

UNIVERSIDADE FEDERAL DE UBERLÂNDIA

Angelo Travizan Neto

**Classificação de traços de personalidade através
de Linguagem Natural aplicada a ambientes
virtuais de aprendizagem**

Uberlândia, Brasil

2017

Resumo

Em um contexto de aprendizagem colaborativa mediada por computador, é relevante a inserção de metodologias de formação de grupos (agrupamento) de estudantes para que as interações entre os mesmos possam conduzir à melhoria do aprendizado. O projeto de pesquisa, no qual se insere este trabalho, consiste na identificação dos traços de personalidade do aprendiz através de Linguagem Natural. A partir destes, será possível prever as melhores formações de grupos. Para tanto, foi feita uma coleta de textos escritos por estudantes, para então classificar seus traços de personalidade, de acordo com o modelo *Big Five*. Assim, após o processamento do texto, utilizando a biblioteca NLTK disponível em *Python*, foi realizada a categorização das palavras úteis mineradas do texto, utilizando o dicionário LIWC. Os percentuais de cada categoria de palavra presentes no texto permitem identificar os traços de personalidade do estudante, utilizando algum algoritmo de classificação. Ressalta-se que o modelo *Big Five* já tem sido associado à formação de grupos e que a obtenção de textos escritos em ambientes virtuais de aprendizagem é viável, já que contém ferramentas como chats e fóruns. Assim, este trabalho poderá ser inserido nestes ambientes a fim de prover informações para a composição de grupos.

Palavras-chave: Algoritmos de classificação, Linguagem Natural, traços de personalidade

Abstract

In a context of collaborative computer-mediated learning, it is important to include groups formation methodologies (grouping) of students so that interactions between them can lead to the improvement of learning. The research project, in which this work is inserted, consists of identifying the personality traits of the learner through Natural Language. From these, it will be possible to predict the best group formations. To do so, a collection of texts written by students was made, to then classify their personality traits, according to the Big Five model. Thus, after processing the text using the NLTK library available in Python, the categorization of the useful words mined in the text was performed using the LIWC dictionary. The percentages of each category of word present in the text allow to identify the traits of personality of the student, using some classification algorithm. It should be emphasized that the Big Five model has already been associated with the formation of groups and that obtaining texts written in virtual learning environments is feasible, since they contain tools such as chats and forums. Thus, this work can be inserted in these environments in order to provide information for group composition.

Keywords: *Classify Algorithm, Natural Language, Personality Traits*

Lista de ilustrações

Figura 1 – Processamento básico de um texto	17
Figura 2 – Treinamento da base	19
Figura 3 – Fluxograma do processamento geral utilizando a base de treinamento .	23
Figura 4 – Tradução da linguagem informal para a formal	28

Lista de tabelas

Tabela 1 – Análise do método de Regressão para o traço Extroversão	20
Tabela 2 – Análise do método de Regressão para o traço Neuroticismo	21
Tabela 3 – Análise do método de Regressão para o traço Abertura	21
Tabela 4 – Análise do método de Regressão para o traço Conscienciosidade	21
Tabela 5 – Análise do método de Regressão para o traço Afabilidade	22
Tabela 6 – Resultados	24
Tabela 7 – Erros	25
Tabela 8 – Correlações entre traços de personalidade e categorias de palavras afetivas	25
Tabela 9 – Correlações entre traços de personalidade e categorias de preocupações pessoais	26

Lista de abreviaturas e siglas

AVA	Ambiente Virtual de Aprendizagem
IA	Inteligência Artificial
PLN	Processamento de Linguagem Natural
NLTK	Natural Language Toolkit
LIWC	Linguistic Inquiry and Word Count
PT-BR	Português do Brasil

Sumário

1	INTRODUÇÃO	7
1.1	Objetivo Geral	7
1.2	Objetivos Específicos	8
1.3	Organização do Texto	8
2	REVISÃO DA LITERATURA E REFERENCIAL TEÓRICO	9
2.1	Trabalhos Correlatos	9
2.2	Teorias e Técnicas	10
2.2.1	Modelo Big Five	10
2.2.2	Processamento de Linguagem Natural	13
2.2.3	Identificação de Palavras Significativas	13
2.2.4	Categorização das Palavras	14
3	CLASSIFICAÇÃO DOS TRAÇOS	17
3.1	Processamento do texto	17
3.1.1	Contagem das frequências das categorias do LIWC	18
3.1.2	Categorias do LIWC e Modelo <i>Big Five</i>	18
3.2	Processamento dos textos para a inferência dos traços de personalidade	22
3.3	Integração com o Ambiente Virtual de Aprendizagem Moodle	23
4	RESULTADOS	24
5	CONCLUSÕES	27
	REFERÊNCIAS	30

1 Introdução

Na aprendizagem mediada por computador, existe um certo impedimento a ser tratado: não há interação entre os alunos com os professores, dificultando o processo de aprendizagem. Uma das inúmeras soluções para isso é a aprendizagem colaborativa. Em um contexto em que isso é levado em consideração, é relevante a inserção de metodologias interativas entre estudantes para que seja melhorado o aprendizado, através da formação de grupos ([ALAVI, 1994](#)).

No projeto de pesquisa é estudada a otimização da aprendizagem em Ambientes Virtuais de Aprendizagem (AVAs) utilizando-se Inteligência Artificial. Esses ambientes analisam, classificam e objetivam melhorar o nível de aprendizado dos estudantes. Tendo como contexto a aprendizagem colaborativa, que baseia-se na inserção de atividades com interações entre estudantes, com o auxílio da IA, são identificadas características dos estudantes, utilizando algumas diferentes métricas ([MAGNISALIS; DEMETRIADIS; KAKOSTAS, 2011](#)). Classificações essas, úteis para obter um resultado que decida qual grupo o aluno melhor se adaptará para que haja a otimização da experiência ([KUMAR; ROSÉ, 2014](#)).

Uma hipótese de solução e de apoio à aprendizagem colaborativa neste ambiente é fazer a classificação e extração das características de personalidade dos alunos para que sejam úteis em um futuro processamento, e que sejam entradas em algoritmos que decidam quais as melhores interações para aquele perfil, pois é relevante a maneira como a personalidade afeta na aprendizagem ([ALTANOPOULOU; TSELIOS, 2015](#)). Como o modo que o aluno tem de entrar em contato com o material é através de um AVA, existem dificuldades ao se tentar identificar suas características. Portanto, é proposto neste trabalho inferir estes perfis utilizando os textos escritos pelos alunos no Ambiente Virtual de Aprendizagem, pois a escolha das palavras em uma escrita é influenciada pelo perfil psicológico do aluno ([ALTANOPOULOU; TSELIOS, 2015](#)).

1.1 Objetivo Geral

O objetivo deste trabalho é classificar os traços de personalidade de estudantes, através de textos escritos em um Ambiente Virtual de Aprendizagem (AVA). Esta classificação será utilizada para formação de grupos em atividades colaborativas.

1.2 Objetivos Específicos

- Aplicar um questionário de 44 itens do modelo *Big Five*, juntamente com perguntas adicionais para a coleta de dados escritos dos alunos;
- Contabilizar a frequência das categorias do dicionário LIWC;
- Treinar a base de dados para a previsão dos traços de personalidade;
- Realizar testes utilizando 20 alunos;
- Avaliar os erros dos resultados, em relação à aplicação do questionário de 44 itens, para a medição da precisão da proposta;

1.3 Organização do Texto

Este documento é dividido em 5 capítulos. No Capítulo 2 é abordada toda a revisão da literatura necessária para a conclusão desta pesquisa, juntamente com o referencial teórico. Nele são descritas as teorias e técnicas utilizadas, bem como uma breve explicação de como foram necessários os referenciais mencionados. O Capítulo 3 descreve a classificação dos traços de personalidade. Ele é dividido em três seções. A primeira expõe o processamento do texto, que é baseado na contagem da frequência de categorias, juntamente com a relação desta com os traços de personalidade (treinamento da base utilizando Regressão Linear); a segunda descreve como é realizado o processamento dos diálogos do AVA para que seja feita classificação dos alunos (o fluxo resumido do processamento, utilizando a base treinada anteriormente); e a terceira parte é destinada à integração deste estudo com o Ambiente Virtual de Aprendizagem *Moodle*. O quarto capítulo é destinado aos resultados do trabalho. Em seguida, no Capítulo 5, é descrita a conclusão da pesquisa, juntamente com os trabalhos futuros.

