UNIVERSIDADE DE SÃO PAULO
ESCOLA DE ARTES, CIÊNCIAS E HUMANIDADES
PROGRAMA DE PÓS-GRADUAÇÃO EM SISTEMAS DE INFORMAÇÃO

BÁRBARA BARBOSA CLAUDINO DA SILVA

Reconhecimento de traços de personalidade com base em textos

São Paulo

2018

BÁRBARA BARBOSA CLAUDINO DA SILVA

Reconhecimento de traços de personalidade com base em textos

Dissertação apresentada à Escola de Artes, Ciências e Humanidades da Universidade de São Paulo para obtenção do título de Mestre em Ciências pelo Programa de Pós-graduação em Sistemas de Informação.

Área de concentração: Metodologia e Técnicas da Computação

Versão corrigida contendo as alterações solicitadas pela comissão julgadora em 27 de Fevereiro de 2018. A versão original encontra-se em acervo reservado na Biblioteca da EACH-USP e na Biblioteca Digital de Teses e Dissertações da USP (BDTD), de acordo com a Resolução CoPGr 6018, de 13 de outubro de 2011.

Orientador: Prof. Dr. Ivandré Paraboni

São Paulo

2018

Autorizo a reprodução e divulgação total ou parcial deste trabalho, por qualquer meio convencional ou eletrônico, para fins de estudo e pesquisa, desde que citada a fonte.

CATALOGAÇÃO-NA-PUBLICAÇÃO

(Universidade de São Paulo. Escola de Artes, Ciências e Humanidades. Biblioteca)
CRB-8 4936

Silva, Bárbara Barbosa Claudino da
Reconhecimento de traços de personalidade com base em
textos / Bárbara Barbosa Claudino da Silva ; orientador Ivandré
Paraboni. – 2018.
98 f. : il.

Dissertação (Mestrado em Ciências) - Programa de Pós-
Graduação em Sistemas de Informação, Escola de Artes,
Ciências e Humanidades, Universidade de São Paulo.
Versão corrigida

1. Linguagem natural - Processamento. 2. Personalidade -
Análise. 3. Facebook. I. Paraboni, Ivandré, orient. II. Título.

CDD 22.ed.– 006.35

Dissertação de autoria de Bárbara Barbosa Claudino Silva, sob o título “**Reconhecimento de traços de personalidade com base em textos**”, apresentada à Escola de Artes, Ciências e Humanidades da Universidade de São Paulo, para obtenção do título de Mestre em Ciências pelo Programa de Pós-graduação em Sistemas de Informação, na área de concentração Metodologia e Técnicas da Computação, aprovada em 27 de Fevereiro de 2018 pela comissão julgadora constituída pelos doutores:

Prof. Dr. Ivandré Paraboni

Instituição: Universidade de São Paulo (USP)

Presidente

Prof. Dr. Fernando Fagundes Ferreira

Instituição: Universidade de São Paulo (USP)

Profa. Dra. Arianne Machado Lima

Instituição: Universidade de São Paulo (USP)

Profa. Dra. Profa. Dra. Ariadne Maria B. Rizzoni Carvalho

Instituição: Universidade de Campinas (Unicamp)

Dedico este trabalho a todas as pessoas que acreditaram em mim e que me ajudaram, de alguma maneira, a concluí-lo.

Agradecimentos

Agradeço a minha família, principalmente aos meus pais Laureci Barbosa e Ricardo Claudino da Silva, e meus irmãos, Raiza e Lucas, por me apoiarem durante todo o período do mestrado.

Ao meu orientador, Prof. Dr. Ivandré Paraboni, pela paciência, revisões infinitas, confiança e por me guiar nesses anos de mestrado.

Ao meu noivo, Guilherme Figueirôa, pela compreensão, por me incentivar e apoiar todos os dias, não só no mestrado, mas em todos os meus desafios.

A todos os professores do Programa de Pós-graduação em Sistemas de Informação especialmente a Profa. Dra. Sarajane Marques Peres, pelo apoio desde o trabalho de conclusão de curso no bacharelado e por ter me incentivado a iniciar o programa de mestrado.

Às amigas de projetos de inclusão de mulheres na tecnologia agradeço pela compreensão e ajuda. Às amigas Jéssica, Priscilla e Vivian, que me ajudaram a “manter a sanidade” durante as disciplinas e a pesquisa.

Por fim, meus agradecimentos a banca de qualificação pelas contribuições e a participação dos membros da banca examinadora da defesa.

“A nossa personalidade deve ser indevassável, mesmo por nós próprios: daí o nosso dever de sonharmos sempre, e incluírmo-nos nos nossos sonhos, para que nos não seja possível ter opiniões a nosso respeito.”
(Fernando Pessoa, in ‘Livro do Desassossego’)

Resumo

SILVA, Bárbara Barbosa Claudino da. **Reconhecimento de traços de personalidade com base em textos**. 2018. 98 f. Dissertação (Mestrado em Ciências) – Escola de Artes, Ciências e Humanidades, Universidade de São Paulo, São Paulo, 2018.

Apresentamos uma pesquisa na área de Processamento de Linguagem Natural, para reconhecimento de personalidade com base em textos da língua portuguesa. Neste trabalho utilizamos textos provenientes da rede social *Facebook*, em conjunto com o modelo de personalidade dos Cinco Grandes Fatores, para construir um corpus rotulado com as personalidades de seus autores e, após a identificação das características mais relevantes para o reconhecimento de personalidade, construir modelos computacionais utilizando essas características. Utilizando-se métodos provenientes de léxicos, como o dicionário LIWC ou atributos psicolinguísticos, e métodos provenientes do próprio texto, como *bag of words*, representação distribuída de palavras e de documentos foram desenvolvidos modelos para reconhecimento de personalidade sem a necessidade de outros métodos mais comumente utilizados para essa tarefa, como inventários ou entrevistas com psicólogos. Os resultados dos métodos de representação distribuída são ligeiramente superiores do que os resultados utilizando o dicionário LIWC, com a vantagem de não exigirem recursos dependentes de um idioma específico.

Palavras-chaves: processamento de linguagem natural, modelo dos cinco grandes fatores, personalidade, *Big Five*

Abstract

Silva, Bárbara Barbosa Claudino da. **Personality traits recognition through texts**. 2018. 98 p. Dissertation (Master of Science) – School of Arts, Sciences and Humanities, University of São Paulo, São Paulo, 2018.

We present a research proposal in the Natural Language Processing field, to recognize personality through texts in the portuguese language. Using texts from the social network Facebook we built a corpus labeled with authors Big-5 personality traits, and after identifying the most relevant atributes to recognize personality, we built computational models based on those attributes. The model was expected to recognize personality without the help of any other methods commonly used in this task, such as inventories or interviews with psychologists. Using lexical methods such as the LIWC dictionary or psycholinguistic attributes, and methods from the text itself, such as bag of words, distributed representation of words and documents, we obtained models for personality recognition without the need of other methods most commonly used for this task. The results of distributed representation methods are slightly better than the results using the LIWC dictionary, with the advantage of not requiring features dependent on a specific language.

Keywords: Natural language processing, big five model, personality, big five

Lista de figuras

Figura 1 – Modelo de uma rede CBOW	25
Figura 2 – Modelo de uma rede skip-gram	25
Figura 3 – Exemplo de operação vetorial possível em um word2vec	27
Figura 4 – Modelo de aprendizado da representação distribuída de um documento	27
Figura 5 – Bloco de memória com apenas uma célula de rede LSTM com as portas analógicas de esquecimento, entrada e saída	29
Figura 6 – Divisão para classificação binária	40
Figura 7 – Divisão para multi-classificação	40
Figura 8 – Tela de autorização para o aplicativo no <i>Facebook</i>	59
Figura 9 – Trecho do IGFP-5 da aplicação <i>web</i>	59
Figura 10 – Tela ilustrativa de resultados	60
Figura 11 – Distribuição da dimensão Extroversão no conjunto de dados	61
Figura 12 – Distribuição da dimensão Agradabilidade no conjunto de dados	61
Figura 13 – Distribuição da dimensão Meticulosidade no conjunto de dados	62
Figura 14 – Distribuição da dimensão Instabilidade no conjunto de dados	62
Figura 15 – Distribuição da dimensão Liberalidade no conjunto de dados	62

Lista de quadros

Quadro 1 – Características e comportamentos de cada uma das dimensões de personalidade	20
Quadro 2 – Resumo dos principais trabalhos revisados, apresentando a subseção que ele se encontra, método utilizado, tipo de conhecimento, número de participantes e dimensões de personalidade	56
Quadro 3 – Resumo das estratégias apresentadas neste capítulo, apresenta-se o nome da estratégia, o método utilizado, o conhecimento e detalhes (se houverem)	77
Quadro 4 – Resumo das hipóteses referentes as estratégias 1 à 14	90
Quadro 5 – Resumo dos melhores resultados para as estratégias complementares .	90

Lista de tabelas

Tabela 1 – Divisão em positivo e negativo de acordo com a média da dimensão . . .	66
Tabela 2 – Personalidade separadas em extremos - usando a média e desvio padrão	67
Tabela 3 – Resultados para o uso de SVM para a divisão binária ao meio	68
Tabela 4 – Resultados para o uso de SVM para a divisão binária ao extremo . . .	69
Tabela 5 – Medida F dos experimentos 1 à 14 para as cinco dimensões de persona- lidade	80
Tabela 6 – Medida F média das estratégias complementares para classificação . . .	86
Tabela 7 – Medida F das estratégias de classificação ternária	87
Tabela 8 – Medida F das estratégias de classificação ternária	88
Tabela 9 – Erro quadrático médio das estratégias de regressão	89

Sumário

1	Introdução	15
1.1	<i>Objetivo . . . . .</i>	17
1.2	<i>Hipótese . . . . .</i>	17
1.3	<i>Organização do trabalho . . . . .</i>	17
2	Conceitos fundamentais	18
2.1	<i>O modelo dos Cinco Grandes Fatores (CGF) . . . . .</i>	18
2.2	<i>Inventários CGF . . . . .</i>	21
2.3	<i>LIWC - Linguistic Inquiry Word Count . . . . .</i>	22
2.4	<i>Propriedades Psicolinguísticas . . . . .</i>	23
2.5	<i>Representação distribuída de palavras . . . . .</i>	24
2.6	<i>Representação distribuída de sentenças e documentos . . . . .</i>	27
2.7	<i>Redes LSTM (Long-Short Term Memory) . . . . .</i>	28
3	Trabalhos relacionados	32
3.1	<i>Abordagens que utilizam textos de ensaios . . . . .</i>	32
3.1.1	Abordagem lexical usando gramáticas funcionais sistêmicas	32
3.1.2	Uso de algoritmos de classificação, regressão e <i>ranking</i>	33
3.1.3	Detecção de personalidade utilizando redes neurais profundas	36
3.2	<i>Abordagens que utilizam textos de blogs . . . . .</i>	38
3.2.1	Abordagem baseada em modelos n-gramas	38
3.2.2	Treino em um corpus pré-processado e teste em um corpus com pouco pré-processamento	42
3.2.3	Uso de regressão logística para seleção de atributos LIWC	43
3.2.4	Detecção de atributos relevantes para reconhecimento de personalidade	44
3.3	<i>Abordagens que utilizam textos de redes sociais . . . . .</i>	46
3.3.1	Reconhecimento de personalidade independente de língua	46
3.3.2	Reconhecimento de personalidade em atualizações de <i>status</i> do <i>Facebook</i>	48
3.3.3	Uso do modelo de personalidade <i>Myers-Briggs</i>	50
3.4	<i>Iniciativas para o português brasileiro . . . . .</i>	52

3.4.1	Mapeamento de personalidade para o português brasileiro	52
3.4.2	Léxico em Português-Brasileiro para mineração de personalidade em textos	53
3.4.3	Estudos apresentados na competição PAN-2015	54
3.5	<i>Considerações</i>	55
4	Estudo exploratório	58
4.1	<i>Implementação da aplicação de coleta de dados</i>	58
4.2	<i>Coleta de dados</i>	60
4.3	<i>Procedimento</i>	63
4.4	<i>Resultados</i>	67
4.5	<i>Considerações</i>	70
5	Reconhecimento de personalidade com base em textos	71
5.1	<i>Modelos de reconhecimento de personalidade propostos</i>	71
5.1.1	Estratégia baseada em Bag of Words - BoW	72
5.1.2	Estratégia baseada em conhecimento psicolinguístico	73
5.1.3	Estratégias baseadas em modelos de representação distribuída de palavras	73
5.1.4	Estratégia baseada em modelos de representação distribuída de documentos	75
5.1.5	Estratégias baseadas em redes LSTM	76
5.2	<i>Discussão</i>	76
6	Avaliação	78
6.1	<i>Hipóteses</i>	78
6.2	<i>Descrição dos dados</i>	80
6.3	<i>Resultados</i>	80
6.4	<i>Análise de atributos relevantes para reconhecimento de personalidade</i> .	84
6.5	<i>Outras análises</i>	85
6.5.1	Estratégias complementares	85
6.5.2	Estratégias baseadas em classificação individual de postagens . . .	87
6.5.3	Estratégia baseada em classificação ternária	88
6.5.4	Estratégias baseadas em regressão	89

6.6	<i>Considerações</i>	89
7	Conclusões	91
	Referências ¹	93

---

¹ De acordo com a Associação Brasileira de Normas Técnicas. NBR 6023.

1 Introdução

A personalidade representa as dimensões mais fundamentais de variação do comportamento humano. Cada traço de personalidade e suas diversas combinações mostram perfis psicológicos diferentes e, portanto, preferências distintas para cada indivíduo (ALLPORT; ODBERT, 1936).

Conhecer a personalidade de uma pessoa traz diversas vantagens, como facilitar o processo de seleção de um candidato para uma vaga de emprego, possibilitar uma fácil identificação de líderes, segmentar um mercado com o uso de propagandas personalizadas e muitas outras. Além disso, é possível utilizar esse conhecimento para aplicações de *e-learning*, mudando a estratégia de ensino para cada tipo de personalidade, ou até mesmo para modelagem de discurso gerado automaticamente (WALKER et al., 2014).

Uma forma de representar a personalidade humana que é amplamente aceita na psicologia e em diversas áreas da Ciência da Computação é por meio do modelo dos Cinco Grandes Fatores (CGF). Este modelo categoriza a personalidade humana em cinco dimensões (extroversão, agradabilidade, meticulosidade, instabilidade e liberalidade) e foi proposto a partir de estudos do uso da língua natural que indivíduos empregam para descrever a si mesmos e aos outros (JOHN; SRIVASTAVA, 1999).

As cinco dimensões do modelo CGF compõem a personalidade de um sujeito. Para a obtenção destas dimensões é necessário que um indivíduo responda um inventário de personalidade, que pode ser longo, ou que tenha um especialista para avaliá-lo. Assim, o reconhecimento da personalidade está sujeito à disposição de um indivíduo em responder a um questionário ou à contratação de um especialista para cuidar da avaliação. Com o intuito de simplificar essa tarefa, estudos na área de processamento de linguagem natural buscaram criar modelos de reconhecimento automático de personalidade humana com base em textos. (ARGAMON et al., 2005; OBERLANDER; NOWSON, 2006; NOWSON, 2007; MAIRESSE, 2008; GILL; NOWSON; OBERLANDER, 2009; IACOBELLI et al., 2011; CELLI, 2012; FARNADI et al., 2013; NUNES; TELES; SOUZA, 2013; PLANK; HOVY, 2015; MACHADO et al., 2015; MAJUMDER et al., 2017).

Em especial, diversos estudos têm procurado implementar esta tarefa a partir de textos provenientes de redes sociais como *Facebook* (CELLI, 2012; FARNADI et al., 2013), *Twitter* (PLANK; HOVY, 2015; NUNES; TELES; SOUZA, 2013), blogs (OBERLANDER;

NOWSON, 2006; NOWSON, 2007; GILL; NOWSON; OBERLANDER, 2009; IACOBELLI et al., 2011) e outras (MACHADO et al., 2015). O presente trabalho também está incluído nesta categoria de pesquisa, mais especificamente utilizando textos provenientes do *Facebook* de usuários brasileiros.

Embora o número de trabalhos de reconhecimento de personalidade tenha aumentado, eles são majoritariamente dedicados a língua inglesa (apenas em Celli (2012) e em Nunes, Teles e Souza (2013) os textos estão em outra língua). Estudos deste tipo, de modo geral, tendem a ser concentrados em duas categorias: (1) modelos que utilizam bases de conhecimento externas ao texto (MAIRESSE, 2008; GILL; NOWSON; OBERLANDER, 2009; FARNADI et al., 2013), como dicionários psicolinguísticos do tipo LIWC (PENNEBAKER; FRANCIS; BOOTH, 2001) ou MRC (COLTHEART, 1981), ou (2) modelos que fazem uso de representações puramente textuais baseadas em contagens de n-gramas (OBERLANDER; NOWSON, 2006; NOWSON, 2007; PLANK; HOVY, 2015). De modo geral, a primeira categoria (baseada em conhecimento lexical) tende a apresentar melhores resultados (MAIRESSE, 2008), porém é mais dependente da língua-alvo em questão.

De especial interesse para este trabalho, observamos que diversas tarefas de processamento de linguagem natural têm obtido avanços com uso de representação distribuída de palavras, ou *word embeddings* (MIKOLOV et al., 2013a). Modelos deste tipo, que mapeiam as palavras de um vocabulário em um espaço vetorial de números reais, começaram a ser empregados em alguns estudos iniciais de reconhecimento de personalidade para o inglês (MAJUMDER et al., 2017).

Considerando-se o caso específico do português brasileiro, os estudos de reconhecimento de personalidade a partir de texto existentes são relativamente incipientes e fazem uso de conjuntos de dados em volume reduzido (NUNES; TELES; SOUZA, 2013). Além disso, esses trabalhos não fazem uso de modelos de representação distribuída de palavras.

Com base nestas observações, sugere-se assim a oportunidade de desenvolvimento de uma pesquisa em nível de mestrado tendo como foco o reconhecimento automático (i.e., computacional) de personalidade humana a partir de textos em português.

1.1 *Objetivo*

O objetivo dessa pesquisa é desenvolver e avaliar modelos computacionais de reconhecimento de personalidade a partir de textos em português brasileiro, incluindo alternativas baseadas em representação distribuída de palavras porém não limitadas a estas, procurando identificar quais métodos apresentam melhores resultados no domínio do *Facebook* para usuários brasileiros.

1.2 *Hipótese*

O uso de modelos de reconhecimento de personalidade a partir de texto baseados em representação distribuída de palavras proporcionará resultados comparáveis aos de métodos dependentes de conhecimento lexical, como o dicionário LIWC. Esta hipótese será verificada comparando-se os diversos modelos aplicados à tarefa de classificação binária de personalidade com base no modelo CGF em um corpus de textos proveniente do *Facebook* para usuários brasileiros. Espera-se que os modelos baseados em representação distribuída de palavras apresentem resultados iguais ou superiores ao dos métodos de base considerados, incluídos aqui modelos baseados no conhecimento disponibilizado pelo dicionário LIWC.

1.3 *Organização do trabalho*

Este trabalho está organizado da seguinte forma: o capítulo 2 apresenta alguns conceitos básicos necessários para a compreensão da presente dissertação e dos trabalhos relacionados. O capítulo 3 apresenta as pesquisas relacionadas ao tema de reconhecimento computacional de personalidade com o uso de textos. O capítulo 4 descreve como foi feita a coleta de dados e o estudo exploratório realizado. O capítulo 5 apresenta o trabalho realizado, dividido em subseções para cada um dos experimentos, e a hipótese principal de pesquisa e hipóteses dos experimentos criados. O capítulo 6 trata dos resultados finais de cada um dos experimentos apresentados no capítulo anterior. Finalmente, o capítulo 7 apresenta as conclusões finais do presente trabalho.

2 Conceitos fundamentais

Esta seção trata de conceitos básicos da psicologia e ciência da computação, necessários para uma melhor compreensão deste trabalho. A seção 2.1 apresenta um breve histórico do desenvolvimento do modelo CGF e de suas dimensões fundamentais. A seção 2.2 descreve alguns dos inventários usados para estimar as dimensões do CGF e motiva o uso do inventário IGFP-5 (Inventário dos cinco Grandes Fatores de Personalidade) que utilizaremos neste trabalho. A seção 2.3 contém uma breve apresentação sobre o dicionário LIWC, léxico utilizado em parte considerável dos trabalhos relacionados apresentado no capítulo 3. A seção 2.4 apresenta o dicionário de propriedades psicolinguísticas, ambos os léxicos serão utilizados neste trabalho. A seção 2.5 apresenta o conceito de representações distribuídas de palavras e seu uso para a resolução de problemas de processamento de linguagem natural. A seção 2.6 trata da representação distribuída de documentos, outra técnica utilizada neste trabalho. Por fim a seção 2.7 apresenta um tipo específico de rede neural recorrente, a rede LSTM (Long Short Term Memory) que é utilizada em um dos experimentos apresentados na seção 5.1.5.

2.1 O modelo dos Cinco Grandes Fatores (CGF)

O modelo dos Cinco Grandes Fatores de personalidade (CGF) (GOLDBERG, 1990) é uma taxonomia para a personalidade, ou seja, um modelo usado na psicologia para representar a personalidade humana por meio de cinco dimensões. Em sua forma atual, tal qual apresentado por John e Srivastava (1999), este modelo contempla cinco dimensões ou fatores, a saber: agradabilidade, meticulosidade, extroversão, instabilidade e liberalidade. Graças a essa taxonomia, é possível estabelecer um domínio comum, que facilita a comunicação entre pesquisadores para novas descobertas e permite o estudo de todos os milhares de atributos que tornam um indivíduo único, agrupando características relacionadas (JOHN; SRIVASTAVA, 1999).

As dimensões CGF foram propostas a partir de análises de termos que as pessoas usam para se auto descrever, ou para descrever outros, usando linguagem natural. Partindo do pressuposto de que grande parte dos atributos socialmente relevantes para caracterizar a personalidade estão presentes na linguagem natural, o estudo de Allport e Odbert

(1936) selecionou 18.000 termos do dicionário que poderiam ser usados para descrever um indivíduo. Com base nesse levantamento foram classificadas as seguintes categorias: traços de personalidade, estados temporais, julgamentos de conduta e reputação e características físicas (JOHN; SRIVASTAVA, 1999).

Como a lista de categorias obtida por Allport e Odbert (1936) era muito extensa, o trabalho de Cattell (1943) a utilizou apenas como ponto de partida, e extraiu um subconjunto de 4.500 traços de personalidade. Usando tanto agrupamento empírico quanto o conhecimento da literatura disponível na época, conseguiu-se reduzir esse conjunto para 35 variáveis. Utilizando uma análise de fatores, a partir dessas 35 variáveis iniciais foram identificados 12 fatores, que se tornaram parte do questionário final de 16 fatores de personalidade apresentado por Cattell (1943). Apesar de afirmar-se em Cattell (1943) que os fatores descobertos em seu trabalho obtinham os mesmos resultados, tanto em autoavaliação como em avaliações de terceiros, as reanálises do modelo proposto não confirmaram os fatores obtidos em sua pesquisa devido às limitações para se trabalhar com matrizes de muitas variáveis da época.

Apesar de não ser possível reproduzir o estudo inicial em Cattell (1943), este possibilitou uma nova análise das matrizes por Tupes e Christal (1961). Nesta segunda análise foram identificados ‘cinco fatores relativamente fortes e recorrentes’, estudo este que foi posteriormente replicado por outros autores.

Os fatores descobertos em Tupes e Christal (1961) foram rotulados como extroversão, agradabilidade, meticulosidade, instabilidade e liberalidade. Esses fatores foram denominados Cinco Grandes Fatores (CGF) para enfatizar a ideia que cada um desses fatores é bastante amplo, e que cada um deles resume um grande número de características mais específicas de personalidade. Os fatores são vistos de forma bipolar, ou seja, sempre há um fator e o reverso do mesmo, capturando a noção de que cada pessoa pode apresentar uma variação positiva ou negativa em cada fator. Estas dimensões são sumarizadas no quadro 1.

