os quais o rótulo da classe associada é conhecido.

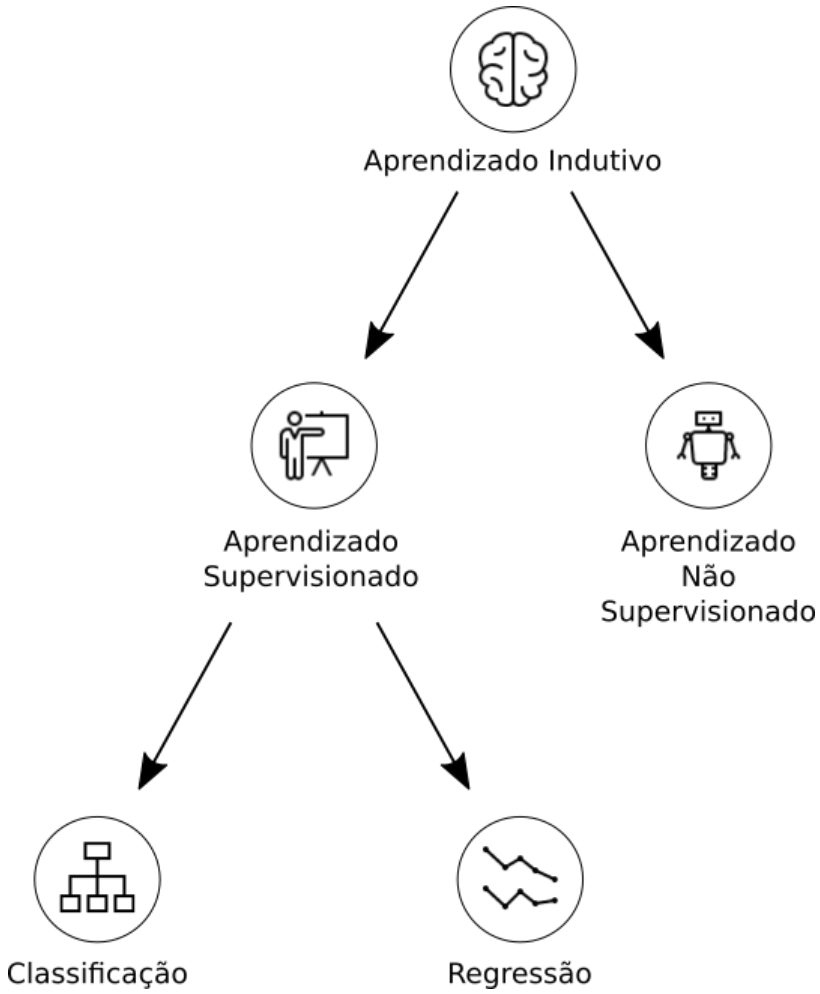

Figura 1: Árvore hierárquica do aprendizado indutivo, a qual é dividida em algoritmos supervisionado e não- supervisionado.

De acordo com (MOTTA, 2016), classificadores são utilizados para a predição de classes de objetos e pode ser dita como o processo de generalização dos dados a partir de diferentes instâncias. Existe uma tendência de se referir a problemas com uma resposta quantitativas como problemas de regressão e aqueles com uma resposta qualitativa como problemas de classificação.

Dado um conjunto de exemplos como ilustrado na Figura 2, os classificadores devem encontrar uma função geral capaz de prever adequadamente as saídas para novos exemplos, após o treinamento, o classificador é avaliado e se necessário o processo de classificação pode ser ajustado usando o conhecimento sobre o domínio do problema para escolher os dados de entrada ao algoritmo de aprendizado.