2 Revisão da Literatura e Referencial Teórico

Este capítulo apresenta alguns trabalhos que fizeram uso de Processamento de Linguagem Natural (PLN) para identificação de traços personalidade, bem como as teorias e técnicas utilizadas para classificação destes traços.

2.1 Trabalhos Correlatos

Os avanços recentes em análise de textos encaminham a possibilidade de que sistemas computacionais sejam capazes de reconhecer e detectar características e traços de personalidade em texto. Porém, esses métodos ainda estão em fase de desenvolvimento e são, em sua grande maioria, para a língua inglesa, existindo poucos estudos para língua portuguesa. A variabilidade de aplicações é notável, já que existem estes estudos aplicados em redes sociais como *Facebook*, *Twitter*, Ambientes Virtuais de Aprendizagem (AVAs), dentre outros.

O primeiro e similar trabalho correlato é o de (MACHADO et al., 2015), que consiste na construção de um léxico em português do Brasil que contém características de personalidade nas palavras, para extrair traços de personalidade em textos. Léxicos são recursos criados para processamento em texto objetivando um posterior tratamento computacional. Este trabalho se baseia em (PENNEBAKER; FRANCIS; BOOTH, 2001), que constitui-se de um software léxico para a contagem de palavras, sentenças, dentre outras coisas em um texto, de acordo com o dicionário LIWC (*Linguistic Inquiry and Word Count*).

Outro trabalho relacionado é (ALAM; RICCARDI, 2014), que faz um estudo semelhante à este, mas relacionado à conversas em áudio. Nele, são extraídos traços de personalidade através de dados falados, utilizando-se diferentes tipos de recursos: acústico, linguístico, e de psicolinguístico. A base do processamento é a seguinte: a fala é diretamente repassada para um analisador acústico, sendo também transcrita para um analisador chamado *Token*. Da saída destes analisadores é feita uma combinação de recursos para que possa ser feita a avaliação. A classificação desta análise também é feita baseada no modelo *Big Five* descrito em (GOLDBERG, 1981), dividida entre os recursos acústico, linguístico e psicolinguístico.

Em (FILHO; PARDO; ALUISIO, 2013) é feita uma avaliação do dicionário em português do Brasil LIWC, utilizado para análise de sentimentos. Essa avaliação é feita através da comparação deste dicionário com outros dois recursos disponíveis: o *Opinion Lexicon* (NA et al., 2009) e *SentiLex* (CARVALHO; SILVA, 2015). Para classificação

de opiniões, o dicionário LIWC obteve melhor performance em textos positivos. Já em textos negativos, o *SentiLex* possui melhor desempenho. Em classificações de sentenças, o dicionário LIWC ainda possui uma dificuldade em textos classificados como negativos, portanto conclui-se que o dicionário LIWC possui uma melhor performance indicando positividade do que negatividade, em textos aplicados, através da melhor precisão obtida como resultado na conclusão do artigo.

As premissas descritas na introdução deste trabalho foram também utilizadas para iniciar uma pesquisa preliminar e relacionada a esta (NETO; FERREIRA; FERNANDES, 2017). Nesta também é utilizada IA para apoiar a aprendizagem colaborativa, e consequentemente a formação de grupos de alunos. Assim como nesse trabalho, os estudantes são classificados utilizando a linguagem, pretendendo obter como resultado seus traços. Para isso, a relação entre as palavras escritas pelos alunos e suas personalidades foi observada. O foco dessa pesquisa era encontrar um grau numérico desta relação, que pudesse ser utilizado como apoio para a execução e implementação deste. Nos resultados encontrados em (NETO; FERREIRA; FERNANDES, 2017) foram verificadas relações positivas e influentes das palavras escolhidas pelos alunos, com seus traços de personalidade, validando a utilização de textos escritos para inferência de traços de personalidade.

2.2 Teorias e Técnicas

Como dito anteriormente, este trabalho utiliza textos escritos em linguagem natural para identificação de características psicológicas do estudante. Por isso, esta seção apresenta os principais aspectos do modelo *Big Five* relativo aos traços de personalidade, que têm sido utilizados para formação de grupos, e descreve as ferramentas utilizadas para processamento de linguagem natural.

2.2.1 Modelo Big Five

O modelo *Big Five* utilizado nesse trabalho é fundamentado em (GOLDBERG, 1981), que define os cinco fatores da personalidade descritos pelo método lexical (baseado em uma análise linguística). Eles são fundamentados entre as seguintes perguntas a serem feitas: o sujeito é ativo e dominante, ou passivo e submissivo? Ele é agradável (presente) ou desagradável (frio e distante)? Posso contar com ele (ele é responsável), ou ele é irresponsável? Ele é neurótico ou estável? É aberto à novas experiências, ou inflexível (ignorante)? As respostas dessas perguntas caracterizam os cinco fatores da personalidade mencionados anteriormente.

Os cinco traços de personalidade são definidos como extroversão, neuroticismo, conscienciosidade, abertura à novas experiências e afabilidade. Extroversão caracteriza pessoas que são mais comunicativas, e usualmente mostram emoções positivas. Neuroti-

cismo é caracterizado por pessoas que possuem baixa tolerância ao estresse. Conscienciosidade caracteriza pessoas organizadas, que fazem planejamentos diante de um projeto, e que geralmente são responsáveis. Abertura caracteriza a tendência em ser uma pessoa curiosa, inquisitiva, e interessada em novas ideias. Por último, afabilidade é relacionada à habilidade de ser cooperativa, agradável, ser uma pessoa que se preocupa com o próximo. Cada traço possui seu escore, que pode ser alto ou baixo. Escore baixo para o traço extroversão pode caracterizar um indivíduo introvertido, assim como escore baixo para neuroticismo pode caracterizar um indivíduo mais tolerante ao estresse, e assim por diante (GOLDBERG, 1981).

Uma maneira de classificar indivíduos, a fim de obter seus traços de personalidade, é através da aplicação de questionários. Esta metodologia é utilizada para a inferência inicial dos perfis de personalidade (conjunto de escores de todos os traços) adequado na preparação da base de treinamento. É empregado o inventário de 44 itens (perguntas), em que cada uma refere-se positivamente a um traço (JOHN; DONAHUE; KENTLE, 1991). Os itens são:

1. É conversador, comunicativo.
2. Às vezes é frio, distante.
3. Tende a ser crítico com os outros.
4. É minucioso, detalhista no trabalho.
5. É assertivo, não teme expressar o que sente.
6. Insiste até concluir a tarefa ou o trabalho.
7. É depressivo, triste.
8. Gosta de cooperar com os outros.
9. É original, tem sempre novas ideias.
10. É temperamental, muda de humor facilmente.
11. É inventivo, criativo.
12. É reservado.
13. Valoriza o artístico, o estético.
14. É emocionalmente estável, não se altera facilmente.
15. É prestativo e ajuda os outros.

16. É, às vezes, tímido, inibido.
17. Pode ser um tanto descuidado.
18. É amável, tem consideração pelos outros.
19. Tende a ser preguiçoso.
20. Faz as coisas com eficiência.
21. É relaxado, controla bem o estresse.
22. É facilmente distraído.
23. Mantém-se calmo nas situações tensas.
24. Prefere trabalho rotineiro.
25. É curioso sobre muitas coisas diferentes.
26. É sociável, extrovertido.
27. É geralmente confiável.
28. É, às vezes, rude (grosseiro) com os outros.
29. É cheio de energia.
30. Começa discussões, disputas com os outros.
31. É um trabalhador de confiança.
32. Faz planos e os segue a risca.
33. Tem uma imaginação fértil.
34. Fica tenso com frequência.
35. É engenhoso, alguém que gosta de analisar profundamente as coisas.
36. Fica nervoso facilmente.
37. Faz tudo com muito entusiasmo.
38. Tende a ser desorganizado.
39. Gosta de refletir, brincar com as ideias.
40. Tem capacidade de perdoar, perdoa fácil.
41. Preocupa-se muito com tudo.