Quadro 1 – Características e comportamentos de cada uma das dimensões de personalidade

	Características	Comportamentos
Extroversão	Comumente entusiasmados, falantes, gostam de companhia e são bastante sociáveis.	Normalmente tomam a liderança de um projeto, enquanto os introvertidos podem ficar quietos mesmo quando não concordam com outros.
Agradabilidade	Agradáveis, cooperativos, amáveis, generosos e gostam de manter um bom relacionamento com outros.	Consolam amigos em momentos difíceis e dificilmente enganam outras pessoas. Os ‘não agradáveis’ pensam apenas em si mesmos e comumente bajulam outros para obter vantagens.
Meticulosidade	Meticulosos, normalmente são bastante focados, possuem autodisciplina, são organizados e gostam de planejamentos.	Chegam cedo para compromissos e trabalham duro para realizar uma tarefa bem feita. Os não meticulosos postergam tarefas e perdem muito tempo para conseguir atingir um objetivo.
Instabilidade	Neuróticos, são instáveis, maldispostos, preocupados e estressados.	Ficam chateados quando outros ficam bravos com eles e são facilmente intimidados. Os emocionalmente estáveis aceitam as coisas boas e ruins da vida sem reclamar.
Liberalidade	Liberais, são curiosos, imaginativos, criativos e aventureiros.	Buscam atividades que quebram rotinas e aprendem algo novo apenas por prazer. Os não-liberais raramente gostam de mudanças e evitam discussões filosóficas.

Fonte: (JOHN; SRIVASTAVA, 1999; ANDRADE, 2008)

Outros estudos para validar os fatores foram conduzidos em Norman (1963) e Goldberg (1990), onde aborda-se um conjunto de variáveis iniciais diferentes das apresentadas em Cattell (1943), além de pesquisas com auto-avaliação, avaliação por terceiros e diversas combinações de variáveis. Esses estudos mostraram que os fatores do CGF estavam presentes em todas as abordagens, sugerindo que esta é uma estrutura replicável para as grandes dimensões de personalidade para a língua inglesa.

Após os resultados apresentados em Norman (1963) e Goldberg (1990) restava ainda a necessidade de garantir a universalidade do CGF e, para isso, era necessário que testes com outras línguas além do inglês fossem realizados. A partir de 1990 começaram a surgir estudos para o CGF com esse intuito, de modo que observou-se que o modelo apresentou resultados positivos quando replicado para línguas germânicas, também observou-se que é

bastante complexo interpretar seus resultados em outras línguas. Por exemplo, em John e Srivastava (1999) citam-se estudos como Szirmák e Raad (1994) e Caprara e Perugini (1994) que mostram alguns dos problemas presentes na análise do húngaro e do italiano, nas quais uma seleção de variáveis iniciais e problemas de tradução trouxeram dificuldades na interpretação e geração de cinco fatores com as mesmas características do CGF. Apesar disso, quando há mais de cinco dimensões, foi demonstrado que os cinco fatores do CGF estão entre eles e que as outras dimensões que surgem em determinadas línguas são bastante similares às dimensões do CGF. Em John e Srivastava (1999) afirma-se que não existem dados que sugerem que cada cultura e língua possua o seu próprio e único conjunto de dimensões de personalidade. Ao menos no nível de dimensões amplas, as culturas são mais parecidas entre si do que era esperado inicialmente na pesquisa.

2.2 Inventários CGF

A psicologia possui diversas formas de avaliar a personalidade utilizando o modelo CGF, e uma delas é através de inventários (ou questionários) que utilizam certas palavras, em especial adjetivos, ligadas aos cinco fatores. Com a popularização do CGF na comunidade de psicologia, surgiu assim a necessidade de padronizar instrumentos desse tipo. O primeiro questionário desenvolvido com esse intuito foi o inventário *NEO Personality Inventory* (NEO PI), assim denominado por contemplar os fatores *neuroticism* (instabilidade), *extraversion* (extroversão) e *openness* (liberalidade) (MCCRAE; COSTA, 1989). Posteriormente foi proposta uma versão revisada chamada NEO PI-R (*NEO Personality Inventory Revised*) descrita em Costa e McCrae (1992) onde propõe-se uma versão que inclui as dimensões *agreeableness* (agradabilidade) e *conscientiousness* (meticulosidade) e consiste de 240 itens que medem seis facetas de cada uma das dimensões do CGF.

Devido à extensão do inventário NEO PI-R, em Costa e MacCrae (1992) apresenta-se também o inventário NEO-FFI (*NEO Five Factor Inventory*), que utiliza os itens que melhor representam cada um dos cinco fatores. O inventário NEO-FFI é composto de 60 itens, não necessariamente expressando igualmente o conjunto de facetas presentes em cada uma das dimensões do CGF. Contudo, como esse novo inventário ainda tomava bastante tempo e gerava dúvidas devido à ambiguidade dos seus itens, em John, Donahue

e Kentle (1991) propôs-se o inventário BFI (*Big Five Inventory*) - inventário dos cinco grandes fatores de personalidade.

Diferentemente dos outros inventários, o inventário BFI não utiliza apenas itens na forma de características, (por exemplo, o item “original”), mas sim frases curtas contendo a palavra-traço que o item pretende avaliar, (por exemplo, apresentando o item “é original, sempre tem novas ideias”). O BFI é composto de 44 perguntas e as respostas estão em uma escala de Likert de cinco pontos, que variam de 1 (‘Discordo totalmente’) até 5 (‘Concordo totalmente’). Para gerar a pontuação de uma dimensão de personalidade, é necessário conhecer quais perguntas possuem os itens negativos e positivos para cada dimensão, para então computar o valor ou pontuação do indivíduo para cada dimensão de personalidade, conforme detalhado em John, Naumann e Soto (2008).

Por ser mais curto que o NEO-FFI e o NEOPI-R, e por fornecer um contexto mais amplo para o indivíduo que realiza o teste, o inventário BFI tornou-se bastante popular entre os psicólogos (BENET-MARTINEZ; JOHN, 1998). Esse inventário é considerado também mais indicado caso haja pouco tempo para resposta ou não haja ninguém para esclarecer possíveis dúvidas dos indivíduos (JOHN; SRIVASTAVA, 1999), que é o caso do presente trabalho.

A adaptação para o português brasileiro a ser considerada no presente trabalho é o inventário BFI denominado IGFP-5 em Andrade (2008). O IGFP-5 foi elaborado tendo o inventário espanhol como referência e a versão em inglês para validação. Optou-se por usar o IGFP-5 porque, segundo John e Srivastava (1999), não há uma tradução do inventário para o português brasileiro. Além disso, o inventário foi validado em Andrade (2008) para 5.089 indivíduos das cinco regiões brasileiras.

2.3 LIWC - *Linguistic Inquiry Word Count*

O dicionário *Linguistic Inquiry Word Count* (LIWC) é um recurso linguístico computacional amplamente utilizado em abordagens de reconhecimento de personalidade, que categoriza as palavras em um grande conjunto de textos (PENNEBAKER; FRANCIS; BOOTH, 2001).

Inicialmente, o dicionário LIWC foi construído para identificar um grupo de palavras pertencentes a dimensões básicas e cognitivas, mas com o tempo as categorias de palavras

aumentaram de forma considerável. Estas categorias incluem emoções positivas, emoções negativas, palavras funcionais, estratégias cognitivas, vários tipos de conteúdo temático e vários elementos de composição linguística. Por exemplo, a palavra “chorou” faz parte das categorias do dicionário LIWC que representam tristeza, emoção negativa, verbo e passado.

As categorias do dicionário LIWC foram criadas com base nas opiniões de especialistas humanos que usaram dicionários para elencar as palavras que mais tinham relação com uma determinada categoria. Quando os especialistas terminavam uma lista de palavras, um outro grupo de especialistas avaliava cada uma das palavras e decidia se ela traduzia o conceito da categoria. Uma palavra só era mantida no dicionário caso os dois grupos de especialistas tivessem concordado.

O dicionário LIWC foi traduzido para o português brasileiro, e em Filho, Pardo e Aluisio (2013) faz-se uma comparação do dicionário LIWC para análise de sentimentos, com o *Opinion Lexicon* (SOUZA et al., 2011) e o dicionário *Sentilex* (SILVA; CARVALHO; SARMENTO, 2012). O dicionário LIWC obtém um desempenho superior comparado aos outros léxicos, indicando especialmente a representação de emoções positivas. O dicionário LIWC é amplamente empregado no reconhecimento de personalidade humana a partir de textos (MAIRESSE, 2008; CELLI, 2012; GILL; NOWSON; OBERLANDER, 2009; FARNADI et al., 2013; IACOBELLI et al., 2011) e sua versão em português será considerada no presente trabalho.

2.4 *Propriedades Psicolinguísticas*

As propriedades psicolinguísticas das palavras são aquelas que dependem das experiências pessoais dos indivíduos que fazem uso dessas palavras, sendo portanto diferente de outras propriedades, como significado ou forma. O conjunto de dados de propriedades psicolinguísticas mais conhecido para o inglês é o *MRC Psycholinguistic database* (COLTHEART, 1981) que contém estatísticas para mais de 150.000 palavras com 27 propriedades psicolinguísticas. O MRC foi utilizado em Mairesse (2008) para reconhecimento de personalidade em textos em inglês.

Exemplos de propriedades disponibilizadas na base MRC incluem capacidade de imaginação, ou seja, a velocidade com que uma palavra traz uma imagem à sua mente;

concretude, que indica o grau em que uma palavra refere-se a coisas, pessoas, ou lugares; familiaridade, que indica o grau em que as pessoas usam e conhecem certas palavras; frequência subjetiva que é uma estimativa da frequência das palavras escritas ou faladas por um indivíduo e idade de aquisição ou (aoa - *age of acquisition*) que é uma estimativa da idade na qual uma palavra foi aprendida.

No caso do Português brasileiro, estas mesmas cinco propriedades foram disponibilizadas pela base de atributos psicolinguísticos (SANTOS et al., 2017) e dispõe de 26.874 palavras anotadas com essas propriedades.

2.5 Representação distribuída de palavras

Assim como em outras tarefas de PLN, o reconhecimento de personalidade a partir de texto pode fazer uso de uma representação explícita do seu conteúdo. Dentre diversas estratégias possíveis de representação textual, avanços recentes em diversas áreas têm demonstrado as vantagens de modelos baseados em representações distribuídas de palavras do tipo *word embeddings*. Estes modelos são descritos brevemente nesta seção.

Na representação distribuída de palavras, as palavras de um vocabulário são mapeadas em um espaço vetorial de números reais (*word embeddings*) que capturam um grande número de relações sintáticas e semânticas (MIKOLOV; YIH; ZWEIG, 2013). A representação distribuída de palavras pode ser usada para diversas tarefas, incluindo experimentos para ranqueamento de propagandas e correspondência de conteúdo como em Bai et al. (2009); para complementar conhecimento e relacionamentos em bases de conhecimento ou ontologias, como em Weston et al. (2013); ou ainda para a categorização de imagens, como apresentado em Srivastava e Salakhutdinov (2012). Um estudo sobre o uso de modelos deste tipo em diferentes tarefas de PLN para o português brasileiro é apresentado em Hartmann et al. (2017).

Existem vários tipos de implementações das arquiteturas de representação distribuída de palavras, mas aqui trataremos apenas das que são implementadas na ferramenta word2vec¹ (MIKOLOV et al., 2013a) que implementa as arquiteturas skip-gram e CBOW (*continuous bag-of-words*) para computar a representação distribuída das palavras.

CBOW é uma arquitetura de rede neural que prediz uma palavra à partir das outras palavras-contexto que estão em torno dela. O número de palavras a serem consideradas no

¹ <https://code.google.com/archive/p/word2vec/>

contexto, denominadas palavras-contexto, vai depender do tamanho da janela configurado inicialmente. Resumidamente, a arquitetura CBOW consiste de uma rede que, dado um número de palavras-contexto, tenta prever qual outra palavra poderia aparecer junto deste conjunto de contexto. Por exemplo, a palavra ‘travesseiro’ a partir das palavras-contexto ‘o gato sentou no’. A figura 1 mostra um exemplo de arquitetura dessa rede.

Figura 1 – Modelo de uma rede CBOW

Fonte: Mikolov, Yih e Zweig (2013)

Diferentemente da arquitetura CBOW, a arquitetura skip-gram prediz o contexto de uma palavra baseado na palavra que é passada (a palavra-alvo). Assim, dado uma palavra a rede prevê quais outras palavras podem aparecer junto desta. A figura 2 mostra um exemplo de rede dessa arquitetura.

Figura 2 – Modelo de uma rede skip-gram

Fonte: Mikolov, Yih e Zweig (2013)

No word2vec faz-se uso de uma rede neural que utiliza o princípio da máxima verossimilhança para maximizar a probabilidade da próxima palavra w_t (palavra-alvo)

dado algumas palavras anteriores h (histórico) em uma função *softmax*, onde o $score(w_t, h)$ computa a compatibilidade da palavra w_t com o contexto h , como mostra a equação 1 (MIKOLOV et al., 2013b). Este modelo é treinado maximizando-se o log da verosimilhança no conjunto de treino, ou seja, maximizando-se a equação 2. Esse modelo contudo é muito custoso, já que é necessário computar e normalizar cada uma das probabilidades usando a pontuação (*score*) para cada uma das palavras w' no contexto h a cada passo do treinamento (MIKOLOV et al., 2013b).

$$P(w_t|h) = softmax(score(w_t, h)) = \frac{exp(score(w_t, h))}{\sum_{palavra-w'-no-vocab} exp(score(w', h))} \quad (1)$$

$$logP(w_t|h) = score(w_t, h) - log(\sum_{palavra-w'-no-vocab} exp(score(w', h))) \quad (2)$$

Todavia, para adquirir os vetores de palavras não é necessário todo esse processo. Para que isso seja possível, os modelos CBOW e skip-gram utilizam uma classificação binária considerando as palavras-alvo verdadeiras (w_t) de k palavras-ruído imaginárias, inseridas no mesmo contexto. O objetivo então é maximizado quando o modelo descreve probabilidades altas para as palavras-alvo verdadeiras e probabilidades baixas para as palavras-ruído. Essa técnica é chamada de *negative sampling* e, no limite, as atualizações que essa função fornece se aproximam da função de uma *softmax*. Contudo, diferentemente da *softmax*, essa função escala com o número de palavras-ruído k e não com o tamanho do vocabulário, tornando o treinamento muito mais rápido (MIKOLOV et al., 2013b).

Uma das propriedades da representação distribuída de palavras é a de operações entre os vetores. O exemplo mais popular é a equação $vetor(Rei) + vetor(Homem) - vetor(Mulher) \approx vetor(Rainha)$, ilustrado na figura 3. Para observar esse tipo de associação no espaço vetorial é necessário que os modelos sejam treinados em um conjunto de dados suficientemente grande, já que a qualidade dos vetores aumenta conforme o número de dados disponíveis (MIKOLOV et al., 2013a).

Como visto em Mikolov et al. (2013a), a combinação de hiper-parâmetros de um modelo de representação distribuída de palavras depende bastante da tarefa, já que para cada parâmetro existe uma combinação ótima de arquitetura, tamanho de vetores, taxa de sub amostragem e tamanho da janela ideais. Algumas destas combinações serão exploradas nos modelos de reconhecimento de personalidade propostos no presente trabalho.

Figura 3 – Exemplo de operação vetorial possível em um word2vec

Fonte: Mikolov, Yih e Zweig (2013)

2.6 Representação distribuída de sentenças e documentos

A representação distribuída de documentos é uma extensão do que acontece na representação distribuída de palavras, para a criação de uma representação numérica de um documento, usando blocos maiores de palavras como sentenças, parágrafos ou até mesmo documentos inteiros. Para isso, em Le e Mikolov (2014) utiliza-se o próprio conceito da representação distribuída de palavras, como apresentado na seção 2.5, mas adicionando-se outro vetor que, como podemos ver na figura 4, é representado pelo quadrado escuro (que contém o **D** - *Paragraph ID*). Nesse modelo, cada parágrafo é mapeado para um único vetor, representado por uma coluna na matriz D (documento) e cada palavra também é mapeada em um vetor único, representado por uma coluna na matriz W. A seguir, é feita a média ou a concatenação do vetor de parágrafo e os vetores de palavras para prever a próxima palavra dentro de um contexto.

Figura 4 – Modelo de aprendizado da representação distribuída de um documento

Fonte: Le e Mikolov (2014)

Nota-se que a estrutura apresentada na figura 4 é muito semelhante à apresentada na figura 1, à exceção do *Paragraph ID* que atua como se fosse uma outra palavra. O parágrafo atua como uma memória que registra o que está faltando no contexto atual,

representando o conceito de um documento. Essa estrutura é chamada *Distributed Memory version of Paragraph Vector (PV-DM)*.

Os vetores de parágrafos possuem vantagens em relações a modelos *bag of words* (BoW), pois preservam a semântica das palavras. Por exemplo, o modelo BoW não distingue, ‘forte’, ‘poderoso’ e ‘Paris’ pois estas palavras podem estar muito próximos em um espaço vetorial, o que não acontece nos vetores de parágrafos onde ‘Paris’ estará bem mais distante de ‘forte’ e este mais próximo de ‘poderoso’. Além disso, os vetores de parágrafos levam em consideração a ordem das palavras, o que de outra forma só seria possível com um modelo de n-gramas de ordem elevada, o que criaria uma representação com uma alta dimensão e que possui pouco poder de generalização (LE; MIKOLOV, 2014).

2.7 Redes LSTM (Long-Short Term Memory)

As Redes LSTM (*Long-Short Term Memory*) (HOCHREITER; SCHMIDHUBER, 1997) são um tipo de rede neural recorrente que mantém uma “memória” de longo prazo, permitindo a ela utilizar informações recebidas em outro momento para auxiliar na tomada de decisão com a informação atual. LSTMs têm sido utilizadas de diferentes maneiras, como para reconhecimento de fala (DAHL et al., 2012; GRAVES; MOHAMED; HINTON, 2013); reconhecimento de imagens (KRIZHEVSKY; SUTSKEVER; HINTON, 2012; DONAHUE et al., 2015); tradução (WU et al., 2016); dentre muitas outras tarefas computacionais.

Utilizando uma Rede Neural Recorrente (RNR) já é possível conectar informações anteriores com a informação atual, contudo uma RNR utiliza apenas informação recente para, por exemplo, realizar uma classificação. Quando é necessário utilizar dados que foram passados em um espaço muito longo de tempo à rede, uma rede LSTM geralmente possui um melhor desempenho (HOCHREITER; SCHMIDHUBER, 1997).

Uma RNR se diferencia de uma rede *feedforward* pelo processo de carregar informações pelo tempo através das suas camadas escondidas. Matematicamente, essa característica é expressa por: $h_t = \phi(Wx_t + Uh_{t-1})$, onde h_t é a função da camada escondida no tempo t , x_t é a entrada neste mesmo tempo t , W é a matriz de pesos que modifica a entrada, como em uma rede *feedforward*, e h_{t-1} é o resultado da camada escondida do estado anterior, multiplicado pela matriz U , a matriz de transição, similar a uma cadeia

de Markov (NORRIS, 1998). A soma da entrada com a camada escondida é normalizada por uma função ϕ , uma sigmoide ou $\tanh$ (CHUNG et al., 2015).

Como esse ciclo de *feedback* ocorre em cada passo da série, cada camada escondida contém traços não apenas da camada escondida anterior, mas também de todas aquelas que precederam h_{t-1} enquanto a memória puder persistir. Por esse motivo, a informação rapidamente se transforma e desaparece, tornando difícil para essa configuração de RNR obter uma “memória” de longo prazo. Esse problema é conhecido como Desaparecimento (e Explosão) do Gradiente (*Vanishing (and Exploding) Gradients*) (HOCHREITER; SCHMIDHUBER, 1997).

Ao contrário das RNRs, redes LSTMs ajudam a preservar o erro que é propagado através do tempo e das camadas da rede. Através de uma arquitetura que possui uma célula de memória dotada de três portas analógicas sintonizáveis, essa arquitetura permite decisões sobre o que armazenar, e quando permitir leitura, escrita e deleções. As portas analógicas são implementadas com operações elemento a elemento por sigmóides, que variam de 0 a 1.

Figura 5 – Bloco de memória com apenas uma célula de rede LSTM com as portas analógicas de esquecimento, entrada e saída

Fonte: Gers, Schmidhuber e Cummins (1999)

O elemento principal de uma rede LSTM é a célula de memória, que pode ser vista na figura 5. As equações 3 à 11 foram extraídas de Gers, Schmidhuber e Cummins (1999) e auxiliam na compreensão do funcionamento da rede. Na porta de entrada (*input gate* na figura) os dados de entrada são avaliados através de uma função sigmoide dada pela equação 3, que decide se o novo dado será mantido na célula (caso a resposta seja 1) ou

removido (caso a resposta seja 0). A função de ativação da porta de entrada é dada pela equação 4.

$$f(x) = \frac{1}{1 + e^{-x}} \quad (3)$$

$$net_{in}(t) = w_{in}y(t-1); y_{in}(t) = f_{in}(net_{in}(t)) \quad (4)$$

Na porta de esquecimento, serão combinados os valores de entrada net_c , representados pela equação 5, com os valores que serão esquecidos, caso haja algum. A decisão de esquecer os valores é feita por meio da equação 6.

$$net_c = w_c y(t-1) \quad (5)$$

$$net_{\varphi}(t) = w_{\varphi} y(t-1); y_{\varphi}(t) = f_{\varphi}(net_{\varphi}(t)) \quad (6)$$

Para combinar o resultado das duas portas, soma-se o resultado do portão de entrada com o resultado do portão do esquecimento como vemos na equação 7, e os dados de entrada são normalizados pela equação 8.

$$s_c(t) = y_{\varphi}(t)s_c(t-1) + y_{in}(t)g(net_c(t)) \quad (7)$$

$$g(x) = \frac{4}{1 + e^{-x}} - 2 \quad (8)$$

Finalmente, a informação de entrada chega até a porta de saída, onde escolhe-se quais informações serão usadas na saída da célula. Essa decisão também é fornecida por uma função sigmoide, referenciada na equação 9.

$$net_{out}(t) = w_{out}y(t-1); y_{out}(t) = f_{out}(net_{out}(t)) \quad (9)$$

$$y(t) = y_{out}(t)h(s_c(t)) \quad (10)$$

$$h(x) = \frac{2}{1 + e^{-x}} - 1 \quad (11)$$

A saída é baseada na informação que sai da porta de saída, mas ela é filtrada para só expor a informação relevante para o tempo t , como mostra a equação 10, que é normalizada pela equação 11.

3 Trabalhos relacionados

Neste capítulo são apresentados alguns trabalhos importantes para a área de reconhecimento de personalidade, divididos pelo tipo de texto utilizado. Essa divisão foi escolhida pois cada tipo de texto possui métodos específicos para coleta de dados e momentos diferentes em que foram escritos, além de outras particularidades discutidas a seguir.

Os trabalhos que utilizam textos de ensaios são apresentados na seção 3.1. Os trabalhos que utilizam postagens de blogs encontram-se na seção 3.2. Os trabalhos que utilizam textos de redes sociais são apresentados na seção 3.3. Ao final, na seção 3.4 têm-se algumas iniciativas específicas para o português brasileiro.

3.1 *Abordagens que utilizam textos de ensaios*

Nesta seção serão tratados trabalhos em que os autores dos textos analisados puderam escrever ensaios sobre um tema livre em um ambiente que se destinava apenas ao experimento.

3.1.1 Abordagem lexical usando gramáticas funcionais sistêmicas

O trabalho de Argamon et al. (2005) trata do problema de reconhecimento das dimensões extroversão e instabilidade em ensaios, utilizando uma abordagem lexical baseada na gramática *Systemic Functional Grammar* (SFG). Gramáticas como o SFG podem ser vistas como sistemas de escolhas de significados em que todas as escolhas léxicas e estruturais são representadas em termos de funções semânticas (CHAPELLE, 2006).