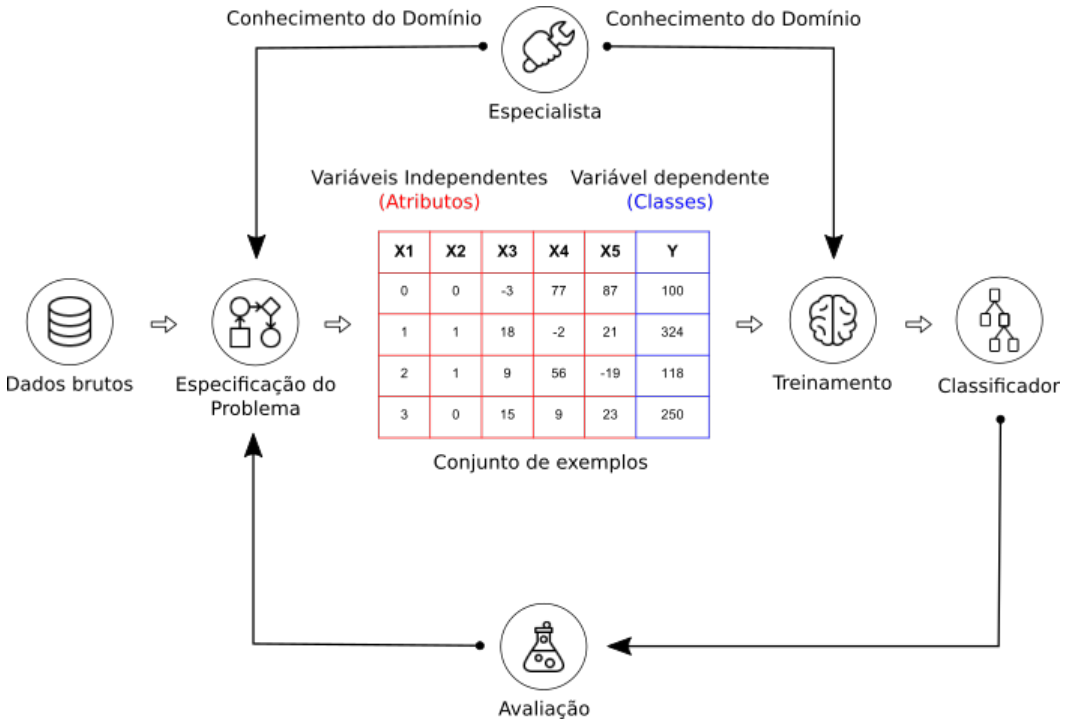

Figura 2: Fluxo do processo de classificação, o modelo encontra uma função geral capaz de prever as saídas, a especificação do problema pode ser reajustada com o conhecimento do domínio para obter um melhor resultado.

2.3 Ferramenta para mineração de dados

Diversas ferramentas disponíveis para exploração de dados dispõem de soluções para o processamento e a análise das informações de forma ágil e simples. Em uma análise comparativa (BOSCARIOLI; VITERBO; TEIXEIRA, 2014) demonstra que não existe uma única ferramenta com características melhores para todas as aplicações em mineração de dados.

Em um estudo que comparou quatro ferramentas (KMINE, *Orange*, Tanagra, Weka), todas de código aberto, gratuitas e muito utilizadas na pesquisa e na academia, (WAHBEH *et al.*, 2011) concluiu que a ferramenta Weka apresentou o melhor desempenho, seguido pelo *Orange*, e, depois, pelo KMINE e Tanagra.

Para este trabalho, foi escolhida a ferramenta *Orange* (DEMŠAR *et al.*, 2013) por ser muito utilizada no meio acadêmico, ter sido bem avaliada quando comparada a outras, ser utilizada como uma biblioteca na linguagem Python (ROSSUM; DRAKE, 2003) e utiliza a conceituada biblioteca *Scikit-learn* (PEDREGOSA *et al.*, 2011) internamente para Aprendizado de Máquina.

A ferramenta *Orange* na atual versão 3.4 desenvolvida pelo laboratório de Inteligência Artificial da Faculdade de Computação e Ciência da Informação da Universidade de *Ljubljana* na Eslovênia sob a licença GPL, possui uma interface gráfica denominada *Orange Canvas*. Por meio de sua interface ilustrada na Figura 3 é possível conectar e interligar os objetos montando um fluxo de trabalho para o desenvolvimento de modelos de classificação, incluindo *Adaboost*, *Naive Bayes*, Regras de Decisão, Árvores de Decisão, etc..

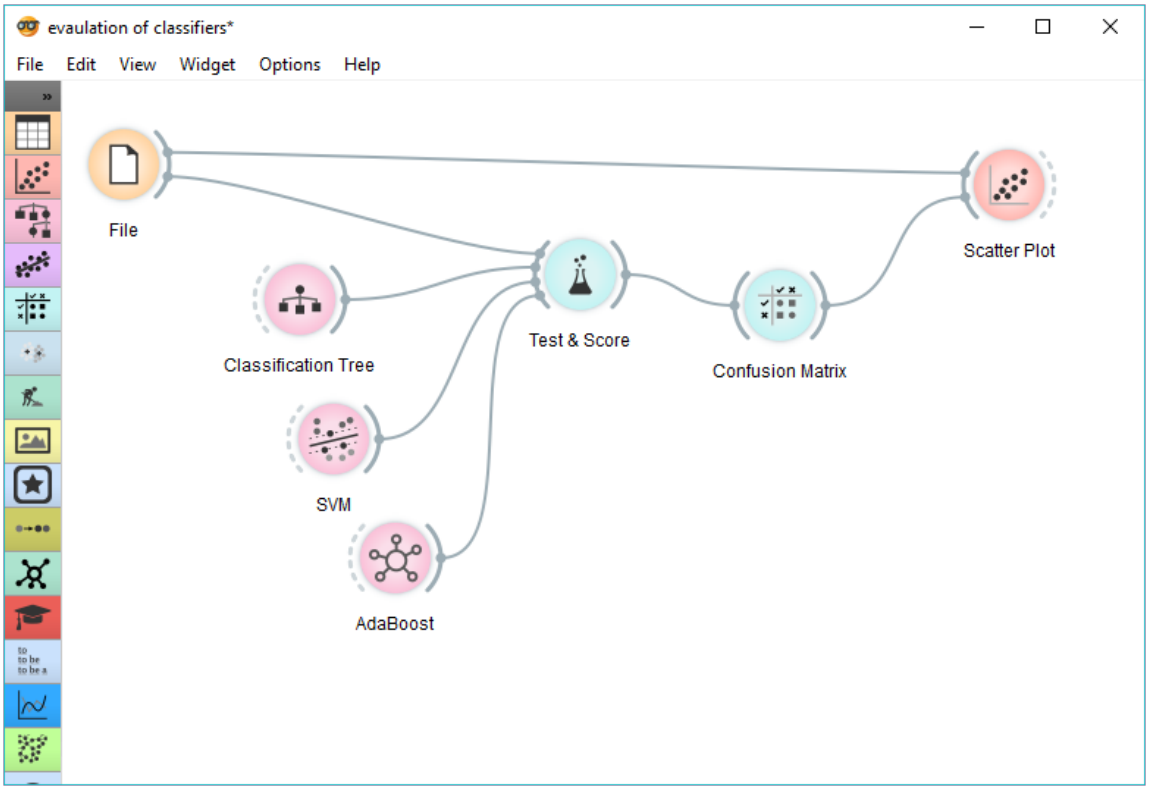

Figura 3: Ferramenta de mineração de dados *Orange Canvas* executando teste de desempenho dos classificadores AdaBoost, SVM e *Classification Tree* na matriz de confusão.