42. Tende a ser quieto, calado.
43. Tem poucos interesses artísticos.
44. É sofisticado em artes, música ou literatura.

Como dito anteriormente, cada pergunta do inventário refere-se a um determinado traço de personalidade. Por exemplo, a pergunta número 1, avalia se o indivíduo é introvertido ou extrovertido. Portanto, a partir da resposta que ele der, é possível calcular a influência neste traço.

A maneira escolhida para os alunos responderem cada item é através de uma representação numérica de 1 à 5, tendo os seguintes significados: 1 equivale à "discordo totalmente"; 2 à "discordo"; 3 à "neutro"; 4 à "concordo"; e 5 à "concordo totalmente". Deste modo é possível que se faça com que o aluno tenha respostas mais detalhadas comparado ao modelo de respostas "sim" ou "não".

2.2.2 Processamento de Linguagem Natural

De acordo com (RUSSELL et al., 2003), PLN é uma sub-área de IA e da linguística que estuda os problemas de geração e compreensão das línguas humanas por máquinas. Este processamento é relevante para este trabalho, pois a identificação das características de personalidade será feita através da análise de textos escritos. Mas, além disto, é necessário dar interpretação às palavras significativas do texto a fim de relacioná-las aos traços, ou seja, é necessário categorizá-las. Assim, uma das subseções descreve o dicionário LIWC, que permite tal categorização.

2.2.3 Identificação de Palavras Significativas

A filtragem do texto é realizada através de um programa na linguagem de programação *Python*. A escolha desta linguagem se justifica através da existência da biblioteca chamada NLTK (*Natural Language Toolkit*), que é uma plataforma para a criação de programas *Python* para tratar com dados de linguagem natural (BIRD, 2006).

A biblioteca NLTK possui uma lista de *stopwords* (palavras irrelevantes para compreensão do texto), além de ser disponibilizada para a linguagem PT-BR. Alguns exemplos de *stopwords* seriam artigos e preposições, ou seja, pequenas palavras tais como "os", "as", "de", "em", "para", "com", dentre outras. No processamento, estas devem ser descartadas. Esta operação é facilitada quando utiliza-se esta lista, juntamente com a linguagem *Python*. Ao extrair do texto todas as palavras, basta fazer uma operação de remoção com a lista de *stopwords*, que resultará em um conjunto que contém todas as palavras significativas para interpretação do texto.

2.2.4 Categorização das Palavras

O dicionário LIWC é incorporado no software *Linguistic Inquiry and Word Count*. Este é um programa que pré-processa um texto para futuras análises, utilizando outro dicionário, denominado léxico. O software objetiva auxiliar o processamento de texto, relacionando as palavras e categorias.

O dicionário contém palavras com polaridades positivas e negativas em relação às suas categorias (TAUSCZIK; PENNEBAKER, 2010). As categorias selecionadas para serem utilizadas no trabalho são: "I", "we", "negate", "quant", "swear", "family", "friend", "humans", "posemo", "anx", "anger", "sad", "insight", "cause", "discrep", "tentat", "certain", "see", "hear", "feel", "bio", "body", "health", "sexual", "ingest", "relativ", "motion", "space", "time", "work", "achieve", "leisure", "home", "money", "relig", "death", "assent", "nonfl" e "filler". Dentre todas as categorias disponíveis do dicionário, foram excluídas somente as referentes a artigos, verbos, pronomes, dentre outros. Isso se justifica pois nelas não contém palavras relevantes ao processamento (palavras que serão eliminadas no pré-processamento do texto descrito na Seção 3.1). O dicionário é traduzido para a linguagem PT-BR, porém os nomes das categorias não. No entanto, não é impactante no processamento feito neste trabalho, pois a inferência dos traços de personalidade é realizada ainda assim.

As categorias estão descritas na língua inglesa. Porém o dicionário é disponível na língua PT-BR. A utilização do LIWC é justificada através dos resultados do estudo (FILHO; PARDO; ALUISIO, 2013), que elabora uma avaliação deste, comparando-o com outros dois métodos utilizados em análise de sentimentos através da linguagem. Esta pesquisa é melhor esclarecida na Seção 2.1.

O LIWC é um software léxico, cujo diferencial é a categorização de palavras, conforme descrito anteriormente. Essas categorias podem denotar influência no perfil psicológico de quem redige o texto, pois através das palavras presentes no texto, pode-se observar características referentes aos traços de personalidade (TAUSCZIK; PENNEBAKER, 2010). Uma maneira de relacionar essas palavras com o traço de personalidade é fazendo a contagem da frequência em que as categorias aparecem, para que seja feita uma análise.

Uma maneira de inferir os traços de personalidade de um indivíduo é através da aplicação de questionários estudados pela psicologia. Então se existe um modo de computar a frequência das categorias do LIWC no texto, e um de deduzir os traços de personalidade, o proposto neste trabalho é o treinamento de uma base com estas duas entradas. Este treinamento objetiva a automação na dedução das características de personalidade, que é feita nas Seções 3.1.1 e 3.1.2. Assim como no trabalho de (MACHADO et al., 2015), esta pesquisa é baseada no modelo do *Big Five* (GOLDBERG, 1981). Este

modelo refere-se aos cinco fatores da personalidade descritos pelo método lexical (baseado em uma análise linguística), melhor descrito na Seção 2.2.1. A saída do léxico são as porcentagens que o indivíduo possui em cada traço de personalidade.

Além dos softwares descritos acima, utilizou-se também *Konstanz Information Miner* (KNIME), um ambiente que organiza o fluxo de dados de maneira simples e visual, habilitando a integração de novos algoritmos e ferramentas para a manipulação dos dados (BERTHOLD et al., 2009). Sua parte visual é dividida em módulos. Existem módulos de operação em banco ("*MySQL Connector*", "*Database Reader*"); operações com tabelas de banco ("*Inner Join*", "*Group by*"), módulos para salvar e ler um modelo em arquivos ("*Model writer*", "*Model reader*"), dentre outros. Nele é possível instalar pacotes de ferramentas de várias áreas para que possa ser feito processamento no fluxo de dados, juntamente com a organização. Um exemplo de pacote que foi útil para este trabalho é o do software WEKA (WITTEN et al., 2016), que disponibiliza vários algoritmos de classificação. Dentre estes, a Regressão Linear foi útil para o desenvolvimento desta pesquisa para prever as porcentagens dos traços de personalidade dos estudantes.

O estudo de Regressão Linear é utilizado para prever uma variável dependente, diante de variáveis independentes em uma base de dados. Neste trabalho, as frequências das categorias do dicionário LIWC são as variáveis independentes, e as porcentagens dos traços de personalidade (que é o resultado) são as dependentes. No KNIME existe um módulo para calcular a Regressão Linear, que é do software WEKA. Este baseia-se no critério de informação de *Akaike* (AIC) para a seleção do melhor modelo de predição. Este critério admite a existência de um modelo dito como "real" que descreve os dados que são desconhecidos, e tenta escolher dentre um grupo de modelos avaliados, o que minimiza a divergência de *Kullback-Leibler* (K-L). Esta divergência está relacionada à informação perdida por se utilizar um modelo aproximado, e não o "real" (BOZDOGAN, 1987). Este módulo disponibiliza os seguintes parâmetros de configuração para a execução:

- Método de seleção de atributos: maneira com que a regressão vai selecionar os atributos relevantes para a predição. É possível selecionar três modos:
 - Sem seleção de atributos;
 - M5;
 - *Greedy*.
- Eliminar atributos colineares: é útil para eliminar os atributos que não influenciam na variável dependente. É possível habilitá-lo ou desabilitá-lo.
- Minimal: Opção para economizar memória (não mantém armazenado o cabeçalho do conjunto de dados).