Os ensaios formavam dois corpúsculos de redações de estudantes, compostos por 1.157 ensaios livres e 1.106 ensaios de autoanálise profunda. Os autores dos ensaios responderam ao inventário NEO-FFI (COSTA; MACCRAE, 1992). As dimensões de personalidade foram divididas em três partes iguais e apenas o primeiro terço e o último terço de cada dimensão de personalidade foram considerados, classificando as pessoas em positivas ou negativas para cada uma das dimensões.

Com o auxílio do SFG foram extraídos dos textos expressões conjuntivas, indicadores modais e adjetivos valorativos. Além disso, palavras funcionais também foram consideradas como atributos.

Para criação do modelo, utilizou-se o algoritmo SVM (*Support Vector Machine*) (CORTES; VAPNIK, 1995) com uma validação cruzada de dez conjuntos. Utilizou-se o peso atribuído às variáveis do algoritmo SVM para entender qual atributo era mais valioso para cada uma das dimensões, de forma similar ao que posteriormente seria feito em Gill, Nowson e Oberlander (2009).

Na análise de atributos, constatou-se que os adjetivos de valoração foram os que obtiveram melhor resultado para a dimensão instabilidade, mostrando que a diferença entre indivíduos neuróticos e não-neuróticos é, principalmente, a forma como eles interagem com as pessoas e o ambiente. A acurácia do classificador para essa dimensão foi de 58%. Já para a dimensão extroversão, os melhores resultados foram obtidos apenas com as palavras funcionais e a acurácia obtida foi de 57%.

Nota-se que foi usado apenas um texto de cada um dos autores, o que segundo Argamon et al. (2005) torna a tarefa mais difícil já que os métodos mais bem-sucedidos para prever a personalidade fazem uso de perguntas diretas ou o uso de vários textos de um mesmo autor ao longo do tempo.

3.1.2 Uso de algoritmos de classificação, regressão e *ranking*

Em Mairesse (2008) propõe-se vários métodos de reconhecimento de personalidade utilizando algoritmos de classificação, regressão e *ranking*. Para isso, além de textos compostos por ensaios de estudantes, utilizou-se também a transcrição de áudio de conversas para o reconhecimento da personalidade. Além da tarefa de reconhecimento, foi apresentado um modelo de geração de linguagem natural para projetar as dimensões de personalidade, que não será discutido nesta seção que tratará apenas da etapa do reconhecimento de personalidade utilizando o *corpus* de ensaios.

O *corpus* de ensaios utilizado contém 2.479 textos, com 1,9 milhões de palavras, produzidos por meio de uma tarefa na qual os participantes podiam escrever qualquer coisa que viesse à mente, durante vinte minutos. Além de produzir os ensaios, os participantes

responderam ao inventário BFI (subseção 2.2) para que as dimensões de personalidade fossem obtidas.

Como atributos para o treinamento dos modelos usou-se o dicionário LIWC (PENNEBAKER; FRANCIS; BOOTH, 2001) e a base MRC (COLTHEART, 1981), que contém estatísticas de aproximadamente 150.000 palavras, como idade de aquisição, frequência de uso e familiaridade. Do dicionário LIWC usou-se a frequência com que cada uma das palavras presentes nas 88 categorias selecionadas apareceram nos textos. Outros 14 atributos foram extraídos da base MRC, sendo estes a média do valor do atributo obtido no conjunto de ensaios e não a frequência, como é usado no dicionário LIWC.

Observou-se a correlação de cada um dos atributos do dicionário LIWC e MRC com os dados para verificar a importância das variáveis individualmente, independentemente do modelo usado. É interessante observar que em Mairesse (2008) é examinada a correlação dos atributos com a língua natural, tanto na modalidade escrita como falada, e que no entanto resultados distintos foram obtidos em cada uma. Por exemplo, para o fator extroversão o atributo ‘número de palavras’ obtém uma correlação alta para corpus de áudio, mas isso não acontece com o corpus de ensaios, mostrando que os atributos mais importantes para o reconhecimento da personalidade de um indivíduo podem variar a conforme o tipo de comunicação.

Em Mairesse (2008) também compara-se algoritmos de classificação com algoritmos de regressão e de *ranking*. Para os algoritmos de classificação, optou-se por dividir o conjunto de dados em duas porções de tamanhos iguais, em que a primeira é composta por indivíduos positivos para uma determinada dimensão e a segunda é composta de indivíduos negativos para esta dimensão. Os métodos avaliados foram árvore de decisão (ROKACH; MAIMON, 2008), k-vizinhos mais próximos (k-NN) (ALTMAN, 1992), *Naive Bayes* (NB) (RUSSELL; NORVIG; INTELLIGENCE, 1995), *Ripper rule learning* (COHEN, 1995), *Adaboost* (ADA) (FREUND; SCHAPIRE, 1995) e *Support Vector Machine* (SVM). Para todos, usou-se uma validação cruzada de dez conjuntos.

A dimensão de personalidade classificada com melhores resultados foi agradabilidade, que obteve uma acurácia de 62,1% utilizando SVM. A dimensão instabilidade possuiu a segunda melhor acurácia, com 57,4% também utilizando SVM. Para extroversão, meticulosidade e agradabilidade o modelo com SVM obteve uma acurácia de 55%, ligeiramente acima do *baseline* de 50%. Os modelos com SVM obtiveram o melhor desempenho

para todos os fatores, exceto extroversão, que obteve o melhor resultado com o algoritmo ADA.

Como as dimensões de personalidade são valores escalares, é possível tratar o problema de reconhecimento de personalidade com o uso de métodos de regressão (MAIRESSE, 2008). Para essa tarefa avaliou-se os seguintes métodos: regressão linear, árvore de regressão M5, árvore de regressão M5 retornando um modelo linear e uma árvore de decisão REPTree baseada em SVMs com núcleo linear. Os resultados foram comparados com a média de cada dimensão no conjunto de treino.

O melhor modelo foi a árvore de regressão M5, que obteve um erro relativo de 93,3% para liberalidade e 96,4% para neuroticismo. Usando a regressão linear também fez-se uma comparação dos atributos do dicionário LIWC com os atributos da base MRC. Os resultados mostraram que, em geral, o uso de atributos LIWC tem um desempenho melhor que o uso de atributos MRC quando usados separadamente, exceto para extroversão, em que os atributos MRC são superiores até mesmo aos atributos dos dois dicionários combinados.

Como alternativa aos modelos de classificação e regressão, em Mairesse (2008) é discutido ainda o uso de algoritmos de *ranking* para ordenar indivíduos de acordo com uma dimensão de personalidade. Para esse fim, escolheu-se o algoritmo *RankBoost* (FREUND et al., 2003). O *baseline* utilizado modela o *rank* dos textos aleatoriamente, possuindo uma perda de *ranking* de (0,5) na média (quanto menor a perda, melhor o resultado). O melhor modelo de *ranking* foi obtido para a dimensão liberalidade, com uma perda de 0,39. A dimensão instabilidade possui o segundo melhor resultado, com 0,42 de perda, seguido de meticulosidade e extroversão. Apesar de todas as dimensões estarem acima do *baseline*, a melhora encontrada é bastante baixa. Ainda usando *ranking*, fez-se um estudo de atributos, comparando-se um modelo que utilizou apenas os atributos LIWC com um que utilizou apenas os atributos MRC. O modelo com os atributos MRC não mostrou nenhum aumento na perda de *ranking*, o que sugere que, para esse modelo, apenas os atributos LIWC foram relevantes (MAIRESSE, 2008).

Em Mairesse (2008) também propõe-se que os métodos de regressão e *ranking* são estratégias superiores para modelar a personalidade quando comparadas com métodos de classificação, argumentando ainda que a divisão em multiclases, como a apresentada em Oberlander e Nowson (2006) (em que há a divisão em três ou cinco classes para uma mesma dimensão) não considera que as classes façam parte de uma ordenação geral, e que

esses modelos são obrigados a ignorar a importância de atributos que se correlacionam com a ordenação de todas as classes. Além disso, há ainda uma comparação do *ranking* com a classificação binária, comparando-se o modelo criado com *RankBoost* e o algoritmo NB. O método de *ranking* obtém um resultado melhor para quatro das cinco dimensões de personalidade, não sendo estatisticamente superior ($p < 0,05$) apenas para neuroticismo.

Por fim, a dimensão liberalidade provou-se a dimensão mais fácil de modelar, mostrando desempenho superior em todas as tarefas de reconhecimento. Isso pode ser um indício de que essa dimensão seja mais frequente em textos do tipo ensaios (MAIRESSE, 2008).

O trabalho de Mairesse (2008) é bastante importante para a área de reconhecimento de personalidade e referência para muitos outros estudos. A estratégia de comparação de algoritmos de regressão, classificação e *ranking*, assim como a comparação de atributos LIWC e MRC e a diferente percepção entre a autoavaliação e a avaliação por terceiros (diferença entre o método de obtenção de personalidade do corpus de ensaios e do corpus de áudio) são também potencialmente úteis para o presente trabalho.

3.1.3 Detecção de personalidade utilizando redes neurais profundas

No trabalho de Majumder et al. (2017) desenvolve-se uma técnica para reconhecimento de personalidade baseado em atributos extraídos de uma representação distribuída de palavras. Nesta abordagem, estes atributos são combinados para se tornarem vetores de documentos com uma rede neural profunda do tipo CNN (*Convolutional Neural Networks*) (LECUN; BENGIO et al., 1995). Os vetores de documentos são então combinados com os atributos apresentados em Mairesse (2008) para a classificação binária de cada um dos cinco traços de personalidade do CGF, utilizando uma rede neural totalmente conectada com uma camada escondida. Com essa estratégia obteve-se resultados melhores que os apresentados no trabalho de Mairesse (2008).

O trabalho utilizou o mesmo corpus de ensaios apresentado em Mairesse (2008), e fez uso de dois tipos de atributos, o primeiro deles consistindo de uma representação distribuída de palavras, e o segundo sendo o conjunto de atributos apresentados originalmente no trabalho de Mairesse (2008), ou seja, atributos provenientes do dicionário LIWC e da base MRC.

Após a extração dos atributos dos dicionários LIWC e MRC, fez-se a extração da representação distribuída de palavras. Contudo, foram consideradas apenas as sentenças que carregavam alguma palavra/emoção que estivesse contida no léxico *NRC Emotion Lexicon* (MOHAMMAD; TURNEY, 2010). Este léxico contém 6.468 palavras, e portanto caso a sentença não contivesse alguma palavra do léxico, ela era removida. Esse processo reduziu o número de palavras para a representação distribuída de palavras, contudo os autores não mencionam o tamanho final obtido. Além desta técnica de remoção de palavras, também testou-se remover aleatoriamente metade das sentenças do corpus e não remover nenhuma sentença. A eliminação aleatória trouxe melhores resultados do que manter todas as sentenças, e a eliminação por sentenças que contivessem palavras com alguma emoção do léxico trouxe os melhores resultados. Não há explicação para o melhor desempenho da eliminação de sentenças aleatórias em detrimento de utilizar todas as sentenças (MAJUMDER et al., 2017).

Após a eliminação de sentenças, utilizou-se a ferramenta *word2vec* para transformar as palavras em vetores de palavras de tamanho 300. Os vetores compõem a entrada para uma rede do tipo CNN, que transformou os vetores de palavras em vetores de sentenças e depois de documentos, resultando em uma matriz de tamanho 600. Por fim, essa matriz de documentos foi combinada com os atributos dos dicionários LIWC e MRC, gerando um vetor final de 684 atributos para cada documento.

Para a tarefa de classificação utilizou-se uma rede MLP (*multi layer perceptron*) e uma SVM de núcleo polinomial, onde os dados de entrada para os dois algoritmos correspondem às saídas da rede CNN. Foram feitos diversos experimentos, com várias configurações de parâmetros e entradas para a rede, como com ou sem filtro de sentenças, diferentes classificadores, diferentes filtros de convolução e atributos diferentes para os classificadores (por exemplo, utilizando apenas a saída da rede CNN ou utilizando a saída da rede combinada com os atributos LIWC e MRC). Os resultados mostram que o filtro de sentenças utilizando o léxico *NRC Emotion Lexicon* e a estratégia de adicionar os atributos dos dicionários LIWC e MRC foram essenciais para que os melhores resultados ultrapassassem o apresentado em Mairesse (2008), já que os atributos da rede CNN sozinhos não foram o suficiente para melhorar os resultados originais. Os autores atribuem esse fato ao pequeno volume de dados de treinamento, já que o conjunto de dados original possui apenas 1,9 milhão de palavras, sem considerar as palavras eliminadas pelo filtro de sentenças.

É interessante frisar que os melhores resultados para cada dimensão de personalidade estão apresentados em diferentes estratégias, ilustrando o quão difícil é obter um modelo único que seja o melhor para todas as dimensões de personalidade. A melhor acurácia para extroversão foi de 58,1, utilizando uma rede MLP, com filtro de sentenças e a saída da rede CNN, contra 55,8 do trabalho original apresentado em Mairesse (2008). O resultado apenas com CNN sem os filtros de frases e atributos de léxicos apresentou acurácia de 55,7. Já para liberalidade, agradabilidade e meticulosidade, apesar da configuração do modelo que possui a melhor acurácia ser a mesma da apresentada para extroversão, os filtros de convolução são diferentes.

O trabalho de Majumder et al. (2017) mostra o uso de estratégias mais recentes para reconhecimento de personalidade, como o uso de representações vetoriais de palavras e redes neurais profundas para essa tarefa. Tais estratégias são utilizadas no presente trabalho onde utiliza-se redes profundas do tipo LSTMs, que é inclusive uma das linhas de investigação sugeridas como trabalho futuro em Majumder et al. (2017).

3.2 *Abordagens que utilizam textos de blogs*

Esta seção trata de trabalhos em que os participantes das pesquisas nem sempre responderam um inventário de personalidade listado dentre os inventários válidos para o CGF. Os textos analisados foram escritos pelos participantes antes de se saber que eles participariam das pesquisas e normalmente apenas textos de um curto período de tempo foram coletados para as análises (e.g., apenas textos de um mês específico do *blog*).

3.2.1 *Abordagem baseada em modelos n-gramas*

O trabalho de Oberlander e Nowson (2006) objetivou classificar a personalidade de autores de *blogs* em quatro dimensões de personalidade utilizando modelos de n-gramas aliado aos algoritmos de classificação NB e SVM. O trabalho contou com 71 participantes, totalizando 410.000 palavras relativas aos textos pessoais do *cópus*. O pré-processamento consistiu de algumas substituições nos textos de modo que os substantivos foram substituídos pelo rótulo *NP1*, toda pontuação foi substituída pelo rótulo *p* e alguns rótulos foram adicionados para marcar o início e o fim de um texto. Além desse pré-

processamento, o tamanho dos textos foi truncado para evitar a influência de indivíduos muito prolixos na classificação. Os textos eram limitados ao tamanho da média do número de palavras mais dois desvios padrões. Os participantes responderam a um questionário de 41 itens com uma escala *Likert* de 5 pontos derivado do NEO-PI-R (COSTA; MACCRAE, 1992), que guarda semelhança com o inventário BFI.

Avaliou-se apenas quatro dimensões, pois optou-se por desconsiderar a dimensão liberalidade que, ao contrário das outras dimensões do CGF, não apresentou uma distribuição normal. Apesar da escolha do melhor algoritmo para classificar personalidade não ser o foco do trabalho de Oberlander e Nowson (2006), foi realizada uma comparação entre SVM e NB usando uma validação cruzada de dez conjuntos.

Um dos destaques da abordagem de Oberlander e Nowson (2006) é o método utilizado para agrupar as pontuações em cada uma das dimensões. Foram definidos dois tipos de tarefas de classificação. Nas tarefas de classificação binárias, indivíduos são divididos em extremos de alta e baixa pontuação (e, quando existem, os intermediários são desconsiderados). Nas tarefas de multi-classificação, todos os indivíduos são considerados. Estas decisões têm impacto direto sobre a acurácia da classificação, já que a tarefa pode se tornar mais difícil conforme aumenta-se o número de classes a serem consideradas.

Para a classificação binária foram propostos os três primeiros critérios a seguir e, para a multi-classificação, os últimos quatro. Estes critérios são também ilustrados nas figuras 6 e 7.

1. Os indivíduos com valores positivos para uma dimensão são aqueles que estão um desvio padrão acima da média e os que possuem valores negativos estão um desvio padrão abaixo da média. Todos os outros são deixados de fora, formando-se assim dois grupos relevantes, como pode ser visto na linha (1) da figura 6.
2. Para incluir mais pessoas, a classe positiva possui indivíduos com meio desvio padrão acima da média e classe negativa meio desvio padrão abaixo da média, formando dois grupos maiores que o item acima, como pode ser visto na linha (2) da figura 6.
3. Por fim, os autores dividiram o corpus ao meio, formando dois grupos. Entretanto, em Oberlander e Nowson (2006) aponta-se um problema em relação a esse método, pois algumas instâncias positivas estão muito mais próximas da classe negativa do que de outras dentro da mesma classe, como pode ser visto na linha (3) da figura 6.

4. Por esse critério são incluídos os indivíduos contemplados no critério (1), ou seja, extremos da dimensão, mais os indivíduos excluídos do critério (2). Assim exclui-se alguns indivíduos entre o extremo da dimensão e o ponto médio desta, formando três grupos, como pode ser visto na linha (4) da figura 7.
5. Por esse critério são incluídos os indivíduos do critério (1), mas dessa vez todos os indivíduos eliminados daquele critério são incluídos neste como grupo intermediário, formando três grupos, como pode ser visto na linha (5) da figura 7.
6. Esse critério considera o mesmo grupo positivo e negativo do critério (2), contudo os que foram antes excluídos agora compõem o grupo intermediário, formando três grupos, como pode ser visto na linha (6) da figura 7.
7. O critério mais complexo é composto de um conjunto de cinco grupos onde os extremos do critério (1) são separados dos grupos positivos e negativos do critério (2), criando um grupo de extremos positivos e extremos negativos, positivos e negativos (não extremos), e o grupo intermediário, que é formado da mesma maneira que no critério (4), formando cinco grupos, como pode ser visto na linha (7, como pode ser visto na linha (4) da figura 7) da figura 7.

	Lowest		...	Highest	
1	L		–		H
2	LL		–	HH	
3	LLL			HHH	

Figura 6 – Divisão para classificação binária

Fonte: Oberlander e Nowson (2006)

	Lowest		...	Highest	
4	L	–	m	–	H
5	L		M		H
6	Ll		m	hH	
7	L	l	m	h	H

Figura 7 – Divisão para multi-classificação

Fonte: Oberlander e Nowson (2006)

As tarefas de classificação consideradas usam cinco grupos de modelos de n-gramas:

1. Todos os n-gramas mais frequentes no corpus.

2. Os n-gramas mais comuns para cada extremo de uma dimensão, como apresentado no critério (1), com frequência maior do que cinco.
3. Os n-gramas que aparecem no mínimo 50% das vezes no grupo (2) em cada um dos extremos.
4. Os n-gramas que estão no grupo (2) com uma significância de $p < 0,001$, que também estão no grupo (3) e têm uma correlação de $p < 0,05$ com a personalidade de um indivíduo.
5. Os n-gramas que estão no grupo (2), mas que foram selecionados automaticamente com o uso dos métodos *BestFitSearch* e *CfsSubEval evaluator* da ferramenta Weka (WITTEN; FRANK, 2005).

Nos resultados, observou-se que o algoritmo NB obteve um resultado superior ao SVM em 41 dos 60 testes. Os resultados para o algoritmo SVM não foram informados. O conjunto de modelo de n-gramas que obtém os melhores resultados é o conjunto selecionado automaticamente (grupo 5), lembrando que a comparação foi feita levando-se em consideração o *baseline* da classe majoritária. Em Oberlander e Nowson (2006) sugere-se que o melhor desempenho do grupo selecionado automaticamente pode ter sido causado por *overfitting* já que existem apenas 71 indivíduos no conjunto de dados.

As tarefas de classificação binária apresentam o comportamento esperado, ou seja, quanto mais difícil a tarefa, pior o desempenho do classificador. Por outro lado, a acurácia se manteve bastante similar entre as três tarefas. Na análise dos resultados, observou-se também como as dimensões de personalidade se comportam de modo distinto em cada um dos tipos de classificação: a acurácia da dimensão extroversão foi de 32,2% para tarefa com cinco classes, em contraste com a dimensão agradabilidade, que obteve uma acurácia de 77,2% para a mesma tarefa. Como os resultados foram apresentados apenas com base na acurácia, não é possível analisar de forma mais detalhada qual das estratégias obteve o melhor resultado.

Em Oberlander e Nowson (2006) apresentam-se diversas contribuições potencialmente relevantes para a presente pesquisa, como o método usado para dividir as classes das dimensões de personalidade, o uso de modelos que consideram a ordem das palavras e o modo como os autores analisam as tarefas e os conjuntos de atributos por dimensão de personalidade. Em especial, uma estratégia de divisão de classes semelhante à proposta neste trabalho foi utilizada no estudo exploratório a ser apresentado no capítulo 4.

3.2.2 Treino em um corpus pré-processado e teste em um corpus com pouco pré-processamento

O trabalho em Nowson (2007) propõe usar um modelo treinado em um corpus de pequenas proporções, com 410.000 palavras, no qual foi realizado um pré-processamento cuidadoso, para então prever quatro dimensões de personalidade de indivíduos de um corpus maior, de 4.8 milhões de palavras, obtido de forma semi-automática.

O corpus de pequenas proporções, denominado OC (*Old Corpus*) é o mesmo corpus apresentado por Oberlander e Nowson (2006). O corpus maior, denominado NC (*New Corpus*) é um corpus muito semelhante ao utilizado posteriormente em Gill, Nowson e Oberlander (2009), usando o mesmo pré-processamento e o mesmo tipo de inventário não padrão para o CGF, porém com um número menor de palavras (4.8 milhões).

Em Nowson (2007), assim como em Oberlander e Nowson (2006), não foi considerada a dimensão agradabilidade, já que tanto para o corpus NC quanto para o OC não existe uma distribuição normal para essa dimensão. Outra semelhança com Oberlander e Nowson (2006) está na utilização de contagens de n-gramas na definição das tarefas de classificação binária e ternária, utilizando as mesmas técnicas de divisões para cada dimensão que foram utilizadas em Oberlander e Nowson (2006) (subseção 3.2.1). Para essas tarefas de classificação, optou-se pelo algoritmo NB.

O grande desafio deste trabalho foi criar os modelos para as diversas estratégias com base nos dados obtidos em OC e testá-los em NC, já que o corpus OC corresponde a apenas 4% dos dados totais, e normalmente treina-se o modelo com aproximadamente 2/3 dos dados (NOWSON, 2007).

Inicialmente, comparou-se a habilidade do modelo em classificar textos do corpus NC com as respostas do inventário. A tarefa foi realizada da maneira clássica: treinou-se o modelo com os dados do corpus e testou-se com outro grupo dos mesmos dados. Os resultados foram comparáveis aos apresentados em Oberlander e Nowson (2006), já que a técnica utilizada é a mesma, embora os corpus sejam diferentes. A comparação mostrou que os resultados foram inferiores aos apresentados no trabalho anterior, mas ainda conseguiram ultrapassar o *baseline* de 50% de acerto na classificação binária, chegando a 66,4% de acurácia para meticulosidade, dimensão com melhor acurácia.

Considerando a tarefa mais desafiadora, ou seja, treinar usando dados de OC e testar com NC, o classificador ternário obteve mais sucesso, com um desempenho melhor

que o classificador binário quando comparado ao *baseline* aleatório. Segundo Nowson (2007), apesar do ruído contido no corpus NC e dos resultados abaixo do esperado, ainda é possível obter informações de personalidade de indivíduos utilizando-se um corpus pouco processado e, segundo os mesmos autores, o uso de um corpus com mais textos de um mesmo autor e novos algoritmos são estratégias que podem facilitar essa tarefa.