CAPÍTULO 3

FUNDAMENTAÇÃO TEÓRICA

Este capítulo irá abordar alguns conceitos que serão necessários durante o desenvolvimento deste trabalho para que o projeto tenha fundamentos concretos.

3.1 Processamento de linguagem natural

O desenvolvimento de modelos computacionais para a realização de tarefas que dependem de informações expressas em uma língua natural também conhecido como processamento de linguagem natural, cresce desde o início da década de 1990, segundo (VIEIRA; LOPES, 2010) "o crescimento da internet e a profusão de textos disponíveis direcionaram os esforços do PLN para o tratamento de textos mais do que para o discurso falado". Ainda segundo o autor neste mesmo período iniciou as pesquisas sobre conjuntos de textos sobre um domínio de conhecimento onde cada uma das suas palavras foram identificadas segundo sua função sintática.

O processamento de linguagem natural segundo o trabalho de (TEIXEIRA; AZEVEDO, 2011) pode envolver diversas etapas: *tokenizer* ou divisão do texto em termos mais simples, *phrase chunking* ou análise sintática e *part-of-speech tagging* ou identificação da classe gramatical das palavras. Existem também outras etapas mais específicas, como a identificação de entidades (datas, nomes, número, etc), entretanto algumas destas etapas são essenciais para o processamento correto do texto, como o *tokenizer*.

3.2 *Bag of words*

Segundo (ALVES, 2010) o BOW (*Bag of Words*) é o modelo mais utilizado em aplicações de classificação de texto. Com baixo custo em termos de processamento este modelo transforma a cadeia de caracteres de um documento num conjunto de

palavras, registrando além da presença de uma palavra, a sua frequência.

Entretanto ainda segundo o autor, propriedades básicas do texto, como a ordem em que as palavras ocorrem e a pontuação, são ignoradas, além da incapacidade em capturar a semântica do texto, isto é, há palavras com significados distintos que apesar de serem exatamente iguais têm significados diferentes, dependendo do contexto em que são utilizadas.

Obviamente, termos que aparecem em todos os documentos são denominados *stop words* e não serão analisados, geralmente são os pronomes, artigos e as preposições. Estes termos não são úteis, visto que têm uma semântica fraca e somente desempenham um papel funcional no texto. Para melhorar os métodos de processamento normalmente são removidas, em vários casos a remoção das *stop words* não traz consequências graves segundo o autor.

Na Tabela 2, $\mathbf{w}_i$ representa uma palavra, $\mathbf{d}_j$ representa um documento e $\mathbf{p}_{ij}$ o peso atribuído a cada palavra no documento.

Tabela 2: Modelo *Bag of Words* é maneira mais comum de representar coleções de documentos no qual cada documento é representado por um vetor e cada palavra da coleção representa uma dimensão do vetor.

	$\mathbf{w}_1$	$\mathbf{w}_2$	$\mathbf{w}_3$	$\mathbf{w}_{...}$	$\mathbf{w}_n$
$\mathbf{d}_1$	p_{11}	p_{12}	p_{13}	$p_{...}$	p_{1n}
$\mathbf{d}_2$	p_{21}	p_{22}	p_{23}	$p_{...}$	p_{2n}
$\mathbf{d}_3$	p_{31}	p_{32}	p_{33}	$p_{...}$	p_{3n}
$\mathbf{d}_4$	p_{41}	p_{42}	p_{43}	$p_{...}$	p_{4n}
$\mathbf{d}_n$	p_{n1}	p_{n2}	p_{n3}	$p_{...}$	p_{nn}

Ainda segundo (ALVES, 2010) existem várias medidas para calcular os valores dos pesos de p_{ij}. Essas medidas podem ser classificadas em dois tipos distintos: baseadas em frequências e binárias. Os pesos baseados em frequência visam contabilizar o número de ocorrências de um dado termo num determinado documento, servindo como base para diversas medidas estatísticas e os pesos binários indicam a ocorrência ou não de um dado termo num determinado documento.

3.3 Aprendizado de Máquina supervisionado

De acordo com (FREITAS *et al.*, 2005) no decorrer da última década, o Aprendizado de Máquina tem atestado ser uma ferramenta eficiente para realizar tarefas linguísticas, que de outro maneira seria impossível devido à enorme quantidade de mão-de-obra e tempo necessário.

Segundo o trabalho de (BATISTA *et al.*, 2003), na área de Aprendizado de Máquina foram propostos diversos paradigmas aptos a aprender a partir de um conjunto de exemplos. Uma premissa básica para todos os paradigmas de Aprendizado de Máquina supervisionado é que o conceito a ser induzido deve ser referente ao caso observado, ou seja, cada exemplo deve estar denominado com a classe a qual pertence.