Na Seção 3.1.2 é feita uma análise da combinação desses parâmetros no intuito de maximizar a precisão da predição. Na plataforma KNIME existe um módulo chamado "*Numeric Scorer*". Ao interligar a saída da regressão nele, são calculados:

- Erros quadráticos médios: $1/n * \sum (p_i - r_i)^2$
- Erros absolutos médios: $1/n * \sum |p_i - r_i|$
- Desvios quadráticos médios: $\sqrt{1/n * \sum (p_i - r_i)^2}$
- R^2 s: $1 - \sum (p_i - r_i)^2 / \sum (r_i - 1/n * \sum r_i)^2$

onde r_i são os valores numéricos das colunas, p_i são os valores previstos, e n o número de testes feitos.

A análise é feita baseada nestas saídas deste módulo. De antemão, é possível notar que o objetivo desta análise é minimizar os erros e os desvios quadráticos médios, e maximizar os R^2 s (que é o coeficiente que indica o quanto o modelo consegue explicar os valores observados) (CAMERON; WINDMEIJER, 1997).

3 Classificação dos Traços

O desenvolvimento da classificação dos traços foi dividido em três etapas. A primeira é o processamento básico do texto. Na segunda etapa, é feita a categorização das palavras, através do dicionário LIWC e as características dos traços personalidade, que foi dividida em duas partes: contagem das frequências de cada categoria do dicionário LIWC e associação dessas frequências com os traços de personalidade do modelo *Big Five* (GOLDBERG, 1981). A Seção 3.2 é destinada à aplicação do estudo realizado, ou seja, nela é descrito um resumo de como é feito o processamento desde a coleta do texto, até a classificação dos traços de personalidade. Na Seção 3.3 é exposto como foi feita a integração de todo esse estudo com o Ambiente Virtual de Aprendizagem chamado *Moodle*.

3.1 Processamento do texto

Para que possa ser efetuado o estudo em um texto, é necessário que seja feito um processamento prévio. Este consiste em coletar todas suas palavras, eliminar as que não são relevantes, chamadas *stopwords*, eliminar a pontuação e processar as restantes. Isto é, de um texto completo, no final do processamento são extraídas somente as principais palavras para a análise. Alguns exemplos de *stopwords* são artigos, preposições e pequenas palavras tais como "os", "as", "de", "em", "para" e "com". A Figura 1 ilustra a técnica básica de processamento de texto utilizada nesta etapa.

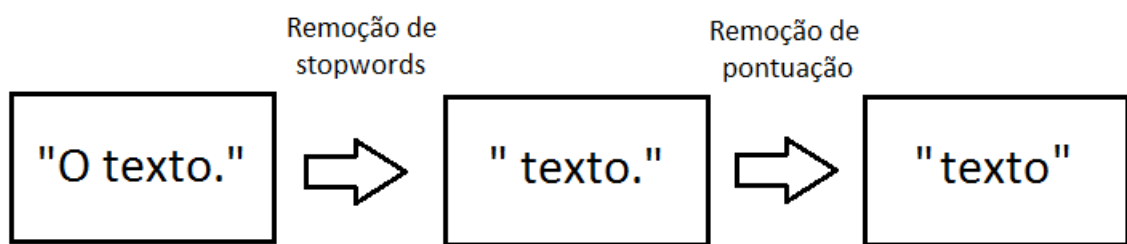


Figura 1 – Processamento básico de um texto

Após a separação e coleta das palavras relevantes, obtém-se uma lista de palavras separadas de um contexto. Na Seção 3.1.1 essas palavras serão classificadas de acordo com o dicionário LIWC para que sejam feitas as contagens das frequências de cada categoria deste dicionário presentes no texto. Isto é feito com o objetivo de relacionar estas categorias com os traços de personalidade do autor do texto.

3.1.1 Contagem das frequências das categorias do LIWC

Esta sub-seção é destinada à contagem da frequência de cada categoria do dicionário LIWC contida no texto, para que possa ser inferida a característica de personalidade. Desde que o dicionário é disponibilizado como tabela de banco de dados, e os textos são extraídos do banco de dados do AVA, a contagem é feita através de operações básicas de junção de tabelas em SQL.

Tendo a frequência com que cada categoria do LIWC aparece no texto, é possível fazer uma análise para inferir a influência destas em cada traço, o que é feito na sub-seção [3.1.2](#).

3.1.2 Categorias do LIWC e Modelo *Big Five*

Através da seção [3.1.1](#), foi obtido como resultado uma lista das categorias do dicionário LIWC, juntamente com a frequência de cada uma no texto. Baseando-se na própria definição de algumas categorias, espera-se que exista uma relação entre estas e os traços de personalidade do modelo *Big Five*, por possuírem definições semelhantes. Por exemplo, a categoria "negemo" tem por definição palavras que esboçam emoções negativas. Disso podemos esperar que as palavras que a pertencem sejam influentes no traço "neuroticismo" do *Big Five*, pois ele define sentimentos de negação, causando baixa tolerância ao estresse ([GOLDBERG, 1981](#)).

Como dito anteriormente, tendo as frequências das categorias no texto, deseja-se relacioná-las aos traços de personalidade. Estes podem ser inferidos com o auxílio de questionários, como por exemplo, o de 44 itens exposto na Seção [2.2.1](#). Mas a aplicação de questionários é inviável, pois demanda tempo, e os alunos podem escolher não responder. Então a aplicação de questionários foi realizada apenas para treinar uma base para, posteriormente, inferir os traços através desta base treinada. Esta preparação é descrita nos seguintes passos:

1. Aplicação do questionário de 44 itens nos alunos.
2. Coleta de texto dos estudantes.
3. Pré-processamento dos textos coletados.
4. Contagem das frequências das categorias do LIWC contidas nos textos coletados.
5. Aplicação da técnica para relacionar os traços (resultantes do passo [1](#)) com as frequências das categorias (resultantes do passo [4](#)).

Como pode ser observado, o inventário foi aplicado no passo [1](#). Mas como descrito na Seção [2.2.1](#), as respostas são numéricas (graduadas de 1 à 5). Através disto conclui-se

que é necessária a criação de novas perguntas abertas (respondidas utilizando texto) para que haja dados escritos dos estudantes (que são coletados no passo 2). Portanto, no passo 4, tem-se os escores dos traços de personalidade dos alunos (resultantes do passo 1) e as frequências das categorias do LIWC. No passo 5, foi proposta a utilização da técnica de Regressão Linear, pois o objetivo era prever o valor de uma variável, dadas variáveis independentes em uma certa base de dados. Nesse caso, as variáveis que seriam previstas seriam os traços do modelo *Big Five* (extroversão, neuroticismo, conscienciosidade, afabilidade, abertura) e as independentes seriam as frequências das categorias do dicionário LIWC. Esse treinamento pode ser feito utilizando outros algoritmos classificadores, como Redes Neurais (ou Neuronais), Algoritmo *Naive Bayes*, dentre outros citados na Seção 5.

Como é possível notar na Figura 2, são inseridos como parâmetros para a Regressão Linear do software WEKA (WITTEN et al., 2016) as respostas dos usuários ao questionário, o dicionário LIWC, e os escores dos traços dos usuários resultantes da aplicação do inventário. Neste momento, os textos das respostas já foram processados, para que sejam extraídas somente as palavras úteis para a classificação dos traços, através do método descrito na Seção 3.1.