3.2.3 Uso de regressão logística para seleção de atributos LIWC

Em Gill, Nowson e Oberlander (2009) objetiva-se investigar as razões pelas quais indivíduos com diferentes personalidades escrevem textos de *blogs*. Para isso, fez-se uso do dicionário LIWC com o auxílio de regressão logística. Optou-se por não utilizar nenhum inventário padrão do CGF, mas sim um inventário de teor não especificado, de vinte e cinco perguntas (cinco para cada dimensão) que guardava certa similaridade com as perguntas e dimensões do modelo CGF.

O corpus apresentado possui 14,8 milhões de palavras (média de 2.941,4 e desvio padrão de 1.374,6) produzidas por 2.393 voluntários. Foram incluídos apenas usuários que publicaram mais de mil palavras por mês, e indivíduos com mais de 5.000 palavras tiveram as suas publicações truncadas da mesma forma que foi feita em Oberlander e Nowson (2006). O único pré-processamento foi feito de forma automática, utilizando alguns rótulos HTML para remoção de atributos como *links*, para garantir que apenas postagens redigidas pelos voluntários da pesquisa fossem consideradas.

Como os resultados dos testes possuíam valores em uma escala de zero a cinco, fez-se necessário a divisão desses grupos para a tarefa de classificação. Os resultados dos testes foram divididos em três categorias: negativo com a pontuação de 0 a 1, intermediário com pontuação de 2 a 3 e positivo com pontuação de 4 a 5. Utilizou-se o dicionário LIWC (subseção 2.3) para obter as seguintes categorias: mecanismos cognitivos; processos sensoriais e perceptuais; tempo; ocupação; atividades de lazer e estados físicos e funções.

Como o objetivo do trabalho era apresentar o que motiva as diversas dimensões de personalidade a escreverem em seus respectivos *blogs*, utilizou-se uma regressão logística para mostrar a relação que existe entre cada um dos atributos que foram utilizados, palavras agrupadas de acordo com o dicionário LIWC e as dimensões de personalidade. Dessa relação, afirmou-se que postagens de blogueiros neuróticos tendem a ser mais

autoterapêuticas, que os extrovertidos tendem a falar mais de si mesmos e sobre emoções vividas por eles, que indivíduos liberais discorrem mais sobre a vida cotidiana e que indivíduos agradáveis escrevem publicações que se adaptam a vários contextos.

Conclui-se que é possível identificar a personalidade dos autores e descobrir temas de escrita de maior afinidade com cada dimensão de personalidade. Contudo, os autores esperam que, além dos atributos de conteúdo, os atributos de estilo também possam ser relevantes para reconhecer as dimensão de personalidade.

3.2.4 Detecção de atributos relevantes para reconhecimento de personalidade

Em Iacobelli et al. (2011) pretendeu-se descobrir qual o melhor conjunto de atributos para o reconhecimento de personalidade. Para esse fim, foram utilizados modelos de bigramas, *tf-idf* e *stopwords* aliado a atributos do dicionário LIWC para a classificação da personalidade de um autor de texto de *blog*.

O trabalho utilizou um conjunto de dados composto de três mil participantes e um teste de personalidade simplificado, em que os participantes deveriam responder as perguntas com ‘sim’ ou ‘não’. Dado que a proporção de personalidade era diferente entre os polos positivos e negativos de uma dimensão, algumas das respostas foram descartadas aleatoriamente para que houvesse uma proporção de 50% em todas as dimensões.

Todas as palavras presentes no *córpus* foram submetidas ao processo de *stemming*. A tarefa de reconhecimento de personalidade foi modelada com o uso de diversos tipos de conhecimento, como frequência de bigramas e palavras usadas por mais de um autor, a inclusão ou omissão de *stopwords*, importância de uma palavra dado o seu *tf-idf*, presença ou ausência de uma palavra e os atributos do dicionário LIWC. Estes diversos atributos foram combinados em novos experimentos, sendo que o dicionário LIWC não foi usado nas combinações e considerado apenas como um experimento isolado. Um exemplo de experimento considerado foi a combinação de presença de frequência das palavras individuais, bigramas e uso de *stopwords*. Utilizou-se o algoritmo SVM com os valores padrões da ferramenta Weka (WITTEN; FRANK, 2005). Uma vez que o intuito do trabalho era a comparação de atributos, apenas esse algoritmo foi utilizado.

Como o dicionário LIWC foi tratado isoladamente, é mais fácil perceber o impacto que este tipo de conhecimento teve sobre a acurácia da classificação. O uso do dicionário

LIWC sozinho não obteve acurácia maior do que nenhuma das combinações dos outros atributos. A combinação que obteve a melhor acurácia foi a que fazia uso de bigramas, *stopwords* e a presença/ausência das palavras. Em Iacobelli et al. (2011) usou-se a redução e o aumento de acurácia em cada experiência (combinações dos conjuntos de atributos) para explicar como um atributo pode ser capaz de classificar uma dimensão de personalidade quando considerado de forma isolada. Por exemplo, remover as *stopwords* para agradabilidade reduz a acurácia do modelo, o que mostra a importância dessas palavras nessa dimensão.

É apresentado também um estudo dos bigramas utilizados para entender quais deles são relevantes para cada dimensão de personalidade. Para isso, utilizou-se o conjunto de bigramas que melhor classificou cada dimensão e, a partir dos documentos que os contêm, foi possível saber se o autor encontrava-se no extremo positivo ou negativo de uma determinada dimensão de personalidade. O bigrama é considerado preditor da classe positiva caso a maioria dos autores dos documentos tenha a classe positiva para uma dimensão, ou negativo em caso contrário. Assim, foram avaliadas quais palavras uma determinada dimensão de personalidade mais utiliza, e se essas palavras têm relação com a dimensão.

O trabalho de Iacobelli et al. (2011) mostra a importância da estrutura do texto para classificar a personalidade, ilustrado pelo desempenho superior do uso de bigramas em relação a palavras isoladas e se comparado com o uso dos atributos do dicionário LIWC isoladamente. Além disso, a importância da estrutura do texto também pode ser percebida no aumento da acurácia obtida pelo uso de *stopwords*, já que a remoção destas reduziu a acurácia do modelo. Além do mais, o estudo de cada bigrama mais utilizado pelas diferentes dimensões de personalidade mostra mais uma vez como bigramas são atributos fortes para o reconhecimento de personalidade de um indivíduo. As descobertas apresentadas em Iacobelli et al. (2011) reforçam o uso de modelos que consideram atributos estruturais do texto, já abordado por Oberlander e Nowson (2006) e servem de motivação para possível utilização destes no presente trabalho.

3.3 Abordagens que utilizam textos de redes sociais

Esta seção trata de trabalhos que possuem uma maior semelhança com o presente estudo, discutindo aqueles que fazem uso de textos provenientes de redes sociais. Nesse ambiente, o autor dos textos nem sempre está preocupado com a correção do que é escrito, faz uso de neologismos, *hashtags* e nem sempre escreve a intervalos regulares, podendo assim produzir textos de teor muito diferente de uma publicação para outra. Por outro lado, a quantidade de textos produzidos por um mesmo indivíduo ao longo do tempo pode ser bem maior do que os apresentados nas seções 3.1 e 3.2, o que é potencialmente útil para a tarefa de reconhecimento automático de personalidade.

3.3.1 Reconhecimento de personalidade independente de língua

Em Celli (2012) apresentou-se uma abordagem para resolver dois grandes problemas para a descoberta de personalidade por meio de textos: a descoberta de personalidade independente do idioma e a redução do número de dados rotulados para treinamento. Para isso, foram utilizados textos em italiano e em inglês com aprendizado semi-supervisionado e atributos que não dependiam da língua.

O modelo de reconhecimento de personalidade proposto por Celli (2012) faz uso de dois conjuntos de dados: um conjunto de ensaios utilizados em Mairesse (2008) para o inglês e um conjunto de textos do *Facebook* em italiano.

No trabalho de Celli (2012) têm-se quatro conjuntos de atributos. Contudo, ao invés de usar a frequência do atributo no corpus, utilizou-se a correlação entre o atributo e as dimensões de personalidade. Os tipos de atributos utilizados foram:

- **Psicológico (M):** atributos MRC (COLTHEART, 1981) presentes no trabalho de Mairesse (2008). Estes incluem contagem de caracteres, contagem de sílabas, familiaridade, etc.
- **Linguístico (I):** palavras em inglês e n-gramas associados a dimensões positivas e negativas, como apresentado em Iacobelli et al. (2011).
- **Multi-linguístico (C):** atributos independentes de língua extraídos dos dicionários LIWC e MRC, como pontos de exclamação, números, parênteses, taxa de repetição, etc.

- **Psicolinguístico (L)**: atributos LIWC presentes no trabalho de Mairesse (2008), como palavras associadas a afeto, tristeza, sentir, ouvir, positivas, negativas, etc.

Usou-se revocação, precisão e a medida F para comparar o desempenho dos conjuntos de atributos separadamente em um classificador binário do tipo SVM. Os resultados sugeriram que o conjunto (L) possuiu a melhor revocação e medida F com 0,96 e 0,68, respectivamente. O conjunto (C) obteve a maior precisão, com 0,54.

Ainda utilizando as mesmas métricas de desempenho, ao invés de comparar os atributos separadamente, fez-se uso de todas as possíveis combinações dos conjuntos de atributos. A melhor medida F é obtida com os conjuntos (I), (L) e (M) combinados. Em nenhuma das combinações o conjunto (M) aumenta a precisão, exceto quando utilizado com o conjunto (L).

Além da classificação, fez-se um experimento para prever a pontuação de cada uma das dimensões com regressão. O conjunto de atributos que possuiu o menor RMSE (*Root Mean Square Error*) é o (C).

Estes resultados foram obtidos levando-se em conta os dados do corpus de ensaios presentes no trabalho de Mairesse (2008), que está em inglês. Contudo, utilizando os dados do *Facebook* em italiano, obteve-se resultados diferentes para a classificação, em que o conjunto (C) possuiu melhor precisão (0,55) e revocação (0,76). Para a tarefa em italiano não existia o conjunto (M), já que o MRC não possui adaptação para o italiano, e nem o conjunto (I), já que não foram gerados os n-gramas para esse conjunto de dados. Restaram apenas o conjunto (L), que foi gerado com base na versão italiana do dicionário LIWC, e o conjunto (C), que apresentou uma diferença média de correlação muito baixa em relação ao inglês (de apenas 0,0062).

Além do modelo de classificação binária, foi utilizado também um método de classificação ternária, em que além da classe positiva e negativa existia uma classe intermediária. O melhor conjunto de atributos para a classificação ternária foi o (C), que apresentou a melhor medida F (0,54). Os resultados da classificação ternária, como esperado, não foram melhores que os resultados apresentados para classificação binária.

Além dos estudos de reconhecimento, o estudo em Celli (2012) apresenta também algumas outras análises de interesse para o presente trabalho. Uma delas tratou de como os seres humanos preveem a personalidade uns dos outros. Nesta análise enfatizou-se que reconhecer a personalidade é uma tarefa bastante difícil para uma pessoa comum, e

portanto há uma enorme vantagem em um método de obtenção da personalidade por meio de textos. Também realizou-se uma análise utilizando regras de associação para entender quais dimensões de personalidade apareceram juntas com maior frequência. Um exemplo é a regra de que meticulosidade e a dimensão reversa de instabilidade aparecem juntas. No entanto, a confiança é relativamente baixa (0,67), o que mostra a alta variabilidade dessa regra.

O trabalho aponta ainda várias dificuldades ao minerar dados de personalidade, especialmente em redes sociais. Dentre as dificuldades apontadas está a de obter os textos que nem sempre estão públicos, e a dificuldade de obter dados rotulados e filtrar textos que podem se apresentar em múltiplos idiomas.

No trabalho de Celli (2012) é realizado também um comparativo com trabalhos importantes da área, como os de Mairesse (2008) e Oberlander e Nowson (2006), em que há inclusive a replicação do experimento do primeiro. Além disso, o trabalho apresenta uma técnica inédita até então de classificação da personalidade com atributos independente da língua, que obteve resultados superiores ao uso de atributos extraídos do dicionário LIWC, além de comparar diferentes modelos e técnicas de pré-processamento de textos em diferentes domínios, como os ensaios e atualizações de *status* do *Facebook*.

3.3.2 Reconhecimento de personalidade em atualizações de *status* do *Facebook*

Em Farnadi et al. (2013) apresenta-se uma abordagem para reconhecimento de personalidade utilizando atualizações de *status* do *Facebook*. O objetivo do trabalho era classificar um *corpus* no domínio *Facebook* utilizando o *corpus* de ensaios presentes no trabalho de Mairesse (2008) para treinamento de um modelo computacional.

O *corpus* de Farnadi et al. (2013) possuía 9.917 atualizações de *status* de 250 indivíduos. O trabalho não é específico quanto ao inventário de personalidade utilizado. Um diferencial deste trabalho em relação aos outros vistos até aqui é o uso de atributos utilizados para o modelo. Além do dicionário LIWC (L), utilizou-se atributos de rede social (S) como tamanho da rede, densidade, transitividade, etc., atributos temporais (T) como frequência de atualizações por dia, número de atualizações, e outros (O) como número de palavras repetidas, número de atualizações por usuário, número de URLs, etc.

Com estes atributos, criou-se um modelo de classificação binária para cada dimensão de personalidade utilizando-se a ferramenta Weka (WITTEN; FRANK, 2005). Comparou-se o desempenho dos algoritmos SVM com núcleo linear, KNN com $k = 1$ e NB. Cada um dos modelos foi comparado com a classe majoritária medindo-se a revocação, precisão e medida F.

Assim como em Oberlander e Nowson (2006), nos dados considerados por Farnadi et al. (2013) também houve um grande número de indivíduos com a dimensão liberalidade positiva (aproximadamente 70%). Isso, segundo Farnadi et al. (2013), fez com que nenhum modelo computacional tivesse um resultado superior ao *baseline*. Considerando-se apenas os conjuntos de atributos individualmente, o conjunto (S) foi o que obteve o melhor desempenho em relação ao *baseline*.

Foi realizada também uma combinação de todos os atributos possíveis para descobrir qual o melhor conjunto de atributos para cada dimensão. Para extroversão e agradabilidade o conjunto (S) isoladamente alcançou a melhor medida F. Já para instabilidade, os atributos mais relevantes são os do conjunto (O). Meticulosidade obteve a melhor medida F utilizando os conjuntos de atributos (L), (S) e (T), e liberalidade obteve a melhor medida F utilizando o conjunto (L) e o conjunto (S).

Para entender quais atributos foram mais importante para cada dimensão de personalidade dentro de um conjunto de atributos, utilizou-se a correlação de Pearson entre os atributos e as respostas dos inventários de personalidade. A dimensão extroversão teve uma boa correlação com todos os atributos do conjunto social. Já para meticulosidade, a correlação positiva de tamanho da rede aliada à correlação negativa com densidade sugeriu que a rede desses sujeitos era grande e esparsa.

Em Farnadi et al. (2013), assim como em Nowson (2007), além da avaliação do *cópus* original, fez-se uso de dois *cópus* distintos para treinamento e teste do modelo computacional de reconhecimento de personalidade. Utilizou-se o conjunto de ensaios disponível no trabalho de Mairesse (2008) para realizar uma tarefa de aprendizado entre diferentes domínios, já que os textos presentes em Mairesse (2008) são ensaios e os textos presentes em Farnadi et al. (2013) são atualizações de status de *Facebook*. Como os textos de ensaios de Mairesse (2008) não possuem nenhuma característica de redes sociais, foi possível utilizar apenas atributos do dicionário LIWC (L). Diferentemente do que foi apresentado por Nowson (2007), foram apresentadas três estratégias de treinamento e teste com os diferentes domínios de textos, listadas abaixo:

- **FB-Ensaio** - Treino do classificador no *cópus* do *Facebook* e teste no *cópus* de ensaios.
- **Ensaio-FB** - Treino do classificador no *cópus* de ensaios e teste no *cópus* do *Facebook*.
- **Ensaio+FB-FB** - Treino do classificador em um *cópus* que combina os textos de ensaios e do *Facebook* e teste no *cópus* do *Facebook*.

Os resultados dessa comparação entre domínios mostraram que os textos de ensaios não apresentam resultados positivos quando usa-se o modelo treinado com os textos do *Facebook*. Contudo, o modelo treinado usando os textos de ensaios obteve melhores resultados para o *Facebook*, assim como o modelo treinado com os dois tipos de texto.

Assim como no trabalho de Celli (2012), em Farnadi et al. (2013) acredita-se que não há diferença entre diferentes domínios de texto para a classificação da personalidade. Apresentou-se também um estudo de outros atributos como tempo e atributos provenientes de uma análise da rede social do indivíduo, além de atributos linguísticos para a classificação da personalidade. Ao fim do trabalho, apresentou-se um breve experimento para realizar uma seleção de atributos mais fina para previsão de uma dimensão específica, como utilizar apenas o atributo *anger* do dicionário LIWC para o reconhecimento da dimensão instabilidade. Nesta tarefa, foi obtido um resultado superior a qualquer uma apresentada nesta abordagem.

Do ponto de vista da presente pesquisa, o trabalho é relevante porque apresenta um estudo do domínio *Facebook*, que é o tipo de texto a ser considerado na presente pesquisa.

3.3.3 Uso do modelo de personalidade *Myers-Briggs*

No trabalho de Plank e Hovy (2015), utilizou-se um conjunto extenso de *tweets* para reconhecimento de personalidade. Diferentemente de outros trabalhos, utiliza-se o indicador *Myers-Briggs* para medir a personalidade (BRIGGS, 1962). Este indicador modela quatro dimensões - Introvertido-Extrovertido (I-E), Intuitivo-Sensorial (I-S), Racional-Sentimental (T-F) e Julgamento-Percepção (J-P) - que, combinadas, geram 16 tipos de personalidade.

Ao todo, foram coletados 1,2 milhão de *tweets*, e fez-se uma análise comparativa entre a população do experimento e a população em geral para verificar a distribuição

das personalidades. Como atributos para o modelo, usou-se bigramas aliados a atributos provenientes da rede social, *e.g.*, gênero e número de seguidores.

Comparando-se a distribuição das dimensões da população norte-americana em geral com a população obtida pelo estudo, foi observado que há mais introspectivos do que na população em geral. Os tipos mais raros na população eram os que estavam presentes em maior número no conjunto de dados de Plank e Hovy (2015). Por este motivo, acreditou-se que um modelo gerado por tais dados poderia não ser tão fiel à distribuição de personalidade da população. Apesar desse fato, em Plank e Hovy (2015) os autores sugerem que o texto de *tweets* faz uso real da língua e que, portanto permite uma análise em maior escala e está menos enviesado do que uma entrevista pessoal para a obtenção da personalidade.

O pré-processamento consistiu em substituir *hashtags*, URLs e nomes de usuários por *tokens* que pudessem ser identificados nas próximas etapas. O treinamento foi feito com uma regressão logística da biblioteca *scikit-learn*¹. Os melhores resultados foram obtidos para I-E e T-F, com acurácia de 72,5 e 61,2, respectivamente. As demais dimensões não conseguiram superar o *baseline* da classe majoritária. Segundo Plank e Hovy (2015), esse resultado deve-se à falta de atributos que representem bem essas dimensões ou até mesmo à possibilidade de que não haja evidências linguísticas para essas dimensões, já que elas estão fortemente correlacionadas com a percepção. Fez-se também um experimento exploratório removendo-se as *stopwords* do cópys, mas o desempenho do modelo com as *stopwords* foi superior e os resultados apresentados anteriormente incluem essas palavras.

O trabalho de Plank e Hovy (2015) enfatiza o grande conjunto de dados que estava disponível para análise e o fato de que poucos trabalhos da área possuem essas proporções. Além disso, foi utilizado um modelo de personalidade diferente do CFG e afirma-se que algumas dimensões não obtiveram uma boa classificação, o que não acontece com trabalhos que já conseguiram resultados melhores para todas as dimensões. Em Plank e Hovy (2015) destaca-se que o uso do modelo CGF é muito mais difundido e que a construção desse modelo, por ser baseada em uma análise fatorial da própria língua natural, contribui para que o modelo CGF seja preferível a abordagens baseadas no modelo *Myers Briggs*.

---

¹ <http://scikit-learn.org/stable/>

3.4 Iniciativas para o português brasileiro

Nesta seção são listados trabalhos que consideram o caso específico de reconhecimento de personalidade para o português brasileiro.

3.4.1 Mapeamento de personalidade para o português brasileiro

Em Nunes, Teles e Souza (2013) tem-se uma abordagem de reconhecimento de personalidade para o português brasileiro utilizando técnicas para a língua inglesa (MAIRESSE et al., 2007) e postagens da rede social *Twitter*. Para cumprir o objetivo do trabalho, usou-se dois inventários de personalidade para realizar a coleta explícita dos dados. Nessa coleta, os participantes do experimento respondem aos inventários TIPI (*Ten-Item Personality Inventory*) e NEO-IPIP (*NEO-International Personality Item Pool*). Para possibilitar que mais pessoas respondessem aos inventários, desenvolveu-se um aplicativo para o sistema operacional *Android*. Além da possibilidade de responder pelo aplicativo, também era possível utilizar uma plataforma *web* que já estava preparada para esse fim. O trabalho não esclarece como os questionários foram traduzidos para o português brasileiro, nem como eles lidaram com a diferença de idiomas, já que era possível respondê-los em português e em inglês.

A coleta implícita dos dados consistiu em coletar *tweets* dos indivíduos que responderam aos inventários. Ao todo foram coletados *tweets* de 28 participantes. Como o intuito do trabalho era obter expressões baseadas nas características do texto de um indivíduo, como as utilizadas por Mairesse (2008), optou-se por adaptar o *Thesaurus* eletrônico para o português do Brasil (TEP 2.0) (MAZIERO et al., 2008) para os atributos LIWC. A versão do LIWC para o português só foi criada em 2013, ano dessa publicação, possível motivo pelo qual os autores fizeram essa adaptação.

Baseando-se no trabalho de Mairesse et al. (2007), foram utilizadas expressões de mapeamento para cada dimensão, como soma dos termos importantes destas. Um exemplo é a expressão para agradabilidade, para a qual as palavras negativas foram subtraídas das palavras positivas, e destas subtraíram-se os artigos, somou-se 100 e dividiu-se tudo por dois.

Para verificar se as expressões de mapeamento eram válidas, fez-se o cálculo da correlação dos resultados das expressões matemáticas de cada dimensão e os resultados dos inventários de personalidade. A correlação dos resultados com as respostas do TIPI é bastante fraca, abaixo de 0,3, assim como a correlação destes com as respostas do NEO-IPIP.

Como a amostra de participantes foi muito pequena, é difícil afirmar se a técnica foi mal sucedida ou se o mau resultado ocorreu devido ao tamanho da amostra. Além disso, o trabalho utiliza as propostas do trabalho de Mairesse (2008), que utilizou os textos em inglês e um corpus de ensaios, enquanto no trabalho de Nunes, Teles e Souza (2013) o corpus é composto por textos de redes sociais em português.

Existem diversas melhorias possíveis para o trabalho de Nunes, Teles e Souza (2013). Uma delas, citada pelos próprios autores, é o teste em um conjunto maior de dados. Apesar das limitações, o trabalho de Nunes, Teles e Souza (2013) é importante porque representa uma das poucas iniciativas de reconhecimento de personalidade para o português brasileiro que pudemos identificar até o presente momento.

3.4.2 Léxico em Português-Brasileiro para mineração de personalidade em textos

O estudo de Machado et al. (2015) objetivou mapear a ocorrência de palavras indicadoras de personalidade em textos para a construção de um léxico que incorpore características de personalidade das palavras. Isso foi feito em três etapas: avaliação do dicionário LIWC; construção e adaptação de um dicionário chamado *Personalitatem Lexicon*; e teste em uma disciplina de graduação utilizando inventários de personalidade. Apenas a última etapa foi descrita no trabalho.