Entretanto o autor cita que “Se todos os casos são memorizados, o classificador pode se tornar lento e difícil de manusear. O ideal é reter casos prototípicos que juntos resumem toda uma informação importante.”.

3.4 Classificador *Adaboost*

Boosting é um conjunto de métodos e procedimentos de Aprendizado de Máquina que mescla vários classificadores fracos para aperfeiçoar a *acurácia* geral. O algoritmo atualiza os pesos dos exemplos a cada iteração e cria um classificador adicional. Os classificadores são combinados por um esquema simples de votação. O algoritmo mais famoso baseado em Boosting é o Adaboost (Algoritmo Adaptive Boosting) que atualiza os pesos dos exemplos em que os classificadores anteriores cometeram erros, focando o classificador adicional nos exemplos mais difíceis (DUARTE, 2009).

Merjildo et al. (2013) mostra as propriedades que simplificam o uso do classificador *AdaBoost*. O autor cita em seu trabalho que os parâmetros empregados para analisar dados de grandes proporções e as margens entre as classes podem ser mais precisas do que em outros métodos e o custo computacional é baixo, dado que corresponde a um programa de complexidade linear e evita o uso de componentes computacionais pesados.

Por fim, este classificador escolhido para essa pesquisa, tem como objetivo reconhecer padrões complexos a partir de combinações de características simples

baseadas em formas que podem estar presentes em textos de redação do tipo dissertativo-argumentativo.

3.5 Métricas utilizadas

Avaliar o desempenho de um modelo é um dos estágios principais deste estudo, a avaliação se baseia nas predições que são produzidas pelo modelo induzido.

Na comparação dos resultados experimentais serão utilizadas sete métricas de avaliações sobre modelo supervisionado *AdaBoost*, que são:

Matriz de confusão - A matriz de confusão de uma hipótese H oferece uma medida efetiva do modelo de classificação, ao mostrar o número de classificações corretas versus as classificações preditas para cada classe, sobre um conjunto de exemplos verdadeiros como descrito na Tabela 3.

		Valor verdadeiro	
		positivos	negativos
Valor previsto	positivos	VP	FP
	negativos	FN	VN

Tabela 3: Matriz de confusão ou tabela de contingência

Acurácia - porcentagem de amostras positivas e negativas classificadas corretamente sobre a soma de amostras positivas e negativas dada pela fórmula da equação 3.1:

$$acuracia = \frac{TP+TN}{VP+VN+FP+FN} \tag{3.1}$$

Sensitividade - ou *recall* é a porcentagem de amostras positivas classificadas corretamente sobre o total de amostras positivas dada pela fórmula da equação 3.2:

$$sensitividade = \frac{TP}{TP+FN} = \frac{TP}{Positivo} \tag{3.2}$$

Precisão - ou *precision* é a porcentagem de amostras positivas classificadas corretamente sobre o total de amostras classificadas como positivas dada pela fórmula da equação 3.3:

$$precisao = \frac{TP}{TP+FP} \tag{3.3}$$

Especificidade - ou *specificity* é a porcentagem de amostras negativas identificadas corretamente sobre o total de amostras negativas dada pela fórmula da equação 3.4:

$$especificidade = \frac{TN}{TN+FP} = \frac{TN}{Negativo} \tag{3.4}$$

F1 - ou *F-Measure* é uma média ponderada de precisão e sensitividade dada pela fórmula da equação 3.5:

$$f1 = \frac{2*(precisao*sensitividade)}{precisao+sensitividade} \tag{3.5}$$

Curva ROC - a Curva ROC é um gráfico da porcentagem de amostras corretamente classificadas como positivas dentre todas as positivas reais versus a porcentagem de amostras erroneamente classificadas como positivas dentre todas as negativas reais ou um *trade-off* entre TPR e FPR, dadas pelas fórmulas da equações:

$$TPR = \frac{TP}{TP+FN} \tag{3.6}$$

$$FPR = \frac{FP}{FP+TN} \tag{3.7}$$

Obviamente, um modelo de classificação ideal teria TPR = 1 e FPR = 0 (taxa de acerto = 1 e taxa de erro = 0).

As principais métricas da literatura citadas neste estudo tem o objetivo de avaliar diretamente, de forma independente, o desempenho do classificador induzido.

CAPÍTULO 4

MÉTODO PROPOSTO

Para concluir com êxito o desenvolvimento deste trabalho e consequentemente os objetivos propostos, o método utilizado para solução do problema é composto das seguintes etapas sequenciais:

Como já foi dito o banco de redações UOL foi desenvolvido e armazenado em páginas HTML, o que permite o uso de um *Web Crawler*, um algoritmo que explora a estrutura de grafo da *Web* para navegar de uma página para outra. A Figura 4 ilustra a etapa que o *Web Crawler* recupera as páginas, filtra as redações avaliadas e coleta cada uma para um repositório local.

Figura 4: Um *Web Crawler*, navega entre as páginas HTML do banco de redações UOL de forma metódica e automatizada indexando textos de redações que posteriormente serão filtrados e coletados.

Na etapa subsequente a Figura 5 ilustra a normalização dos textos, que consiste em uma técnica de remoção de caracteres não alfa-numéricos presentes no HTML e espaços desnecessários, tal que o valor textual ainda seja o mesmo que o original. Após a normalização será organizado as as diversas partes que compõem a redação (tema, título, texto e nota) em uma estrutura JSON para armazenamento e uso futuro.