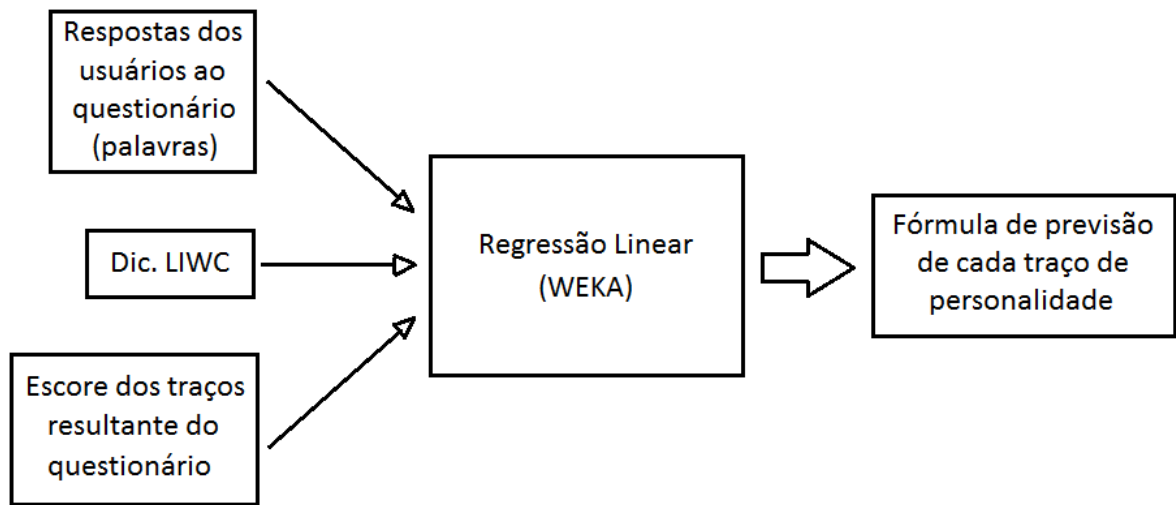


Figura 2 – Treinamento da base

O resultado da Regressão Linear é apresentado através de uma fórmula conforme equação 3.1.

$$Escore_t = \sum_{i=1}^n Coeficiente_i * Categoria_i \quad (3.1)$$

onde t é o traço, $Escore_t$ é o escore dado em porcentagem do traço t , n é a número de categorias disponíveis no dicionário LIWC, $Coeficiente_i$ é o valor do i -ésimo coeficiente relacionado ao treinamento e $Categoria_i$ é a frequência da i -ésima categoria no texto. Para inferir as porcentagens dos traços de personalidade, basta substituir as variáveis pelos valores disponíveis através do processamento da contagem de frequência das cate-

Tabela 1 – Análise do método de Regressão para o traço Extroversão

Extroversão						
Entrada			Saída			
Método de seleção de atributo	Eliminar atributos colineares	Minimal	R^2	Erro absoluto médio	Erro quadrático médio	Desvio quadrático médio
Sem atributo	Não	Não	0,516	9,301	143,637	11,985
Sem atributo	Não	Sim	0,516	9,301	143,637	11,985
Sem atributo	Sim	Não	0,504	9,414	147,013	12,125
Sem atributo	Sim	Sim	0,504	9,414	147,013	12,125
M5	Não	Não	0,432	10,128	168,534	12,982
M5	Não	Sim	0,432	10,128	168,534	12,982
M5	Sim	Não	0,427	9,943	169,864	13,033
M5	Sim	Sim	0,427	9,943	169,864	13,033
Greedy	Não	Não	0,303	11,282	206,719	14,378
Greedy	Não	Sim	0,303	11,282	206,719	14,378
Greedy	Sim	Não	0,38	10,66	183,882	13,56
Greedy	Sim	Sim	0,38	10,66	183,882	13,56

gorias. Note que quanto mais dados são inseridos no treinamento, mais os coeficientes desta fórmula são precisos para a previsão do traço. Conforme dito anteriormente, cinco fórmulas neste formato (uma para cada traço de personalidade) foram obtidas.

Conforme Seção 2.2.4, para o auxílio da organização dos dados e do fluxo, foi utilizado a plataforma de análise de dados chamada KNIME, onde foram instalados os módulos do software WEKA, que contém o método de Regressão Linear. Para que houvesse melhor precisão na aplicação deste método, foi feita uma análise dos parâmetros disponíveis que poderiam ser alterados para a execução. Dentre estes estão o "Método de seleção de atributos", que pode ser sem atributo, M5 ou *Greedy*; o parâmetro "Eliminar atributos colineares", que pode ser ativo ou não; e "Minimal", que também pode estar ativo ou não. Os valores analisados para decidir qual a melhor combinação de parâmetros para a aplicação foram: R^2 , os erros absolutos médios e quadráticos médios, e o desvio quadrático médio. Para que hajam melhores resultados da previsão, deseja-se maximizar o valor de R^2 , e minimizar os valores dos erros e do desvio médio quadrático. As combinações de parâmetros (entrada da análise) e seus respectivos resultados (saída) podem ser observados nas Tabelas 1, 2, 3, 4 e 5.

Através de todas as tabelas pôde-se concluir que a melhor combinação para todos os traços foi a seguinte:

- Sem método de seleção de atributos;
- Não eliminar atributos colineares;
- Não utilizar minimal;

Assim sendo, esta foi a forma que os parâmetros da Regressão Linear foram escolhidos e utilizados. Tendo como resultado desta seção uma fórmula útil para a extração

Tabela 2 – Análise do método de Regressão para o traço Neuroticismo

Neuroticismo						
Entrada			Saída			
Método de seleção de atributo	Eliminar atributos colineares	Minimal	R^2	Erro absoluto médio	Erro quadrático médio	Desvio quadrático médio
Sem atributo	Não	Não	0,51	9,99	190,139	13,789
Sem atributo	Não	Sim	0,51	9,99	190,139	13,789
Sem atributo	Sim	Não	0,413	11,301	227,986	15,099
Sem atributo	Sim	Sim	0,413	11,301	227,986	15,099
M5	Não	Não	0,418	11,568	225,827	15,028
M5	Não	Sim	0,418	11,568	225,827	15,028
M5	Sim	Não	0,252	13,859	290,328	17,039
M5	Sim	Sim	0,252	13,859	290,328	17,039
Greedy	Não	Não	0,353	12,475	251,145	15,848
Greedy	Não	Sim	0,353	12,475	251,145	15,848
Greedy	Sim	Não	0,181	14,495	317,978	17,832
Greedy	Sim	Sim	0,181	14,495	317,978	17,832

Tabela 3 – Análise do método de Regressão para o traço Abertura

Abertura						
Entrada			Saída			
Método de seleção de atributo	Eliminar atributos colineares	Minimal	R^2	Erro absoluto médio	Erro quadrático médio	Desvio quadrático médio
Sem atributo	Não	Não	0,44	8,159	104,792	10,237
Sem atributo	Não	Sim	0,44	8,159	104,792	10,237
Sem atributo	Sim	Não	0,328	8,854	125,768	11,215
Sem atributo	Sim	Sim	0,328	8,854	125,768	11,215
M5	Não	Não	0,369	8,701	118,056	10,865
M5	Não	Sim	0,369	8,701	118,056	10,865
M5	Sim	Não	0,18	9,754	153,581	12,393
M5	Sim	Sim	0,18	9,754	153,581	12,393
Greedy	Não	Não	0,207	9,607	148,509	12,186
Greedy	Não	Sim	0,207	9,607	148,509	12,186
Greedy	Sim	Não	0,113	9,941	165,972	12,883
Greedy	Sim	Sim	0,113	9,941	165,972	12,883

Tabela 4 – Análise do método de Regressão para o traço Conscienciosidade

Conscienciosidade						
Entrada			Saída			
Método de seleção de atributo	Eliminar atributos colineares	Minimal	R^2	Erro absoluto médio	Erro quadrático médio	Desvio quadrático médio
Sem atributo	Não	Não	0,571	7,007	78,685	8,87
Sem atributo	Não	Sim	0,571	7,007	78,685	8,87
Sem atributo	Sim	Não	0,514	7,658	89,046	9,436
Sem atributo	Sim	Sim	0,514	7,658	89,046	9,436
M5	Não	Não	0,417	7,964	106,909	10,34
M5	Não	Sim	0,417	7,964	106,909	10,34
M5	Sim	Não	0,44	8,091	102,727	10,135
M5	Sim	Sim	0,44	8,091	102,727	10,135
Greedy	Não	Não	0,395	8,198	111,031	10,537
Greedy	Não	Sim	0,395	8,198	111,031	10,537
Greedy	Sim	Não	0,36	8,858	117,277	10,829
Greedy	Sim	Sim	0,36	8,858	117,277	10,829

Tabela 5 – Análise do método de Regressão para o traço Afabilidade

Afabilidade						
Entrada			Saída			
Método de seleção de atributo	Eliminar atributos colineares	Minimal	R^2	Erro absoluto médio	Erro quadrático médio	Desvio quadrático médio
Sem atributo	Não	Não	0,589	6,977	87,718	9,366
Sem atributo	Não	Sim	0,589	6,977	87,718	9,366
Sem atributo	Sim	Não	0,534	7,522	99,297	9,965
Sem atributo	Sim	Sim	0,534	7,522	99,297	9,965
M5	Não	Não	0,557	7,27	94,383	9,715
M5	Não	Sim	0,557	7,27	94,383	9,715
M5	Sim	Não	0,461	8,258	114,896	10,719
M5	Sim	Sim	0,461	8,258	114,896	10,719
Greedy	Não	Não	0,403	8,647	127,254	11,281
Greedy	Não	Sim	0,403	8,647	127,254	11,281
Greedy	Sim	Não	0,372	8,579	133,955	11,574
Greedy	Sim	Sim	0,372	8,579	133,955	11,574

da porcentagem dos traços de personalidade, é possível fazer o processamento dos textos de cada estudante, a fim de realizar a inferência de seus traços de personalidade.