O corpus utilizado era composto de dois textos de bate-papos de dois estudantes de um curso de pós-graduação em Gestão Pública. De forma manual, um dos autores classificou cada uma das palavras do bate-papo, considerando-se as palavras que aparecem no dicionário LIWC, especialmente as que estão presentes na categoria de emoções positivas e emoções negativas. O autor classificou as palavras em quatro classes (não significativa, pouco significativa, significativa ou muito significativa) de acordo com cada dimensão de personalidade presente no CGF. Por exemplo, dada a palavra ‘tristeza’, o autor poderia classificá-la como pouco significativa para extroversão. Para que se determinasse a

dimensão de personalidade de um indivíduo, foi necessário somar todas as palavras que eram relevantes para cada dimensão de personalidade e dividir pela quantidade de palavras do texto.

A análise dos textos mostrou que um dos indivíduos tende mais a Socialização e o outro para Realização, sem considerar as demais dimensões de personalidade um indivíduo. Como os participantes não responderam nenhum inventário de personalidade, não foi possível fazer uma comparação dos resultados com suas personalidades.

O trabalho de Machado et al. (2015) traz poucos detalhes sobre a metodologia usada para escolher a dimensão para uma palavra, já que quem o definia era um dos pesquisadores. Além disso, como foram analisados apenas dois textos, não é possível saber se o método usado é efetivo. Os próprios autores comentam ao fim do trabalho que a avaliação de um ou poucos textos pode não refletir os fatores psicológicos de um indivíduo com precisão.

3.4.3 Estudos apresentados na competição PAN-2015

A competição PAN-CLEF é composta de uma série de eventos científicos e tarefas compartilhadas para descobertas com base em textos que acontece desde 2007. A competição possui três modalidades principais: identificação de autoria, originalidade e confiabilidade de textos.

O objetivo da competição do ano de 2015, na modalidade de identificação de autoria de textos, era identificar idade, gênero, e as dimensões de personalidade do CGF de usuários do *Twitter*. Utilizando um teste de personalidade *online*, conseguiu-se coletar um conjunto de dados anotado com idade, gênero e as dimensões de personalidade do modelo CGF para quatro línguas: inglês, espanhol, italiano e holandês. Neste trabalho a tarefa de reconhecimento de personalidade é tratada como regressão computacional e os resultados são avaliados levando-se em conta o erro quadrático médio (RANGEL et al., 2015).

O melhor resultado da edição de 2015 do evento foi obtido por Álvarez-Carmona et al. (2015), que está entre os três melhores resultados independente da língua. Neste trabalho, combinam-se representações de segunda ordem (*Second Order Representation*), com a técnica LSA (*Latent Semantic Analysis*) (LANDAUER; DUMAIS, 1997). Também com bons resultados, o trabalho de González-Gallardo et al. (2015) utilizou-se de combinações

de n-gramas de caracteres e n-gramas de *part-of-speech*, enquanto que em Grivas, Krithara e Giannakopoulos (2015) combinou-se contagens tf-idf com modelos de n-gramas e atributos de estilo textual.

Nota-se que, apesar de o sistema de Álvarez-Carmona et al. (2015) ser o vencedor global da competição, este sistema não apresenta o melhor desempenho para todas as dimensões de personalidade, e novamente observamos que não há uma estratégia única que proporcione resultados ótimos para todas as cinco dimensões do modelo CGF.

3.5 Considerações

Neste capítulo apresentamos uma visão geral de estudos sobre o reconhecimento computacional da personalidade humana. Os trabalhos apresentam diferentes maneiras de dividir os resultados dos inventários de personalidade para a tarefa de classificação, além de apresentarem outras possibilidades para o reconhecimento de personalidade, como o uso de regressão ou *ranking*. Apesar do estudo em Mairesse (2008) defender o uso dessas duas técnicas, elas são usadas apenas em Mairesse (2008) e Celli (2012). A técnica de classificação, por outro lado, permanece a mais popular dentre os estudos consultados, presente em todos eles.

O quadro 2 sumariza os principais trabalhos relacionados que foram apresentados neste capítulo.

Com base nos resultados do quadro 2, observamos que o uso do dicionário LIWC é frequentemente utilizado do dicionário LIWC: dos dez trabalhos presentes, seis utilizam o dicionário. O mesmo também ocorre com o uso de contagens de n-gramas, presentes em cinco dos trabalhos estudados. Esse fato ilustra a importância desses dois tipos de conhecimento para o reconhecimento de personalidade a partir de textos. Percebe-se também que apenas nos trabalhos de Iacobelli et al. (2011), Celli (2012) e Majumder et al. (2017) existe uma comparação de conjuntos de atributos textuais, enquanto nos outros trabalhos apenas um conjunto de atributos é utilizado isoladamente ou os outros atributos provêm de outras fontes.

Dos trabalhos apresentados, o único para a língua portuguesa é o de Nunes, Teles e Souza (2013). Este trabalho possui um número muito reduzido de participantes, sendo inferior apenas ao *corpus* de *Facebook* apresentado por Celli (2012) que utiliza o *corpus*

Quadro 2 – Resumo dos principais trabalhos revisados, apresentando a subseção que ele se encontra, método utilizado, tipo de conhecimento, número de participantes e dimensões de personalidade

Sec	Trabalho	Método	Conhecimento	Domínio	Part.	Dimensões
3.1.1	ARGAMON et al.	SVM	Gramática SFG	Ensaaios	2.263	extroversão e instabilidade
3.2.1	OBERLANDER; NOWSON	NB e SVM	N-gramas	Blogs	71	Todas, exceto liberalidade
3.2.3	GILL; NOWSON; OBERLANDER	Regressão logística	LIWC	Blogs	2.393	Todas
3.2.2	NOWSON	NB	N-gramas	Blogs	1.672	Todas, exceto liberalidade
3.1.2	MAIRESSE	Classificação: árvore de decisão, k-NN, NB, Ripper rule learning, ADA e SVM. Regressão e ranking.	MRC e LIWC	Ensaaios	2.479	Todas
3.3.1	CELLI	SVM	LIWC, n-gramas, MRC e atr. independentes da língua	Ensaaios e Facebook	23	Todas
3.3.2	FARNADI et al.	KNN, SVM e NB	LIWC, attr. de redes sociais e atr. de tempo	Facebook	250	Todas
3.3.3	PLANK; HOVY	Regressão logística	Bigramas, gênero e meta dados do twitter	Twitter	1.500	Não usa CGF
3.2.4	IACOBELLI et al.	SVM	Frequência de bigramas e palavras usadas por mais de um autor, stopwords, tf-idf, presença ou ausência de uma palavra e LIWC	Twitter	3.000	Todas
3.4.1	NUNES; TELES; SOUZA	Cálculo de correlações	TEP 2.0	Twitter	28	Todas
3.1.3	MAJUMDER et al.	CNN	LIWC e representação distribuída de palavras	Ensaaios	2.393	Todas

Fonte: Bárbara Barbosa Claudino Silva, 2018

do trabalho de Mairesse (2008) para treinar um modelo classificador de personalidade. Na seção anterior foi apresentado ainda um segundo trabalho para o português brasileiro, mas por se tratar da construção de um léxico, este não foi acrescentado à tabela final. A

existência de apenas dois trabalhos sobre esse tema nesse capítulo, ilustra a necessidade de mais estudos de personalidade para a língua portuguesa e reforçam importância do presente trabalho para a área.

4 Estudo exploratório

Este capítulo descreve a metodologia utilizada para a coleta de dados e um estudo exploratório realizado com os dados obtidos. Este estudo consistiu em realizar a tarefa de classificação para as cinco dimensões de personalidade do modelo CGF. Para isso, desenvolveu-se um aplicativo de coleta de dados, apresentado na seção 4.1, e utilizou-se duas estratégias de divisão de personalidade para a classificação, além de diferentes conjuntos de atributos, apresentados na seção 4.3. Os resultados desse estudo se encontram na seção 4.4.

4.1 Implementação da aplicação de coleta de dados

Para obtenção de um *cópus* de atualizações de *status* do *Facebook* rotulado com as personalidades do modelo CGF, fez-se necessário a implementação de um aplicativo de coleta de dados de personalidade e de postagens do *Facebook*. Antes do início do desenvolvimento do aplicativo, foi feita a solicitação para o Comitê de Ética em Pesquisa em Sujeitos Humanos da Escola de Artes, Ciências e Humanidades (EACH) da Universidade de São Paulo (USP) e o termo de consentimento livre e esclarecido para participação em pesquisa é apresentado ao participante na primeira página do aplicativo¹.

O aplicativo desenvolvido é uma aplicação *web*, desenvolvida em *Ruby on Rails*, e seu código fonte está disponível em um repositório no *github*². A primeira página consiste de uma breve explicação sobre o CGF, um *link* para a política de uso dos dados e um *link* para o termo de consentimento da pesquisa. O indivíduo pode clicar “*Entrar com o Facebook*” para acessar o inventário CGF. Ao clicar nesse botão, o indivíduo dará permissão para que o aplicativo acesse seus dados básicos e o histórico de atualizações do *Facebook*. Caso o indivíduo não aceite essa solicitação, ele recebe uma mensagem de erro e os seus dados não são coletados, sendo portanto desconsiderado na presente pesquisa.

Quando o indivíduo aceita a solicitação do próprio *Facebook*, conforme pode ser visualizado na figura 8, ele consegue responder às quarenta e quatro perguntas do IGFP-5 (ANDRADE, 2008), como mostra a figura 9. Ao clicar em ‘Mostrar resultados’ o indivíduo é redirecionado para uma página ilustrativa, que mostra seus resultados de personalidade

¹ https://teste-personalidade.herokuapp.com/research_terms

² <https://github.com/bahbbc/masters-big5>

Figura 8 – Tela de autorização para o aplicativo no *Facebook*

Fonte: Bárbara Barbosa Claudino Silva, 2018

com a porcentagem de cada dimensão e uma breve explicação destas, como pode ser visto na figura 10. O objetivo dessa página é apenas oferecer alguma recompensa para o participante, ou seja, informar detalhes sobre a sua personalidade, possivelmente motivando-o a divulgar ainda mais o inventário para seu círculo social.

1 = Discordo totalmente, 2 = Discordo parcialmente, 3 = Não concordo nem discordo, 4 = Concordo parcialmente, 5 = Concordo totalmente

1 - É conversador, comunicativo.	☐ 1	☒ 2	☐ 3	☐ 4	☐ 5
2 - Às vezes é frio e distante	☐ 1	☒ 2	☐ 3	☐ 4	☐ 5
3 - Tende a ser crítico com os outros.	☐ 1	☒ 2	☐ 3	☐ 4	☐ 5
4 - É minucioso, detalhista no trabalho.	☐ 1	☒ 2	☐ 3	☐ 4	☐ 5
5 - É assertivo, não teme expressar o que sente.	☐ 1	☒ 2	☐ 3	☐ 4	☐ 5
6 - Insiste até concluir a tarefa ou o trabalho.	☐ 1	☒ 2	☐ 3	☐ 4	☐ 5
7 - É depressivo, triste.	☐ 1	☒ 2	☐ 3	☐ 4	☐ 5
8 - Gosta de cooperar com os outros.	☐ 1	☒ 2	☐ 3	☐ 4	☐ 5
9 - É original, tem sempre novas idéias.	☐ 1	☐ 2	☒ 3	☐ 4	☐ 5

Figura 9 – Trecho do IGFP-5 da aplicação *web*

Fonte: Bárbara Barbosa Claudino Silva, 2018

A coleta de atualizações do *Facebook* foi feita de forma assíncrona. Isso é importante pois indivíduos com muitas atualizações ficariam ‘travados’ na página do inventário e poderiam fechar a página do navegador antes que a coleta fosse finalizada. Quando o indivíduo está na página final da aplicação, a requisição para coleta de atualizações fica em uma fila e, quando executada, realiza uma requisição para a *GraphAPI*³, que é a API responsável por coletar todas as atualizações daquele indivíduo, e as armazena em um banco de dados na aplicação *web*.

O aplicativo gera vários arquivos CSV (*Comma Separated Values*) como saída de dados: um arquivo que resume os dados básicos de todos os participantes e a pontuação

³ <https://developers.facebook.com/docs/graph-api/overview>

Teste de Personalidade

Parabéns, Bárbara! Você completou o teste!

A esquerda estão os traços principais do Big-5, e a direita os traços 'reversos'.

Por favor, compartilhe esse teste no seu facebook ou peça para as pessoas responderem para ajudar ainda mais na pesquisa. =)

Compartilhar 590

Uma breve explicação sobre cada traço

Agradável

Agradabilidade é a tendência para ser compassivo e cooperante em vez de suspeito e antagonista face aos outros. Este traço reflete diferenças individuais na preocupação com a harmonia social. Indivíduos "agradáveis" ou "amáveis" valorizam a boa relação com os outros. Eles são geralmente respeitosos, amigáveis, generosos, prestáveis e dispostos a fazer compromissos. Pessoas "agradáveis" têm também uma visão otimista da natureza humana. Elas acreditam que as pessoas são basicamente honestas, decentes e dignas de confiança. Indivíduos "não-amigáveis" ou mais independentes põem o interesse próprio acima da boa relação com os outros. Eles normalmente não se preocupam com o bem-estar dos outros, e por vezes o seu ceticismo acerca dos motivos dos outros fá-los ser desconfiados e pouco cooperativos.

Figura 10 – Tela ilustrativa de resultados

Fonte: Bárbara Barbosa Claudino Silva, 2018

em cada uma das dimensões do CGF; e arquivos individuais contendo os textos de cada participante. Nestes arquivos, cada linha corresponde à atualização de *status* de um participante e o nome do arquivo é o identificador daquele indivíduo. Os arquivos foram extraídos por meio de um procedimento de importação, da aplicação original, que gerava os CSVs e os compartilhava em uma pasta do *Dropbox*⁴.

4.2 Coleta de dados

O aplicativo foi divulgado em listas de alunos da USP e pelos próprios participantes da pesquisa através de suas páginas do *Facebook*. A coleta de dados foi iniciada no dia 5 de agosto de 2015 e encerrada no dia 5 de outubro de 2016. Foram coletados dados de 1.039 indivíduos, totalizando 194.400 atualizações de *status* do *Facebook* e mais de 2 milhões de palavras. Os indivíduos avaliados possuem em média 24 anos, contendo pessoas com no mínimo 18 anos e no máximo 61 anos. Quanto ao gênero dos participantes, 56% pertencem ao gênero masculino.

⁴ <https://www.dropbox.com>

Conforme discutido em (MAIRESSE et al., 2007), este método de autoavaliação possui precisão reconhecidamente inferior à avaliação por terceiros (*i.e.*, feita por especialistas em psicologia). Entretanto, em virtude dos custos de avaliação profissional em grande escala, a autoavaliação costuma ser o método mais comum em trabalhos desta área (MAIRESSE, 2008; ARGAMON et al., 2005; OBERLANDER; NOWSON, 2006; GILL; NOWSON; OBERLANDER, 2009; CELLI, 2012; IACOBELLI et al., 2011), e pode ser considerada suficiente para os propósitos do presente estudo.

Os valores de personalidade para cada uma das dimensões do CGF é um número escalar que varia de 1 a 5. As figuras 11 à 15 mostram a distribuição desses valores no corpus coletado para as cinco dimensões do CGF.

Figura 11 – Distribuição da dimensão Extroversão no conjunto de dados

Fonte: Bárbara Barbosa Claudino Silva, 2018

Figura 12 – Distribuição da dimensão Agradabilidade no conjunto de dados

Fonte: Bárbara Barbosa Claudino Silva, 2018

Pelas figuras nota-se que a distribuição de cada dimensão de personalidade é diferente, principalmente a dimensão Liberalidade, com a média em 3,7 a média mais distante comparada com as outras dimensões. Outra característica do conjunto abordado é o número baixo de pessoas com Agradabilidade e Liberalidade. Em todas as dimensões -

Figura 13 – Distribuição da dimensão Meticulosidade no conjunto de dados

Fonte: Bárbara Barbosa Claudino Silva, 2018

Figura 14 – Distribuição da dimensão Instabilidade no conjunto de dados

Fonte: Bárbara Barbosa Claudino Silva, 2018

Figura 15 – Distribuição da dimensão Liberalidade no conjunto de dados

Fonte: Bárbara Barbosa Claudino Silva, 2018

exceto Liberalidade - percebe-se que são poucos os indivíduos que ficam no extremo da dimensão, com poucos valores no 1 ou 5.

Para o experimento propriamente dito, optou-se por remover todos os participantes com menos de 780 palavras, o que corresponde ao primeiro quartil do número total de palavras do conjunto de dados. Do total de 1.039 participantes do corpus, foram identificados 778 indivíduos com textos de mais de 780 palavras. Mesmo com esta redução,

até onde temos conhecimento este é o maior conjunto de dados do gênero já produzido para o português brasileiro.

4.3 Procedimento

O objetivo deste estudo foi o desenvolvimento de modelos de reconhecimento da personalidade humana segundo postagens da rede social *Facebook*. De forma mais específica, o presente estudo irá abordar duas questões principais, a saber:

1. como modelar o reconhecimento de personalidade na forma de um problema de classificação.
2. quais tipos de conhecimento são mais úteis nesta tarefa.

Estas questões serão tratadas na forma de um experimento envolvendo modelos de aprendizagem de máquina supervisionada baseada no *cópus* descrito na seção 4.2.

Para que os atributos de aprendizagem sejam computados, o texto coletado passa por um extenso procedimento de pré-processamento e normalização. Este procedimento é descrito brevemente a seguir, e faz uso de dicionários de português e inglês e diversas bases auxiliares anotadas manualmente a partir de uma amostra dos dados coletados (PARABONI, 2016). Para cada linha de texto do arquivo de entrada, os seguintes passos são executados:

- **Remoção de caracteres especiais.** Símbolos não pertencentes ao conjunto de caracteres ocidentais são substituídos por espaço.
- **Remoção de URLs.** Linhas iniciadas por ‘http’ ou ‘www’ são desconsideradas.
- **Segmentação.** Espaços são inseridos antes e depois de cada palavra ou símbolos de pontuação, se necessário, e espaços consecutivos são eliminados. O resultado é uma série de termos (que podem ser palavras, pontuação ou termos compostos, como *hashtags*, a serem processados um a um).
- **Tratamento de *hashtags*.** Os símbolos # são eliminados. O termo resultante é tratado normalmente.
- **Linhas tracejadas.** Caracteres do tipo hífen no início do termo são eliminados. O termo resultante é tratado normalmente.

- **Termos enfatizados por hífen.** Se metade ou mais dos caracteres do termo for um hífen (*e.g.*, ‘f-é-r-i-a-s’), os hifens são removidos, e o termo corrigido é tratado normalmente.
- **Termos compostos hifenizados.** Se, por outro lado, o termo possui mais de dois hifens, é considerado um termo composto (*e.g.*, ‘hoje-eu-vou-dormir-muito’). Neste caso o termo é desmembrado em termos individuais, a serem tratados normalmente.
- **Termos compostos anotados.** Os termos que constam em uma lista de termos compostos anotados manualmente são desmembrados em termos individuais segundo regra da própria lista (*e.g.*, ‘nãoovou’ é desmembrado em ‘não’ e ‘vou’) Os termos resultantes são tratados normalmente.
- **Falsos homógrafos, termos reescritos e erros ortográficos.** Os termos que constam em uma lista de falsos homógrafos [15], de formas de escrita alternativa (*e.g.*, ‘véii’ como substituto de ‘velho’) e erros ortográficos anotados manualmente (*e.g.*, ‘paralizacao’) são substituídos pelo termo correto. O termo corrigido é tratado normalmente.
- **Termos compostos delimitados por maiúsculas.** Termos na forma de sequência de caracteres minúsculos com separadores maiúsculos (*e.g.*, ‘hojeTemFesta’). Os termos resultantes são tratados normalmente.
- **Reiteração.** Os termos resultantes das operações acima podem eles próprios ser termos compostos, e por isso o processo é repetido até resultar em uma lista de termos atômicos (*i.e.*, palavras ou sinais de pontuação) que não podem mais ser subdivididos.

O resultado deste procedimento é uma lista de termos atômicos a serem processados individualmente como segue. Para cada termo atômico da linha de texto (sentença), os seguintes passos são executados:

- **Pontuação e Numeração.** Se o termo for um sinal de pontuação ou se iniciar por um dígito, o termo é desprezado. O processamento segue então analisando o próximo termo.
- **Classificação.** A palavra é consultada em uma série de dicionários e listas de anotação manual. Primeiramente, é consultado o dicionário DELAF-Pt , seguido de listas anotadas manualmente representando emoções, palavras adicionadas manual-

mente, nomes próprios, palavras estrangeiras, palavras desconhecidas, sequências de itens a serem ignorados e por fim o dicionário inglês DELAF-En, nesta ordem.

- **Tratamento de ênfase por repetição de vogais.** Palavras que não foram reconhecidas até este ponto são examinadas para verificar se é um caso de ênfase por repetição de letras exceto R ou S (*e.g.*, "siiiiiiiiim"). Se uma versão sem letras repetidas existe no dicionário português, a palavra é corrigida, e a palavra é reclassificada repetindo-se o passo Classificação.
- **Acentuação automática.** Palavras que não foram reconhecidas até este ponto são examinadas para verificar se podem ser reescritas com certos tipos de acentuação. Por exemplo, palavras terminadas com 'cao' são reescritas com 'ção' e consultadas no dicionário. Se identificada, a palavra é convertida na sua forma normal, e a palavra corrigida é reclassificada repetindo-se o passo Classificação.

Após o pré-processamento textual, foram computados diversos tipos de atributos lexicais que serviram de apoio a este experimento exploratório. Cada atributo corresponde à proporção deste sobre o total de itens (palavras, símbolos de pontuação etc.) de cada texto. Após a extração de atributos conta-se com 164 atributos para utilizar de entrada em um algoritmo de classificação. Estes atributos são descritos brevemente a seguir.

- **Atributos de pré-processamento:** atributos obtidos à partir da correção ortográfica e pré-processamento dos textos, como número de palavras com grafia incorreta, indicativos de risadas, emoções positivas e negativas típicas da linguagem da rede social, etc. (PARABONI, 2016).
- **Atributos lexicais do dicionário Unitex-PB:** atributos obtidos do dicionário Unitex para o português brasileiro (MUNIZ, 2004), representando o número de ocorrências de cada característica representada no dicionário, como quantidade de itens em cada tempo verbal, quantidade de palavras femininas, etc.
- **Atributos psicolinguísticos do dicionário LIWC:** atributos que representam o número de ocorrências de cada categoria LIWC para o português brasileiro (FILHO; PARDO; ALUISIO, 2013) no texto em questão.

Dada a questão de pesquisa (1), neste trabalho seguimos boa parte dos estudos prévios sobre o tema, como (OBERLANDER; NOWSON, 2006; CELLI, 2012; ARGAMON et al., 2005; MAIRESSE, 2008; MAJUMDER et al., 2017), e optamos pela definição de cinco

tarefas de classificação independentes, isso é, uma para cada dimensão de personalidade do modelo CGF. Esta decisão é justificada pelo tipo de aplicação pretendida (*i.e.*, desejamos reconhecer traços de personalidade exclusivamente a partir de texto, e não a partir de outros traços já conhecidos), e também pelo fato de que os fatores de personalidade do modelo CGF são, por definição, altamente independentes (ANDRADE, 2008).

Por se tratar de uma tarefa de classificação, e pelo fato de que as pontuações de personalidade do CGF serem valores escalares, fez-se necessário utilizar uma estratégia que definisse quando um indivíduo pertenceria a uma determinada dimensão de personalidade dado sua pontuação. Para isso no presente experimento investigamos duas estratégias de divisão de classes de aprendizagem, propostas em Oberlander e Nowson (2006).

A primeira estratégia considerada consiste em dividir os indivíduos em positivos ou negativos para uma determinada dimensão de personalidade de acordo com a média desta. Esta estratégia é denominada divisão ao meio. A tabela 1 apresenta o resultado dessa divisão para todas as dimensões do conjunto de dados.