Figura 5: Os textos são submetidos aos algoritmos de normalização e posteriormente estruturados e armazenados no padrão JSON.

Na terceira etapa ilustrada pela Figura 6 será utilizada a ferramenta de mineração de dados *Orange* (DEMŠAR *et al.*, 2013). Será necessário realizar estudo e análise para obter o conhecimento necessário para desenvolvimento de um fluxo de trabalho, seleção e treinamento dos modelos classificadores, concluindo todos os objetivos propostos nesta etapa.

Figura 6: O *corpus* será utilizado em um fluxo de trabalho da ferramenta *Orange* para treinar os modelos classificadores.

A quarta e última etapa é ilustrada pela Figura 7, onde o classificador previamente ajustado e treinado será submetido aos testes de Acurácia, Sensitividade, Precisão, Especificidade, Curva ROC e Matriz de Confusão. Os resultados serão representados graficamente e comparados para analisar o desempenho do classificador.

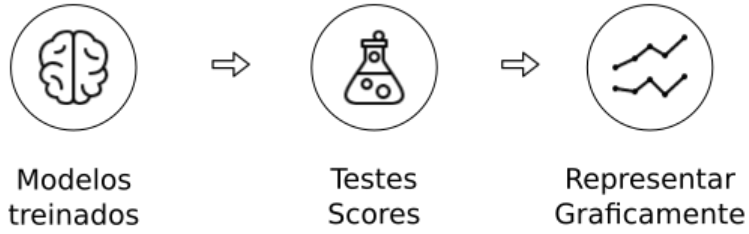

Figura 7: O classificador ajustado e treinado será submetido a testes, e os resultados comparados graficamente com o objetivo de analisar o desempenho do classificador induzido.

CAPÍTULO 5

RESULTADOS PRELIMINARES

Este capítulo é dedicado a apresentar os resultados preliminares e adversidades obtidas na indução do classificador *AdaBoost*, um algoritmo que utiliza *Boosting* como método de aprendizagem, flexível para se combinar com os vários algoritmos base de aprendizagem disponíveis.

5.1 Dados Desbalanceados

Para a indução do classificador AdaBoost sobre a primeira competência exigida em um texto de redação, isto é, “Demonstrar domínio da norma padrão da língua escrita.”, a ferramenta *Orange* selecionou de forma aleatória uma amostra de 30% do corpus de redações, ou seja, aproximadamente 131 redações de 436.

A qualidade dos dados no *dataset* é uma fator fundamental neste processo de indução, e trabalhar com dados desbalanceados tende à produzir regras de classificação que beneficiam as classes majoritárias, isto é, com maior probabilidade de ocorrência, resultando em uma baixa taxa de predição para o grupo minoritário.

Como demostrado no Gráfico 8, a amostra de 30% das redações selecionadas no *dataset* apresentou dados desbalanceados, ou seja, 70% da amostra selecionada tende para as classes 0.50 e 1.00, os demais 30% restante, para as classes 0.00, 1.50 e 2.00.

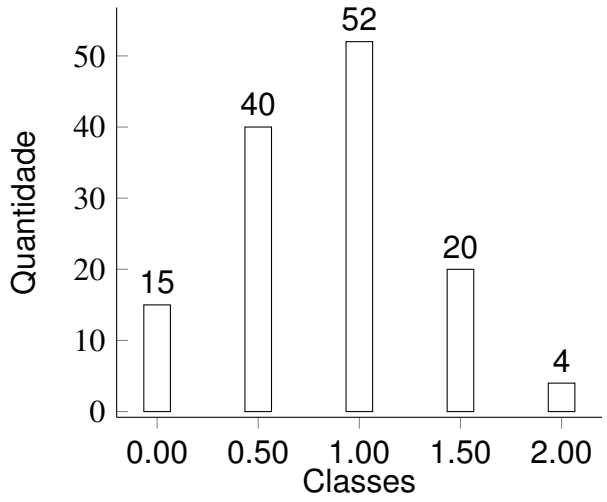

Figura 8: Distribuição das classes em uma amostra de 131 redações selecionadas aleatoriamente no *dataset*.

O desbalanceamento de dados presente na amostra era uma condição esperada, no processo de valoração, a competência é avaliada por, pelo menos, dois avaliadores, de forma independente, se ocorrer diferenças nas notas da competência inferior a 20% entre elas, é calculado o valor médio, fator que colabora efetivamente para tendência de uma ou mais classes presentes na amostra.

5.2 Métricas de Desempenho

Nos resultados preliminares, este estudo utilizou as principais métricas da literatura para análise do desempenho de classificadores, tendo como foco as métricas: Curva ROC, Acurácia, *F-Score*, *Precision* e *Recall*.

A Tabela 4 exibe os resultados das principais métricas de desempenho de classificadores sobre cada classe induzida e também a média geral de cada métrica.

	Resultado da avaliação				
Classes	**ROC**	**Acurácia**	***F-Score***	***Precision***	***Recall***
0.00	0.498	0.828	0.096	0.845	0.828
0.50	0.552	0.640	0.349	0.653	0.640
1.00	0.499	0.509	0.422	0.506	0.509
1.50	0.549	0.579	0.222	0.755	0.759
2.00	0.541	0.915	0.140	0.899	0.915
Média	**0.529**	**0.694**	**0.246**	**0.737**	**0.730**

Tabela 4: Resultado das métricas de desempenho do classificador AdaBoost.