3.2 Processamento dos textos para a inferência dos traços de personalidade

O processamento básico do texto é feito utilizando como base o método citado na Seção 3.1. Como pode-se observar na Figura 3, através do tratamento do texto, é extraída uma lista das palavras importantes. Com elas é realizada a contagem das frequências das categorias do dicionário LIWC, assim como na Seção 3.1.1. Estas frequências foram utilizadas na base de treinamento para que sejam inferidos os escores dos traços de personalidade. Para inferir tais porcentagens, basta substituir as variáveis da fórmula gerada como na Seção 3.1.2 pelos valores disponíveis através do processamento da contagem de frequência das categorias.

Para o auxílio do controle do fluxo de dados observado na Figura 3, foi utilizado o software KNIME (exposto na Seção 2.2.4). Nele é possível incorporar todos os programas feitos no *Python*, juntamente com a biblioteca NLTK. Portanto, o processamento completo é feito nesta plataforma. A Regressão Linear utilizada para a análise feita na Seção 3.1.2 faz parte do pacote de ferramentas do software WEKA (WITTEN et al., 2016), que foi instalado separadamente no KNIME.

Como resultado do processamento geral ilustrado na Figura 3, obtêm-se todas as porcentagens dos traços dos estudantes.

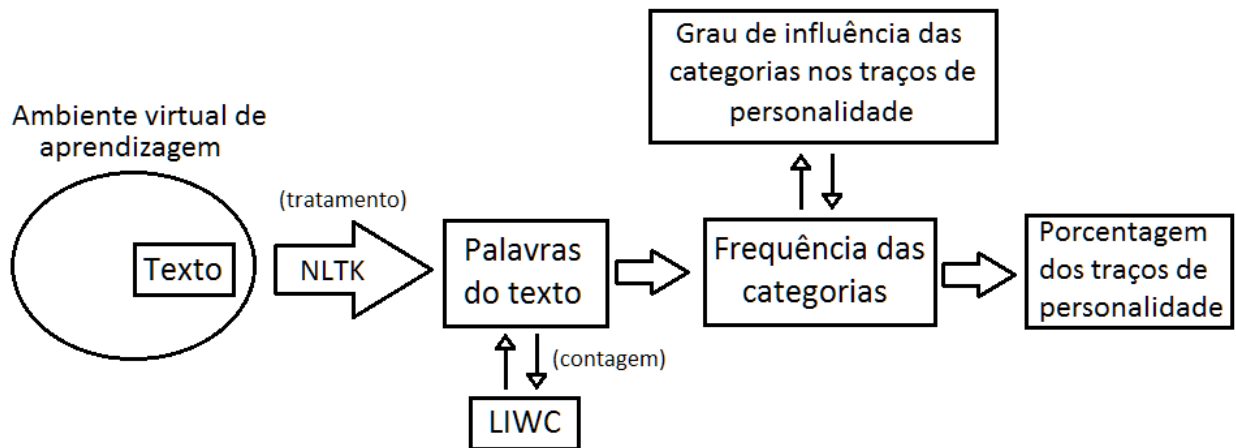


Figura 3 – Fluxograma do processamento geral utilizando a base de treinamento

3.3 Integração com o Ambiente Virtual de Aprendizagem Moodle

O AVA utilizado neste estudo, o *Moodle*, possui um banco de dados que armazena todos os diálogos e mensagens de todos os usuários, como também o histórico de login, fórum, cursos disponíveis, grupos de alunos, usuários e senhas, dentre outros. Como o objetivo deste trabalho é somente o processamento dos textos encontrados nos chats (escritos por alunos) e mensagens, a integração é feita com foco nas tabelas de mensagens: "message" e "message_read". Essas tabelas são referentes às mensagens trocadas pelos usuários de maneira privada. O banco de dados deste AVA funciona da seguinte maneira: primeiramente, o que foi digitado pelo estudante é inserido na tabela "message". Assim que esta mensagem é lida pelo destinatário, ela é transferida para a tabela "message_read" (DOUGIAMAS; TAYLOR, 2003). Portanto, extração dos textos do AVA serão nas tabelas citadas anteriormente. Esta possuem dois atributos relevantes: "useridfrom", que é o ID do usuário autor, e "fullmessage", que é a mensagem digitada pelo aluno. Ambos são do tipo *String*. Assim como mencionado na Seção 3.2, foi utilizado o software KNIME para a organização do fluxo de dados e para o processamento. Nele é possível fazer a extração diretamente do banco de dados do *Moodle* através do módulo chamado *Mysql Connector* (BERTHOLD et al., 2009), buscando o conteúdo das tabelas "message" e "message_read".

4 Resultados

O processamento dos diálogos descrito na Seção 3.2 foi aplicado em 100 indivíduos, com a finalidade de fazer uma análise de qualidade da predição do estudo deste trabalho. Foi executado também o inventário de 44 itens do modelo *Big Five* descrito na Seção 2.2.1 nestes mesmos indivíduos, para se ter um resultado considerado válido, juntamente com as questões extras abertas inseridas para coletar texto escrito. O intuito é fazer uma comparação entre as porcentagens calculadas através do processamento da Regressão Linear e as que foram resultantes da aplicação do inventário. O conjunto de treinamento foi formado por 80 indivíduos e os 20 restantes formaram o conjunto de teste.

A Tabela 6 é composta pelos valores previstos pela Regressão Linear e pelos reais (resultantes da aplicação do questionário de 44 itens) dos cinco traços de personalidade referentes à cada um dos 20 indivíduos testados.

Tabela 6 – Resultados

Indivíduo	Extroversão		Neuroticismo		Afabilidade		Abertura		Conscienciosidade	
	Previsto	Real	Previsto	Real	Previsto	Real	Previsto	Real	Previsto	Real
1	10.198	50.0	47.405	62.5	54.557	59.38	34.419	50.0	56.400	42.5
2	54.122	28.13	74.446	59.38	34.220	62.5	76.057	42.5	50.495	87.5
3	31.665	46.88	60.769	43.75	47.057	62.5	54.166	40.0	65.142	47.5
4	38.918	65.63	82.166	37.5	55.937	84.38	31.334	72.5	53.054	80.0
5	35.563	31.25	28.495	84.38	72.780	68.75	50.975	82.5	61.513	85.0
6	35.563	31.25	28.495	84.38	72.780	68.75	50.975	82.5	61.513	85.0
7	-1.180	59.38	46.930	53.13	50.256	50.0	56.649	75.0	67.324	47.5
8	62.417	56.25	27.928	21.88	60.472	56.25	50.352	80.0	62.171	60.0
9	50.401	40.63	31.076	40.63	59.445	81.25	63.363	67.5	62.824	87.5
10	60.372	65.63	38.079	40.63	60.422	68.75	62.414	62.5	61.034	62.5
11	30.431	53.13	63.079	40.63	66.701	37.5	66.341	57.5	62.134	65.0
12	49.992	43.75	43.024	50.0	119.257	65.63	69.520	72.5	72.873	55.0
13	41.121	75.0	42.713	53.13	51.492	65.63	68.526	72.5	77.905	90.0
14	75.060	31.25	25.048	93.75	58.027	50.0	30.542	77.5	67.519	37.5
15	61.947	87.5	35.909	25.0	71.645	59.38	68.075	92.5	63.903	87.5
16	43.242	75.0	57.267	43.75	63.726	62.5	68.659	75.0	60.785	70.0
17	18.364	43.75	60.911	34.38	73.522	71.88	20.266	60.0	53.541	60.0
18	44.685	78.13	53.281	37.5	63.357	87.5	55.516	70.0	65.341	85.0
19	62.667	46.88	85.324	62.5	74.598	68.75	74.355	60.0	56.052	52.5
20	60.037	40.63	-34.068	53.13	70.500	68.75	45.882	70.0	71.011	60.0

Conforme a Tabela 6 existem previsões próximas do esperado, e predições com valores muito fora do real. O Indivíduo 1, por exemplo, obteve como resultado da aplicação do questionário um escore de 50% no traço chamado extroversão. O valor previsto pela regressão neste caso foi de 10.198%. Analisando psicologicamente, este indivíduo é considerado neutro em relação à extroversão, mas foi inferido um valor baixo para o escore deste traço. Isto permite identificá-lo como introvertido, sendo uma conclusão equivocada. Em contrapartida, existem indivíduos que foram classificados quase que exatamente com

o valor esperado. Um exemplo deste caso é o Indivíduo 10, que possui um escore de 62.5% (através do questionário), e a Regressão previu 62.41%.