Tabela 1 – Divisão em positivo e negativo de acordo com a média da dimensão

Personalidade	positivo		negativo		média
extroversão	412	52,9%	366	47,1%	3,15
agradabilidade	406	52,1%	372	47,9%	3,54
meticulosidade	381	48,9%	397	51,1%	3,25
instabilidade	389	50,0%	389	50,0%	3,17
liberalidade	408	52,4%	370	47,6%	3,78

Fonte: Bárbara Barbosa Claudino Silva, 2018

A segunda estratégia é denominada divisão em extremos e usa a média e o desvio padrão para obter os indivíduos com o extremo de uma dimensão. Assim, para os indivíduos positivos para uma dimensão d tem-se: $\mu(d) + \sigma(d)$ e para os negativos: $\mu(d) - \sigma(d)$ onde σ representa o desvio padrão e μ a média para uma determinada dimensão. Os indivíduos que não atendem esses critérios são removidos do conjunto de dados considerado. A tabela 2 apresenta o resultado dessa divisão para todas as dimensões do conjunto de dados, com a sua proporção em positivo e negativo para uma determinada dimensão, além dos indivíduos excluídos, ou seja, que não possuem uma pontuação tão alta ou tão baixa para a dimensão e portanto serão desconsiderados na tarefa de classificação.

Com relação a questão (2) do experimento, (*i.e.*, quais tipos de conhecimento são mais úteis à tarefa), foram realizados experimentos com cada um dos tipos de atributos (*i.e.*, pré-processamento, lexicais/Unitex-PB e psicolinguísticos/LIWC) além de um experimento

Tabela 2 – Personalidade separadas em extremos - usando a média e desvio padrão

Personalidade	positivo		negativo		medianos	
extroversão	161	20,6%	143	18,3%	474	60,1%
agradabilidade	144	18,5%	107	13,7%	527	68,8%
meticulosidade	117	15,0%	134	18,0%	527	67,0%
instabilidade	151	19,4%	120	15,4%	507	65,2%
liberalidade	144	18,5%	99	12,7%	535	68,8%

Fonte: Bárbara Barbosa Claudino Silva, 2018

que levava em conta todos os atributos combinados, totalizando quatro estratégias de combinação de atributos.

Em virtude da divisão do conjunto de atributos em blocos, cabe assim destacar que o presente experimento não tratará do papel individual de cada atributo para a solução do problema, mas apenas dos modelos gerais correspondentes a cada fonte de conhecimento como um todo. Os quatro modelos propostos foram aplicados às dez tarefas propostas (2 alternativas de divisão em classes binárias vezes 5 dimensões de personalidade), resultando em 40 classificadores binários.

Para a construção do classificador utilizou-se uma estratégia de classificação binária do tipo SVM com núcleo radial com todos os valores padrões (*e.g.* γ padrão é $1/n$, onde n é o número de atributos do conjunto de dados e o $c = 1$) da biblioteca *e1071* da linguagem *R*⁵. Escolhemos utilizar a SVM por ser um algoritmo bastante popular entre vários dos trabalhos vistos (vide quadro 2). O objetivo (2) é independente do tipo de algoritmo utilizado, já que queremos descobrir quais tipos de conhecimento são mais úteis ao modelo.

4.4 Resultados

Para cada uma das dimensões do CGF, vê-se que classes positivas e negativas já estavam aproximadamente balanceadas (tabelas 1 e 2). Os resultados finais foram obtidos através do uso de uma validação cruzada de dez partes, e para avaliar os modelos computacionais gerados utilizou-se a medida F, precisão, revocação e acurácia.

Os resultados da avaliação com precisão, revocação e medida F são apresentados nas tabelas 3 e 4 para cada uma das dimensões de personalidade do CGF. Em ambas tabelas, os melhores resultados para cada dimensão CGF são destacados.

⁵ <https://www.r-project.org/>

Tabela 3 – Resultados para o uso de SVM para a divisão binária ao meio

	Medida F	Revocação	Precisão	Acurácia
Extroversão				
Pré-proces.	0,56	0,53	0,61	0,52
Unitex-PB	0,57	0,55	0,59	0,55
LIWC	0,61	0,58	0,64	0,57
Todos	0,59	0,55	0,64	0,54
Agradabilidade				
Pré-proces.	0,26	0,45	0,28	0,49
Unitex-PB	0,41	0,32	0,32	0,50
LIWC	0,41	0,57	0,45	0,54
Todos	0,31	0,62	0,35	0,50
Meticulosidade				
Pré-proces.	0,56	0,55	0,61	0,50
Unitex-PB	0,51	0,49	0,55	0,45
LIWC	0,64	0,59	0,71	0,58
Todos	0,60	0,56	0,68	0,52
Instabilidade				
Pré-proces.	0,32	0,44	0,31	0,47
Unitex-PB	0,39	0,47	0,35	0,50
LIWC	0,43	0,46	0,43	0,50
Todos	0,40	0,45	0,38	0,48
Liberalidade				
Pré-proces.	0,38	0,44	0,34	0,47
Unitex-PB	0,36	0,45	0,31	0,48
LIWC	0,44	0,50	0,40	0,51
Todos	0,39	0,45	0,36	0,48

Fonte: Bárbara Barbosa Claudino Silva, 2018

Com relação à questão de pesquisa (1), relacionada à divisão de classes de aprendizagem, observamos que, de modo geral, a estratégia de divisão em extremos apresenta melhores resultados, como sugere Oberlander e Nowson (2006), obtendo resultados similares apenas para a dimensão meticulosidade. No entanto, nota-se pela tabela 2 que ela descarta um grande número de dados intermediários e, portanto, em termos práticos, a estratégia de divisão ao meio é mais atraente.

Com relação à questão de pesquisa (2), relacionada ao tipo de conhecimento mais útil para reconhecimento de personalidade, observamos que, de modo geral, o uso de conhecimento proveniente do dicionário LIWC, ainda que por vezes combinado com outras fontes de conhecimento, apresenta melhores resultados. Esse fato pode ser comprovado verificando-se a acurácia das dimensões instabilidade, meticulosidade e agradabilidade

Tabela 4 – Resultados para o uso de SVM para a divisão binária ao extremo

	Medida F	Revocação	Precisão	Acurácia
Extroversão				
Pré-proces.	0,59	0,54	0,66	0,53
Unitex-PB	0,60	0,61	0,61	0,59
LIWC	0,62	0,62	0,63	0,60
Todos	0,63	0,62	0,63	0,61
Agradabilidade				
Pré-proces.	0,73	0,65	0,85	0,63
Unitex-PB	0,70	0,63	0,79	0,60
LIWC	0,73	0,68	0,82	0,65
Todos	0,72	0,64	0,83	0,63
Meticulosidade				
Pré-proces.	0,41	0,52	0,42	0,52
Unitex-PB	0,45	0,56	0,42	0,53
LIWC	0,59	0,71	0,56	0,65
Todos	0,48	0,66	0,45	0,59
Instabilidade				
Pré-proces.	0,69	0,59	0,86	0,57
Unitex-PB	0,67	0,58	0,79	0,55
LIWC	0,71	0,63	0,82	0,61
Todos	0,69	0,59	0,82	0,58
Liberalidade				
Pré-proces.	0,67	0,62	0,73	0,60
Unitex-PB	0,67	0,57	0,80	0,55
LIWC	0,71	0,64	0,82	0,64
Todos	0,72	0,64	0,82	0,65

Fonte: Bárbara Barbosa Claudino Silva, 2018

para a divisão em extremos e para todas as dimensões de personalidade na divisão ao meio.

Por outro lado, os atributos computados na fase de pré-processamento dos textos parecem ser os que menos contribuem para a solução do problema. Uma possível explicação para o baixo desempenho do conjunto de atributos de pré-processamento pode estar relacionada ao gênero de texto considerado. A linguagem adotada nas redes sociais tende a possuir características próprias (*e.g.*, o uso frequente de palavras abreviadas como “vc” para “você”) que talvez sejam utilizadas de forma generalizada por indivíduos com diversos tipos de personalidade. Em outras palavras, é possível que o uso de certas abreviaturas típicas da linguagem das redes sociais não sofra influência significativa da personalidade.

4.5 *Considerações*

De modo geral, estes resultados sugerem que a divisão ao meio é uma estratégia de divisão mais adequada para a tarefa, já que além de não excluir nenhum indivíduo da tarefa de classificação, como faz a estratégia de divisão em extremos, obtém um desempenho de modo geral razoável para algumas dimensões. Com relação à natureza do conhecimento a ser empregado no reconhecimento de personalidade a partir de textos em português brasileiro, o estudo exploratório confirma o papel do conhecimento psicolinguístico proveniente do dicionário LIWC, resultado este já observado em estudos do gênero aplicados aos idiomas italiano (CELLI, 2012) e inglês (MAIRESSE, 2008), dentre outros. Estas observações serão levadas na proposta dos modelos de reconhecimento de personalidade a serem discutidos no próximo capítulo.

5 Reconhecimento de personalidade com base em textos

A revisão bibliográfica apresentada no capítulo 3 confirma que é possível reconhecer a personalidade de um indivíduo com base em seus textos. Além disso, apesar de grande parte dos trabalhos vistos serem específicos para outras línguas, segundo (JOHN; SRIVASTAVA, 1999), o modelo CGF também deve ser válido para o português brasileiro.

No caso específico do português brasileiro, o estudo exploratório discutido no capítulo 4 sugere que, neste caso, esta abordagem também pode ser utilizada. Contudo, o estudo apresentado ainda é incipiente, baseado apenas no conhecimento lexical proveniente dos dicionários LIWC e Unitex-PB. Não foram exploradas outras formas de conhecimento empregadas na literatura, como o uso de técnicas que utilizam conhecimento proveniente do próprio texto, como utilizado em Oberlander e Nowson (2006), ou Majumder et al. (2017) que utiliza representação distribuída de palavras. Assim, é possível que outras formas de conhecimento proporcionem uma melhora nos resultados obtidos neste primeiro momento.

Este capítulo propõe uma série de modelos para a tarefa de reconhecimento de personalidade do modelo CGF com base em textos. A seção 5.1 apresenta as estratégias de entrada (*input*) de dados mais relevantes para a tarefa de classificação, utilizando diferentes formas de representação de texto como BoW, representação distribuída de palavras (*word embeddings*), de documentos e léxicos como LIWC e propriedades psicolinguísticas. Para avaliar essas diferentes estratégias utilizou-se majoritariamente o algoritmo *Random Forest*, além de algumas alternativas baseadas em redes LSTMs. Os resultados desses experimentos encontram-se no capítulo 6.

5.1 Modelos de reconhecimento de personalidade propostos

Para a tarefa de reconhecimento de personalidade a partir de texto, propomos uma série de modelos utilizando o algoritmo *Random Forest* com base em diferentes formas de representação textual, e complementados por alternativas que utilizam redes LSTMs. Como visto na seção 4.5 do estudo exploratório, optamos por usar a estratégia de divisão ao meio, que não exclui nenhum indivíduo da pesquisa. Mantendo a estratégia de utilizar

todos os dados disponíveis, optamos também por utilizar os dados dos indivíduos com poucas postagens, identificados na seção 4.2 como o primeiro quartil do conjunto de dados.

O modelo *Random Forest* foi escolhido especialmente para viabilizar a comparação de atributos. Por esse motivo, a configuração das árvores é mantida fixa, mantendo os hiper-parâmetros sempre os mesmos em todos os experimentos. O algoritmo *Random Forest* permite observar a importância de cada um dos atributos de um modelo, característica que foi explorada para compreender melhor quais atributos do dicionário LIWC e atributos psicolinguísticos são mais relevantes para cada dimensão de personalidade. Para a construção dos modelos utilizou-se o pacote *scikit-learn* da linguagem *Python*.

Os modelos baseados em *Random Forest* utilizam cinco formas de representação textual, a saber: representação distribuída de palavras; representação distribuída de documentos; dicionário LIWC; atributos psicolinguísticos; e *bag of words*. Estes modelos foram criados utilizando-se Essa configuração de modelos *Random Forest* foi feita em uma validação cruzada de 10 partes e os resultados dos experimentos desta seção são apresentados na seção 6.3.

Como forma de complementar a avaliação dos modelos de reconhecimento de personalidade baseados no método *Random Forest*, implementou-se um modelo de reconhecimento de personalidade baseado em redes LSTM. O modelo de rede LSTM foi escolhido devido à sua capacidade de computar dependências de longo prazo, ou seja, de associar dados que entraram em momentos diferentes na rede para a tomada de decisão final.

Os modelos baseados em redes LSTM utilizam três estratégias de modelos de representação distribuída de palavras como entrada para a rede. Contudo, neste capítulo apresentamos apenas a estratégia que utiliza o texto de postagens do *Facebook* para a criação dos vetores de palavras. Devido ao grande número de hiper-parâmetros dessa rede maiores detalhes da implementação são explicados na seção 5.1.5. Para essa estratégia utilizou-se um conjunto de treino e validação, composto por 70% (727 indivíduos) e 30% (312 indivíduos) dos dados respectivamente.

5.1.1 Estratégia baseada em Bag of Words - BoW

A primeira estratégia de reconhecimento de personalidade proposta é baseada em um modelo de representação textual do tipo Bag of Words (BoW). Aqui utilizamos o

cópus original de textos, ou seja, o cópus apresentado no capítulo 4 com as mesmas etapas de pré-processamento textual. Além destas etapas descritas na seção 4.3, utilizamos o processo de lematização textual para reduzir a dimensionalidade da matriz de frequência de palavras. Ao final do processo de geração dos unigramas obteve-se uma matriz de 22.612 termos. Esta estratégia será aqui denominada **E1-BoW**.

5.1.2 Estratégia baseada em conhecimento psicolinguístico

Visto que o uso de conhecimento psicolinguístico é uma forma amplamente difundida de reconhecer personalidade (MAIRESSE, 2008; GILL; NOWSON; OBERLANDER, 2009; CELLI, 2012; FARNADI et al., 2013; IACOBELLI et al., 2011; MAJUMDER et al., 2017), consideramos uma estratégia de reconhecimento de personalidade baseada nas categorias LIWC provenientes de Filho, Pardo e Aluisio (2013) combinadas com as propriedades psicolinguísticas estimadas em Santos et al. (2017).

Como já apresentado no capítulo 4, para o cálculo dos atributos do dicionário LIWC, considerou-se que cada atributo representa a proporção de cada atributo do dicionário sobre o total de itens.

Já para o cálculo das propriedades psicolinguísticas, para obter um valor de cada um dos atributos para um conjunto de postagens faz-se necessário o uso de um método específico para combinação do valor de cada palavra. No presente trabalho utilizou-se a média de cada atributo para todas as palavras presentes na postagem para combinar os atributos.

A estratégia geral que utiliza conhecimento das propriedades psicolinguísticas combinadas com os atributos do dicionário LIWC será denominada **E2-LIWC+psico**.

5.1.3 Estratégias baseadas em modelos de representação distribuída de palavras

O principal conjunto de estratégias de reconhecimento de personalidade propostos no presente estudo faz uso de modelos de representação vetorial de palavras ou *word embeddings* (MIKOLOV et al., 2013a).

Assim como acontece com a estratégia que utiliza o modelo *BoW*, nestas estratégias o próprio texto é transformado para ser utilizado como atributo, conforme discutido na

seção 2.5. Em Mikolov et al. (2013a) vê-se que existem diversas combinações possíveis de arquiteturas de representações vetoriais, e que cada problema possui uma arquitetura ótima diferente. Por esse motivo, neste trabalho consideramos uma série de modelos de *word embeddings* baseados tanto na arquitetura CBOW como *skip-gram*.

Com relação à arquitetura CBOW, em virtude de restrições de memória computacional disponível, o estudo limitou-se a sete modelos de *word embeddings* deste tipo, em todos os casos considerando uma frequência mínima de cinco palavras. Estes modelos levaram à proposta de quatro estratégias fundamentais de reconhecimento de personalidade, aqui denominadas E3, E4, E5 e E6 e duas variações destas a serem discutidas no capítulo 6. As estratégias fundamentais baseadas na arquitetura CBOW são as seguintes:

- **E3-cbow-300:** Vetor de tamanho 300 treinado em um conjunto de 50.000 *tweets* (CASIMIRO; PARABONI, 2014). Como cada palavra é um vetor de tamanho 300, para combinar as palavras que formam um post de *Facebook* utilizou-se a média vetorial das palavras que compunham o conjunto de *posts* de cada sujeito.
- **E4-cbow-600:** Vetor de tamanho 600 treinado em um cópulus de 50.000 *tweets*. Para combinar as palavras, utilizou-se a média vetorial de cada palavra que compunha o conjunto de *posts* de cada sujeito.
- **E5-cbow-600-norm:** Vetor de tamanho 600 treinado em um cópulus de 50.00 *tweets*. Para combinar as palavras, assim como na estratégia **E4-cbow-300** utilizou-se a média vetorial. No entanto, o cópulus de *posts* foi normalizado de uma maneira diferente, apenas convertendo as letras maiúsculas em minúsculas, sem nenhum outro dos pré processamentos textuais descritos no capítulo 4 de pré-processamento textual.
- **E6-cbow-600-stop:** Vetor de tamanho 600 treinado em um cópulus de 50.00 *tweets*. Para combinar as palavras, assim como na estratégia **E4-cbow-300** utilizou-se a média vetorial. No entanto, as *stopwords* foram removidas da combinação vetorial, ou seja, palavras da lista de *stopwords* em português da biblioteca *nltk*¹ não foram consideradas para a média vetorial.

Com relação à arquitetura *skip-gram*, foram considerados seis modelos de *word embeddings* desta arquitetura, em todos os casos também utilizando-se de uma frequência

---

¹ <http://www.nltk.org/>

mínima de cinco palavras. Estes modelos levaram à proposta de quatro estratégias fundamentais de reconhecimento de personalidade, aqui denominadas E7, E8, E9 e E10 e duas variações adicionais destas a serem discutidas no capítulo 6. As estratégias fundamentais baseadas na arquitetura *skip-gram* são as seguintes:

- **E7-skip-300**: Tamanho 300 treinado em um conjunto de 50.000 *tweets*. Como cada palavra é um vetor de tamanho 300, para combinar as palavras que formam um post de *Facebook* utilizou-se a média vetorial de cada palavra que compunha o conjunto de *posts* de cada sujeito.
- **E8-skip-600**: Tamanho 600 treinado em um cópua de 50.000 *tweets*. Para combinar as palavras, utilizou-se a média vetorial de cada palavra que compunha o conjunto de *posts* de cada sujeito.
- **E9-skip-600-stop**: Vetor de tamanho 600 treinado em um cópua de 50.00 *tweets*. Para combinar as palavras, utilizou-se a média vetorial, como na (**E8-skip-600**). No entanto, as *stopwords* foram removidas da combinação vetorial.
- **E10-skip-600-norm**: Vetor de tamanho 600 treinado em um cópua de 50.00 *tweets*. Para combinar as palavras, assim como na estratégia **E8-skip-600** utilizou-se a média vetorial. No entanto, o cópua de *posts* foi normalizado de uma maneira diferente, apenas convertendo as letras maiúsculas em minúsculas, sem nenhum outro dos pré processamentos textuais descritos no capítulo 4 de pré-processamento textual.

5.1.4 Estratégia baseada em modelos de representação distribuída de documentos

Em complemento ao uso de modelos de representação distribuída de palavras da seção 5.1.3, consideramos também uma estratégia baseada na representação distribuída de documentos do tipo PV-DM (LE; MIKOLOV, 2014) discutido na seção 2.6. Para a estratégia considerou-se o tamanho de oito para os vetores de atributos, e vinte épocas (números de iterações) no cópua. Esta estratégia será aqui denominada **E11-doc2vec**.

5.1.5 Estratégias baseadas em redes LSTM

Diferentemente de todas as estratégias anteriores, consideramos ainda no presente estudo uma estratégia geral de reconhecimento de personalidade baseada em redes neurais do tipo LSTM (cf. seção 2.7). O estudo limitou-se a três configurações para as redes LSTMs, em todas as estratégias utilizando-se no máximo 10.000 palavras, que foram compostas das primeiras palavras únicas do conjunto total de palavras, que compôs a matriz de entrada do algoritmo.

Configurou-se todas as redes com o núcleo = 5, função de ativação sigmóide, número de épocas = 5, tamanho do *batch* = 64 e número da janela = 100. Essa configuração levou a implementação de três estratégias: **E12-LSTM-600**, **E13-LSTM-skip-100** e **E14-LSTM-skip-150-media**.

- **E12-LSTM-600**: Não utiliza nenhuma representação distribuída de *tweets*, ao invés disso, utiliza o próprio conjunto de dados do *Facebook* para criar a representação distribuída que será usada de entrada para a rede. Para isso utilizamos uma representação de tamanho 600.
- **E13-LSTM-skip-100**: Utiliza uma representação distribuída de palavras com arquitetura *skip-gram* de tamanho 100 que foi treinada com 150.000 *tweets* do mesmo conjunto de *tweets* utilizado na seção 5.1.3. Nenhuma combinação de palavras foi feita.
- **E14-LSTM-skip-150-media**: Utiliza uma representação distribuída de palavras com arquitetura *skip-gram* de tamanho 100 que foi treinada com 150.000 *tweets* e para combinar as palavras das postagens iniciais faz uso da média.

5.2 Discussão

Neste capítulo foram propostas 14 estratégias computacionais de reconhecimento de personalidade a partir de texto. Estas estratégias são resumidas no quadro 3.

Estas estratégias, e diversas variações destas, são avaliadas em uma série de experimentos descritos no próximo capítulo.

Quadro 3 – Resumo das estratégias apresentadas neste capítulo, apresenta-se o nome da estratégia, o método utilizado, o conhecimento e detalhes (se houverem)

Estratégia	Método	Conhecimento	Detalhes
E1-BoW	RF	unigramas	-
E2-LIWC+psico	RF	LIWC e atr psico.	-
E3-cbow-300	RF	<i>word embeddings</i>	tamanho 300
E4-cbow-600	RF	<i>word embeddings</i>	tamanho 600
E5-cbow-600-norm	RF	<i>word embeddings</i>	dados sem pré-processamento
E6-cbow-600-stop	RF	<i>word embeddings</i>	stopwords
E7-skip-300	RF	<i>word embeddings</i>	tamanho 300
E8-skip-600	RF	<i>word embeddings</i>	tamanho 600
E9-skip-600-stop	RF	<i>word embeddings</i>	stopwords
E10-skip-600-norm	RF	<i>word embeddings</i>	dados sem pré-processamento
E11-doc2vec	RF	<i>document embeddings</i>	-
E12-LSTM-600	LSTM	<i>word embeddings</i>	tamanho 600
E13-LSTM-skip-100	LSTM	<i>word embeddings</i>	tamanho 100
E14-LSTM-skip-150-media	LSTM	<i>word embeddings</i>	tamanho 150

Fonte: Bárbara Barbosa Claudino Silva, 2018

6 Avaliação

Este capítulo descreve a avaliação das estratégias de reconhecimento de personalidade propostas no capítulo anterior, além de experimentos secundários relativos à tarefa de reconhecimento de personalidade. Na seção 6.1 apresenta-se a hipótese de pesquisa e as hipóteses geradas a partir das estratégias do capítulo 5. Os resultados destas estratégias podem ser verificados na seção 6.3. Além das análises fundamentais apresentadas no capítulo anterior também foram desenvolvidas outras estratégias para a tarefa, suas descrições e resultados encontram-se na seção 6.5.

6.1 Hipóteses

Conforme enunciado no capítulo 1 a hipótese geral dessa pesquisa é a de que o uso de modelos de reconhecimento de personalidade a partir de texto baseados em representação distribuída de palavras proporcionará resultados comparáveis aos de métodos dependentes de conhecimento lexical, como o dicionário LIWC. Esta hipótese será verificada comparando-se os diversos modelos aplicados à tarefa de classificação binária de personalidade com base no modelo CGF.

De forma mais específica, esta hipótese será verificada com uso do conjunto de dados de postagens do *Facebook* obtido neste trabalho, comparando-se o modelo baseado em representação distribuída de palavras com alternativas que utilizem outras forma de conhecimento¹. Espera-se que o modelo proposto apresente resultados superiores ou equivalentes aos modelos que usam somente léxicos, para o reconhecimento de uma ou todas as dimensões de personalidade do CGF.