A adversidade de classes desbalanceadas, pode produzir um modelo com elevadas taxas de acurácia global para determinadas classes, como o ocorrido nas classes 0.00 e 2.00 de 0.828 e 0.915 respectivamente, entretanto frequentemente tende a prejudicar a identificação de exemplos pertencentes a grupos minoritários.

Ilustrada na Figura 9, a representação gráfica da Curva ROC de cada classe induzida pelo classificador.

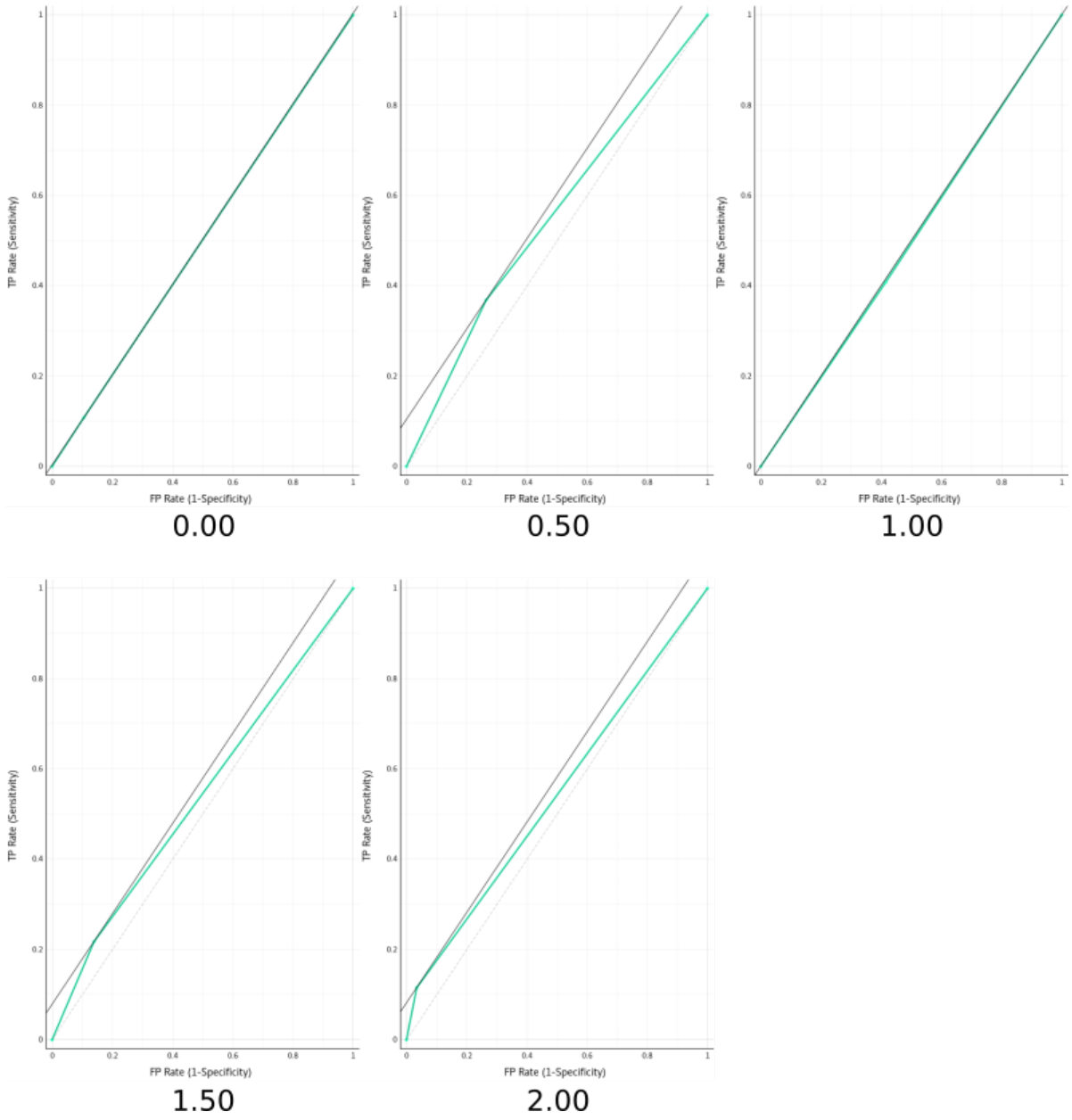

Figura 9: Representação gráfica da Curva ROC para cada classe (0.00, 0.50, 1.00, 1.50 e 2.00) induzida no modelo AdaBoost.

O comportamento esperado para a curva, é que a mesma, se aproxime o máximo possível de 1 em cada classe. Entretanto a métrica se apresenta como uma reta nas classes 0.00 e 1.00, nas demais classes tende sutilmente a 1. Conclui-se que a predição das classes 0.00 e 1.00 nos testes estão ocorrendo de forma aleatório pelo classificador.

Por fim a tabela de contingência ou matriz de confusão, que através da discriminação dos erros ou acertos preditos para cada classe demonstra o desempenho do classificador, uma das métricas mais eficiente de se analisar um

classificador.