A existência de casos muito bons e outros muito ruins exige que seja feito o cálculo dos erros médios (mencionados na Seção 2.2.4) para que possa tirar uma conclusão geral da análise desses 20 indivíduos. A Tabela 7 é resultante dos cálculos dos erros referidos acima, na qual é possível observar que os valores foram relevantes, principalmente do traço neuroticismo. Isso implica em dificuldade em atingir alta eficiência na previsão da Regressão Linear. Portanto, os resultados não foram muito eficazes. Os valores de R^2 foram todos menores do que zero. Isso significa que para esse modelo de Regressão, não é eficaz analisar estes valores, ou seja, R^2 não é um coeficiente de determinação que deve ser considerado (CAMERON; WINDMEIJER, 1997).

Tabela 7 – Erros

	Extroversão	Neuroticismo	Afabilidade	Abertura	Conscienciosidade
R^2	-1.562	-2.418	-1.751	-2.481	-0.280
Erro absoluto médio	22.803	25.663	13.576	20.297	16.347
Erro quadrático médio	737.122	1195.936	361.668	600.273	364.532
Desvio quadrático médio	27.150	34.582	19.017	24.500	19.092

Como resultado deste estudo pode ser citado também o artigo (NETO; FERREIRA; FERNANDES, 2017), que foi feito paralelo a esta pesquisa (conforme dito na Seção 2.1). Neste, é estudada a relação entre categorias do LIWC com os traços do *Big Five*. Então, para verificar isso, foi utilizado Correlação Linear. O resultado desta é um número real entre 0 e 1, onde 0 é pouco e 1 muito relacionado. Então, foram criadas tabelas de relações em que as linhas referem-se às categorias do LIWC, as colunas referem-se aos traços de personalidade e o conteúdo é a correlação entre categoria e traço.

A Tabela 8 correlaciona os traços neuroticismo, extroversão e afabilidade com as categorias *negative emotion*, *anger*, *anxiety*, *sadness* e *positive emotion*. A justificativa da escolha destas categorias e traços de personalidade é devido à definição destes, pois referem-se às palavras afetivas.

Tabela 8 – Correlações entre traços de personalidade e categorias de palavras afetivas

	Neuroticismo	Extroversão	Afabilidade
Neg. emotion	0.19	-0.16	-0.77
Anger	-0.49	0	-0.58
Anxiety	0.29	0.35	0
Sadness	0.29	0	0
Pos. emotion	-0.05	0.37	0.76

Através da Tabela 8, nota-se a alta relação do traço afabilidade com a categoria *positive emotion* e relação negativa com *negative emotion*. Isto era esperado, pois indivíduos afáveis tendem a ser cooperativos, bondosos e fáceis de lidar, justificando estes

resultados. Da mesma forma, foi criada a Tabela 9, que correlaciona os traços conscienciosidade e abertura com as categorias *insight*, *work* e *achieve*. É notada a alta relação do traço conscienciosidade com *work* e *achieve*. Isso se verifica, pois indivíduos conscienciosos tendem a se organizar, e estabelecerem metas para a conclusão de um trabalho.

Tabela 9 – Correlações entre traços de personalidade e categorias de preocupações pessoais

	Conscienciosidade	Abertura à novas experiências
Insight	-0.32	0.02
Work	0.61	-0.66
Achieve	0.61	-0.22

5 Conclusões

Através da pesquisa realizada, foram alcançados os principais objetivos de classificação de traços, mesmo não sendo exatamente como o esperado. A categorização não foi precisa, tendo com base os altos graus de erros calculados na Seção 4. Com base nos resultados, foi possível notar a grande inexatidão da previsão do método de Regressão Linear aplicado neste estudo. Isso se deve ao fato de que os textos utilizados nos testes não foram escritos na linguagem formal, os usuários de Internet preferem escrever de uma maneira que se baseia em abreviações de palavras. Isto ocorre para facilitar e agilizar a escrita nas ferramentas da Internet. Por exemplo: a palavra "hoje", em um conversa na Internet, será escrita do seu modo abreviado: "hj". Isso é um fator prejudicial para o processamento de texto realizado neste estudo. Como no dicionário LIWC não contém essas abreviações, as palavras que se encaixam neste modo de escrita não são reconhecidas pelo processamento. Portanto, o texto é reduzido somente às que foram inseridas exatamente como são definidas na língua portuguesa.

É importante notar que a partir do aumento do tamanho dos textos, a qualidade da classificação é melhorada. Da mesma forma, o crescimento do treinamento da base ocasiona a melhora do grau de precisão da predição. Isto pode ocorrer quando esta classificação estiver inserida em um AVA, pois com a dinâmica do uso de ferramentas como chat e fóruns durante o processo de aprendizagem, será possível coletar mais textos e, conseqüentemente, melhorar a acurácia dos algoritmos de classificação.

No decorrer da pesquisa, algumas dificuldades foram notadas. Dentre elas, destaca-se a instalação do AVA *Moodle*. A documentação encontrada que auxilia na instalação não esclareceu as dúvidas geradas neste processo. Um fator dificultante para a implementação foi a instalação das bibliotecas "*numpy*", "*pandas*" e "*google.protobuf*" no sistema operacional *Windows*. Elas são necessárias para que os códigos escritos em *Python* sejam executados com ferramentas de análise de dados (MCKINNEY, 2012).

Os objetivos gerais e específicos do trabalho foram alcançados, tendo sido possível utilizar Linguagem Natural e processamento de textos para a classificação de traços de personalidade. Essa classificação foi aplicada a AVA's, com sucesso, neste caso o *Moodle*. A aceitação do artigo relacionado à parte desta pesquisa é relevante. Os resultados foram compatíveis com as condições e promissores, indicando que a pesquisa deve continuar. Esta expansão pode ter como objetivo a melhora da acurácia da predição.

Como trabalhos futuros pode-se estudar e implementar os seguintes pontos:

- O estudo de outras técnicas e algoritmos classificadores para realizar a

inferência dos traços de personalidade:

Assim como mencionado na Seção 3.1.2, existem outras maneiras de realizar o treinamento de uma base para a inferência dos traços de personalidade do modelo *Big Five* neste estudo. A proposta, no primeiro momento do trabalho, foi a utilização da Regressão Linear. Mas na literatura encontram-se diversos modos úteis para a classificação, como Redes Neurais (ou Neuronais), Algoritmo *Naive Bayes*, árvores de decisão, dentre outros.

Um possível trabalho futuro seria a análise da classificação dos traços de personalidade, utilizando outras técnicas de inferência. Essa análise seria constituída pelo cálculo de eficiência, juntamente com a performance destes diferentes algoritmos.

- **A tradução da linguagem informal da Internet para o PT-BR formal utilizado pelo dicionário LIWC:**

De acordo com a Seção 4, os resultados não obtiveram muita acurácia, através do alto índice dos erros calculados. Acredita-se que a linguagem informal influenciou na precisão da inferência. Uma maneira de solucionar este problema é fazer esta tradução.