Considerando-se entretanto que há diferentes representações distribuídas ideais para diferentes problemas (MIKOLOV et al., 2013a), realizamos diversos experimentos com base nas estratégias propostas. A partir de diferentes combinações de parâmetros, atributos e modelos, enunciamos seis hipóteses mais específicas com o objetivo de identificar a melhor estratégia para o problema proposto como segue:

- **H1** - A arquitetura *skip-gram* proporciona melhores resultados de reconhecimento de personalidade do que a arquitetura CBOW. Está hipótese será verificada comparando-

---

¹ Uma destas alternativas, baseada exclusivamente no dicionário LIWC, seria semelhante ao modelo já implementado e discutido no capítulo 4

se o valor médio de medida F das estratégias que utilizam a arquitetura *skip-gram* (estratégias E7 a E10) com as estratégias que utilizam a arquitetura CBOW (estratégias E3 a E6). Espera-se que as estratégias E3 a E6 apresentem maior resultado médio do que as estratégias E7 a E10.

- **H2** - O uso de um vetor de tamanho 600 proporciona melhores resultados de reconhecimento de personalidade do que a um vetor de tamanho 300. Essa hipótese pode ser verificada comparando-se o valor médio de medida F das estratégias E3-cbow-300 e E4-cbow-600 com as estratégias E7-skip-300 e E8-skip-600. Espera-se que as estratégias E4-cbow-600, E8-skip-600 apresentem maior resultado médio do que as estratégias E3-cbow-300, E7-skip-300.
- **H3** - O texto que não foi pré-processado proporciona melhores resultados de reconhecimento de personalidade do que o texto que foi pré-processado. Essa hipótese pode ser verificada comparando-se o valor médio de medida F das estratégias E4-cbow-600, E8-skip-600 com as estratégias E5-cbow-600-norm, E10-skip-600-norm. Espera-se as estratégias E5-cbow-600-norm e E10-skip-600-norm apresentem maior resultado médio do que as estratégias E4-cbow-600 e E8-skip-600.
- **H4** - A remoção de *stopwords* proporciona melhores resultados de reconhecimento de personalidade do que estratégias que consideram todas as palavras do texto. Essa hipótese pode ser verificada comparando-se o valor médio de medida F das estratégias E4-cbow-600, E8-skip-600 com as estratégias E6-cbow-600-stop, E9-skip-600-stop. Espera-se que as estratégias E6-cbow-600-stop e E9-skip-600-stop apresentem maior resultado médio do que as estratégias E4-cbow-600, E8-skip-600.
- **H5** - A representação vetorial de documentos proporciona melhores resultados de reconhecimento de personalidade do que a representação vetorial de palavras. Essa hipótese pode ser verificada comparando-se o experimento E11-doc2vec com os experimentos E3 a E10. Espera-se que a estratégia E11-doc2vec apresente maior resultado médio do que as estratégias E3 a E10.
- **H6** - Utilizar o conjunto de dados sem nenhuma estratégia de combinação trará melhores resultados para a LSTMs. Essa hipótese poderá ser verificada comparando-se as estratégias E12-LSTM-600 com as estratégias E13-LSTM-skip-100 e E14-LSTM-skip-150-media. Espera-se que a estratégia E12-LSTM-600 apresente maior resultado médio do que as estratégias E13 e E14.

6.2 Descrição dos dados

Os dados utilizados para a tarefa de reconhecimento de personalidade, tanto para classificação, quanto para regressão, são os mesmos apresentados na seção 4.2. Contudo, nesta etapa da pesquisa, dos 1.039 participantes apenas os 11 participantes que não tinham nenhuma postagem no *Facebook* foram descartados, restando 1.028 participantes com textos. A média de palavras deste conjunto de dados é 2.624, o número máximo de palavras geral é 26.537 e o mínimo 20.

Na seção 4.3 vemos que a divisão em extremos exclui mais da metade do conjunto de dados atuais, e portanto optamos por usar a divisão ao meio para todos os experimentos de classificação binária descritos daqui para frente.

6.3 Resultados

A tabela 5 apresenta os resultados médios de medida F das catorze estratégias propostas, obtidos com a partir da validação cruzada de dez partes.

Tabela 5 – Medida F dos experimentos 1 à 14 para as cinco dimensões de personalidade

	EXT	AGR	MET	INS	LIB
<i>baseline</i>	0.33	0.34	0.33	0.32	0.34
E1-BoW	0.60	0.53	0.57	0.53	0.55
E2-LIWC+psico	0.61	0.53	0.57	0.54	0.53
E3-cbow-300	0.60	0.55	0.58	0.54	0.58
E4-cbow-600	0.61	0.55	0.58	0.55	0.55
E5-cbow-600-norm	0.61	0.56	0.58	0.55	0.59
E6-cbow-600-stop	0.61	0.56	0.57	0.55	0.56
E7-skip-300	0.61	0.55	0.57	0.54	0.57
E8-skip-600	0.60	0.54	0.57	0.54	0.58
E9-skip-600-stop	0.61	0.55	0.58	0.54	0.57
E10-skip-600-norm	0.61	0.55	0.59	0.55	0.57
E11-doc2vec	0.60	0.58	0.57	0.55	0.55
E12-LSTM-600	0.58	0.53	0.45	0.54	0.59
E13-LSTM-skip-100	0.50	0.50	0.50	0.42	0.46
E14-LSTM-skip-150-media	0.45	0.47	0.45	0.45	0.45

Fonte: Bárbara Barbosa Claudino Silva, 2018

Com relação à hipótese H1 - de que arquiteturas CBOW apresentam resultados superiores arquiteturas skip-gram - observamos que os modelos com arquiteturas CBOW

possuem medidas F superiores aos modelos com arquiteturas skip-gram. Esta observação oferece suporte à hipótese H1.

Para a arquitetura CBOW obtém-se as melhores medidas F para extroversão, instabilidade e liberalidade, enquanto a arquitetura skip-gram obteve os melhores resultados para duas dimensões: extroversão e meticulosidade.

Com relação à hipótese H2 - de que arquiteturas de tamanho 600 apresentam resultados superiores a arquiteturas de tamanho 300 - observamos que os modelos com arquiteturas de tamanho 600 possuem medidas F superiores aos modelos de tamanho 300. Esta observação oferece suporte à hipótese H2.

Pode-se ver na tabela 5 que as estratégias E3-cbow-300, E4-cbow-600, E7-skip-300 e E8-skip-600 possuem resultados bem parecidos para todas as dimensões de personalidade e a média geral dos resultados das duas estratégias considerando-se todas as dimensões de personalidade é a mesma: 0,56.

Considerando-se a hipótese H3 - de que um corpus que não teve quase nenhum pré-processamento apresenta resultados superiores a um corpus pré-processado - observamos que o modelo E5-cbow-600-norm possui medida F superior ao modelo E4-cbow-600. Esta observação oferece suporte à hipótese H3.

Observa-se que a estratégia E5-cbow-600-norm obtém um dos melhores resultados para extroversão, instabilidade e liberalidade. Considerando-se apenas representação distribuída de palavras, esse experimento obtém o melhor resultado para liberalidade. Já na arquitetura skip-gram, (experimento E10-skip-600-norm) nota-se que os resultados para o texto sem pré-processamento também são um dos melhores para extroversão e a estratégia obtém o melhor resultado para meticulosidade.

Além dos resultados positivos em diferentes dimensões de personalidade, considerando-se o resultado médio para todas as dimensões de personalidade o experimento E5-cbow-600-norm obtém a melhor média global de personalidades (0,578), mostrando a importância desse processo para o reconhecimento da dimensão liberalidade, que é o resultado que apresenta maior diferença entre a estratégia E5-cbow-600-norm e a estratégia E4-cbow-600.

Com relação à hipótese H4 - de que a remoção de *stopwords* apresenta resultados superiores ao invés de considerar todas as palavras do texto - observamos que os modelos E6-cbow-600-stop e E9-skip-600-stop possuem medida F superior aos modelos E4-cbow-600 e E8-skip-600. Esta observação oferece suporte à hipótese H4.

Percebe-se que a estratégia E6-cbow-600-stop tem um dos melhores resultados para as dimensões extroversão e instabilidade. A diferença da média das dimensões de personalidade entre as estratégias é pequena, sendo a média geral de 0,568 para E4-cbow-600 e 0,570 para E6-cbow-600-stop. Obtemos resultados semelhantes quando consideramos a arquitetura skip-gram. A estratégia E8-skip-600 tem média geral de 0,566 contra 0,568 para E9-skip-600-stop. Notamos que a remoção de *stopwords* fez uma diferença pequena, mas ainda assim, obteve melhor desempenho que o experimento que utiliza todas as palavras.

Com relação à hipótese H5 - que considera que a representação distribuída de documentos apresenta resultados superiores às demais - observamos que o modelo E5-cbow-600-norm possui medida F superior ao modelo E11-doc-2vec. Esta observação não oferece suporte à hipótese H5.

Como não há uma só estratégia que obteve o melhor resultado para todas as dimensões de personalidade utilizaremos o resultado médio de todas as dimensões de personalidade para considerar uma estratégia vencedora. A estratégia que atende esses requisitos é a estratégia E5-cbow-600-norm, que obteve média de 0,578. A média do experimento E11-doc-2vec é 0,570, um dos melhores resultados de média junto com os experimentos: E6-cbow-600-stop e E10-skip-600-norm, que obtiveram as médias 0,570 e 0,574 respectivamente.

Apesar da estratégia E11-doc-2vec obter o melhor resultado para a dimensão agradabilidade, diferentemente do experimento E5-cbow-600-norm, ela só possui dois dos melhores resultados (agradabilidade e instabilidade) sendo que o experimento E5-cbow-600-norm possui três dos melhores resultados (extroversão, instabilidade e liberalidade).

Com relação à hipótese H6 - de que a rede LSTM sem nenhuma estratégia de combinação apresenta resultados superiores as redes com estratégias de combinação ou utilizando um conjunto de dados externo - observamos que o modelo E12-LSTM-600 possui medida F superior aos modelos E13-LSTM-skip-100 e E14-LSTM-skip-150. Esta observação oferece suporte à hipótese H6.

Quanto à hipótese H6 que trata da mudança do algoritmo de classificação, utilizando uma rede neural recorrente do tipo LSTM, vemos que a melhor estratégia é a E12-LSTM-600, que obtém média geral de 0,53. As outras estratégias E13 e E14 têm médias gerais de 0,47 e 0,45, respectivamente. O experimento E12-LSTM-600 também obtém um dos

melhores resultados para a dimensão liberalidade, apesar dos resultados mais baixos para as outras dimensões.

Vemos que os resultados da LSTM são inferiores aos apresentados nos experimentos E1 à E11. Uma das possibilidades do baixo desempenho das redes LSTM é devido à configuração de hiper-parâmetros da rede. Poderíamos ter utilizado estratégias de *GridSearch* a fim de encontrar a melhor configuração para a rede. Apesar deste ponto, nota-se também que em Majumder et al. (2017) as redes neurais CNNs sozinhas não obtiveram melhor desempenho do que a estratégia que utiliza léxicos e apenas a combinação da representação distribuída de palavras com o léxico LIWC obteve um desempenho superior ao apresentado em Mairesse (2008).

Apesar dos resultados de medida F ficarem parecidos em várias das estratégias apresentadas, notamos pela tabela 5 que não existe uma única estratégia que seja a melhor para todas as dimensões de personalidade considerando-se a classificação binária.

Considerando apenas a dimensão extroversão podemos notar que não há grande diferença entre modelos CBOW e skip-gram, já que os resultados em todos os experimentos são semelhantes.

Considerando-se os resultados para a dimensão agradabilidade, a estratégia que obteve o melhor desempenho foi a que utiliza representação distribuída de documentos. É interessante notar que diferentemente do que acontece com a dimensão extroversão, onde várias estratégias possuem os melhores resultados, nesta dimensão só essa estratégia foi a campeã. A diferença da estratégia vencedora foi um pouco mais relevante do que a vista para a extroversão, chegando a 0,05 quando comparada com a estratégia E2-LIWC+psico.

Os resultados para meticulosidade apontam que a melhor estratégia é a E10-skip-600-norm, uma arquitetura skip-gram com os dados originais com pouco pré-processamento. A semelhança do que acontece com a dimensão extroversão, a diferença entre os experimentos não é muito alta, sendo de 0,04 quando consideramos apenas experimentos com representação distribuída de palavras.

Já para instabilidade as melhores estratégias são E4-cbow-600, E5-cbow-600-norm e E6-cbow-600-stop, experimentos com a arquitetura CBOW e a estratégia E11-doc2vec, que utiliza representação distribuída de documentos.

Assim como a dimensão instabilidade, a dimensão liberalidade também obtém o melhor resultado para o conjunto de dados que teve pouco pré-processamento, também utilizando a arquitetura CBOW. Semelhante com o que aconteceu com meticulosidade, os

resultados com a representação distribuída de palavras foram um pouco melhores do que os obtidos pelo dicionário LIWC+psico ou BoW.

A exceção da dimensão extroversão, onde os resultados foram iguais com o uso de léxicos (E2-LIWC+psico) ou de outras estratégias de representação distribuída de palavras, as demais dimensões apresentam seus melhores resultados utilizando técnicas mais recentes, como a representação distribuída de palavras ou de documentos. A estratégia que utiliza *word embeddings* também possui outra vantagem: diferente do dicionário LIWC ela não depende que exista um dicionário pronto para a língua do estudo em questão, possibilitando que línguas que não possuam esse dicionário também se beneficiem de modelos computacionais de reconhecimento de personalidade.

Apesar das estratégias vencedoras à primeira vista terem um resultado insatisfatório, comparando-se com o *baseline* de classe majoritária vemos que as estratégias vencedoras para cada uma das dimensões de personalidade conseguem quase o dobro de melhora em relação a sempre utilizar a classe majoritária para uma determinada personalidade.

6.4 *Análise de atributos relevantes para reconhecimento de personalidade*

Como a estratégia E2-LIWC+psico combina atributos que são interpretáveis, ou seja, categorias LIWC e atributos psicolinguísticos, realizamos uma análise para verificar a importância de cada um dos atributos para cada uma das dimensões de personalidade. Os atributos psicolinguísticos são destacados em *itálico* e os resultados observados foram os seguintes:

- **Extroversão** - emoções positivas, palavras de amizade, palavras de afeto, palavras de família, palavras relacionadas a humanos.
- **Agradabilidade** - emoções positivas, palavras de conquista, palavras de amizade, palavras relacionadas a humanos, palavras de negação.
- **Meticulosidade** - emoções positivas, palavras de raiva, palavras relativas a dinheiro, palavras relativas a saúde, palavras de afeto.
- **Instabilidade** - primeira pessoa do singular, palavras de percepção, palavras de amizade, palavras no presente, palavras causais.
- **Liberalidade** - *idade de aquisição*, palavras relativas a dinheiro, *frequência subjetiva das palavras*, palavras sobre movimento, palavras no passado.

Nota-se que para a dimensão liberalidade as palavras mais importantes fazem sentido com a descrição dessa personalidade, como pode ser visto no quadro 1.

Um estudo similar a este foi realizado para a língua inglesa em Mairesse (2008). Assim como ocorreu no estudo, aqui também são poucos os atributos psicolinguísticos que aparecem como atributos mais importantes para os modelos.

6.5 Outras análises

Além do teste das seis hipóteses principais discutido na seção anterior, nesta seção descrevemos brevemente uma série de testes complementares que foram realizados ao longo do desenvolvimento deste trabalho, e algumas definições alternativas do problema de reconhecimento de personalidade a partir de textos. Na seção 6.5.1, apresentamos uma série de variações das estratégias discutidas nos capítulos anteriores. Na seção 6.5.2, o problema de reconhecimento de personalidade é modelado com base em postagens individuais, e não coleções completas de textos (ou *timelines*) de cada autor. Na seção 6.5.3, é considerada também a classificação ternária de personalidade, ou seja, baseada em indivíduos com valores baixo, alto e intermediário para as dimensões CGF. Finalmente, na seção 6.5.4, o reconhecimento de personalidade é implementado como um problema de regressão.

6.5.1 Estratégias complementares

Nesta seção apresentamos cinco variações das estratégias apresentadas e seus resultados. Em primeiro lugar, como base de comparação, utilizou-se uma estratégia que utiliza unicamente o dicionário LIWC. Os dados utilizados passaram pelo mesmo pré-processamento apresentado no capítulo do estudo, contudo os resultados são um pouco diferentes dos apresentados na seção 6.3. Essa mudança é esperada, já que usamos um algoritmo de classificação diferente e temos mais 250 indivíduos do que no experimento anterior, pois optamos por não excluir nenhum indivíduo com postagens da pesquisa. Esta estratégia será denominada **E15-LIWC**.

Além das estratégias apresentadas na seção 5.1.3, aqui apresentamos mais quatro estratégias de representação distribuída de palavras, duas para a arquitetura CBOW e

duas para a arquitetura skip-gram. As configurações para a criação dos modelos são as mesmas apresentadas no capítulo 5.

- **E16-cbow-600-soma**: Arquitetura CBOW. Vetor de tamanho 600 que foi treinado em um corpus de 50.000 *tweets*. Para combinar as palavras utilizou-se a soma vetorial de cada palavra que compunha o conjunto de *posts* de um sujeito.
- **E17-cbow-600-tfidf**: Arquitetura CBOW. Vetor de tamanho 600 que foi treinado em um corpus de 50.000 *tweets*. Para combinar as palavras utilizou-se o *tf-idf* do corpus de treino para dar um peso diferente para cada palavra, multiplicando-se cada vetor de palavra pela seu *tf-idf* e depois dividindo esse vetor pelo número de palavras em um conjunto de *posts* de um indivíduo.
- **E18-skip-600-soma**: Arquitetura skip-gram. Tamanho 600 que foi treinado em um corpus de 50.000 *tweets*. Para combinar as palavras utilizou-se a soma vetorial de cada palavra que compunha o conjunto de *posts* de um sujeito.
- **E19-skip-600-tfidf**: Arquitetura skip-gram. Tamanho 600 que foi treinado em um corpus de 50.000 *tweets*. Para combinar as palavras utilizou-se o *tf-idf* do corpus de treino para dar um peso diferente para cada palavra, multiplicando-se cada vetor de palavra pela seu *tf-idf* e depois dividindo esse vetor pelo número de palavras em um conjunto de *posts* de um indivíduo.

A tabela 6 apresenta os resultados das estratégias complementares 15 a 19.

Tabela 6 – Medida F média das estratégias complementares para classificação

	EXT	AGR	MET	INST	LIB
E15-LIWC	0.61	0.53	0.57	0.53	0.55
E16-cbow-600-soma	0.60	0.53	0.55	0.54	0.53
E17-cbow-600-tfidf	0.58	0.52	0.55	0.46	0.52
E18-skip-600-soma	0.59	0.53	0.56	0.53	0.53
E19-skip-600-tfidf	0.59	0.52	0.54	0.46	0.52

Fonte: Bárbara Barbosa Claudino Silva, 2018

Percebe-se que levando-se em conta apenas os experimentos de representação distribuída de palavras os experimentos com tfidf têm o pior desempenho médio geral, além de não obterem nenhum dos melhores resultados individualmente. O resultado médio geral de todas as personalidades das estratégias E15-cbow-600-tfidf e E17-skip-600-tfidf é 0,52.

A estratégia de soma também obtém resultados mais baixos do que as estratégias apresentadas no capítulo 5 apresentando média geral de 0,53 para E14-cbow-600-soma e 0,54 para E16-skip-600-soma. Como essas estratégias também não apresentaram nenhum dos melhores resultados individuais elas passaram a configurar entre as estratégias complementares.

O objetivo destes experimentos foi descobrir qual a melhor forma de combinação da representação vetorial de palavras, apesar da pequena diferença percebemos que a melhor configuração é a que utiliza a média, em ambas as arquiteturas, CBOW e skip-gram. Como apontado na seção 6.3 o resultado médio das duas arquiteturas é similar e as maiores diferenças estão nas variações de pré-processamento.

6.5.2 Estratégias baseadas em classificação individual de postagens

Todas as estratégias apresentadas até esta seção consideram todas as postagens de um indivíduo como um único documento e todo o seu texto (independentemente de tempo) é utilizado para a tarefa de reconhecimento de personalidade. Apesar da personalidade ser única para cada um, queríamos verificar se à partir de uma única postagem é possível identificar a personalidade de uma pessoa.

Para isso replicamos a resposta dada por um participante da pesquisa para cada postagem feita por ele. Essa estratégia fez com que o número de documentos fosse de 1.039 para 238.970 o que possibilitou apenas testes com LIWC e *BoW*, já que para o uso da representação distribuída seria necessária muita memória computacional.

Tabela 7 – Medida F das estratégias de classificação ternária

	EXT	AGR	MET	INST	LIB
<i>baseline</i>	0.32	0.34	0.33	0.32	0.34
E20-single-LIWC	0.60	0.53	0.57	0.54	0.53
E21-single-BoW	0.48	0.50	0.49	0.47	0.49

Fonte: Bárbara Barbosa Claudino Silva, 2018

Nota-se que a melhor estratégia desta seção é a que utiliza o dicionário LIWC. Podemos perceber que os resultados com o uso do dicionário LIWC não são muito diferentes dos que utilizam LIWC considerando todos as postagens (tabela 6) . Contudo, os resultados entre postagens individuais e completas não ficaram tão semelhantes com o uso de BoW, que obtém resultados baixos para a dimensão extroversão.

6.5.3 Estratégia baseada em classificação ternária

A divisão em extremos apresentada no capítulo 4 possibilita a criação da tarefa de classificação ternária, como utilizada em Oberlander e Nowson (2006). Nesta estratégia a classe intermediária, ao invés de ser descartada, entra como terceira classe do conjunto de dados classificando indivíduos como intermediários para uma determinada dimensão.

Utilizando-se da mesma configuração de *Random Forest* apresentada no capítulo 5, fez-se necessário o uso de uma técnica para tratar conjuntos de dados desbalanceados, já que essa divisão possui muito mais indivíduos na classe intermediária. Para esse fim, utilizou-se a técnica que dá maior peso para as classes que aparecem menos no conjunto de dados, de forma que estas tenham maior importância para o classificador.

A tabela 8 apresenta os resultados da classificação ternária para o dicionário LIWC, BoW e CBOW de tamanho 600 (configurado da mesma maneira que o experimento E4-cbow-600).

Tabela 8 – Medida F das estratégias de classificação ternária

	EXT	AGR	MET	INST	LIB
<i>baseline</i>	0.25	0.26	0.27	0.26	0.26
LIWC	0.29	0.23	0.27	0.31	0.29
BoW	0.32	0.34	0.39	0.34	0.29
CBOW-600	0.30	0.26	0.28	0.30	0.27

Fonte: Bárbara Barbosa Claudino Silva, 2018

A partir dos resultados da tabela vemos que diferentemente do que aconteceu com a classificação binária a estratégia que utiliza BoW obteve o melhor desempenho e os melhores resultados estão nas dimensões meticulosidade e instabilidade. Percebemos que para a classificação ternária o desempenho da estratégia vencedora não é tão superior em comparação com o *baseline* da classe majoritária, aqui sendo sempre a classe intermediária.

Também nota-se que diferentemente do que acontece com a classificação binária, existe uma estratégia que é a melhor para todas as dimensões de personalidade, contudo esse fato também pode ser explicado devido ao número limitado de experimentos para esse estudo.

6.5.4 Estratégias baseadas em regressão

Apesar do problema de reconhecimento de personalidade ter sido apresentado como um problema de classificação no capítulo 4, modelos de regressão ou ranking têm um desempenho melhor na tarefa de reconhecimento de personalidade (MAIRESSE, 2008), já que não é necessário fazer nenhuma normalização na variável original, que varia de 1 a 5 e possui uma ordenação própria. Sendo assim, nesta seção optamos por tratar o problema de reconhecimento de personalidade como um problema de regressão estatística.