A Tabela 5 exibe ao longo da diagonal em tons de cinza as decisões corretas: número de verdadeiros positivos TP e verdadeiros negativos TN; já os elementos fora dessa diagonal representam os erros cometidos: número de falsos positivos FP e falsos negativos FN. É notável que o valor ideal fora da diagonal seja sempre igual a 0.

		Predição					
		0.00	**0.50**	**1.00**	**1.50**	**2.00**	∑
Atual	**0.00**	4	18	13	2	0	**37**
	0.50	13	42	44	14	1	**114**
	1.00	18	51	78	32	11	**190**
	1.50	7	14	31	15	2	**69**
	2.00	4	2	14	3	3	**26**
	∑	**46**	**127**	**180**	**66**	**17**	**436**

Tabela 5: Tabela de contingência ou Matriz de confusão resultante da indução do classificador AdaBoost.

5.3 Considerações Finais

E notável que adversidade de classes desbalanceadas influenciou consideravelmente nos resultados preliminares. A próxima etapa deste estudo merece destaque em uma seção exclusiva para discussão do tema e a análise dos principais métodos na literatura para balanceamento de classes.

As dificuldades observadas no estudo do problema proposto motivam melhorias e o surgimento de novas estratégias pra a continuidade do trabalho.

CAPÍTULO 6

PLANO DE TRABALHO

Para que seja possível a realização deste estudo, o plano de trabalho consiste em um conjunto de objetivos e processos que visa a conclusão do trabalho de modo adequado. Os pontos relevantes para a realização do trabalho foram ordenados e enumerados nas atividades a serem realizadas:

6.1 Plano de atividade

1. Revisão Bibliográfica: a primeira atividade enumerada deste estudo, parte da introdução ao problema proposto até a sustentação do estudo por meio de leitura de artigos e trabalhos sobre Aprendizado de Máquina;

2. Coleta de Dados: após a realização da revisão e embasamento dos fundamentos de sustentação do estudo, a etapa de coleta de dados será executada para formação de um *corpus* de redações;

3. Tratamento dos dados: posteriormente a formação da base de conhecimento, todos os textos de redações serão submetidos individualmente a um processo de normalização, este processo visa a remoção de caracteres não alfa numéricos sem alterar o seu valor textual;

4. Indução do modelo: logo depois de organizar uma base de conhecimento concreta e normalizada, um fluxo de trabalho pode ser projetado e ajustado para induzir um modelo de Aprendizado de Máquina sobre uma porcentagem da base conhecimento que se deseja aprender;

5. Testes e *Score*: imediatamente a indução do modelo, será realizado a etapa de testes de *acurácia* geral, *overfitting* e *noise*, caso algum dos testes falhem ou não atendam aos padrões esperados a etapa anterior será repetida com o objetivo de melhorar os resultados do modelo induzido;

6. Análise dos resultados: após os teste apresentarem um valor aceitável, o restante da base de conhecimento não utilizada na indução do modelo pode ser submetido a predição para comparação gráfica do conhecimento adquirido pelo modelo induzido;

7. Escrita da Monografia: embasado no resultado das etapas anteriores deste trabalho, este processo corrente tem o objetivo de transcrever todos os métodos envolvido para a resolução do problema proposto, considerando as revisões do orientador.

8. Defesa do projeto: por fim, a preparação final para a defesa à banca examinadora.

6.2 Cronograma gráfico do Plano de Atividades

	Atividade	Fevereiro - 2017	Março - 2017	Abril - 2017	Maio - 2017	Junho - 2017	Julho - 2017	Agosto - 2017	Setembro - 2017	Outubro - 2017	Novembro - 2017
1	Revisão Bibliográfica	✓	✓	✓							
2	Coleta de Dados				✓						
3	Tratamento dos dados				✓						
4	Indução do modelo					✓	✓	✓	✓		
5	Testes e *Score*					✓	✓	✓	✓		
6	Análise dos resultados					✓	✓	✓	✓	✓	
7	Escrita da Monografia	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
8	Entrega da Monografia										✓

Referências Bibliográficas

ALVES, Alexandra Isabel Magalhães. **Modelo de representação de texto mais adequado à classificação**. Dissertação (Mestrado) — Instituto Superior de Engenharia do Porto, 11 2010. Online; acessado 06 Junho 2017.

BATISTA, GEAPA; CARVALHO, ACPLF; MONARD, Maria C; BRASIL, Silicon Graphics. Aplicando seleção unilateral em conjuntos de exemplos desbalanceados: Resultados iniciais. In: **XIX CONGRESSO NACIONAL DA SOCIEDADE BRASILEIRA DE COMPUTAÇÃO "EDUCAÇÃO E APRENDIZAGEM NA SOCIEDADE DA INFORMAÇÃO**. [S.l.: s.n.], 1999. v. 20, p. 327–340.

BATISTA, Gustavo Enrique de Almeida Prado *et al.* **Pré-processamento de dados em aprendizado de máquina supervisionado**. Tese (Doutorado) — Universidade de São Paulo, 2003.

BOSCARIOLI, Clodis; VITERBO, José; TEIXEIRA, Mateus Felipe. Avaliação de aspectos de usabilidade em ferramentas para mineração de dados. **Anais da I Escola Regional de Sistemas de Informação do Rio de Janeiro**, v. 1, n. 1, p. 107–114, 2014.