Conforme ilustrado na Figura 4, a linguagem da Internet é caracterizada por abreviações das palavras. Neste exemplo, a palavra "Onde" da linguagem formal é substituída por "Ond", "você" é substituída por "vc", e "hoje" por "hj". Uma maneira simples de resolver este problema, sem precisar fazer o tratamento do texto completo (que é caro computacionalmente), é inserir no dicionário LIWC as palavras abreviadas, com as categorias das palavras originais. Por exemplo: a palavra "hoje" está contida nas categorias "relative" e "time" do dicionário. Então para que a abreviação "hj" seja válida (tenha o mesmo sentido do que a formal) basta inserí-la no LIWC com as mesmas categorias da palavra original ("relative" e "time"). Desta forma, ela influenciará da mesma forma no processamento do que a sua respectiva palavra formal ("hoje"). Assim sendo, é necessário que seja feito um estudo utilizando dados de Internet para inferir as palavras originais, dada as abreviações; ou é possível que seja feito através da análise aparente da palavra, sem o apoio de um processamento computacional.

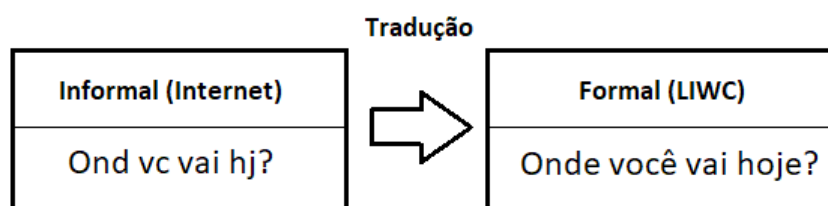


Figura 4 – Tradução da linguagem informal para a formal

- **O estudo de bigramas e trigramas:**

Neste trabalho, foi utilizado para todo o processamento cada palavra de maneira isolada. Como resultado do pré-processamento do texto descrito na Seção 3.1, tem-se as palavras isoladas para que seja feito todo o estudo. Porém, existem expressões com duas ou três palavras - chamadas bigramas e trigramas, respectivamente - que têm sentidos diferentes do que estas mesmas isoladas e, conseqüentemente, influenciarão de maneira equivocada nos traços de personalidade. Por exemplo, a expressão "não aceito" pode influenciar de maneira contrária do que o sentido que lhes é dado ao isolá-las ("não" e "aceito"). Ao categorizá-las no dicionário LIWC, a palavra "não" pertence à categoria "*negate*", e "aceito" às categorias "*swear*", "*humans*", "*posemo*", "*insight*", "*tentat*" e "*achieve*". Pode-se observar que as categorias das duas palavras separadas são completamente diferentes do que se elas estivessem juntas. A expressão "não aceito" poderia ser categorizada somente na categoria *negate* - que caracteriza palavras negativas - e provavelmente influenciaria de maneira negativa no traço extroversão. Deste modo, pode-se concluir que esta pesquisa pode aumentar ainda mais o grau de precisão da predição estudada.

Referências

- ALAM, F.; RICCARDI, G. Fusion of acoustic, linguistic and psycholinguistic features for speaker personality traits recognition. In: IEEE. *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. [S.l.], 2014. p. 955–959. Citado na página 9.
- ALAVI, M. Computer-mediated collaborative learning: An empirical evaluation. *MIS quarterly*, JSTOR, p. 159–174, 1994. Citado na página 7.
- ALTANOPOULOU, P.; TSELIOS, N. How does personality affect wiki-mediated learning? In: IEEE. *Interactive Mobile Communication Technologies and Learning (IMCL), 2015 International Conference on*. [S.l.], 2015. p. 16–18. Citado na página 7.
- BERTHOLD, M. R. et al. Knime-the konstanz information miner: version 2.0 and beyond. *AcM SIGKDD explorations Newsletter*, ACM, v. 11, n. 1, p. 26–31, 2009. Citado 2 vezes nas páginas 15 e 23.
- BIRD, S. Nltk: the natural language toolkit. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. *Proceedings of the COLING/ACL on Interactive presentation sessions*. [S.l.], 2006. p. 69–72. Citado na página 13.
- BOZDOGAN, H. Model selection and akaike’s information criterion (aic): The general theory and its analytical extensions. *Psychometrika*, Springer, v. 52, n. 3, p. 345–370, 1987. Citado na página 15.
- CAMERON, A. C.; WINDMEIJER, F. A. An r-squared measure of goodness of fit for some common nonlinear regression models. *Journal of Econometrics*, Elsevier, v. 77, n. 2, p. 329–342, 1997. Citado 2 vezes nas páginas 16 e 25.
- CARVALHO, P.; SILVA, M. J. Sentilex-pt: Principais características e potencialidades. *Oslo Studies in Language*, v. 7, n. 1, 2015. Citado na página 9.
- DOUGIAMAS, M.; TAYLOR, P. Moodle: Using learning communities to create an open source course management system. 2003. Citado na página 23.
- FILHO, P. P. B.; PARDO, T. A.; ALUISIO, S. M. An evaluation of the brazilian portuguese liwc dictionary for sentiment analysis. In: *Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology (STIL)*. [S.l.: s.n.], 2013. p. 215–219. Citado 2 vezes nas páginas 9 e 14.
- GOLDBERG, L. R. Language and individual differences: The search for universals in personality lexicons. *Review of personality and social psychology*, v. 2, n. 1, p. 141–165, 1981. Citado 6 vezes nas páginas 9, 10, 11, 14, 17 e 18.
- JOHN, O. P.; DONAHUE, E. M.; KENTLE, R. L. *The big five inventory—versions 4a and 54*. [S.l.]: Berkeley, CA: University of California, Berkeley, Institute of Personality and Social Research, 1991. Citado na página 11.

- KUMAR, R.; ROSÉ, C. P. Triggering effective social support for online groups. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, ACM, v. 3, n. 4, p. 24, 2014. Citado na página 7.
- MACHADO, A. A. et al. Personalitatem lexicon: um léxico em português brasileiro para mineração de traços de personalidade em textos. In: UNIVERSIDADE FEDERAL DE ALAGOAS-UFAL. *Conferência Latino-americana de Objetos e Tecnologias de Aprendizagem, X; Congresso Brasileiro de Informática na Educação, IV; Simpósio Brasileiro de Informática na Educação, XXVI*. [S.l.], 2015. Citado 2 vezes nas páginas 9 e 14.
- MAGNISALIS, I.; DEMETRIADIS, S.; KARAKOSTAS, A. Adaptive and intelligent systems for collaborative learning support: A review of the field. *IEEE transactions on Learning Technologies*, IEEE, v. 4, n. 1, p. 5–20, 2011. Citado na página 7.
- MCKINNEY, W. *Python for data analysis: Data wrangling with Pandas, NumPy, and IPython*. [S.l.]: "O'Reilly Media, Inc.", 2012. Citado na página 27.
- NA, S.-H. et al. Improving opinion retrieval based on query-specific sentiment lexicon. In: SPRINGER. *European Conference on Information Retrieval*. [S.l.], 2009. p. 734–738. Citado na página 9.
- NETO, A. T.; FERREIRA, T. B.; FERNANDES, M. A. Development of a big-five personality traits classification approach via analysis of texts in brazilian portuguese. XIII Simpósio Brasileiro de Sistema de Informação, UFLA - Lavras, MG, 2017. Citado 2 vezes nas páginas 10 e 25.
- PENNEBAKER, J. W.; FRANCIS, M. E.; BOOTH, R. J. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, v. 71, p. 2001, 2001. Citado na página 9.
- RUSSELL, S. J. et al. *Artificial intelligence: a modern approach*. [S.l.]: Prentice hall Upper Saddle River, 2003. v. 2. Citado na página 13.
- TAUSCZIK, Y. R.; PENNEBAKER, J. W. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of language and social psychology*, Sage Publications Sage CA: Los Angeles, CA, v. 29, n. 1, p. 24–54, 2010. Citado na página 14.
- WITTEN, I. H. et al. *Data Mining: Practical machine learning tools and techniques*. [S.l.]: Morgan Kaufmann, 2016. Citado 3 vezes nas páginas 15, 19 e 22.