Considerando-se o problema de reconhecimento de personalidade como uma tarefa de regressão utilizamos também uma *Random Forest*, com o número máximo de atributos definido pelo $\log_2$ dos atributos utilizados. Para essa tarefa considerou-se três estratégias: dicionário LIWC, BoW e uma representação distribuída de palavras como apresentada em E5-cbow-600.

Tabela 9 – Erro quadrático médio das estratégias de regressão

	EXT	AGR	MET	INST	LIB
LIWC	0.66	0.29	0.42	0.67	0.37
BoW	0.65	0.30	0.42	0.68	0.36
CBOW-600	0.67	0.44	0.52	0.68	0.48

Fonte: Bárbara Barbosa Claudino Silva, 2018

Ao contrário do que acontece com a medida F, quanto menor o erro quadrático médio, melhor. Também diferentemente dos experimentos de classificação binária, na regressão os melhores resultados estão na estratégia que utiliza o dicionário LIWC e verifica-se que as dimensões extroversão e instabilidade obtêm os piores resultados, enquanto agradabilidade obtém o melhor resultado.

Neste experimento também vemos que a estratégia que usa LIWC, considerando-se a média de todas as dimensões, obtém o melhor resultado. Essa unanimidade em escolher um modelo para reconhecimento de personalidade nas estratégias que não a binária possivelmente deve-se ao número reduzido de experimentos apresentados.

6.6 Considerações

Neste capítulo foram apresentadas a avaliação das estratégias principais apresentadas no capítulo 5 e 13 experimentos secundários, que com os 14 experimentos apresentados

no capítulo anterior totalizam 27 experimentos para a tarefa de reconhecimento de personalidade.

As hipóteses apresentadas neste capítulo são apresentados em resumo no quadro 4.

Quadro 4 – Resumo das hipóteses referentes as estratégias 1 à 14

Hipótese	Melhor estratégia	Resultado
H1 - Arquitetura CBOW apresentam melhores resultados que a arquitetura skip-gram	E3 a E6	Confirmada
H2 - Arquiteturas de tamanho 600 apresentam melhores resultados que a arquiteturas de tamanho 300	E4 e E8	Confirmada
H3 - Textos não processados apresentam melhores resultados do que textos processados	E5	Confirmada
H4 - Remoção de stopwords traz melhores resultados do que o texto completo	E4 e E8	Confirmada
H5 - Representação vetorial de documentos traz melhores resultados do que representação vetorial de palavras	E5	Não confirmada
H6 - rede LSTM sem nenhuma estratégia de combinação apresenta melhores resultados do que redes que utilizam alguma estratégia combinação	E12	Confirmada

Fonte: Bárbara Barbosa Claudino Silva, 2018

O quadro 5 apresenta os resumo dos resultados para as estratégias complementares.

Quadro 5 – Resumo dos melhores resultados para as estratégias complementares

Estratégia complementar	Melhor estratégia
Classificação individual de postagens	LIWC
Classificação ternária	BoW
Regressão	LIWC

Fonte: Bárbara Barbosa Claudino Silva, 2018

Os resultados apresentados neste capítulo mostram o quanto o léxico LIWC é importante para a tarefa de reconhecimento de personalidade, mas também mostra que as estratégias que utilizaram representação distribuída de palavras, especialmente a estratégia E5-cbow-600-norm, obtém resultados superiores que o léxico para a classificação binária de personalidade.

7 Conclusões

Este documento apresentou uma pesquisa em nível de mestrado com o objetivo de reconhecer a personalidade com base em textos do português brasileiro. A pesquisa consistiu em construir um *cópus* de textos da rede social *Facebook*, rotulados com a personalidade de seus autores, identificar as características mais relevantes para o reconhecimento de personalidade e implementar uma série de modelos computacionais utilizando essas características.

As principais contribuições deste estudo foram a elaboração do *cópus*, que possivelmente é o maior do gênero já desenvolvido para o português brasileiro, e o conjunto de estratégias computacionais implementadas. Na avaliação destas estratégias, comparou-se diversos modelos para a tarefa de reconhecimento de personalidade, que utilizaram conhecimento de léxicos (dicionário LIWC e atributos psicolinguísticos), unigramas, representação distribuída de palavras e representação distribuída de documentos para verificar se os modelos onde utiliza-se representação distribuída de palavras possuíam um desempenho melhor que os métodos que utilizam conhecimento lexical.

Os resultados finais mostram uma diferença pequena entre o uso da representação distribuída de palavras e o uso de léxicos. Todavia, os modelos baseados em representação distribuída de palavras possuem os resultados ligeiramente superiores, com a vantagem de serem independentes da disponibilidade de um dicionário específico para cada língua.

Uma possível melhora para o desempenho da representação distribuída de palavras pode ser alcançada utilizando redes neurais profundas, como a rede LSTM utilizada no presente trabalho. Apesar dos resultados inferiores apresentados pelas estratégias que utilizaram reconhecimento de personalidade, melhores resultados poderiam ser obtidos com uma melhor configuração de hiperparâmetros da rede ou com uma abordagem mista, que combine o uso de léxicos com representação distribuída de palavras.

Resultados parciais deste trabalho foram apresentados nos seguintes artigos:

- Uma versão inicial do *cópus* b5 foi utilizada em um experimento de reconhecimento de personalidade a partir de diferentes fontes textuais em Santos, Paraboni e Silva (2017).
- O *cópus* b5 em sua versão completa, incluindo componentes desenvolvidos em projetos correlatos, foi apresentado em Ramos et al. (2018).

- O experimento-piloto descrito no capítulo 4 foi apresentado em Silva e Paraboni (2018).

Além disso, o córpus b5 foi recentemente reutilizado em um projeto de pesquisa correlato enfocando o problema de caracterização autoral a partir de publicações da rede social *Facebook* (HSIEH; DIAS; PARABONI, 2018).

Os resultados principais desta pesquisa, detalhados nos capítulos 5 e 6, serão objeto de futuras publicações. O córpus b5 ¹ e os modelos implementados ² são disponibilizados para reúso para fins de pesquisa.

---

¹ <https://drive.google.com/file/d/0B-KyU7T8S8bLTHpaMnh2U2NWZzQ/view>

² <https://github.com/bahbbc/masters-big5>

Referências³

- ALLPORT, G. W.; ODBERT, H. S. Trait-names: A psycho-lexical study. *Psychological monographs*, Psychological Review Company, v. 47, n. 1, p. i, 1936. Citado 2 vezes nas páginas 15 e 19.
- ALTMAN, N. S. An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, Taylor & Francis, v. 46, n. 3, p. 175–185, 1992. Citado na página 34.
- ÁLVAREZ-CARMONA, M. et al. INAOE's participation at PAN'15: Author Profiling task—Notebook for PAN at CLEF 2015. In: CAPPELLATO, L. et al. (Ed.). *CLEF 2015 Evaluation Labs and Workshop – Working Notes Papers, 8-11 September, Toulouse, France*. [S.l.]: CEUR-WS.org, 2015. ISSN 1613-0073. Citado 2 vezes nas páginas 54 e 55.
- ANDRADE, J. M. d. *Evidências de validade do inventário dos cinco grandes fatores de personalidade para o Brasil*. Tese (Doutorado), 2008. Citado 4 vezes nas páginas 20, 22, 58 e 66.
- ARGAMON, S. et al. Lexical predictors of personality type. In: *Proceedings of the 2005 Joint Annual Meeting of the Interface and the Classification Society of North America*. [S.l.: s.n.], 2005. Citado 6 vezes nas páginas 15, 32, 33, 56, 61 e 65.
- BAI, B. et al. Supervised semantic indexing. In: *ACM. Proceedings of the 18th ACM conference on Information and knowledge management*. [S.l.], 2009. p. 187–196. Citado na página 24.
- BENET-MARTINEZ, V.; JOHN, O. P. Los cinco grandes across cultures and ethnic groups: Multitrait-multimethod analyses of the big five in spanish and english. *Journal of personality and social psychology*, American Psychological Association, v. 75, n. 3, p. 729, 1998. Citado na página 22.
- BRIGGS, I. B. M. *The myers-briggs type indicator*. [S.l.]: Consulting Psychologists Press Palo Alto, CA, 1962. Citado na página 50.
- CAPRARA, G. V.; PERUGINI, M. Personality described by adjectives: The generalizability of the Big Five to the italian lexical context. *European Journal of Personality*, Wiley Online Library, v. 8, n. 5, p. 357–369, 1994. Citado na página 21.
- CASIMIRO, C. R.; PARABONI, I. Temporal aspects of content recommendation on a microblog corpus. In: *Computational Processing of the Portuguese Language (PROPOR-2014), Lecture Notes in Computer Science 8775*. São Carlos, Brazil: Springer, 2014. p. 189–194. Citado na página 74.
- CATTELL, R. B. The description of personality: basic traits resolved into clusters. *The journal of abnormal and social psychology*, American psychological association, v. 38, n. 4, p. 476, 1943. Citado 2 vezes nas páginas 19 e 20.
- CELLI, F. *Adaptive Personality Recognition from Text*. Tese (Doutorado) — University of Trento, 2012. Citado 13 vezes nas páginas 15, 16, 23, 46, 47, 48, 50, 55, 56, 61, 65, 70 e 73.

³ De acordo com a Associação Brasileira de Normas Técnicas. NBR 6023.

- CHAPELLE, C. A. *What is Systemic Functional Linguistics*. 2006. Citado na página 32.
- CHUNG, J. et al. Gated feedback recurrent neural networks. In: *International Conference on Machine Learning*. [S.l.: s.n.], 2015. p. 2067–2075. Citado na página 29.
- COHEN, W. W. Fast effective rule induction. In: *Proceedings of the twelfth international conference on machine learning*. [S.l.: s.n.], 1995. p. 115–123. Citado na página 34.
- COLTHEART, M. The MRC psycholinguistic database. *The Quarterly Journal of Experimental Psychology*, Taylor & Francis, v. 33, n. 4, p. 497–505, 1981. Citado 4 vezes nas páginas 16, 23, 34 e 46.
- CORTES, C.; VAPNIK, V. Support-vector networks. *Machine learning*, Springer, v. 20, n. 3, p. 273–297, 1995. Citado na página 33.
- COSTA, P. T.; MACCRAE, R. R. *Revised NEO personality inventory (NEO PI-R) and NEO five-factor inventory (NEO FFI): Professional manual*. [S.l.]: Psychological Assessment Resources, 1992. Citado 3 vezes nas páginas 21, 32 e 39.
- COSTA, P. T.; MCCRAE, R. R. Four ways five factors are basic. *Personality and individual differences*, Elsevier, v. 13, n. 6, p. 653–665, 1992. Citado na página 21.
- DAHL, G. E. et al. Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. *IEEE Transactions on audio, speech, and language processing*, IEEE, v. 20, n. 1, p. 30–42, 2012. Citado na página 28.
- DONAHUE, J. et al. Long-term recurrent convolutional networks for visual recognition and description. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. [S.l.: s.n.], 2015. p. 2625–2634. Citado na página 28.
- FARNADI, G. et al. Recognising personality traits using facebook status updates. In: AAAI. *Proceedings of the workshop on computational personality recognition (WCPR13) at the 7th international AAAI conference on weblogs and social media (ICWSM13)*. [S.l.], 2013. Citado 8 vezes nas páginas 15, 16, 23, 48, 49, 50, 56 e 73.
- 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 3 vezes nas páginas 23, 65 e 73.
- FREUND, Y. et al. An efficient boosting algorithm for combining preferences. *Journal of machine learning research*, v. 4, n. Nov, p. 933–969, 2003. Citado na página 35.
- FREUND, Y.; SCHAPIRE, R. E. A decision-theoretic generalization of on-line learning and an application to boosting. In: SPRINGER. *European conference on computational learning theory*. [S.l.], 1995. p. 23–37. Citado na página 34.
- GERS, F. A.; SCHMIDHUBER, J.; CUMMINS, F. Learning to forget: Continual prediction with LSTM. *Neural Computation*, 12(10):2451–2471, IET, 1999. Citado na página 29.
- GILL, A. J.; NOWSON, S.; OBERLANDER, J. What are they blogging about? personality, topic and motivation in blogs. In: *ICWSM*. [S.l.: s.n.], 2009. Citado 9 vezes nas páginas 15, 16, 23, 33, 42, 43, 56, 61 e 73.

GOLDBERG, L. R. An alternative description of personality: the Big Five factor structure. *Journal of personality and social psychology*, American Psychological Association, v. 59, n. 6, p. 1216, 1990. Citado 2 vezes nas páginas 18 e 20.

GONZÁLEZ-GALLARDO, C. et al. Tweets Classification Using Corpus Dependent Tags, Character and POS N-grams—Notebook for PAN at CLEF 2015. In: CAPPELLATO, L. et al. (Ed.). *CLEF 2015 Evaluation Labs and Workshop – Working Notes Papers, 8-11 September, Toulouse, France*. [S.l.]: CEUR-WS.org, 2015. ISSN 1613-0073. Citado na página 54.

GRAVES, A.; MOHAMED, A.-r.; HINTON, G. Speech recognition with deep recurrent neural networks. In: IEEE. *Acoustics, speech and signal processing (icassp), 2013 ieee international conference on*. [S.l.], 2013. p. 6645–6649. Citado na página 28.

GRIVAS, A.; KRITHARA, A.; GIANNAKOPOULOS, G. Author Profiling Using Stylometric and Structural Feature Groupings—Notebook for PAN at CLEF 2015. In: CAPPELLATO, L. et al. (Ed.). *CLEF 2015 Evaluation Labs and Workshop – Working Notes Papers, 8-11 September, Toulouse, France*. [S.l.]: CEUR-WS.org, 2015. ISSN 1613-0073. Citado na página 55.

HARTMANN, N. et al. Portuguese word embeddings: Evaluating on word analogies and natural language tasks. *Symposium in Information and Human Language Technology*, 2017. Citado na página 24.

HOCHREITER, S.; SCHMIDHUBER, J. Long short-term memory. *Neural computation*, MIT Press, v. 9, n. 8, p. 1735–1780, 1997. Citado 2 vezes nas páginas 28 e 29.

HSIEH, F. C.; DIAS, R. F. S.; PARABONI, I. Author profiling from facebook corpora. In: *11th International Conference on Language Resources and Evaluation (LREC-2018) (to appear)*. Miyasaki, Japan: ELRA, 2018. Citado na página 92.

IACOBELLI, F. et al. Large scale personality classification of bloggers. In: *Affective Computing and Intelligent Interaction*. [S.l.]: Springer, 2011. p. 568–577. Citado 10 vezes nas páginas 15, 16, 23, 44, 45, 46, 55, 56, 61 e 73.

JOHN, O. P.; DONAHUE, E.; KENTLE, R. ‘The Big Five’. *inventory—version 4a and*, v. 54, 1991. Citado na página 22.

JOHN, O. P.; NAUMANN, L. P.; SOTO, C. J. Paradigm shift to the integrative big five trait taxonomy. *Handbook of personality: Theory and research*, v. 3, p. 114–158, 2008. Citado na página 22.

JOHN, O. P.; SRIVASTAVA, S. The Big Five trait taxonomy: History, measurement, and theoretical perspectives. *Handbook of personality: Theory and research*, Guilford, v. 2, n. 1999, p. 102–138, 1999. Citado 7 vezes nas páginas 15, 18, 19, 20, 21, 22 e 71.

KRIZHEVSKY, A.; SUTSKEVER, I.; HINTON, G. E. Imagenet classification with deep convolutional neural networks. In: *Advances in neural information processing systems*. [S.l.: s.n.], 2012. p. 1097–1105. Citado na página 28.

LANDAUER, T. K.; DUMAIS, S. T. A solution to plato’s problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, American Psychological Association, v. 104, n. 2, p. 211, 1997. Citado na página 54.

- LE, Q.; MIKOLOV, T. Distributed representations of sentences and documents. In: *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*. [S.l.: s.n.], 2014. p. 1188–1196. Citado 3 vezes nas páginas 27, 28 e 75.
- LECUN, Y.; BENGIO, Y. et al. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, v. 3361, n. 10, p. 1995, 1995. Citado na página 36.
- MACHADO, A. et al. Personalitatem lexicon: Um léxico em português brasileiro para mineração de traços de personalidade em textos. In: *Anais do Simpósio Brasileiro de Informática na Educação*. [S.l.: s.n.], 2015. v. 26, n. 1, p. 1122. Citado 4 vezes nas páginas 15, 16, 53 e 54.
- MAIRESSE, F. *Learning to Adapt in Dialogue Systems: Data-driven Models for Personality Recognition and Generation*. Tese (Doutorado) — University of Sheffield, United Kingdom, 2008. Citado 24 vezes nas páginas 15, 16, 23, 33, 34, 35, 36, 37, 38, 46, 47, 48, 49, 52, 53, 55, 56, 61, 65, 70, 73, 83, 85 e 89.
- MAIRESSE, F. et al. Using linguistic cues for the automatic recognition of personality in conversation and text. *Journal of artificial intelligence research*, v. 30, p. 457–500, 2007. Citado 2 vezes nas páginas 52 e 61.
- MAJUMDER, N. et al. Deep learning-based document modeling for personality detection from text. *IEEE Intelligent Systems*, IEEE, v. 32, n. 2, p. 74–79, 2017. Citado 11 vezes nas páginas 15, 16, 36, 37, 38, 55, 56, 65, 71, 73 e 83.
- MAZIERO, E. G. et al. A base de dados lexical e a interface web do TeP 2.0: thesaurus eletrônico para o português do Brasil. In: ACM. *Companion Proceedings of the XIV Brazilian Symposium on Multimedia and the Web*. [S.l.], 2008. p. 390–392. Citado na página 52.
- MCCRAE, R. R.; COSTA, P. T. The structure of interpersonal traits: Wiggins’s circumplex and the five factor model. *Journal of personality and social psychology*, American Psychological Association, v. 56, n. 4, p. 586, 1989. Citado na página 21.
- MIKOLOV, T. et al. Efficient estimation of word representations in vector space. *ICLR Workshop Papers*, 2013. Citado 6 vezes nas páginas 16, 24, 26, 73, 74 e 78.
- MIKOLOV, T. et al. Distributed representations of words and phrases and their compositionality. In: *Advances in neural information processing systems*. [S.l.: s.n.], 2013. p. 3111–3119. Citado na página 26.
- MIKOLOV, T.; YIH, W.-t.; ZWEIG, G. Linguistic regularities in continuous space word representations. In: *NAACL-HLT-2013*. [S.l.: s.n.], 2013. v. 13, p. 746–751. Citado 3 vezes nas páginas 24, 25 e 27.
- MOHAMMAD, S. M.; TURNEY, P. D. Emotions evoked by common words and phrases: Using mechanical turk to create an emotion lexicon. In: *Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text*. [S.l.: s.n.], 2010. p. 26–34. Citado na página 37.
- MUNIZ, M. C. M. *A construção de recursos linguístico-computacionais para o Português do Brasil: o projeto de Unitex-PB*. Dissertação (Mestrado) — ICMC / USP São Carlos, 2004. Citado na página 65.

NORMAN, W. T. Toward an adequate taxonomy of personality attributes: Replicated factor structure in peer nomination personality ratings. *The Journal of Abnormal and Social Psychology*, American Psychological Association, v. 66, n. 6, p. 574, 1963. Citado na página 20.

NORRIS, J. R. *Markov chains*. [S.l.]: Cambridge university press, 1998. Citado na página 29.

NOWSON, S. Identifying more bloggers: Towards large scale personality classification of personal weblogs. In: *In Proceedings of the International Conference on Weblogs and Social*. [S.l.: s.n.], 2007. Citado 6 vezes nas páginas 15, 16, 42, 43, 49 e 56.

NUNES, M. A. S. N.; TELES, F. R.; SOUZA, J. G. d. Inferindo personalidade via tweets. *GEINTEC-Gestão, Inovação e Tecnologias*, v. 3, n. 3, p. 045–057, 2013. Citado 6 vezes nas páginas 15, 16, 52, 53, 55 e 56.

OBERLANDER, J.; NOWSON, S. Whose thumb is it anyway?: classifying author personality from weblog text. In: *Proceedings of the COLING/ACL on Main conference poster sessions*. [S.l.: s.n.], 2006. p. 627–634. Citado 19 vezes nas páginas 15, 16, 35, 38, 39, 40, 41, 42, 43, 45, 48, 49, 56, 61, 65, 66, 68, 71 e 88.

PARABONI, I. *Tratamento Lexical para Computação de Personalidade a partir de Textos. Relatório técnico PPgSI nro. 000/2016*. [S.l.], 2016. Citado 2 vezes nas páginas 63 e 65.

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 3 vezes nas páginas 16, 22 e 34.

PLANK, B.; HOVY, D. Personality traits on Twitter—or—how to get 1,500 personality tests in a week. In: *Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*. [S.l.: s.n.], 2015. p. 92–98. Citado 5 vezes nas páginas 15, 16, 50, 51 e 56.

RAMOS, R. M. S. et al. Building a corpus for personality-dependent natural language understanding and generation. In: *11th International Conference on Language Resources and Evaluation (LREC-2018) (to appear)*. Miyasaki, Japan: ELRA, 2018. Citado na página 91.

RANGEL, F. et al. Overview of the 3rd Author Profiling Task at PAN 2015. In: *CLEF 2015 Evaluation Labs and Workshop*. Toulouse, France: CEUR-WS.org, 2015. Citado na página 54.

ROKACH, L.; MAIMON, O. *Data mining with decision trees: theory and applications*. [S.l.]: World Scientific, 2008. Citado na página 34.

RUSSELL, S.; NORVIG, P.; INTELLIGENCE, A. A modern approach. *Artificial Intelligence*. Prentice-Hall, Englewood Cliffs, v. 25, p. 27, 1995. Citado na página 34.

SANTOS, L. B. dos et al. A lightweight regression method to infer psycholinguistic properties for Brazilian Portuguese. In: *International Conference on Text, Speech, and Dialogue*. [S.l.]: Springer, 2017. Citado 2 vezes nas páginas 24 e 73.

- SANTOS, V. G. dos; PARABONI, I.; SILVA, B. B. C. Big five personality recognition from multiple text genres. *TSD – International Conference on Text, Speech and Dialogue*, 2017. Citado na página 91.
- SILVA, B. B. C.; PARABONI, I. Learning personality traits from Facebook text. *IEEE Latin America (to appear)*, 2018. Citado na página 92.
- SILVA, M. J.; CARVALHO, P.; SARMENTO, L. Building a sentiment lexicon for social judgement mining. In: *International Conference on Computational Processing of the Portuguese Language*. [S.l.: s.n.], 2012. p. 218–228. Citado na página 23.
- SOUZA, M. et al. Construction of a portuguese opinion lexicon from multiple resources. In: *8th Brazilian Symposium in Information and Human Language Technology*. [S.l.: s.n.], 2011. p. 59–66. Citado na página 23.
- SRIVASTAVA, N.; SALAKHUTDINOV, R. R. Multimodal learning with deep boltzmann machines. In: *Advances in neural information processing systems*. [S.l.: s.n.], 2012. p. 2222–2230. Citado na página 24.
- SZIRMÁK, Z.; RAAD, B. D. Taxonomy and structure of Hungarian personality traits. *European Journal of Personality*, Wiley Online Library, v. 8, n. 2, p. 95–117, 1994. Citado na página 21.
- TUPES, E. C.; CHRISTAL, R. E. *Recurrent personality factors based on trait ratings*. [S.l.], 1961. v. 60, n. 2, 225–251 p. Citado na página 19.
- WALKER, M. A. et al. Does personality matter? expressive generation for dialogue interaction. *Natural Interaction with Robots, Knowbots and Smartphones*, Springer, p. 285, 2014. Citado na página 15.
- WESTON, J. et al. Connecting language and knowledge bases with embedding models for relation extraction. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2013. Citado na página 24.
- WITTEN, I. H.; FRANK, E. *Data Mining: Practical machine learning tools and techniques*. 2nd. ed. San Francisco, USA: Morgan Kaufman Publishers, 2005. Citado 3 vezes nas páginas 41, 44 e 49.
- WU, Y. et al. Google’s neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*, 2016. Citado na página 28.