BRAGA, Bruno Marx de Aquino. **Teoria da resposta ao item: o uso do modelo de Samejima como proposta de correção para itens discursivos**. Dissertação (Mestrado) — Universidade de Brasília Instituto de Ciências Exatas Departamento de matemática, 7 2015. Online; acessado 06 Junho 2017.

CEBRASPE, CESPE UNB. **Relatório de Gestão CEBRASPE**. 2016. 1–20 p. Online; acessado 07 Abril 2017. Disponível em: <http://www.cespe.unb.br/cebraspe/arquivos/Relatorio_de_Gestao_2016.pdf>.

CSF, Ciência sem Fronteiras. **Estudante de Graduação**. 2017. Online; acessado 07 Abril 2017. Disponível em: <http://www.cienciasemfronteiras.gov.br/web/csf/estudante>.

DEMŠAR, Janez; CURK, Tomaž; ERJAVEC, Aleš; GORUP, Črt; HOČEVAR, Tomaž; MILUTINOVIČ, Mitar; MOŽINA, Martin; POLAJNAR, Matija; TOPLAK, Marko; STARIČ, Anže; ŠTAJDOHAR, Miha; UMEK, Lan; ŽAGAR, Lan; ŽBONTAR, Jure; ŽITNIK, Marinka; ZUPAN, Blaž. Orange: Data mining toolbox in python. **Journal of Machine Learning Research**, v. 14, p. 2349–2353, 2013. Disponível em: <http://jmlr.org/papers/v14/demsar13a.html>.

DUARTE, J. **O Algoritmo Boosting at Start e suas aplicações**. Tese (Doutorado) — PUC-Rio, 2009.

FREITAS, Maria Claudia de; UZEDA-GARRÃO, Milena; OLIVEIRA, Claudia; SANTOS, Cícero Nogueira dos; SILVEIRA, Maria Cândida. A anotaçao de um corpus para o aprendizado supervisionado de um modelo de sn. In: **Proceedings of the III TIL/XXV Congresso da SBC**. [S.l.: s.n.], 2005.

INEP. Edital anual do exame nacional do ensino médio, **EDITAL No 10, DE 14 DE ABRIL DE 2016**. 2016. Online; acessado 05 Junho 2017. Disponível em: <http://download.inep.gov.br/educacao_basica/enem/edital/2016/edital_enem_2016.pdf>.

LARA, Glaucia Muniz Proença. A redação como tema de pesquisa. In: **Leitura: Teoria e Prática**. [S.l.]: 1994, 1994. v. 13, n. 24, p. 62–82.

MONARD, Maria Carolina; BARANAUSKAS, José Augusto. Conceitos sobre aprendizado de máquina. **Sistemas Inteligentes-Fundamentos e Aplicações**, v. 1, n. 1, 2003.

MOTTA, Porthos Ribeiro de Albuquerque. **Estudo Exploratório do Uso de Classificadores para a Predição de Desempenho e Abandono em Universidades**. Dissertação (Mestrado) — Universidade Federal de Goiás Instituto de Informática, 11 2016. Online; acessado 06 Junho 2017.

PEDREGOSA, F.; VAROQUAUX, G.; GRAMFORT, A.; MICHEL, V.; THIRION, B.; GRISEL, O.; BLONDEL, M.; PRETTENHOFER, P.; WEISS, R.; DUBOURG, V.; VANDERPLAS, J.; PASSOS, A.; COURNAPEAU, D.; BRUCHER, M.; PERROT, M.; DUCHESNAY, E. Scikit-learn: Machine learning in Python. **Journal of Machine Learning Research**, v. 12, p. 2825–2830, 2011.

ROSSUM, Guido Van; DRAKE, Fred L. **Python language reference manual**. [S.l.]: Network Theory, 2003.

SILVA, Sílvio Ribeiro da; CARVALHO, Taynan Lima. Produção de texto escrito no ensino médio: Competências requeridas pela avaliação de redação do enem em (des)uso no livro didático de português. **Caminhos em linguística aplicada**, 1o sem 2017, v. 16, n. 1, p. 1–25, 2017. Disponível em: <http://periodicos.unitau.br/ojs-2.2/index.php/caminhoslinguistica>.

SISU, Sistema de seleção unificada. **O que é o Sisu**. 2017. Online; acessado 07 Abril 2017. Disponível em: <http://sisu.mec.gov.br/>.

TEIXEIRA, Diogo; AZEVEDO, Isabel. Análise de opiniões expressas nas redes sociais. **RISTI-Revista Ibérica de Sistemas e Tecnologias de Informação**, Associação Ibérica de Sistemas e Tecnologias de Informação (AISTI), n. 8, p. 53–65, 2011.

VIEIRA, Renata; LOPES, Lucelene. Processamento de linguagem natural e o tratamento computacional de linguagens científicas. **EM CORPORA**, p. 183, 2010.

WAHBEH, Abdullah H; AL-RADAIDEH, Qasem A; AL-KABI, Mohammed N; AL-SHAWAKFA, Emad M. A comparison study between data mining tools over some classification methods. **International Journal of Advanced Computer Science and Applications**, v. 8, n. 2, p. 18–26, 2011.

WAYMO. **We're building a safer driver for everyone**. 2017. Online; acessado 07 Abril 2017. Disponível em: <https://waymo.com/>.